



UNIVERSITÀ DEGLI STUDI  
DI TRENTO

---

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE  
ICT International Doctoral School

# STRUCTURAL SELF-SUPERVISED OBJECTIVES FOR TRANSFORMERS

EXPLOITING UNLABELED DATA AND STRUCTURAL SELF-SUPERVISED  
OBJECTIVES TO ALIGN PRE-TRAINING WITH FINE-TUNING

Luca Di Liello

Advisor

Prof. Alessandro Moschitti

University of Trento

Co-Advisor

Prof. Olga Uryupina

University of Trento

Mentor

Siddhant Garg

Amazon Alexa AI

---

July 2023



# Abstract

*In this Thesis, we leverage unsupervised raw data to develop more efficient pre-training objectives and self-supervised tasks that align well with downstream applications.*

*In the first part, we present three alternative objectives to BERT’s Masked Language Modeling (MLM), namely Random Token Substitution (RTS), Cluster-based Random Token Substitution C-RTS, and Swapped Language Modeling (SLM). Unlike MLM, all of these proposals involve token swapping rather than replacing tokens with BERT’s [MASK]. RTS and C-RTS involve predicting the originality of tokens, while SLM tasks the model at predicting the original token values. Each objective is applied to several models, which are trained using the same computational budget and corpora. Evaluation results reveal RTS and C-RTS require up to 45% less pre-training time while achieving performance on par with MLM. Notably, SLM outperforms MLM on several Answer Sentence Selection and GLUE tasks, despite utilizing the same computational budget for pre-training.*

*In the second part of the Thesis, we propose self-supervised pre-training tasks that exhibit structural alignment with downstream applications, leading to improved performance and reduced reliance on labeled data to achieve comparable results. We exploit the weak supervision provided by large corpora like Wikipedia and CC-News, challenging the model to recognize whether spans of text originate from the same paragraph or document. To this end, we design (i) a pre-training objective that targets multi-sentence inference models by performing predictions over multiple spans of texts simultaneously, (ii) self-supervised objectives tailored to enhance performance in Answer Sentence Selection and its Contextual version, and (iii) a pre-training objective aimed at performance improvements in Summarization.*

*Through continuous pre-training, starting from renowned checkpoints such as RoBERTa, ELECTRA, DeBERTa, BART, and T5, we demonstrate that our models achieve higher performance on Fact Verification, Answer Sentence Selection, and Summarization. We extensively evaluate our proposals on different benchmarks, revealing significant accuracy gains, particularly when annotation in the target dataset is limited. Notably, we achieve state-of-the-art results on the development set of the FEVER dataset and results close to state-of-the-art models using much more parameters on the test set. Furthermore, our objectives enable us to attain state-of-the-art results on ASNQ, WikiQA, and TREC-QA test sets, across all evaluation metrics (MAP, MRR, and P@1). For Summarization, our objective enhances summary quality, as measured by various metrics like ROUGE and BLEURT. We maintain that our proposals can be seamlessly combined with other techniques from recently proposed works, as they do not require alterations to the internal structure of Transformer models but only involve modifications to the training tasks.*

## Keywords

self-supervised pre-training, structural objectives, transformer models, answer sentence selection, fact verification, summarization

# Acknowledgments

*I would like to express my deepest gratitude to my Advisor and Professor, Alessandro Moschitti, for guiding me throughout these years and teaching me invaluable knowledge in the field of Natural Language Processing and related disciplines. I would also like to extend my thanks to my colleagues and friends, Matteo Gabburo and Stefano Campese, who have always been supportive and brought joy into my journey. Additionally, I am grateful to Siddhant Garg, who served as my Mentor and readily assisted me, even when he was not officially assigned as such. Finally, I wish to express my sincere gratitude to my Co-Advisor, Olga Uryupina, for her relentless assistance in both technical and administrative matters.*

*Among my family and friends, I am especially thankful to my parents, Dige and Rosy, who have provided unwavering support, assistance in every possible way, and enabled me to achieve this milestone. Without them, this work would not exist. Furthermore, I would like to express my gratitude to my brother, Daniel, who has been and will always be a reliable source of support and the best companion for enjoyment. A special acknowledgment goes to my grandparents and aunts/uncles, who have shown continuous interest in my journey and provided support along the way. I thank also my girlfriend, Iris, who has supported and loved me throughout the time dedicated to writing this Thesis, enduring my late nights at the computer and always being ready to lend a hand. Lastly, I am grateful to my circle of friends, including Francesco, Marco, Fabio, Matteo, Andrea, Davide, Daniele, and Simone, who have always been eager to spend time together and enjoy each other's company.*

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Contributions	2
1.1.1	Alternative Efficient Pre-Training Objectives	2
1.1.2	Specialized Task-oriented Structural Objectives	3
1.1.2.1	Self-supervised Objectives for Multi-Sentence Inference	3
1.1.2.2	Self-Supervised Objectives for AS2	4
1.1.2.3	Self-supervised objectives for Contextual AS2	4
1.1.2.4	Self-supervised Objectives for Summarization	5
1.2	Implementation & Machines	5
1.3	Publications	6
1.4	Structure of this Thesis	7
<b>2</b>	<b>Background</b>	<b>9</b>
2.1	Natural Language Processing	9
2.1.1	Question Answering	10
2.1.1.1	Passage Retrieval	10
2.1.1.2	Answer Sentence Selection	12
2.1.1.3	Machine Reading Comprehension	13
2.1.2	Fact Verification	14
2.1.3	Summarization	14
2.2	Language Models	15
2.2.1	Transformers	15
2.2.1.1	Tokenization	16
2.2.1.2	Word Embeddings	16
2.2.1.3	Positional Embeddings	17
2.2.1.4	Transformer block	17
2.2.1.5	Transformer Models	19
2.2.2	Pre-Training with objectives	21
2.2.2.1	Token-Level Pre-Training Objectives	21
2.2.2.2	Sentence-Level Pre-Training Objectives	22
2.2.3	Fine-Tuning on downstream tasks	23
<b>3</b>	<b>Related Work</b>	<b>25</b>
3.1	Language Models	25
3.1.1	BERT	25
3.1.2	RoBERTa	25
3.1.3	ELECTRA	26
3.1.4	ALBERT	26

3.1.5	DeBERTa	27
3.1.6	BART	27
3.1.7	T5	27
3.2	Natural Language Processing Tasks	28
3.2.1	Answer Sentence Selection	28
3.2.2	Fact Verification	30
3.2.3	Summarization	31
3.3	Techniques	32
3.3.1	Efficient Objectives for Pre-Training	32
3.3.2	Multi-Sentence Inference	32
3.3.3	Exploiting Document Structure	33
<b>4</b>	<b>Datasets</b>	<b>35</b>
4.1	Pre-Training	35
4.2	Fine-Tuning	36
4.2.1	Answer Sentence Selection	36
4.2.2	Fact Verification	39
4.2.3	Summarization	40
4.3	Benchmark Suites	41
<b>5</b>	<b>Alternative Efficient Objectives</b>	<b>43</b>
5.1	Alternative Objectives	44
5.1.1	Random Token Substitution	45
5.1.2	Cluster-based Random Token Substitution	46
5.1.3	Swapped Language Modeling	47
5.2	Complexity Analysis	47
5.3	Experiments	48
5.3.1	Architectures	48
5.3.2	Pre-Training	49
5.3.3	Fine-Tuning	50
5.4	Results	51
5.4.1	Pre-Training	51
5.4.2	Fine-Tuning	52
5.4.2.1	Base Models	52
5.4.2.2	Small Models	54
5.4.3	RTS vs C-RTS	55
5.5	Pre-Training Computational Trends	55
<b>6</b>	<b>Self-supervised Objectives for Multi-Sentence Inference</b>	<b>57</b>
6.1	Limits of Pairwise Models	57
6.2	Multi-Sentence Transformer Model	58
6.2.1	Joint Encoder Architecture	58
6.2.1.1	Fixed Model	58
6.2.1.2	Flexible Model	59
6.2.2	Multi-Sentence Inference Tasks	60
6.2.3	Training and Inference using Joint Encoders	60
6.2.3.1	Predict a single label	60
6.2.3.2	Predict $k$ labels	61
6.2.4	Continuous Pre-Training of Joint Encoders	62

6.3	Experiments . . . . .	63
6.3.1	Continuous Pre-Training . . . . .	63
6.3.2	Fine-Tuning . . . . .	64
6.4	Results . . . . .	64
6.4.1	Answer Sentence Selection . . . . .	64
6.4.1.1	Re-Rankers Chain . . . . .	66
6.4.2	Fact Verification . . . . .	66
6.5	Latency Analysis . . . . .	67
<b>7</b>	<b>Self-Supervised Objectives for AS2</b>	<b>69</b>
7.1	The issue with common Objectives . . . . .	69
7.2	Alternative Objectives for AS2 . . . . .	70
7.2.1	Spans in the Same Paragraph (SSP) . . . . .	70
7.2.2	Spans in Paragraph (SP) . . . . .	71
7.2.3	Paragraphs in the Same Document (PSD) . . . . .	72
7.3	Experiments . . . . .	72
7.3.1	Continuous Pre-Training . . . . .	72
7.3.1.1	Comparison of objectives hardness . . . . .	74
7.3.2	Fine-Tuning . . . . .	74
7.4	Results . . . . .	75
7.4.1	Base Models . . . . .	75
7.4.2	Large Models . . . . .	77
7.4.3	Only MLM Continuous Pre-Training . . . . .	78
7.4.4	Few-Shot Setting . . . . .	79
7.4.5	General Considerations . . . . .	81
<b>8</b>	<b>Self-supervised objectives for Contextual AS2</b>	<b>83</b>
8.1	Contextual Models . . . . .	83
8.1.1	Injecting Context into Transformers . . . . .	84
8.2	Context-aware Pre-training Objectives . . . . .	84
8.2.1	Static Document-level Context . . . . .	85
8.2.2	Dynamic Paragraph-level Context . . . . .	86
8.2.3	Dynamic Sentence-level Local Context . . . . .	87
8.3	Experiments . . . . .	87
8.3.1	Continuous Pre-Training . . . . .	87
8.3.2	Fine-Tuning . . . . .	88
8.4	Results . . . . .	89
8.4.1	Base Models . . . . .	89
8.4.2	Large Models . . . . .	92
8.4.3	General Considerations . . . . .	93
<b>9</b>	<b>Self-supervised objectives for Summarization</b>	<b>95</b>
9.1	Summarization-oriented Pre-training Objectives . . . . .	95
9.1.1	Static Document-level Summary . . . . .	96
9.2	Experiments . . . . .	97
9.2.1	Continuous Pre-Training . . . . .	97
9.2.2	Fine-Tuning . . . . .	98
9.3	Evaluation . . . . .	98
9.4	Results . . . . .	99

9.4.1	Evaluation with Silver Summaries . . . . .	100
9.4.2	General Considerations . . . . .	101
<b>10</b>	<b>Conclusions</b>	<b>103</b>
10.1	Alternative Efficient Objectives . . . . .	103
10.2	Specialized Objectives . . . . .	104
10.2.1	Multi-Sentence Inference . . . . .	104
10.2.2	Answer Sentence Selection . . . . .	104
10.2.3	Contextual Answer Sentence Selection . . . . .	105
10.2.4	Summarization . . . . .	105
	<b>Bibliography</b>	<b>107</b>
<b>A</b>	<b>Background &amp; Related Work</b>	<b>125</b>
A.1	Data Organization . . . . .	125
A.2	Language Models . . . . .	126
A.2.1	Recurrent Neural Networks . . . . .	126
A.2.2	Convolutional Neural Networks . . . . .	126
A.3	Tokenization methodologies . . . . .	127
A.3.1	Byte-Pair Encoding . . . . .	127
A.3.2	WordPiece . . . . .	128
A.3.3	UnigramLM . . . . .	128
A.3.4	Sentence-Piece . . . . .	129
A.3.5	Loss functions, Back-Propagation and Optimization . . . . .	129
A.4	Machine Reading Comprehension Branches . . . . .	131
A.4.1	Extractive Question Answering . . . . .	131
A.4.2	Generative Question Answering . . . . .	131
A.4.3	Abstractive Question Answering . . . . .	132
<b>B</b>	<b>Datasets</b>	<b>133</b>
B.1	GLUE . . . . .	133
<b>C</b>	<b>Alternative Efficient Objectives</b>	<b>135</b>
C.1	RTS vs C-RTS . . . . .	135
C.2	Negative Results . . . . .	135
C.2.1	C-RTS sampling always from the same Cluster . . . . .	135
C.2.2	Position-based techniques . . . . .	135
C.2.3	LM head-on ELECTRA’s discriminator . . . . .	137
<b>D</b>	<b>Self-Supervised Objectives for AS2</b>	<b>139</b>
D.1	Negative Results . . . . .	139
D.1.1	Additional positives from Large Language Model . . . . .	139



# Chapter 1

## Introduction

In this Thesis, we focus on the development of alternative pre-training objectives for Language Models (LMs), to improve both efficiency and final accuracy on several downstream tasks. Language Models (LMs) are advanced Machine Learning (ML) systems designed for Natural Language Processing (NLP) and Natural Language Understanding (NLU). Nowadays, LMs are predominantly based on deep learning architectures, particularly the Transformer [Vaswani et al., 2017], as seen in models like GPT [Radford and Narasimhan, 2018] and BERT [Devlin et al., 2019]. These models have brought about a revolution in the NLP industry and achieved state-of-the-art performance across a wide range of language-related applications.

The main drawback of LMs regards efficiency. Large Language Models (LLMs) recently reached a remarkable size, with over 100 billion parameters in some cases [Chowdhery et al., 2022; Scao et al., 2022; Zhang et al., 2022a]. The computational requirements for both inference and pre-training of these models are substantial, demanding a significant amount of computational resources [Schwartz et al., 2020; Strubell et al., 2019; Xu et al., 2021]. For example, the carbon footprint of BLOOM, a recently released model with 176B parameters, is estimated to be equivalent to 30 tonnes of CO<sub>2</sub> [Luccioni et al., 2022]. Other models like GPT-3 [Brown et al., 2020], Gopher [Rae et al., 2021] and LLaMa 2 [Touvron et al., 2023], are estimated to require 500, 350 and 540 tonnes of CO<sub>2</sub> to be fully pre-trained, respectively.

The success of Language Models in NLP derives from two key factors: the Transformer architecture and the Transfer-Learning (TL) training technique.

**The Transformer architecture** The Transformer was initially proposed by Vaswani et al. [2017] as a new architecture for Machine Translation (MT). The Transformer enables to capture long-range dependencies and contextual information in text. It consists of stacked neural layers, which allow the model to attend to different parts of the input text and learn meaningful representations and relations. This is performed by comparing multiple times the representation of each token with all the others in the input sequence.

**Transfer-Learning** Transfer-Learning (TL) is a widely used technique to train LMs by leveraging knowledge gained from one task to improve performance on another related task [West et al., 2007]. In traditional Machine Learning (ML) approaches, models are trained directly on specific tasks using large amounts of labeled data. However, TL is based on the assumption that most NLP tasks share common patterns and features, for example the same language and similar sentence/paragraph/document structures. TL involves two main steps: *pre-training* and *fine-tuning* [Peters et al., 2018]. During pre-training, a language model is trained from scratch on a vast amount of unlabeled data from sources like Wikipedia and CommonCrawl to capture

general language understanding. In the fine-tuning step, the pre-trained model is specialized on a specific task or domain using smaller amounts of labeled data. This process allows the model to adapt its learned features to the target task, resulting in improved performance without requiring extensive labeled data for each individual application.

In the subsequent Sections, we present a comprehensive description of the contributions of this Thesis, highlighting both the accomplishments and limitations of the proposed methodologies. Additionally, we provide an in-depth examination of the structure of this work, including the experimental setup and the hardware employed. Lastly, we furnish an overview of the papers published during my Ph.D., relating them to the various chapters of this Thesis.

## 1.1 Contributions

This Thesis examines the pre-training of Language Models from various angles. Initially, we demonstrate that the current state-of-the-art training tasks are not efficient when pre-training time is a constraint. We address this issue by providing alternative pre-training objectives which are both more efficient and lead to higher language model accuracies. Secondly, we design new pre-training tasks that align structurally with the downstream application. This allows us to fine-tune models over downstream datasets using fewer data while achieving better accuracies.

While today’s research is shifting toward Auto-Regressive (AR) LLMs that can generate text for the user, we focus on Auto-Encoders (AE), which still represent an important technique for information embedding, retrieval and classification.

Out of all the downstream tasks examined in this research, special emphasis is placed on Answer Sentence Selection (AS2), a branch of Question Answering (QA) that constitutes the primary focus of my Ph.D. studies.

### 1.1.1 Alternative Efficient Pre-Training Objectives

In this research, we study the pre-training efficiency and time complexity of LMs Auto-Encoders. We develop alternative pre-training tasks that reach higher final accuracies while using less energy. Given the widespread utilization in recent works [He et al., 2020; Lan et al., 2020; Liu et al., 2019b; Sanh et al., 2020], we adopt the Masked Language Modeling (MLM) pre-training task of BERT as the foundation for our experiments. MLM works by masking a small fraction of the input tokens and by tasking the model at predicting their original value. This procedure is expensive because the classification head of the model must span over the whole vocabulary, which usually contains thousands of tokens, to perform predictions. We design our objectives to reduce the classification head size and improve training time. Our proposals that target the inefficiency of MLM can be summarized as:

- **Random Token Substitution (RTS)**, in which input tokens are replaced with others rather than being masked. Subsequently, the model should predict only whether tokens are originals or replacements using a lightweight binary classification head;
- **Cluster-base Random Token Substitution (C-RTS)**, which is similar to RTS but replacements are harder to be detected because we exploit a simple statistical approach to find the more challenging substitutions;
- **Swapped Language Modeling (SLM)**, in which tokens are always replaced with others (and never masked) but the model should predict their original value, as in MLM.

We pre-training different models with our objectives over the same corpora and using a standard experimental setting. Then, we evaluate the effectiveness of our proposals on several downstream tasks. We show that RTS and C-RTS reduce pre-training time by up to 45% when compared with MLM while maintaining a comparable level of accuracy. Finally, it becomes evident that MLM is a sub-optimal technique, as SLM achieves greater accuracy in numerous tasks while utilizing the same computational resources. More details are given in Chapter 5.

## 1.1.2 Specialized Task-oriented Structural Objectives

The four works presented here introduce specialized pre-training techniques designed to adapt the model for specific downstream tasks during the pre-training phase. Our approach involves combining these objectives with the original training task in a continuous pre-training using various publicly available checkpoints. Continuous pre-training is a phase in which a model is trained further on large unlabeled datasets but with different objectives to align better with specific tasks. Remarkably, this continuous pre-training utilizes only 5% to 10% of the computational time and the same data as the original training, ensuring that any observed improvements are not attributed to the introduction of additional knowledge. Furthermore, we emphasize that our techniques are orthogonal to other optimization methods, such as pre-fine-tuning, which is training step performed before the real fine-tuning to further specialize the model using labeled data. We acknowledge that exploring the combination of our objectives with these alternative methods is a promising direction for future research.

### 1.1.2.1 Self-supervised Objectives for Multi-Sentence Inference

Most of the recently proposed LMs are pre-trained with token-level techniques, such as Masked Language Modeling (MLM) or Token Detection (TD), and sometimes also with some sentence-level tasks such as Next Sentence Prediction (NSP) or Sentence Order Prediction (SOP) [Clark et al., 2020; Devlin et al., 2019; Lan et al., 2020]. Token-level objectives task the model at reconstructing a corrupted input while sentence-level objectives task the model at classifying 1 or 2 input sentences, for example by predicting their order in the original document. This creates LM that perform well when the input is composed of at most 2 text spans but struggle at processing more than 3 inputs at a time. We call those models *Pairwise* sentence classifiers.

In this work, we propose a novel architecture we denote as *Jointwise*, that is specifically trained to reason over multiple input text spans at the same time. To train this multi-sentence model, we design a new pre-training task in which the model is provided  $k + 1$  sentences and should predict whether sentence  $\{s_i\}_{0 < i \leq k}$  belongs to the same paragraph as  $s_0$ . We call this objective **Multi-Sentences in the Same Paragraph (MSPP)**. This task is self-supervised and does not require annotated data, because documents from large corpora such as Wikipedia, the BookCorpus or OpenWebText are already divided into several paragraphs by humans. Since every paragraph in a document describes the same general topic from different perspectives, our objective forces the model to reason at the semantic level over multiple text spans.

After the continuous pre-training, which is performed starting from a RoBERTa [Liu et al., 2019b] checkpoint, we adapt our Jointwise architecture to perform Answer Sentence Selection and Fact Verification. In AS2, the LM is given a question and a set of possible answer candidate, and should re-rank them such that correct answers receive the higher score. Fact Verification is instead the task of determining whether a claim is supported or refuted by a set of evidence sentences. For AS2, we provide the model with the question and  $k$  candidates, and it reasons about them jointly to find the best answer. For Fact Verification, we feed the model with the

claim and a set of evidence sentences, and the model takes advantage of the latter to jointly predict whether the claim is supported or not.

By fine-tuning our models over different datasets, we show significant gains in performance over the baseline Pairwise classifier, also outperforming methods that exploit additional labeled data in some scenarios. This work is presented in Chapter 6.

### 1.1.2.2 Self-Supervised Objectives for AS2

There are many cases in which the Jointwise model described before does not adapt well to the final task. For example, in Answer Sentence Selection there may be too few answer candidates to fully exploit the Jointwise architecture effectively. Moreover, the Jointwise inference speed might be slower than an equivalent Pairwise model because of the quadratic complexity growth in the input sequence length. This may be an issue in latency-constrained scenarios, like virtual assistants. For those reasons, here we present self-supervised objectives that target Pairwise model pre-training.

Pairwise models are usually pre-trained with weak supervision using sentence-level objectives such as Next Sentence Prediction or Sentence Order Prediction. We show that those objectives are solved with high accuracies by Transformer models, thus providing weak signals while training. Moreover, those objectives do not force the model at reasoning at the semantic level because are mostly designed only to understand the order of sentences in a given corpus.

We exploit the fact that large unlabeled corpora are divided into documents and paragraphs by humans as a weak form of supervision. We design 3 self-supervised tasks for pre-training that are harder to be solved than previous sentence-level objectives and that force the model to reason at the semantic level. In particular, we propose:

- **Sentences in the Same Paragraph (SSP)**: SSP challenges the model at predicting whether two spans of text are extracted from the same paragraph of a document;
- **Sentences in Paragraph (SP)**: in SP, we extract a span of text from a paragraph and we provide it to the model along with the remaining text in the paragraph. Here, the model should predict if the span belongs to that paragraph in the original document;
- **Paragraphs in the Same Document (PSD)**: PSD tasks the model with predicting whether two entire paragraphs belong to the same original document.

After performing continuous pre-training starting from several checkpoints of RoBERTa, ELECTRA and DeBERTa, we show that our pre-training objectives outperform the baselines on many Answer Sentence Selection datasets and we reach state-of-the-art performance on WikiQA and TREC-QA. Moreover, we test our objectives in the few-shot setting, underlining the optimal performance of our approaches when the data on the target task is scarce. All the details about these experiments are provided in Chapter 7.

### 1.1.2.3 Self-supervised objectives for Contextual AS2

Recently, [Lauriola et al., 2021](#) and [Han et al. \[2021\]](#) showed that context can help re-ranking the best answer candidate in the top position. In Answer Sentence Selection happens that many answers are not well-formed or contain unresolved entities or relations. Additional context can help in solving this issue by adding details about them. In this work, our goal is to design pre-training objectives that help the model in understanding the role of the additional input context. We extend the original architectures of recently released Transformer models such as

RoBERTa and DeBERTa to allow them to tokenize and process 3 input sequences: the question, the answer candidate and the context.

The objectives we propose are extensions of the SSP task described before. We define 3 different methods to gather additional context that will be fed into the model:

- **Static Document-level Context (SDC)**: in SDC, we select the first paragraph of a document as the context. The first paragraph is usually a summary of the document content and thus it can help in solving broken references;
- **Dynamic Paragraph-level Context (DPC)**: DPC sets the context equal to the text that remains in the paragraph after the sampling of the two spans for the SSP objective;
- **Paragraphs in the Same Document (PSD)**: PSD defines the context as the pair of sentences surrounding the second span that is fed to the SSP objective, which corresponds to the answer in the final task. Notice that this aligns well with the downstream AS2 application because the context is related to the answer and not to the question.

As before, we perform continuous pre-training starting from several state-of-the-art pre-trained models and then we evaluate our proposals on many Contextual Answer Sentence Selection datasets. We show that the accuracy of our models is superior to the baselines on all the benchmarks we consider, and we also reach the state-of-the-art on ASNQ when combining our objectives with larger models, such as DeBERTaV3<sub>Large</sub>. We describe in depth these experiments and results in Chapter 8.

#### 1.1.2.4 Self-supervised Objectives for Summarization

Finally, we show that our methodologies can be adapted successfully also to generative tasks. In particular, we analyze specialized pre-training objectives designed for Summarization, in which a model is provided with a document and should generate a coherent and concise summary of the main document’s points. We leverage the weak supervision of large corpora to design an objective that accustoms the model for the Summarization task already while pre-training.

More specifically, we propose **Static Document-level Summary (SDS)**, an objective in which the model receives as input all the paragraphs of a document except the first one, and it must learn to generate the latter. Notice that we chose the first paragraph of the document because it usually acts as a summary of the whole document’s content.

We perform continuous pre-training starting from BART and T5 checkpoints for a few hundred thousand training steps. Then, we fine-tune our models and the baselines on several datasets for Summarization and we show how our methodologies can help to improve performance in Summarization. Furthermore, given the inherent subjectivity of evaluating summaries, we employ a Large Language Model with over 40 billion parameters to generate additional silver summaries. This approach allows for a more robust comparison between models and helps address the challenges associated with subjective evaluations. The full description of those techniques and the results are given in Chapter 9.

## 1.2 Implementation & Machines

We provide the details of the hardware setup used all the experiments of this work:

- CPU: 2 x Intel Xeon Platinum 8275CL @ 3.00GHz (48 cores/96 threads)
- RAM: 1152 GiB DDR4

- GPU: 8 x NVIDIA A100 GPU with 40GB HBM2 memory, connected with NVLink
- Cost: 32\$/h on AWS as of May 2023

Regarding the development of the code, we based our project on different libraries designed for Machine Learning:

- *PyTorch* [Paszke et al., 2019], described by the authors as “Tensors and Dynamic neural networks in Python with strong GPU acceleration”, is a deep learning library that offers high-performance tensors manipulation and auto-differentiation;
- *HuggingFace Transformers & Datasets* [Lhoest et al., 2021; Wolf et al., 2020], a large collection of state-of-the-art architectures, models and datasets for NLP and CV;
- *PyTorch Lightning* [Falcon et al., 2019], a framework that allows researchers to abstract from the hardware and pleasantly train and test Neural Networks in different settings;
- *TorchMetrics* [Detlefsen et al., 2022], a collection of metrics to evaluate models performance.

The framework we developed that can be used to reproduce all the experiments described in this work and the pre-trained checkpoints will be released here: <https://github.com/lucadiliello/transformers-framework>. We train all experiments with *fp16* (AMP) or *bf16*, DeepSpeed and FuseAdam (which equivalent to the AdamW optimizer but more efficient) [Rasley et al., 2020], which decrease GPU memory consumption while improving training speed.

**Evaluation and Statistical Analysis** In the tables displaying results on the downstream datasets, we emphasize our proposed models by highlighting them in bold. Moreover, we include the standard deviation for each fine-tuning on the target datasets. Unless specified otherwise, we perform 5 runs with different initialization seeds. Additionally, for each block in Tables, we underline results that are statistically superior to the corresponding baseline and emphasize the highest overall scores using bold formatting. To determine if there is a significant performance improvement, we conduct a statistical *t*-test with a significance level  $\alpha = 0.05$ .

Finally, notice that we utilize various downstream benchmarks, such as ASNQ and WikiQA, across multiple Chapters of this work. As the experiments were conducted at different times during my Ph.D. pursuit, the baseline results improved over time due to advancements in training frameworks, hardware technologies, and bug fixes. We do not retrain every model with each new software version release, maintaining the original results, even if there are occasional misalignments among different Chapters.

## 1.3 Publications

In this Section, we provide a list of the publications at international conferences that led to the creation of this work:

- **Effective Pretraining Objectives for Transformer-based Autoencoders**  
Luca Di Liello, Matteo Gabburo, Alessandro Moschitti.  
*Findings at The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022*

- **Paragraph-based Transformer Pre-training for Multi-Sentence Inference**  
Luca Di Liello, Siddhant Garg, Luca Soldaini, Alessandro Moschitti.  
*North American Chapter of the Association for Computational Linguistics (NAACL), 2022.*
- **Pre-training Transformer Models with Sentence-Level Objectives for Answer Sentence Selection**  
Luca Di Liello, Siddhant Garg, Luca Soldaini, Alessandro Moschitti.  
*The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022*
- **Context-Aware Transformer Pre-Training for Answer Sentence Selection**  
Luca Di Liello, Siddhant Garg, Alessandro Moschitti.  
*The 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023*

## 1.4 Structure of this Thesis

This Thesis begins with an introduction to Natural Language Processing and Transformer-based models, which is provided in Section 2. We describe in detail the branches of NLP, with a special focus on Question Answering and Answer Sentence Selection. Then, we give an outline of the Transformer architecture, from the tokenization algorithms and the embedding strategies to the structure of the Attention Mechanism. Follows an overview of various categories of Transformer models, such as Auto-Encoders and Auto-Regressive architectures, and a description of the most used training strategies and tasks.

Chapter 3 describes the works related to the discoveries presented in this Thesis. We start by providing an overview of several recently released state-of-the-art models, which are our starting point for many experiments. Then, we describe the most relevant works related to the tasks we address in this Thesis, such as Answer Sentence Selection and Summarization. Finally, we list works related to the methodologies we developed in this publication, such as multi-sentence inference models and specialized continuous pre-training over weakly supervised data.

In Chapter 4, we describe every dataset used in this Thesis, providing details about the data collection process, the number of examples and the annotation procedure.

From Chapter 5 to 9, we elaborate on the main discoveries of this work. In particular, in Chapter 5 we analyze alternative pre-training objectives that reduce training time while increasing final accuracy on several downstream tasks. Then, in Chapter 6 we propose an innovative Transformer architecture that can process many input spans of text at a time while performing a different prediction for each of them. We also design a new training task that, when combined with the proposed architecture, provides benefits in Answer Sentence Selection and Fact Verification settings. In Chapter 7, we study different pre-training objectives to improve models that reason over 2 spans of text by exploiting the weak supervision of large corpora, leading to improvements on various AS2 datasets. Follows Chapter 8, in which we extend the previous idea by adding context to the equation. We design custom pre-training objectives that teach the models at exploiting the context to better re-rank answer candidates while fine-tuning. Finally, in Chapter 9, we show that the weak supervision of unlabeled datasets can be exploited also for tasks different from Question Answering or Fact Verification. We describe self-supervised objectives to improve the performance of Auto-Regressive models in Summarization.

We present the conclusions of this Thesis in Chapter 10, with an overview of the methods used, the best results achieved and a list of the possible future research directions.





# Chapter 2

## Background

In this Section, we start by providing an overview of the Natural Language Processing tasks that we address in this work, with a special focus on Question Answering, Fact Verification and Summarization. Then, we describe the most renowned architectures used for NLP. In particular, we elaborate on the Transformer, which is, at the time of writing, the state-of-the-art for almost all NLP tasks. Finally, we provide an overview of the most common Transformer models referenced in this work by describing their internal architecture, the data and the objectives used for pre-training. The mathematical notation and formulas in this Chapter follow the convention of indexing starting from 1.

### 2.1 Natural Language Processing

Natural Language Processing is the study of algorithms that can process human-generated text. NLP is a mix of computer science and linguistics: while the first field provides the techniques to solve problems such as architectures design, data collection, data processing and models evaluation, linguistics studies the nature and the structure of languages, their syntax, semantics, morphology and their ambiguities [Wikipedia contributors, 2004].

The first forms of symbolic NLP were developed in 1950. They exploited a series of rules to encode the structure of a language and allowed for automatic Machine Translation. However, systems were very limited and performance was far below expectations. Twenty years later, with the development of the first large vocabularies and ontologies, scientists designed the first chatbots that could extract information from knowledge bases. A few years later, in the 80s, many algorithms for automatic parsing, semantic [Lesk, 1986] and morphology extraction [Koskenniemi, 1984] were further developed. However, the main drawback of symbolic NLP was the need of writing complex rules by hand and for every language.

In the 90s, NLP researchers started to use statistical analysis to develop better systems, in particular, using Machine Learning algorithms. ML finally removed the need to manually write rules-based parsers for each language, which was complex and expensive. Great improvements in performance, especially in Machine Translation, were enabled by the growth of the World Wide Web, starting from 1995, which allowed large quantities of raw data to be publicly available. The possibility to use large-scale datasets of unlabeled text pushed on the development of new unsupervised, self-supervised and semi-supervised encoding algorithms, e.g. Word Embeddings [Vinokourov et al., 2002].

Recently, from the 2010s and onwards, algorithms based on Artificial Neural Networks (ANN) have become the standard in NLP. The practice is to encode a sentence by first splitting it into a sequence of tokens, which may be words, sub-words or even single characters. Then,

tokens are encoded in a vector space and further processed together with various Neural Networks architectures. While before 2017 scientists exploited mainly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [dos Santos et al., 2016], recently Transformers [Vaswani et al., 2017] reached state-of-the-art performance in almost all NLP tasks [Devlin et al., 2019; Liu et al., 2019b; Radford and Narasimhan, 2018].

In the next Section, we provide an overview of the NLP task we address in this work. In particular, the focus is on Answer Sentence Selection, which is a branch of Question Answering that is rapidly gaining popularity, especially for virtual assistants such as Amazon Alexa or Google Home.

### 2.1.1 Question Answering

Question Answering (QA) is the task of automatically answer to questions posed in natural language. A QA pipeline is usually composed of multiple modules that deal with (i) the retrieval of data that contain relevant information and (ii) the extraction or generation of the correct answer span that is returned to the user.

The challenges in QA regard the organization of data for the retrieval phase, the parsing of unstructured text, the processing and disambiguation of the question and the elaboration of documents to extract the correct answer. For example, data for the retrieval phase can be organized in ontologies, graphs or simple lists of documents. While the first two methodologies allow fast access to key information, they are hard to create and maintain. On the other hand, a list of documents is easy to compose and update. However, information is stored in an unstructured way and it is not trivial to efficiently search for it.

In this Thesis, we focus mostly on **Open Domain Question Answering** [Voorhees and Tice, 2000], which is a branch of QA where questions are not restricted to a single domain but may potentially span over all human knowledge. In other more concrete words, ODQA is the task of replying to questions for which an answer may be found on the web. Open Domain QA defines new challenges because the system does not know the domain of the question a priori. Common techniques to solve the problem include powerful search mechanisms to retrieve relevant information and Machine Learning models with high generalization capabilities to extract or generate the final answer.

The following Sections describe in depth the components of a QA pipeline. Figure 2.1 shows a high-level perspective of how data are retrieved and processed to find the best answer. In particular, after the gathering of relevant documents through some Information Retrieval system, 4 different Machine Reading Comprehension approaches could be exploited to obtain the final answer. MRC is an umbrella term for methods that, given a user query, process documents and try to extract or generate the correct answer. The main difference between Answer Sentence Selection and other techniques is that in AS2, the retrieved documents are split into smaller entities (e.g. sentences), which are processed separately, while Extractive, Abstractive and Generative QA usually process the whole input document to extract or generated the target answer. While the latter methods have potentially more context to find the correct answer span, Answer Selection techniques can work on much strict latency constraints because input textual sequences are shorter and can be processed in parallel [Garg et al., 2019].

#### 2.1.1.1 Passage Retrieval

Passage Retrieval is a branch of Information Retrieval that studies how to organize, maintain, update and search over large bases of data points. In Question Answering, a passage retrieval system is given a query and should rank the available documents based on their relevance.

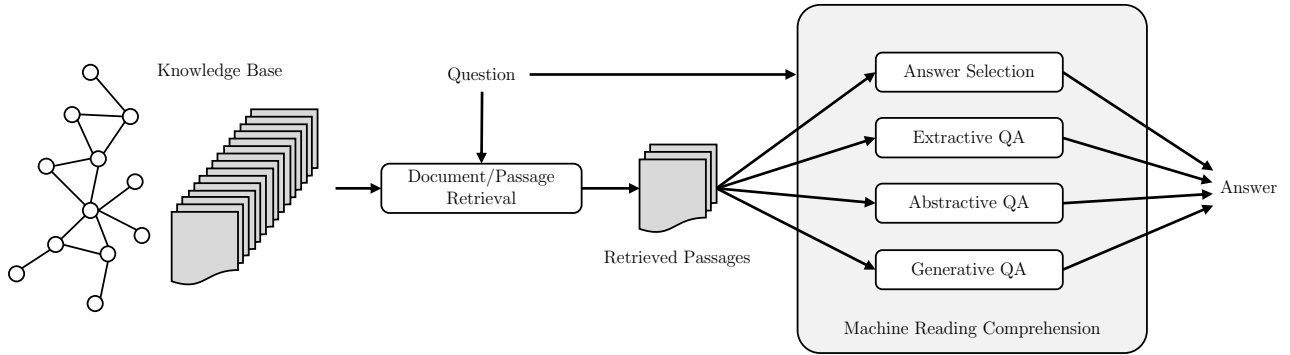


Figure 2.1: Example of the common structure of an ODQA pipeline. Source documents are usually millions or billions of data points organized as graphs or lists. The Retrieval block finds a subset of relevant documents (a few hundred or thousands) related to the question and divides them into smaller entities for easier processing. Finally, Machine Reading Comprehension techniques are used to extract the correct answer span that is returned to the user.

Documents may contain structured or unstructured text, depending on the application. The comparison of the query with each document to compute a relevance score can be performed in different ways. Among the most common, we cite: (i) tokens match, in which the score is computed as the degree of overlap between the query and the document tokens; (ii) similarity score, in which both the query and the document are encoded in a vector space and some similarity metric (e.g. cosine similarity) is used to compute the relevance.

**Token matching techniques** Among the token matching techniques, TF-IDF [Rajaraman and Ullman, 2011] and BM25 [Robertson and Zaragoza, 2009] are two non-parametric approaches that recently gained popularity thanks to their effectiveness and efficiency.

TF-IDF stands for Term Frequency-Inverse Document Frequency, and it works by calculating the importance of a word in a document or corpus by multiplying the frequency of the word (TF) by the inverse frequency of the word in the corpus (IDF). Words that appear frequently in a document but not in the corpus will have a high TF-IDF score and are considered important to the document. In Passage Retrieval, the term frequency (TF) of the query term in the document is multiplied by the inverse document frequency (IDF) of the query term in the corpus. The scores are then summed across all query terms to give a final score for each document. In practice, a map between terms and documents is pre-computed to speed up ranking.

BM25 builds over TF-IDF to improve ranking quality by also taking into account the frequency of query terms in the document, the length of the document, and other factors such as the frequency of query terms in the whole collection of documents.

**Representation similarity techniques** Regarding the techniques based on representation similarity, we cite Dense Passage Retriever (DPR) [Karpukhin et al., 2020], ColBERT [Khattab and Zaharia, 2020] and QUESION-ANSWER DATABASE RETRIEVAL (QUADRo) [Campese et al., 2023]. DPR uses two BERT [Devlin et al., 2019] models to encode both query and each document to a real-valued vector space and applies cosine similarity to compute relevance. The BERT models are fine-tuned to minimize and maximize the encoded distance between positive and negative pairs respectively. In ColBERT, the authors encode the query and the document with a BERT model to obtain vectorial representations, similar to DPR. However, to compute the similarity score, they exploit a cheap interaction module that performs a more fine-grained comparison between the query and the document embeddings, leading to better re-ranking abilities. Finally, QUADRo operates on a database of previously answered question-answer pairs rather than a

pool of documents. The authors encode the input query and millions of question-answer pairs, minimizing the distances between the embeddings when the query and the questions refer to the same concept, improving the retrieval and re-ranking accuracy of dense passage retrieval methods.

Common techniques to speed up lookup in Information Retrieval systems are indexes and inverted indexes, which use hash functions to map similar documents or parts of them to the same hash value. Examples of ready-to-be-used retrieval systems are ElasticSearch<sup>1</sup> and Facebook FAISS [Johnson et al., 2019]. The first allows users to quickly get relevant documents by searching into sharded and distributed indexes by text matching, while the latter allows search by similarity thanks to the vectorial encodings of queries and documents.

For Question Answering, knowledge can be stored in many different modalities depending on the quantity of data and whether they are structured or not. More details about the different data organization methods are given in Appendix A.1.

### 2.1.1.2 Answer Sentence Selection

Answer Sentence Selection (AS2) is a ranking task in which a system is provided with a question and a set of answer candidates, and it should rank the latter based on how likely they answer the given question.

More formally, the system is provided with a question  $q$  and a set of possible answer candidates  $\mathbb{A} = \{a_1, \dots, a_n\}$ . The goal is to select the  $a_k$  that best answers  $q$ . The common practice is to train a binary classifier  $\mathcal{M}$  that is fed with pairs  $(q, a_i), i \in \{1, \dots, n\}$  and learns answers correctness  $s_i = \mathcal{M}(q, a_i)$  by predicting whether  $a_i$  answers  $q$  or not. At inference time, the best answer to  $q$  is chosen by selecting the candidates  $a_k$  that scores the highest probability  $s_k$  of being correct:

$$k = \arg \max_i \mathcal{M}(q, a_i) \quad (2.1)$$

In AS2, there are 2 main common practices used to compute scores  $s_i$  for each answer  $a_i$  given question  $q$ : Bi-Encoders and Cross-Encoders. Bi-Encoders are systems in which the question  $q$  and the answer candidate  $a_i$  are encoded separately, using two different networks  $\mathcal{M}_1$  and  $\mathcal{M}_2$ . If the networks are the same and share the same internal parameters ( $\mathcal{M}_1 = \mathcal{M}_2$ ), we are referring to a “Siamese Networks” system. Bi-Encoders create two embeddings  $e_q = \mathcal{M}_1(q)$  and  $e_{a_i} = \mathcal{M}_2(a_i)$  that are independent from each others because no information from  $q$  is used to compute  $e_{a_i}$  and vice-versa. The main advantage of Bi-Encoders is that every answer candidate embedding can be computed efficiently in advance, even before knowing the question. The final relevance score to determine whether  $a_i$  is a correct answer to  $q$  is typically computed with a similarity metric, such as cosine similarity:

$$s_k = \frac{e_q \cdot e_{a_i}}{\|e_q\| \cdot \|e_{a_i}\|} \quad (2.2)$$

A common example of siamese networks used for Bi-Encoders is represented by Sentence-BERT [Reimers and Gurevych, 2019].

On the other hand, in the Cross-Encoder setting, a model  $\mathcal{M}$  computes a joint representation that depends both on the question  $q$  and the answer candidate  $a_i$ :

$$s_k = \mathcal{M}(q, a_i) \quad (2.3)$$

<sup>1</sup><https://www.elastic.co>

In this Thesis, we focus on Cross-Encoders because they provide the highest level of performance even though they have slightly higher computational requirements in practice. The higher computational cost arises from the encoding of a longer input sequence, which usually is the concatenation of the question and the answer candidate.

The evaluation of AS2 models requires metrics that measure the score quality compared to an ideal ranking of the answers. In an ideal ranking, every positive candidate has a higher score than all the negatives. Common metrics for AS2 include Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Precision@K (P@K). MAP is a soft metric useful to evaluate the whole ranking quality because it measures the Precision at different positions. MRR and P@1 are hard metrics that measure the position of the first correct answer retrieved. In particular, Precision@1 is the fraction of times the model was able to score a correct answer at the top of the ranking.

**Contextual Answer Sentence Selection** Recent studies [Han et al., 2021; Lauriola and Moschitti, 2021] showed that processing every answer candidate separately can lead to suboptimal results because local or global context helps in classifying answer candidates, especially when they are malformed or missing information or when they contain ambiguities or unresolved references to external entities. Contextual AS2 is an extension of AS2 in which the inputs of the language model are augmented with additional context to better disambiguate between answer candidates. More formally, in Contextual AS2 a model  $\mathcal{M}$  receives in input a tuple  $(q, a_i, c_i)$  where  $q$  and  $a_i$  are the original question and the answer candidate while  $c_i$  is the additional context to be processed. As before, the model should predict a ranking score  $s_i = \mathcal{M}(q, a_i, c_i)$ . A typical approach to get the context is to extract the sentences before and after the answer candidate in the original document (local context) or to take the first paragraph of the same document (global context).

### 2.1.1.3 Machine Reading Comprehension

Machine Reading Comprehension (MRC) is a task in NLP that involves building systems that can understand and reason about text in a similar way as humans do. This involves understanding the meaning of words and phrases, as well as the context in which they are used, and using this understanding to answer questions or make decisions based on the text.

There are several approaches to MRC, including rule-based, Information Retrieval-based, and Machine Learning-based systems. Machine Learning-based systems are the most common, and often use deep learning techniques such as neural networks to learn how to understand and reason about text. MRC has a wide range of applications, including Information Retrieval, Question Answering, Summarization, and dialogue systems and is an active area of research.

In the context of Question Answering, the term MRC typically refers to techniques in which the whole document is processed to find the target answer, such as Extractive or Generative Question Answering. While those techniques parse entire documents in a single step and have visibility over a lot of contexts, the elaboration of long sequences is expensive and results in higher latencies.

The three main approaches used to build a Machine Reading Comprehension pipeline are summarized as follows (more details are provided in Appendix A.4):

**Extractive Question Answering** Extractive QA involves the identification and selection of relevant information from a given text to directly answer a question. It relies on algorithms that analyze the text, identify key sentences or phrases, and choose the most appropriate portion

as the answer. Extractive QA models do not generate new content but rather extract existing information.

**Generative Question Answering** Generative QA entails the generation of a response to a question from scratch. It utilizes advanced language models, such as Recurrent Neural Networks or Transformer models, to understand the question and generate a coherent and contextually relevant answer.

**Abstractive Question Answering** Abstractive QA goes beyond simple answer generation by summarizing relevant information and generating a concise response. It involves understanding the question and the source text, identifying key points, and generating a summary that captures the essence of the information.

### 2.1.2 Fact Verification

Fact Verification is a task in NLP that requires verifying the accuracy of a statement or a claim [Liu et al., 2020; Zhong et al., 2020; Zhou et al., 2019]. This is accomplished by checking the sources of the information, comparing them, and evaluating their credibility. In particular, Fact Verification consists of 2 challenges: (i) the retrieval of relevant documents that may contain textual evidence for the claim and (ii) the classification of the claim as supported/refuted/neutral compared to the retrieved evidence.

The first part, which is an Information Retrieval task, is usually performed with reliable search engines such as ElasticSearch or FAISS, as described in Section 2.1.1.1. In this work, we address the second task, in which a model is provided with a claim and a set of evidence texts and should classify whether the claim is supported or not.

More formally, given a claim  $c$  and the set of evidence text spans  $E = \{e_1, \dots, e_n\}$  that were retrieved using some IR technique, the objective is to predict whether  $c$  is supported/refuted/neutral using  $E$ . Specifically, to classify  $c$ , the model should find at least one evidence  $c_i$  supporting/refuting  $c$ .

Fact verification is an important task in NLP because it helps to ensure that information is accurate and reliable, and it helps to prevent the spread of misinformation. It is also an area of active research and development, as new techniques and technologies are developed to improve the efficiency and effectiveness of Fact Verification processes.

### 2.1.3 Summarization

In NLP, Summarization is the task that involves generating concise and coherent summaries that capture the key information from longer documents. The goal is to condense the content while retaining its essential meaning. There two main branches are Extractive Summarization and Abstractive Summarization [Torres-Moreno, 2014].

**Extractive Summarization** In Extractive Summarization, the task is to identify and select the most important sentences or phrases from the original text and arrange them to form a summary. The selected sentences are typically extracted verbatim from the source document without any modification. Extractive methods rely on techniques such as sentence ranking, keyword extraction, and clustering to determine the most salient information. This technique is often the easiest to implement and can preserve the factual accuracy of the original text.

**Abstractive Summarization** In this methodology, the model should generate a summary that may contain new phrases, rephrased sentences, or paraphrased content that captures the meaning of the original text. It involves understanding the source document and generating concise and coherent summaries in a more human-like manner. This task is usually performed by Generative Models, which are given the document in input and generate the summary one token at a time. Abstractive Summarization allows for more flexibility and creativity in generating summaries but can be more challenging due to the need for advanced language understanding and generation capabilities.

## 2.2 Language Models

Language models in NLP are employed as statistical estimators to predict the likelihood of token sequences, assigning probabilities  $P(w_1, w_2, \dots, w_k)$ . These models find applications in tasks like Machine Translation, Text Classification, and Text Generation, enabling the classification or generation of human-like text.

Language Models can be constructed using various approaches, each offering its own set of advantages and drawbacks. One straightforward method involves using  $n$ -gram models [Harris, 1954], which estimate the probability of a word given the preceding  $n - 1$  words in the sequence. For instance, a bigram model ( $n = 2$ ) estimates the probability of a word based solely on the preceding word, while a trigram model ( $n = 3$ ) takes into account the probabilities of the two preceding words. While  $n$ -gram models are relatively simple to implement and can perform well with small datasets, they may exhibit decreased accuracy compared to other approaches when dealing with large datasets or long sequences of text.

An alternative approach involves neural language models, which rely on Artificial Neural Networks (ANNs) and are trained on extensive text datasets. Neural LMs can learn statistical patterns in the data and generate more natural-sounding text compared to  $n$ -gram models. Additionally, they can handle longer sequences and accommodate a larger vocabulary size. Nonetheless, they may necessitate more computational resources for training and may require more data to achieve satisfactory results.

For a long time, Recurrent Neural Networks [Elman, 1990] and Convolutional Neural Networks [Lecun et al., 1998] have been the state-of-the-art for a large variety of NLP tasks. Then, in 2017, Vaswani et al. [2017] proposed the Transformer architecture that quickly became the standard de facto for most NLP tasks such as Machine Translation, Summarization, Question Answering and Natural Language Inference. For a detailed description of RNNs and CNNs, see Appendix A.2

This Section proceeds by describing extensively the Transformer, which is the building block of all the experiments reported in this Thesis. In Appendix A.3.5, we provide an overview of the loss function, the back-propagation and the various optimization algorithms commonly used to train Transformer models.

### 2.2.1 Transformers

The Transformer [Vaswani et al., 2017] is a type of Feed-Forward ANN based on the Attention Mechanism, which allows every token to *attend* to every other token in the text. In particular, the input sequence is processed by a series of Self-Attention layers, which allow the model to directly incorporate information from the entire input sequence at each position. The Transformer also introduces the concept of Multi-Head Attention, which is the usage of multiple Self-Attention layers in parallel, each with its own set of parameters. This allows the model to attend to

different parts of the input sequence simultaneously and to learn more complex relationships between the input tokens.

In more detail, the initial steps to process a sequence of words with a Transformer model involve tokenization and embedding. Tokenization is employed to segment the input text into a sequence of tokens defined in a vocabulary, such as words, sub-words, or characters. Embedding is instead a function that maps these tokens into fixed-size vector representations. Subsequently, the embedded sentence is fed into the Transformer model, which consists of a stack of one or more Transformer blocks. The output of each block, ranging from the first to the last, contains increasingly contextualized representations of each token. Lastly, the token representations outputted by the last layer are passed through a small neural network, typically referred to as a Classification Head, to obtain predictions over a set of predefined labels or to perform regression.

### 2.2.1.1 Tokenization

The tokenization is the process of splitting the input text into a sequence of tokens belonging to a fixed-size vocabulary  $\mathbb{V}$ . Usually, a tokenization algorithm defines both the instructions to create the vocabulary  $\mathbb{V}$  from raw text and the procedure to map the text into a sequence of tokens belonging to  $\mathbb{V}$ . For example, a simple tokenizer may divide the input text into individual words by splitting over whitespaces. It may happen that some input text substrings cannot be tokenized using only the tokens in  $\mathbb{V}$ : in those cases, an OOV (out of vocabulary) token is assigned.

In Appendix A.3, we cover the four main tokenization algorithms used by recent LMs, which are Byte-Pair Encoding (BPE) [Senrich et al., 2016], WordPiece [Schuster and Nakajima, 2012], UnigramLM [Kudo, 2018] and Sentence-Piece [Kudo and Richardson, 2018].

### 2.2.1.2 Word Embeddings

Computers are inherently designed to process numerical data rather than strings. Therefore, every token generated by the tokenizer must be mapped to a vector space before being input to the LM. To achieve this, an embedding layer is utilized to convert categorical variables, such as tokens or integers, into continuous vector representations.

A trivial embedding layer is called *one-hot* encoding, in which every input token is mapped to a binary vector. The binary vector corresponding to a token  $t_i$  has the same length of the vocabulary  $\mathbb{V}$  and contains all 0s but from the position  $i$  corresponding to the token  $t_i$ , which is set to 1. For example, given a vocabulary  $\mathbb{V} = \{t_1, \dots, t_5\}$ , the one-hot encoding of  $t_3$  is:

$$\text{one\_hot}(t_3) = [0, 0, 1, 0, 0] \quad (2.4)$$

The primary issue with one-hot encoding lies in the data sparsity and the encoding size, which corresponds to the size of the vocabulary  $|\mathbb{V}|$ , typically ranging from 30K to 120K and beyond in Transformer models [Devlin et al., 2019; He et al., 2021].

A more sophisticated embedding layer should satisfy the following properties: (i) create a dense representation of tokens and (ii) map related tokens to close positions. Embedding layers are typically implemented as a matrix of weights  $\mathbf{E} \in \mathbb{R}^{|\mathbb{V}| \times d}$ , where each row  $\mathbf{e}_i$  corresponds to the representation of token  $t_i$ . While training, the embedding layer takes as input a token and returns the corresponding row from the matrix as the embedding vector.

While in some architecture the embedding matrix  $\mathbf{E}$  was learned before training, for example with Gensim [Rehurek and Sojka, 2011] or word2vec [Mikolov et al., 2013], Transformer models' embedding layers are learned during the training process of the model, allowing to extract the most relevant information directly from the input data. Embedding layers are useful for dealing



with the large vocabularies of natural languages, allowing them to represent each word with a fixed-size vector, which can be easily processed by the Transformer. They also help to overcome the problem of sparsity in one-hot encoded representation and can effectively capture semantic and syntactic information of words.

### 2.2.1.3 Positional Embeddings

The Transformer architecture, which we will deepen in Section 2.2.1.4, is based on the Attention Mechanism and processes every input token representation at the same time, regardless of the position in the input text. This is different from recurrent architectures such as GRUs [Cho et al., 2014] and LSTMs [Hochreiter and Schmidhuber, 1997], which intrinsically add positional information by processing sequentially one token at a time.

For this reason, Transformer models need an additional technique to inject positional information into the network. The original Transformer implementation used cosine positional embeddings [Vaswani et al., 2017], in which every positional embedding  $\mathbf{e}_i \in \mathbb{R}^d$  to token  $t_i$  is defined as:

$$\mathbf{e}_{i,j} = \begin{cases} \sin \frac{i}{10000^{2j/d}} & \text{if } j \text{ is even} \\ \cos \frac{i}{10000^{2j/d}} & \text{if } j \text{ is odd} \end{cases} \quad (2.5)$$

However, recent models use either learnable absolute or relative positional embeddings.

**Learnable Absolute Positional Embeddings** Learnable absolute positional embeddings are similar to Word Embeddings, in which they are composed of a matrix  $\mathbf{P} \in \mathbb{R}^{s \times d}$ , where each row  $\mathbf{p}_i$  corresponds to the unique representation of a position  $i \in \{1, \dots, L\}$  in the continuous vector space  $\mathbb{R}^d$ . The main advantage is that learnable positional embeddings are easy to implement and work well. However, the sequence length  $L$  must be fixed a priori. Finally, the embedding  $\mathbf{h}_i$  of each token  $t_i$  is computed as the sum of the word and the positional embedding:

$$\mathbf{h}_i = \text{embedding}(t_i) = \mathbf{e}_i + \mathbf{p}_i \quad (2.6)$$

We indicate with  $\mathbf{H} \in \mathbb{R}^{L \times d}$  the embedding matrix of size  $d$  of a sequence of  $L$  tokens that is provided in input to the Transformers architecture. Learnable absolute positional embeddings are used by many well-known models such as BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019b], ALBERT [Lan et al., 2020], ELECTRA [Clark et al., 2020], DeBERTa [He et al., 2020] and BART [Lewis et al., 2020].

**Relative Positional Embeddings** Relative positional embeddings [Shaw et al., 2018] are fixed values that are injected directly into the attention scores matrix to add positional information. The Attention Mechanism, described in Section 2.2.1.4, computes a score for every input tokens pair to estimate the relation between a token and all the others. This information is stored in a matrix  $\mathbf{A} \in \mathbb{R}^{L \times L}$  where  $L$  is the input sequence length. Relative embeddings are contained in a matrix  $\mathbf{P} \in \mathbb{R}^{L \times L}$  that is added to  $\mathbf{A}$ . Every position  $i, j$  in  $\mathbf{P}$  contains a value that represents the distance between token  $t_i$  and  $t_j$ . The actual computation of  $\mathbf{P}$  varies across models and architectures. Relative positional embeddings are used by models such as T5 [Raffel et al., 2020], Transformer-XL [Dai et al., 2019] and XLNet [Yang et al., 2019].

### 2.2.1.4 Transformer block

The original Transformer block is composed of 2 or 3 sub-layers, based on whether it is used as an encoder or as a decoder [Vaswani et al., 2017]. At the core of the Transformer architecture,

there is the Attention Mechanism, which is used to modify the representation of every token based on all the other tokens in the input sequence.

**Attention Mechanism** The Attention Mechanism is an architecture used in neural networks to allow a model to focus on a specific part of the input when processing it by comparing every input token representation with the others. In recent years, several Attention Mechanisms such as Additive Attention [Bahdanau et al., 2015] and Multiplicative Attention [Luong et al., 2015] have been developed. The Transformer uses a modified version of the Multiplicative Attention, called Scaled Dot-Product Attention [Vaswani et al., 2017].

The Scaled Dot-Product Attention works as follows. First, each input hidden state  $\mathbf{h}_i$  containing the representation in  $d$  dimensions of  $t_i$  is projected from  $\mathbb{R}^d$  to  $\mathbb{R}^d$  using three different learnable matrices of weights  $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ . More formally:

$$\begin{aligned}\mathbf{q}_i &= \mathbf{W}_q \mathbf{h}_i \\ \mathbf{k}_i &= \mathbf{W}_k \mathbf{h}_i \\ \mathbf{v}_i &= \mathbf{W}_v \mathbf{h}_i\end{aligned}\tag{2.7}$$

Then, projections are then packed into three matrices  $\mathbf{Q}, \mathbf{K}$  and  $\mathbf{V}$ , called respectively *Queries*, *Keys* and *Values*, and the new hidden states  $\mathbf{H}'$  are computed with the Attention Mechanism as follows:

$$\mathbf{H}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}\tag{2.8}$$

Notice that  $\mathbf{A} = \mathbf{Q}\mathbf{K}^T \in \mathbb{R}^{L \times L}$  contains the attention scores and for every pair of tokens  $t_i, t_j$ ,  $\mathbf{A}_{i,j}$  can be interpreted as the similarity between the two. The scaling factor  $\sqrt{d}$  is used to smooth the *softmax* function and stabilize gradient back-propagation by avoiding very large attention scores. Finally, the multiplication of the softmaxed attention scores with  $\mathbf{V}$  allows to influence every token representation with the attention scores, thus creating a more contextualized hidden state  $\mathbf{H}'$ .

**Multi-Head Attention** In multi-head attention, the input hidden state is split into multiple heads or subsets, and attention is performed separately on each head. Each head performs a linear transformation on the input vectors to project them into a new space, and then applies the Attention Mechanism to the transformed vectors. The results of each head are then concatenated and projected back into the original input space. More specifically, given the number of attention heads  $n$ , the Queries, Keys and Values are computed with:

$$\begin{aligned}\mathbf{q}_i^a &= \mathbf{W}_q^a \mathbf{h}_i \\ \mathbf{k}_i^a &= \mathbf{W}_k^a \mathbf{h}_i \\ \mathbf{v}_i^a &= \mathbf{W}_v^a \mathbf{h}_i\end{aligned}\tag{2.9}$$

with  $a \in \{1, \dots, n\}$  and  $\mathbf{W}_q^a, \mathbf{W}_k^a$  and  $\mathbf{W}_v^a \in \mathbb{R}^{d \times \frac{d}{n}}$ . Then, attention is computed as before for each triple  $(\mathbf{Q}^a, \mathbf{K}^a, \mathbf{V}^a)$ , yielding to a set of new hidden states  $\{\mathbf{H}'^a\}$ . The final hidden state is then computed as a linear projection of the concatenation along dimension  $d$  of  $\{\mathbf{H}'^a\}$ . More specifically,  $\mathbf{H}' = \text{concat}(\mathbf{H}'^1, \dots, \mathbf{H}'^n)\mathbf{W}_0$ , with  $\mathbf{W}_0 \in \mathbb{R}^{d \times d}$ .

The advantage of multi-head attention is that it allows the model to capture different patterns and features from the input sequence, and to learn more complex relationships between the input and output. This can lead to improved performance on several tasks such as Machine Translation, Language Modeling, and Text Classification.

**Feed-Forward Layer** The Feed-Forward layer is a sequence of two affine transformations that are applied to each  $\mathbf{h}_i$  independently. The transformation matrices are  $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d \times f}$  while the biases are  $\mathbf{b}_1 \in \mathbb{R}^f$  and  $\mathbf{b}_2 \in \mathbb{R}^d$ . The computation is performed with:

$$\mathbf{h}_i'' = \mathbf{W}_2 \max(0, \mathbf{W}_1^T \mathbf{h}_i' + \mathbf{b}_1) + \mathbf{b}_2 \quad (2.10)$$

In common Transformer models such as BERT [Devlin et al., 2019] or RoBERTa [Liu et al., 2019b],  $f$  is called *intermediate\_size*. The activation function is not always a *relu* function as in Equation 2.10, but can also be *gelu* or other custom activations depending on the model architecture.

**Encoder** The Transformer block used in encoder layers is composed of an *Self-Attention* and a Feed-Forward layer. Self-Attention means that the Attention Mechanism uses the same hidden state  $\mathbf{H}$  to compute the projections  $\mathbf{Q}$ ,  $\mathbf{K}$  and  $\mathbf{V}$ .

**Decoder** A decoder block is similar to the encoder block, with the addition of an extra attention layer that enables the incorporation of information from the encoder module. This additional block is called *Cross-Attention*. The output of the encoder is used as  $\mathbf{K}$  and  $\mathbf{V}$  of every Cross-Attention layer in the decoder, such that the hidden state is heavily influenced by the information in output from the encoder.

**Classification Head** The Classification Head is a module used to predict a label  $y$  from a set of possible labels  $\mathbb{C}$ . This is usually accomplished by putting a linear layer on top of the output representation of the last Transformer Block [Devlin et al., 2019; Lan et al., 2020]. More complex classification heads composed of multiple linear and activation functions are used in models such as RoBERTa [Liu et al., 2019b] and BART [Lewis et al., 2020].

### 2.2.1.5 Transformer Models

Transformer-based can be categorized into three different classes based on their internal architectures:

- **Encoder-only:** the model is a stack of encoder blocks, such as ELECTRA [Clark et al., 2020], RoBERTa [Liu et al., 2019b] or ALBERT [Lan et al., 2020]. These models are usually trained with self-supervised objectives like BERT’s Masked Language Modeling (MLM) or ELECTRA’s Token Detection (TD). The structure of an Encoder-only model is shown in Figure 2.2a.
- **Encoder-Decoder:** the model comprises two sets of blocks: the encoder and the decoder. The encoder blocks take the input and generate a representation of it, while the decoder blocks generate a response for the user. For instance, if the input is a question, the encoder would process it, and the decoder would generate the answer. Similarly, in Summarization, if the input is a document, the encoder would create its representation, and the decoder would generate a summary. Examples of Encoder-Decoder models are BART [Lewis et al., 2020] and T5 [Raffel et al., 2020]. Those models are usually pre-trained with objectives that challenge the model at reconstructing, from the decoder output, some corrupted input provided to the encoder. Figure 2.2b shows the architecture of an Encoder-Decoder model.
- **Decoder-only:** this class comprehends models composed only of decoder layers, such as the GPT family [Brown et al., 2020; Radford and Narasimhan, 2018; Radford et al.,

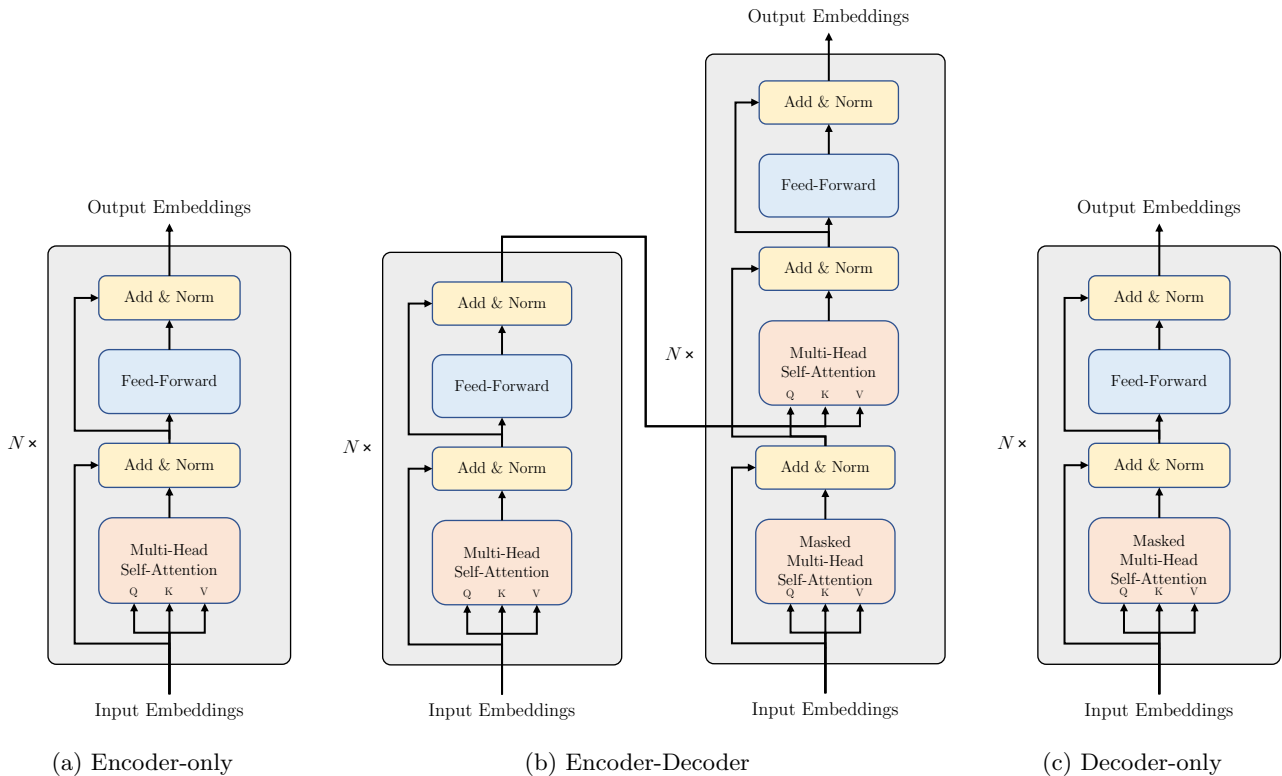


Figure 2.2: The three main Transformer configurations.  $N$  is the number of stacked layers in the encoder or the decoder.

2019], OPT [Zhang et al., 2022a] and Bloom [Scao et al., 2022]. Those models are usually pre-trained with GPT’s Causal Language Modeling (CLM) objective, which consists in predicting the next token in a sequence. Figure 2.2c shows an example of Decoder-only architecture.

Transformers can also be divided between Auto-Encoder and Auto-Regressive models based on the pre-training objective and the approach used for pre-training:

- **Auto-Encoders:** this class of Transformer models is well suited for tasks such as Sequence or Token Classification because each token in the sequence can attend to all the others. These models feature bi-directional attention and can create good vectorial text representations. For example, when trained at predicting some masked token  $t_i$ , the information from all the other tokens in the sequence  $\{t_j\}_{j \neq i}$  can be exploited. Auto-Encoders perform well on classification tasks in which the model should incorporate information from the whole input sequence to predict the target label.
- **Auto-Regressive:** those models only use the tokens preceding the target token  $t_i$  as a source of information. In other words, they are constrained to use only the tokens  $\{t_j\}_{j < i}$  for predicting  $t_i$ . Auto-Regressive models use masked attention layers to prevent the model from exploiting information from future tokens. Given the nature of Auto-Regressive Transformers, those models perform well in generative tasks, such as Summarization and Generative Question Answering, by predicting a token at a time.

Table 2.1 shows a summary of the architecture and type of well-known Transformer models. Most Encoder-only models are Auto-Encoders while Decoder-only models are primarily Auto-Regressive. Mixed models such as T5 and BART use an Auto-Encoder to create a representation of some input text and an Auto-Regressive decoder to output a response for the user.

Type	Encoder-only	Encoder-Decoder	Decoder-only
Auto-Encoder	BERT, RoBERTa, ALBERT, DeBERTa, ELECTRA		
Mixed	XLNet	T5, PEGASUS, BART, BigBird	
Auto-Regressive			GPT, Falcon, Bloom, OPT, Transformer XL

Table 2.1: Configuration of the most renowned Transformer models.

## 2.2.2 Pre-Training with objectives

The pre-training phase of Transformers is a crucial step for accustoming models to understand and generate text. It involves training on a large corpus of text data to learn the statistical patterns and relationships present in natural language. During pre-training, the model is exposed to a vast amount of raw text, such as books, articles, and web pages. The text is typically broken down into smaller units (see Section 2.2.1.1), which allows the model to process and analyze the data more effectively.

Transformers utilize a self-supervised learning approach in pre-training, which means that the training data itself provides the supervision signal for learning. Based on the pre-training objective, such as Masked Language Modeling (MLM) [Devlin et al., 2019], Causal Language Modeling (CLM) [Radford and Narasimhan, 2018] or Token Detection (TD) [Clark et al., 2020], the model is tasked at reconstructing the input which is corrupted in several ways. For example, in MLM some of the tokens are randomly replaced with a special [MASK] and the model’s task is to predict the original value given the surrounding context. In CLM, the model should instead predict the next token in a sequence given all the preceding. By training the model to reconstruct corrupted text, it learns to capture the syntactic and semantic structures of the language.

The corpora used in pre-training consist of large amounts of raw data crawled from the web. Examples are Wikipedia, the BookCorpus and the CommonCrawl.

Pre-training continues for several epochs or iterations until the model achieves a satisfactory level of language understanding and predictions accuracy. Once the pre-training phase is complete, the model can proceed to the fine-tuning stage, where it is further adapted to specific downstream tasks. In the following Sections, we provide an overview of the most cited and used pre-training objective in the literature.

### 2.2.2.1 Token-Level Pre-Training Objectives

Token-level objectives challenge the model to predict missing, replaced or corrupted tokens by exploiting the knowledge of the surrounding context. The most effective token-level pre-training objectives in the literature are Masked Language Modeling, Token Detection, Causal Language Modeling and Denoising.

**Masked Language Modeling** Masked Language Modeling works by randomly replacing some of the input tokens, i.e. 15%, with a special mask token, called [MASK]. The model is then tasked to reproduce the original tokens using the information from the surrounding context. This is usually also referred as the *cloze* task. MLM for pre-training language models was firstly proposed by BERT [Devlin et al., 2019] and later was heavily utilized by many LMs such as RoBERTa [Liu et al., 2019b], ALBERT [Lan et al., 2020] and XLM [Lample and Conneau, 2019]. MLM requires bi-directional attention, thus it targets mainly Auto-Encoder models. The main

drawbacks of MLM are that: (i) original tokens are predicted assuming they are independent of each other; (ii) there is a pre-training/fine-tuning discrepancy because the [MASK] token appears only while pre-training and not while fine-tuning; (iii) is not suited for text generation because tokens can attend to the right.

**Token Detection** Token Detection, which was initially proposed by ELECTRA [Clark et al., 2020], consists in challenging the model at predicting whether input tokens are originals or fakes created by another small Auto-Encoder network trained with MLM. First, a small Auto-Encoder model trained with MLM (the *generator*) is used to replace some masked tokens with plausible replacements. Those new tokens are inserted in the original sequence in place of the [MASK] tokens. Then, a large discriminative model (the *discriminator*) should classify every input token and tell whether it is original or a replacement created by the generator. After pre-training, the generator is discarded and the discriminator is fine-tuned on the downstream task. The advantages of Token Detection are that: (i) there is no discrepancy between pre-training and fine-tuning because the discriminator is never fed with the [MASK] token; (ii) the whole output of the discriminator is used to compute the loss, thus the training is faster because of the more effective error back-propagation. TD has been used to pre-train Auto-Encoder models such as ELECTRA, XLM-E [Chi et al., 2022] and DeBERTaV3 [He et al., 2021].

**Causal Language Modeling** Causal Language Modeling is a simple pre-training objective in which the task is to predict the next token in a sequence. CLM has been extensively used for pre-training Auto-Regressive models such as all the GPT architectures [Brown et al., 2020; Radford and Narasimhan, 2018; Radford et al., 2019], OPT [Zhang et al., 2022a] and Bloom [Scao et al., 2022]. The main drawbacks of CLM are that in inference, predicting one token at a time is expensive, even if the user applies caching techniques for previously computed  $\mathbf{K}$  and  $\mathbf{V}$  matrices.

**Denosing** Denosing is a class of objectives mostly used to train Encoder-Decoder models such as T5 [Raffel et al., 2020] and BART [Lewis et al., 2020]. The task consists in corrupting the input text by removing a subset of tokens (and possibly replacing them with a [MASK] token) before feeding the encoder. Then, the decoder is tasked with generating the original sequence (BART) or just the missing spans (T5). This task helps the model in learning to handle noisy and incomplete input data and in generating coherent output sequences.

### 2.2.2.2 Sentence-Level Pre-Training Objectives

Sentence-level pre-training objectives teach the model to reason at the sentence level and not only on individual tokens. Follows a list of sentence-level objectives commonly used in pre-training of Transformers, usually along with some other token-level task.

**Next Sentence Prediction** NSP was initially proposed by the authors of BERT [Devlin et al., 2019] and consists in providing the model with two consecutive spans of text  $s_1$  and  $s_2$  and tasking the model at predicting whether  $s_2$  appeared after  $s_1$  in the original document. While this objective was shown to improve results on some Pairwise tasks such as Natural Language Inference, Liu et al. [2019b] showed that in longer pre-trainings, the effect of the NSP objective is negligible. The main reason is that in long pre-trainings, NSP becomes a trivial task that can be solved by checking the main topic of  $s_1$  and  $s_2$ , since negatives are created

by sampling both spans from different documents. Thus, the signal that is back-propagated through the model is very weak.

**Sentence Order Prediction** SOP is a harder alternative to NSP that was proposed by [Lan et al. \[2020\]](#). In SOP, the model is fed with two spans of text  $s_1$  and  $s_2$  and should predict their order in the original document. Even if SOP is harder than NSP, large LMs achieve very high accuracy on this task, thus more challenging sentence-level objectives should be researched.

**Sentence Structural Objective** SSO was first proposed by [Wang et al. \[2019b\]](#) and improves over NSP to create a more challenging objective. In SSO, the model is given two spans  $s_1$  and  $s_2$  and should perform a three-way classification to predict whether: (i) the spans were sampled from different documents; (ii)  $s_1$  precedes  $s_2$  in the original document, (iii)  $s_1$  follows  $s_2$  in the original document.

**Sentences Shuffling** Another sentence-level objective is Sentence Shuffling [[Lewis et al., 2020](#)], which consists in shuffling the input sentences before feeding them to the encoder and tasking the decoder at reconstructing the original text.

### 2.2.3 Fine-Tuning on downstream tasks

Fine-tuning is the task of adapting a pre-trained model to perform well on some specific task, such as Natural Language Inference (NLI), Question Answering or Fact Verification.

For sentence-level classification tasks such as Natural Language Inference or Answer Sentence Selection, predictions are performed by plugging a small classification head on top of the Transformer model. For example, BERT [[Devlin et al., 2019](#)] uses a linear layer  $\mathbf{L} \in \mathbb{R}^{h \times |\mathbb{C}|}$  applied to the output embedding of the first token to predict a label  $y \in \mathbb{C}$ , where  $\mathbb{C}$  is the set of possible label classes. RoBERTa [[Liu et al., 2019b](#)] takes instead the average of all token output embeddings before performing the classification.

Regarding the token-level tasks such as POS Tagging, Named Entity Recognition or Extractive Question Answering, a small classification head similar to that described above is applied to all output token embeddings independently.

Finally, *Teacher-Forcing* is used for the fine-tuning of generative architectures. In this case, models are not enhanced with additional parameters but are trained at generating the target sequence using the original language modeling head. In particular, Teacher-Forcing is a training technique where a generative model is trained by providing the correct output at each step. It helps the model learn faster and generate higher-quality sequences. During inference, the model generates output text based on past predictions.

Generally, fine-tuning can be performed in different ways based on the underlying model architecture and the task to be performed. We avoid describing how every model is adapted for every task because this would be very verbose and is out of the scope of this work.





## Chapter 3

# Related Work

This Chapter is divided into three main Sections. First, we describe all the language model architectures we exploit in this work. In the second Section, we report the works related to the tasks we address in this Thesis, such as Answer Sentence Selection. In the last Section instead, we discuss the related state-of-the-art techniques we compare with and from which we took inspiration.

### 3.1 Language Models

The statistics about the model’s number of parameters, the data and the setting used to pre-train the architectures exploited in this work are reported in Table 3.1.

#### 3.1.1 BERT

BERT [Devlin et al., 2019] is an architecture released by Google Research that is built upon the Transformer architecture, which allows it to capture long-range dependencies in text. BERT is initially pre-trained on large amounts of unlabeled text data from sources like books and Wikipedia, using Masked Language Modeling and Next Sentence Prediction objectives. Unlike previous models that mainly relied on left-to-right or right-to-left context, BERT leverages bi-directional context to perform predictions while pre-training or fine-tuning. BERT was initially released in two versions based on the number of parameters, the number of layers and the hidden state size. However, recently the authors released also smaller, larger and multilingual variants.

#### 3.1.2 RoBERTa

RoBERTa’s [Liu et al., 2019b] architecture is the same as BERT. It was developed by Facebook AI Research in which the authors discovered that by training BERT for longer, on more data and with different hyper-parameters, the model could perform much better on a large number of downstream tasks. More specifically, they (i) removed the Next Sentence Prediction objective; (ii) applied dynamic masking in MLM; (iii) increased the batch size and the learning rate; (iv) used a training corpus almost 10 times larger and (v) changed the tokenizer from WordPiece with 30K tokens to BPE with 50K tokens. RoBERTa has been initially released in different variants based on the number of hidden layers, the hidden dimension and the total number of parameters.

Model	Type	# Params	Vocab	# Layers	Hidden size	# Att. heads	# Intern. size	# Tokens
BERT <sub>Small</sub>	AE	13M	30K	12	256	4	1024	65B
BERT <sub>Base</sub>	AE	109M	30K	12	768	12	3072	131B
BERT <sub>Large</sub>	AE	335M	30K	24	1024	16	4096	131B
RoBERTa <sub>Base</sub>	AE	125M	50K	12	768	12	3072	2097B
RoBERTa <sub>Large</sub>	AE	355M	50K	24	1024	16	4096	2097B
ELECTRA <sub>Small</sub>	AE	18M	30K	12	256	4	1024	65B
ELECTRA <sub>Base</sub>	AE	109M	30K	12	768	12	3072	524B♣
ELECTRA <sub>Large</sub>	AE	335M	30K	24	1024	16	4096	1835B♣
ALBERT <sub>Base</sub>	AE	12M	30K	12 (repeating)	768	12	3072	2097B♠
ALBERT <sub>Large</sub>	AE	17M	30K	24 (repeating)	1024	16	4096	2097B♠
DeBERTa <sub>Base</sub>	AE	139M	50K	12	768	12	3072	1048B
DeBERTa <sub>Large</sub>	AE	406M	50K	24	1024	16	4096	1048B
DeBERTaV3 <sub>Base</sub>	AE	184M	128K	12	768	12	3072	2097B
DeBERTaV3 <sub>Large</sub>	AE	434M	128K	24	1024	16	4096	2097B
BART <sub>Base</sub>	AR	139M	50K	6+6	768	12	3072	2097B
BART <sub>Large</sub>	AR	406M	50K	12+12	1024	16	4096	2097B
T5 <sub>Base</sub>	AR	223M	32K	12+12	768	12	3072	1048B
T5 <sub>Large</sub>	AR	738M	32K	24+24	1024	16	4096	1048B

Table 3.1: Comparison of several state-of-the-art Transformer models. “Type” indicates whether the model is an Auto-Encoder or Auto-Regressive model. “Token” is the total number of tokens seen while pre-training. ♣: We report the number of tokens seen in pre-training of ELECTRA<sub>Base++</sub> and ELECTRA<sub>Large-1.75M</sub> respectively, because those are the models that have been released by the authors. ♠: We report the number of tokens seen by the models released in the corresponding repository.

### 3.1.3 ELECTRA

ELECTRA [Clark et al., 2020] is yet another Transformers-based model released by Google Research which is trained as a discriminator rather than as a generator. ELECTRA is composed of a small generator network trained with MLM and a large discriminator model trained with Token Detection. The generator is provided with text in which 15% of the tokens are masked and should find reasonable replacements. On the other hand, the discriminator should identify which tokens are original and which have been introduced by the generator. After the pre-training, the generator is discarded and the discriminator is fine-tuned for the desired task.

ELECTRA offers several advantages over BERT. One significant advantage is its use of the entire discriminator output to compute the loss value, providing a stronger signal for optimization. Additionally, the discriminator in ELECTRA is never exposed to input sequences containing the [MASK] token, as the generator replaces these tokens with challenging alternatives. This addresses a key issue in BERT, where the [MASK] token is present during pre-training but absent during fine-tuning, leading to a discrepancy between the two stages.

ELECTRA was shown to match BERT’s performance using only half of the computational resources required by the latter.

### 3.1.4 ALBERT

ALBERT [Lan et al., 2020] is a Transformer-based model that was developed by Google Research as a more efficient variant of BERT. ALBERT is designed to reduce the memory footprint and computational requirements of BERT while maintaining its effectiveness in various NLP tasks. ALBERT introduces three key differences with BERT. First, the model shares the parameters of all the layers, thus being much leaner in terms of memory footprint and faster in training/inference. Secondly, they refactor the embedding layer into two separate projection matrices that reduce the total number of parameters. Third, they substitute BERT’s NSP

objective with Sentence Order Prediction (SOP), that was shown to be more challenging and more effective when the model is fine-tuned on the downstream tasks. Finally, they pre-train the model for longer and over more data than BERT. Thanks to these optimizations, they achieve state-of-the-art results on a wide range of NLP tasks such as GLUE, SQuAD and RACE.

### 3.1.5 DeBERTa

**DeBERTa** is a Transformer-based architecture developed by Microsoft Research that uses disentangled attention to improve the representation of tokens and positions [He et al., 2020]. In other successful models such as BERT and RoBERTa, the position and the token embeddings are summed together before being fed to the Transformer layers. In DeBERTa, each token is represented with 2 distinct vectors: 1 for the relative positions and 1 for the token representation. In the Transformer layer, the Attention Mechanism is applied to tokens-to-tokens, tokens-to-positions and positions-to-tokens. Position-to-position attention scores are not computed because they are not informative. Moreover, DeBERTa is pre-trained with an enhanced version of MLM that uses absolute positions applied just before the language modeling head to help in predicting the right token in long sequences.

DeBERTa is shown to outperform RoBERTa on various tasks using less computational resources. The largest model released surpassed the human baseline on SuperGLUE [Wang et al., 2019a] for the first time.

**DeBERTaV2** further improves upon the original DeBERTa model<sup>1</sup>. It introduces two key modifications: disentangled attention is now utilized in both Self-Attention and Cross-Attention (for decoding), and it utilizes an additional training objective called document-level language modeling. These enhancements are designed to improve the model’s understanding of long-range dependencies and its generalization capabilities.

**DeBERTaV3** is yet another improvement over DeBERTaV2 in which the model is trained in an ELECTRA-like configuration, i.e. the architecture is composed of a small generator network trained with MLM and a large discriminator trained with TD [He et al., 2021]. At the time of writing, DeBERTaV3 is the most accurate language model among all the architectures with less than one billion parameters.

### 3.1.6 BART

BART (Bidirectional and Auto-Regressive Transformer) is a Sequence-To-Sequence (Encoder-Decoder) language model developed by Facebook AI Research [Lewis et al., 2020]. It is based on the Transformer architecture and is trained on large amounts of unlabeled text data, where it learns to reconstruct corrupted, shuffled or masked input sequences with Teacher-Forcing. This process helps the model to acquire a rich understanding of language patterns and semantics which resulted in state-of-the-art performance across many generative tasks such as Machine Translation, Summarization and Question Answering.

### 3.1.7 T5

T5 (Text-to-Text Transfer Transformer) is a versatile language model developed by Google AI [Raffel et al., 2020]. It is trained using a “text-to-text” framework, where a wide range of

---

<sup>1</sup>More information about DeBERTaV2 can be found here: <https://github.com/microsoft/DeBERTa>

NLP tasks are cast as a text generation problem. T5 is trained to map a source text, such as a prompt or question, to a target text, which can be a summary, a translation, a sentiment analysis, or any other entity. The pre-training objectives of T5 include tasks like Translation, Summarization, Question Answering, Text Classification, and more. By training on several different tasks, T5 learns to generalize across different NLP domains and can be exploited for different applications in a zero-shot setting by only adapting the input prompt, without changes to the model architecture. T5 has achieved state-of-the-art results on various benchmark datasets across multiple tasks such as Machine Translation, outperforming previous models on standard translation benchmarks. T5 also excels in Summarization, Question Answering, and Text Classification, surpassing previous approaches and achieving competitive performance.

## 3.2 Natural Language Processing Tasks

This Section provides an overview of the works related to the tasks we address in this Thesis.

### 3.2.1 Answer Sentence Selection

This Section relates to the experiments described in Chapters 5, 6, 7 and 8. Answer Sentence Selection is a branch of Question Answering in which the task is to re-rank a set of possible answer candidates such that the correct answers to the given question are provided with the highest score.

Previous approaches to AS2 involved using CNNs to learn and score question and answer representations. [Severyn and Moschitti \[2015\]](#) discusses the importance of learning a similarity function between pairs of text spans in ranking tasks such as Information Retrieval and AS2. The authors propose a CNN architecture that learns the optimal representation of text pairs and a similarity function in a supervised way. The model achieves state-of-the-art accuracy on the tasks mentioned before and shows comparable results in tweet re-ranking without the need for manual feature engineering or additional syntactic parsers.

Another relevant improvement in AS2 was achieved thanks to Alignment Networks [[Shen et al., 2017](#)]. Alignment Networks allow the model to measure the similarity between two sentences based on their interaction information. Specifically, the authors exploit a similarity matrix built over the word representation of both input sentences to discover alignments. Moreover, their system automatically computes a weight for every word in the sentences based on estimated importance. By applying this methodology to LSTM Networks, they reach state-of-the-art results in different AS2 benchmarks, such as TREC-QA and WikiQA.

Compare-and-Aggregate architectures for Question Answering have been extensively studied [[Bian et al., 2017a](#); [Wang and Jiang, 2017](#); [Yoon et al., 2019](#)]. Compare-and-Aggregate differs from the previous techniques in the way the question and answer embeddings are used to perform the final prediction. Previous methods usually exploit a single vector representing the question and another vector for the answer to predict the target label. In Compare-and-Aggregate instead, the prediction is performed by comparing smaller units between the question and the answer, such as single-word embeddings. Then, results from all comparisons are aggregated and used for the final decision. It is worth mentioning that the authors of [Yoon et al. \[2019\]](#) mixed Clustering techniques with Compare-and-Aggregate to improve final re-ranking even further.

In [Bonadiman and Moschitti \[2020\]](#), the authors exploit separate encoders for the question and the candidate answer for a more efficient re-ranking. They show it is possible to achieve good accuracy with a training phase lasting only a few minutes. However, the exploitation of

lighter architectures such as LSTMs leads to final lower results on various AS2 datasets when compared with methods based on large Transformer models.

Recently, [Garg et al. \[2019\]](#) achieved state-of-the-art results with their new methodology called Transfer-AND-Adapt. In TANDA, a large Transformers model is first fine-tuned on a large QA corpus (ASNQ) and then it is adapted to the downstream task, which can be a small dataset such as WikiQA and TREC-QA. The advantage is that the model can exploit previously learned relations between the question and the answer to perform better in the final task. This is especially true on small target datasets such as WikiQA and TREC-QA, on which TANDA reaches state-of-the-art accuracy.

Another noteworthy contribution in the field of Answer Sentence Selection is presented by [Gabburo et al. \[2022\]](#). In their work, the authors utilize Auto-Regressive models to generate answers based on the top<sub>k</sub> candidate sentences retrieved by a Pairwise re-ranker. Building upon GenQA [[Muller et al., 2022](#)], they introduce a self-supervised technique that transfers knowledge from AS2 to Generative QA. This technique challenges the model to reconstruct the top answer retrieved by the AS2 system, leveraging only the next *k* candidates in the ranking, thereby eliminating the need for labeled datasets.

**Contextual Answer Sentence Selection** The works described in this Section refer to experiments presented in Chapter 8. Contextual AS2 is an improvement over AS2 in which the model has access to additional knowledge for each answer candidate in order to perform a better re-ranking [[Barlacchi et al., 2022](#)]. The context is a very precious source of information because it can help in solving entities and relations in the answer candidate.

In [Ghosh et al. \[2016\]](#), the authors introduce CLSTM (Contextual LSTM), an extension of the LSTM network that incorporates contextual features into the model. Experimental results over three NLP tasks (Word Prediction, Next Sentence Selection, and Sentence Topic Prediction) show that using both words and topics as features improves CLSTM performance over corresponding baselines. The study demonstrates that enhancing models with additional context is essential to improve accuracy in tasks like Question Answering, Sentence Completion, Paraphrase Generation, and Next Utterance Prediction in dialog systems.

The authors of [Tran et al. \[2018\]](#) enhance AS2 systems with context. To the best of our knowledge, this is the first work targeting specifically Contextual AS2. The authors propose an innovative architecture called Context-dependent Additive Recurrent Neural Network (CARNN) that incorporates strong external signals for a better learning process. The authors show state-of-the-art results on several datasets, including TREC-QA.

In [Tan et al. \[2018\]](#), the authors exploit GRU Networks to model answer candidates and local context, improving performance on two AS2 datasets. Unlike previous approaches, that treated question-answer pairs independently, the authors emphasize the importance of contextual information within the passage. They propose a hierarchical GRU model that incorporates context information at both the word and sentence levels. The model outperforms state-of-the-art methods on the WikiQA and SQuAD datasets, demonstrating the effectiveness of their approach.

[Lauriola et al., 2021](#) propose a Transformer encoder that uses both local and global document-level context to better disambiguate between answer candidates. Moreover, the authors perform experiments over the combination of Transfer-Learning (using the TANDA methodology) and Contextual AS2, reaching state-of-the-art performance on several AS2 datasets, such as WikiQA.

Finally, [Han et al. \[2021\]](#), the authors exploit unsupervised retrieval techniques to incorporate contextual information into Answer Sentence Selection (AS2) models. Traditional AS2 models score question-answer pairs individually without considering the document context. The authors

propose an approach that efficiently incorporates context by extracting relevant sentences using unsupervised similarity techniques and feeding them into the fine-tuned Transformer architecture. Their approach, leveraging a multi-way attention architecture, improves the state-of-the-art in AS2 while maintaining low system latency.

### 3.2.2 Fact Verification

Fact Verification is the task of predicting whether a claim is supported, refuted or neutral compared to a set of evidence sentences. This Section is related to the discoveries of Chapter 6.

An initial significant approach to address the Fact Verification task is represented by GEAR [Zhou et al., 2019]. In GEAR, the authors address the issue that previous methods have not effectively integrated and reasoned over multiple pieces of evidence. The proposed solution, a graph-based evidence aggregating and reasoning framework, allows the transfer of information on a fully-connected evidence graph and the usage of different aggregators to collect evidence information. The authors exploit contextual embeddings or BERT to further enhance performance.

DOMLIN [Stammbach and Neumann, 2019] is another technique that targets Fact Verification. In DOMLIN, the authors design a new system to select evidence sentences that does not only consider the claim but also the previously extracted evidence. They exploit BERT for both retrieval and claims classification and show promising performance gains on the FEVER dataset.

Another relevant work regarding Fact Verification is Kernel Graph Attention Network (KGAT) [Liu et al., 2020]. In this work, the authors build a graph containing the evidence sentences and apply two different kernels to it: (i) node kernels, which measure the relevance of the evidence compared to the given claim and (ii) edge kernels, which propagate evidence in the graph for a more sophisticated Fact Verification. The target is to create a network able to discover claims that are syntactical and semantically acceptable but not strongly supported by the retrieved evidence. KGAT reached state-of-the-art performance on the FEVER benchmark.

In Transformers-XH [Zhao et al., 2020], the authors address the problem that natural languages usually have a complex structure, which can be represented with trees or graphs. Thus, they propose to extend the Attention Mechanism with eXtra Hops, a technique that enables the model to attend also between text sequences than only single tokens. This innovation improves inference over tasks in which the model should reason over several text spans, such as Multi-Hop Question Answering and Fact Verification, reaching state-of-the-art performance on HotpotQA and FEVER.

Tymoshenko and Moschitti [2021] propose two interesting baselines for Fact Verification. They show that when using Transformer-based models, simple transformations of the output embeddings allow them to reach very good accuracies. Specifically, they exploit the evidence sentence embeddings by applying a max-pooling layer or by computing the weighted sum. Results on FEVER show that max-pooling and weighted sum are effective techniques to achieve state-of-the-art performance.

Finally, we describe two related works that use very large language models to retrieve more evidence and to perform Fact Verification. In Stammbach and Ash [2020-10], the authors improve DOMLIN by utilizing a large model for evidence retrieval and by exploiting GPT-3 [Brown et al., 2020] few-shot learning capabilities to summarize the retrieved evidence for each claim in the target dataset. Their new architecture, called DOMLIN++, achieves state-of-the-art results on the FEVER dataset. The other relevant work is DREAM [Zhong et al., 2020], in which the authors propose a new methodology to reason about the semantic structure of evidence sentences. By exploiting pre-trained language models such as XLNet [Yang et al., 2019], they create a graph that describes the semantic similarity of words. Then, they propagate the information through

the graph with Graph Convolutional Networks and Graph Attention Networks to contextualize the representation of words in nodes. Exploiting the described architecture, the authors achieve state-of-the-art results on FEVER on both evaluation metrics, namely Label Accuracy and FEVER score.

### 3.2.3 Summarization

Abstractive Summarization comprehends a wide range of benchmarks, models and techniques. In this Section, we try to summarize the most relevant works to provide an overview of the actual state-of-the-art in this field.

Abstractive Text Summarization is a generative task and the actual state-of-the-art is based on Auto-Regressive Transformer models in the Encoder-Decoder or Decoder-only configuration. For this reason, most of the related works are based on architectures such as BART or T5. For details about those architectures, refer to Section 3.1, while for the differences between the Encoder-Decoder and Decoder-only configurations, read Section 2.2.1.5.

The first work we review is SimCLS [Liu and Liu, 2021], which is a technique that combines contrastive learning with generative models to create high-quality summaries. The authors break down the generation process into two separate stages. First, an Auto-Regressive model such as BART is trained and exploited to generate  $n$  summaries with Beam Search. Secondly, they design a parametric function to evaluate each summary candidate independently. More specifically, they train a RoBERTa model with a contrastive loss to measure the similarity between the original document and each possible candidate. Finally, the authors show that applying SimCLS to BART or PEGASUS improves results on several benchmarks such as CNN/DailyMail and XSum.

PEGASUS [Zhang et al., 2020a] is another model specifically pre-trained for Abstractive Summarization, in which an algorithm selects a subset of sentences that are removed from the original document and used as pseudo-summaries. When released, PEGASUS was shown to outperform previous SOTA on 12 Abstractive Summarization benchmarks.

Another relevant work that addresses Abstractive Summarization is BRIO [Liu et al., 2022]. In this work, the authors address the mismatch between training and inference on generative models, called exposure bias. While training, the model predicts the next token using Teacher-Forcing based on the assumption that all the previous tokens are correct. In inference instead, the previous tokens have been generated by the model itself and there is no guarantee they are correct. Instead of training with maximum likelihood estimation, the authors assume a non-deterministic output distribution of token probabilities and propose an innovative training algorithm in which every summary generated by the model is assigned a probability mass based on the quality. They reached state-of-the-art performance on several Summarization datasets, such as CNN/DailyMail and XSum.

SummaReranker [Ravaut et al., 2022] is a Mixture-of-Experts model that addresses again the exposure bias problem by combining a model specialized for Summarization with a re-ranker that chooses the best summary among those generated by the former. By exploiting PEGASUS as the Summarization architecture, they reached state-of-the-art models on different datasets such as CNN/DailyMail, XSum and Reddit TIFU.

At the time of writing, the state-of-the-art model for Abstractive Summarization on the datasets we consider is MoCa [Zhang et al., 2022b]. Here, the authors address the exposure bias issue in inference, where the model should choose the best summary among possible candidates that deviated from the gold summary. The authors propose an alternative generation algorithm in which the model generates several samples with beam search that are aligned to the gold summaries while training, utilizing the momentum moving average technique.

## 3.3 Techniques

### 3.3.1 Efficient Objectives for Pre-Training

In this Section we analyze recent techniques that try to reduce training time and carbon footprint, to reach the same final accuracy as previous state-of-the-art models. These works are related to our research about efficient and effective objectives presented in Chapter 5.

Recently, many self-supervised pre-training objectives have been proposed to train Transformer-based models. Those objectives are mostly self-supervised because the pre-training phase is performed over large corpora of text without explicit annotation.

An example of a work that targets reduced training time by proposing architectural modifications is ALBERT, described in Section 3.1.4. In ALBERT the parameters of every Transformer layer are tied to save memory, enabling larger batch sizes. However, this approach reduces the expressive power of the models, requiring longer pre-trainings. Another approach used by Sanh et al. [2020] and Turc et al. [2019] is distillation. Distillation exploits a large and already pre-trained teacher model to train a lighter architecture to comparable accuracy, thus compressing the knowledge of the teacher into the smaller model. Distillation reduces the final architecture size, but the pre-training phase remains expensive due to the large teacher model.

Pre-training is typically a time-consuming process, taking weeks and requiring costly machines [Brown et al., 2020; Liu et al., 2019b]. Therefore, it is important to explore alternative training tasks for pre-training Transformers. In Tay et al. [2020], the reader can find a comprehensive survey of recent advancements in Transformer efficiency.

Regarding the learning objectives, SpanBERT [Joshi et al., 2020] introduces two new training tasks: Span Masking (SM) and Span Boundary Objective (SBO). SM masks contiguous spans of text instead of individual tokens, while SBO predicts span content using only the output representation corresponding to the boundary tokens. Another research about Transformers efficiency is proposed by Zhang et al. [2020b], in which the authors suggest to adapt the model to the final task during pre-training to enhance downstream performance, which is related to the methods we propose in Chapter 7, 8 and 9.

In T5 [Raffel et al., 2020], the authors propose the use of Deshuffling [Liu et al., 2019a] to pre-train an Auto-Regressive model. This technique involves shuffling random spans of text and requiring the model to output tokens in the original order. It has shown promising results across various benchmarks.

Although we focus solely on Auto-Encoder architectures, it is worth mentioning the work of Izsak et al. [2021], which discusses the usage of multiple optimizations for faster pre-training in Transformers. They also suggest that using larger models with the same runtime can yield better results. However, inference speed and efficiency decrease as the number of parameters of the model increases.

### 3.3.2 Multi-Sentence Inference

Inference over multiple sentences has already been studied in the past, also for different problems than Answer Sentence Selection. In this work, we do not target new architectures for multi-sentence inference, but effective pre-training strategies to improve models' ability at collecting information from several spans of text.

Answer Sentence Selection is a challenging task because the model needs to effectively capture the semantic connections between questions and answers. In [Bian et al., 2017b], the authors propose two improvements over previous state-of-the-art methods, mostly based on Compare-Aggregate networks, which are described in Section 3.2.1. First, they introduce Dynamic-Clip



Attention, which focuses on reducing noise in the attention matrix to better find relations between the question and the answer tokens. Secondly, they feed their architecture with a question and a set of answer candidates and task the model with predicting the relative order of the latter. They reach state-of-the-art results on WikiQA and TREC-QA.

Another relevant work about inference over multiple text spans is [Ai et al. \[2018\]](#), which addresses the Information Retrieval task. In this work, the authors demonstrate that ranking documents independently for each query is suboptimal because they may follow a different distribution in the feature space. For this reason, they develop a Deep Listwise Context Model, which is applied to the most relevant documents retrieved for each query to improve their final ranking. The proposed architecture is designed to encode multiple documents and to model their interactions, and could be applied on top of other ranking systems. By training with an attention-based loss function, they achieve state-of-the-art results on several benchmarks for Information Retrieval.

The most relevant related works targeting Answer Sentence Selection with multiple answer candidates are [Bonadiman and Moschitti \[2020\]](#) and [Zhang et al. \[2021\]](#). The first work employs cheap and fast neural networks to rank multiple sentences together. It leverages the original order of answer candidates when extracted from source documents for a better ranking. When tested over WikiQA or TREC-QA, the proposed architecture is hundreds of times faster than Transformer-based models such as BERT, and achieves better performance than previous approaches with a similar number of parameters. The second work exploits mutual information and relations between several answer candidates to improve the ranking quality. Specifically, their architecture performs a three-way classification between answer candidate pairs and should predict whether candidates support, refute or are neutral compared to the others, similar to a Fact Verification pipeline. They combine state-of-the-art AS2 systems with their multi-classifier through a joint layer, which outputs scores for the ranking. The proposed architecture obtains state-of-the-art results on several AS2 datasets, such as WikiQA and TREC-QA.

### 3.3.3 Exploiting Document Structure

The intrinsic structure of documents has been leveraged in various studies. For instance, RoBERTa [\[Liu et al., 2019b\]](#) generates input examples for pre-training while avoiding crossing document boundaries. This approach yields a more accurate model across multiple benchmarks because the model focuses on a single topic at a time and can be fine-tuned more effectively for target tasks.

Another work that exploits document structure is REALM [\[Guu et al., 2020\]](#), in which the authors augment models pre-training with a neural retriever that collects additional knowledge from large sources of text, such as Wikipedia. By applying the described technique in both pre-training, fine-tuning and inference, they incorporate useful additional knowledge from millions of semi-structured documents, reaching state-of-the-art on several Open Domain Question Answering benchmarks. This work differs from our pre-training setting because it exploits external knowledge to improve the model’s accuracy.

DeCLUTR [\[Giorgi et al., 2021\]](#) is yet another approach that exploits the document structure to create better sentence-level embeddings. They design a self-supervised objective that challenges the model at predicting whether two sentences are extracted from the same paragraph of a document. However, as opposed to our discoveries, they do not exploit the Cross-Attention of Transformer models and work only at the individual sentence level. The continuous pre-training they apply starting from publicly available checkpoints reduces the performance gap with models trained over labeled datasets. They show improvements in accuracy over several tasks that require reasoning over two input sentences.



# Chapter 4

## Datasets

The architecture and the training objectives proposed in this Thesis have been evaluated on a wide range of benchmarks. In this Section, we describe the domain, the number of examples and the annotation strategy of various datasets. Moreover, we provide details about how data have been collected from the original sources.

### 4.1 Pre-Training

Large corpora of raw text are of fundamental importance for models pre-training. Statistics for all datasets are reported in Table 4.1.

**Wikipedia** The Wikipedia dataset is a collection of text data derived from the Wikipedia online encyclopedia. It encompasses a diverse range of subjects, covering fields like science, history, geography, arts, and more. It contains a substantial amount of text, including introductory paragraphs, detailed explanations, citations, and references, providing a comprehensive and authoritative source of information for a wide array of topics<sup>1</sup>.

**BookCorpus** The BookCorpus dataset [Zhu et al., 2015] is a relatively small collection of text data created by Google Research as a means to provide a diverse and extensive source of textual information for Machine Learning purposes. We use the Open version of this dataset in our experiments, which consists of approximately 17,868 books from a wide range of genres and topics<sup>2</sup>. The books included in the dataset were published between 2007 and 2020 and were released under an open-source license.

**OpenWebText** The OpenWebText dataset is a collection of text data derived from various sources on the internet, excluding the Wikipedia website. It aims to capture a broader range of language patterns, writing styles, and topics compared to datasets focused solely on Wikipedia. The dataset is typically compiled by scraping publicly available web pages, blogs, forums, news articles, and other online sources. The original scraping script was created by OpenAI but it was not publicly released. We use an open version of the OpenWebText, which is based on Reddit dumps that are filtered based on the Karma score ( $\geq 3$ ) of the threads<sup>3</sup>.

---

<sup>1</sup>We publish our pre-processed version of Wikipedia here: [https://huggingface.co/datasets/lucadiliello/english\\_wikipedia](https://huggingface.co/datasets/lucadiliello/english_wikipedia), which corresponds to the dump of November 1, 2021.

<sup>2</sup>We used the version of the BookCorpus released here: <https://huggingface.co/datasets/bookcorpusopen>

<sup>3</sup>We used the version of OpenWebText released here: <https://github.com/jcpeterson/openwebtext>

Dataset	# Docs	# Words	# Words/Doc	# Para/Doc	# Sent/Para	Size
Wikipedia	4,184,712	1,610,571,882	384.8	7.9	2.3	9.9GB
BookCorpus	17,868	1,149,401,064	64,327.3	2005.9	2.5	6.2GB
Open-Web-Text	8,013,769	6,401,507,179	798.8	20.7	1.9	38.8GB
CC-News	149,954,415	62,676,498,100	418.0	16.5	1.3	385.4GB

Table 4.1: Statistics about the pre-training corpora exploited in this work. Cleaning involves filtering HTML tags and JavaScript code from pages, images, tables and other entities which are not raw text.

**CC-News** CC-News is a dataset released by the Common Crawl Foundation that is exclusively centered around news articles gathered from the web. CC-News aims to offer a comprehensive and diverse compilation of news articles, encompassing various topics, languages, and sources. The dataset covers articles from a wide range of news websites, covering domains such as politics, sports, technology, entertainment, and more. We parse the data of the original CC-News dataset to keep a random subset of the English articles published between 2016 and 2020<sup>4</sup> and we filter away artifacts such as HTML code, headers, titles from the WARC files.

## 4.2 Fine-Tuning

### 4.2.1 Answer Sentence Selection

The main task addressed in this Thesis is Answer Sentence Selection. For this task, we evaluate our proposed models and objectives on several public and two industrial datasets. Moreover, we created 2 additional corpora derived from well-known Question-Answering datasets, similar to how the authors in Garg et al. [2019] created ASNQ. With a total of 8 datasets, we present an exhaustive evaluation setting that includes datasets of varying sizes, ranging from very few thousand examples to more than 10 million. Moreover, we include datasets where sentence candidates are extracted from both single and multiple source documents. We created *clean* versions of each dataset, which means we removed questions without at least a positive and a negative answer candidate in the development and test set, which is the standard practice in AS2 [dos Santos et al., 2016; Garg et al., 2019]. This allows us to better compare results because otherwise, there would have been upper bounds to the AS2 metrics dependent on the tested dataset.

**ASNQ** The Answer Sentence Natural Question (ASNQ) is a large dataset for AS2 derived from Google’s Natural Questions corpora [Kwiatkowski et al., 2019]. The questions have been retrieved from Google’s search engine and the system is required to reason over entire Wikipedia articles to find the correct answer. NQ is a challenging dataset because it contains thousands of questions asked by real users and answers extracted from long Wikipedia articles. The annotations include both *long* and *short* answers. The short answers are very small text spans that contain the exact answer to the question. Long answers are instead paragraphs that contain the exact answer but also other sentences and text. Most of the short answers are contained in a long answer. There are very few rare cases in which the short answers are not contained in the long answers, probably due to annotation mistakes. More specifically, NQ contains 307,373 train and 7,830 development user questions and, for each question, the most relevant Wikipedia article is provided. For the train set, a single long answer annotation is available. For the development set instead, the authors asked 5 different annotators to select a long answer. Regarding the short answers, on average there are 0.42 annotations in the train set and 2.10 in the development set.

<sup>4</sup>We release our version of CC-News here: [https://huggingface.co/datasets/lucadiliello/cc\\_news](https://huggingface.co/datasets/lucadiliello/cc_news)

Dataset	Train set				Development set				Test set			
	Q	QA	A/Q	+/Q	Q	QA	A/Q	+/Q	Q	QA	A/Q	+/Q
ASNQ	57242	20377568	356.0	1.1	1336	463914	347.2	1.6	1336	466148	348.9	1.6
WikiQA	2118	20360	9.6	0.5	122	1126	9.2	1.1	237	2341	9.9	1.2
TREC-QA	1158	53310	46.0	5.4	69	1343	19.5	3.5	68	1442	21.2	3.5

Table 4.2: Statistics for ASNQ, WikiQA and TREC-QA. “Q” and “QA” stand for the number of unique question and question-answer pairs. “A/Q” is the average number of candidates per question while “+/Q” is the average number of positive answers per question.

ASNQ is derived from NQ by [Garg et al. \[2019\]](#) with the following rules. First, the authors split every retrieved Wikipedia document into a list of answer sentence candidates with the NLTK tokenizer [[Bird et al., 2009](#)]. Then, they assign a positive label to answer candidates which are entirely contained in a long answer and that contain at least a short answer. Negative answer candidates comprehend instead: (i) sentences in a long answer but not containing a short answer; (ii) sentences containing a short answer but not contained in a long answer; (iii) sentences both not in a long answer and not containing a short answer. Notice that the formers are very hard negatives because they are very close to the short answer (in the same paragraph) and contained in a long answer. Moreover, the second case happens only because of a few mistakes in the annotation process. Finally, the authors discard questions not having at least a positive answer candidate.

The statistics for the resulting ASNQ dataset are reported in [Table 4.2](#). For the experiments in [Chapter 5](#), we report results on the original ASNQ development set, because the dataset does not provide a test set. The results in [Chapters 6, 7 and 8](#) are instead on the test set released by [Soldaini and Moschitti \[2020\]](#), where the authors split the original development set for both validation and testing.

**WikiQA** WikiQA is a small and challenging dataset for Answer Sentence Selection released by Microsoft and based on queries asked to the Bing search engine [[Yang et al., 2015](#)]. Based on research results, each question is associated with a Wikipedia page. Sentences from the summary (first paragraph) are then used as answer candidates and are annotated by crowd-workers. To ensure a high annotation quality, each sentence is independently analyzed by three different workers. Sentences with inconsistent labels were further parsed by a different set of workers and finally, labels were assigned based on majority vote. Statistics about the number of questions and answer candidates are given in [Table 4.2](#).

**TREC-QA** TREC-QA is yet another popular dataset for Answer Sentence Selection created from the TREC-8 to TREC-13 tracks of Question Answering [[Wang et al., 2007](#)]. The corpus is composed of factoid questions while answers are extracted from web documents by participating teams. In particular, the authors of [Wang et al. \[2007\]](#) use the tracks TREC-8 to TREC-12 for training and the track TREC-13 for development and testing. Manual annotation of positive candidates is performed on both the development and test sets. In the training set, only the answers to 100 questions received gold labels while the others were automatically annotated, thus being more noisy. The split containing also the manually annotated candidate answers is called *train-all*. We train over train-all, which is the standard practice [[dos Santos et al., 2016](#)], because even if it is noisier, it contains much more question-answer pairs. Negative answer candidates are automatically extracted by selecting sentences with overlapping words with the question. Statistics for TREC-QA are given in [Table 4.2](#).

Dataset	Train set				Development set				Test set			
	Q	QA	A/Q	+/Q	Q	QA	A/Q	+/Q	Q	QA	A/Q	+/Q
IQAD	221334	3894129	17.6	-	2434	43369	17.8	-	2252	38587	17.1	-
WQA	9984	149513	15.0	0.23	5000	74805	15.0	0.21	2088	33498	16.0	-
					5000	74712	14.9	0.21				

Table 4.3: Statistics for IQAD and WQA.

**IQAD** IQAD (Internal Question Answering Dataset) is a large-scale internal dataset of the Alexa Search team used to evaluate QA systems. The dataset was built from Alexa traffic by de-identifying user questions asked to Alexa, which is a popular virtual assistant. The de-identification process removes all connections between questions and original Alexa customers. The dataset contains about 220K general topic questions and a large set of answer candidates extracted with ElasticSearch from a web index containing more than 1 billion pages. More detailed statistics are reported in Table 4.3. IQAD contains 2 different test sets that differ in the period in which data were gathered and in the annotation procedure. The first set contains 2.2K questions, with about 17 answer candidates per question, annotated manually by crowd-workers. The second test split contains 2K questions instead and about 16 answer candidates per question annotated by crowd-workers following strict fact verification guidelines. A manual analysis of a subsample of 100 questions for each dataset showed higher annotation quality in the second set. In the experimental Sections, we refer to the first test set as “IQAD Bench 1” while the second is named “IQAD Bench 2”. In summary, IQAD is a very challenging dataset because it reflects real user questions, which may be vague, ambiguous or malformed.

**WQA** WQA (Web Question Answering) is yet another AS2 dataset created sampling questions asked to the Alexa virtual assistant in 2019 [Hsu et al., 2021; Zhang et al., 2021]. Questions are non-representative and are anonymized to remove connections to the original users and geographical locations. Moreover, questions are not filtered by topic, so they span several different arguments. For each question, 500 pages were extracted from an index of more than 100M documents crawled from the web using ElasticSearch. Then, answer candidates has been ranked with a state-of-the-art AS2 system and the top 100 were annotated by crowd-workers. Statistics for this dataset are given in Table 4.3. Similarly to IQAD, WQA is a challenging corpus because questions mirror real Alexa customer queries, which usually are noisy, malformed or misleading.

**NewsAS2** NewsAS2<sup>5</sup> is a dataset derived from NewsQA [Trischler et al., 2017], which was originally designed for Machine Reading.

NewsQA is a dataset composed of over 100K question-answer pairs generated by crowd-workers. First, the authors extracted 12.7K articles from the CNN/Daily Mail corpus with random selection. Those articles cover several topics, such as politics and business. Then, crowd-workers have been divided into three groups. The first group task was to generate questions by letting the workers see only the article’s title and a short summary, thus promoting curiosity about the article’s content. The second group was then asked to highlight answers in the documents, if present, or to reject questions when nonsensical or malformed. A positive label was assigned to answer on which at least two crowd-workers agreed. Finally, to improve the dataset quality, the authors asked the third group of workers to review the questions without an agreement on the answer and to make the final decision, or to reject them.

<sup>5</sup>We release NewsAS2 here: [https://huggingface.co/datasets/lucadiliello/news\\_as2](https://huggingface.co/datasets/lucadiliello/news_as2)

Dataset	Train set				Development set				Test set			
	Q	QA	A/Q	+/Q	Q	QA	A/Q	+/Q	Q	QA	A/Q	+/Q
NewsAS2	71561	1840533	25.7	1.7	2070	50895	24.6	1.7	2059	50842	24.7	1.7
TriviaAS2	61688	1843349	29.9	1.7	1201	37470	31.2	5.3	1190	37691	31.7	5.4

Table 4.4: Statistics for NewsAS2 and TriviaAS2.

We downloaded NewsQA from the MRQA [Fisch et al., 2019] competition repository<sup>6</sup> and split the development set in both dev and test because the original test set is not labeled. The conversion from MR to Answer Sentence Selection has been performed with the following algorithm. First, each sample  $(q, \mathbf{D})$  from the original dataset was transformed in a tuple  $(q, \{s_1, s_2, \dots, s_n\})$  by splitting the document  $\mathbf{D}$  into multiple sentences with the NLTK tokenizer [Bird et al., 2009]. Then, each sentence  $s_i$  was labeled as positive if it entirely contained one of the original answer spans, as negative otherwise. NewsAS2 contains more than 70K questions and 1.8M question-answer pairs. The number of questions without at least a positive answer candidate is 1.5% in the development set and 1.1% for the test set. We remove those questions from the corresponding sets (clean setting). Statistics for NewsAS2 are reported in Table 4.4.

**TriviaAS2** TriviaAS2<sup>7</sup> is a dataset for AS2 derived from TriviaQA [Joshi et al., 2017], which was originally designed for Machine Reading.

TriviaQA is a challenging dataset for Reading Comprehension composed of questions and answers retrieved from 14 trivia and quiz-league websites. In particular, about 95K trivia questions authored by enthusiasts were extracted, and for each question, an average of 6 documents were crawled using the Bing search engine. Questions were filtered to have a minimum length of 4 tokens while documents were filtered when containing keywords like trivia, question and answer or when they contained ill-formatted text, such as PDFs. In addition, for each query, the authors extracted an additional set of about 2 documents from Wikipedia by searching the entities contained in the questions. Finally, documents were filtered when not containing the original answer to promote learning with distant supervision.

As for NewsQA, we download TriviaQA from MRQA and split the development into both development and test to have non-hidden annotations. The conversion from MR to AS2 is performed the same algorithm used for NewsAS2. Statistics for the resulting TriviaAS2 are given in Table 4.4.

## 4.2.2 Fact Verification

Fact Verification is the task of predicting whether a claim is entailed, neutral or contradicted by a series of evidence sentences. Given the nature of this task, which requires the model to reason over multiple sentences, it is well suited for our Multi-Sentence Inference architecture, see Section 6.

**FEVER** FEVER [Thorne et al., 2018], an acronym for “Fact Extraction and VERification”, is a challenging dataset composed of about 185K claims and 920K evidence sentences extracted from Wikipedia.

The data has been generated with the following procedure. First, the initial section of 50K popular pages from Wikipedia has been extracted and cleaned. Then, a set of crowd-workers

<sup>6</sup><https://github.com/mrqa/MRQA-Shared-Task-2019>

<sup>7</sup>We release TriviaAS2 here: [https://huggingface.co/datasets/lucadiliello/trivia\\_as2](https://huggingface.co/datasets/lucadiliello/trivia_as2)

Dataset	Train set			Development set			Test set		
	C	E	E/C	C	E	E/C	C	E	E/C
FEVER	145405	722473	4.97	19998	98915	4.94	19998	98839	4.94

Table 4.5: Statistics for the FEVER dataset. “C” and “E” are the total number of claims and evidence sentences, while “E/C” is the average number of evidence sentences per claim.

was tasked to generate claims containing a single statement related to the page topic. To avoid the creation of too complex claims, the crowd-workers were provided with a limited set of words collected from linked pages that could be used for the generation. Moreover, crowd-workers were allowed to modify existing claims, thus changing whether they were supported by the available evidence or not.

In the second phase, crowd-workers were provided with claims and were asked to find a set of evidence texts to support or refute the claim. If the claim could not be supported or contradicted, the workers should have provided evidence for their decision.

Since FEVER’s original task involves both Information Retrieval of the evidence sentences and Recognizing Textual Entailment, we used the evidence collected by [Thorne et al. \[2018\]](#) using a BERT-based DocIR<sup>8</sup>. The final dataset statistics are reported in Table 4.5. On average, for each claim, a set of 5 evidence candidates is available to predict the target label. To evaluate our models, we submitted them to the FEVER shared task website<sup>9</sup> because labels for the test set are hidden.

### 4.2.3 Summarization

The statistics for all datasets about Abstractive Summarization used as benchmarks in this Thesis are reported in Table 4.6.

**CNN/DailyMail** The CNN/DailyMail dataset is a widely used benchmark dataset for Abstractive Summarization which was developed by [Hermann et al. \[2015\]](#). The dataset is composed of news articles scraped from the CNN and The Daily Mail news websites between 2010 and 2015. Every article of the dataset is paired with summaries that were written by humans, providing a concise and informative representation of the article’s main points. The dataset covers a wide range of topics, including politics, technology, science, sports, and entertainment. The articles are written in English and have a mix of short and long sentences, making the dataset challenging to summarize.

**XSum** XSum is a large-scale dataset for text Summarization that was created by [Narayan et al. \[2018\]](#). The XSum dataset consists of approximately 227K news articles and their corresponding summaries. The articles were sourced from a wide range of news outlets, including The Guardian, The Independent, and The Daily Mail and cover a variety of topics, including politics, business and sports. The summaries in the XSum dataset are typically one or two sentences in length and provide a concise description of the main points of the article. The dataset was built with a two-stage process. First, a group of human annotators wrote a summary for each article. Then, a second group of human annotators reviewed the summaries and edited them to ensure they provided an accurate and comprehensive summary of the article, thus ensuring high-quality summaries.

<sup>8</sup>We release our version of the FEVER dataset here: <https://huggingface.co/datasets/lucadiliello/fever>

<sup>9</sup><https://competitions.codalab.org/competitions/18814>



Dataset	Set	Examples	Document Length				Summary Length			
			Mean	Std	Min	Max	Mean	Std	Min	Max
CNN/DailyMail	Train	287113	4033.6	1954.3	48	15925	294.7	120.1	14	7388
	Dev	13368	3923.8	1992.3	245	11412	328.0	146.2	52	8541
	Test	11490	3967.0	2024.7	293	11991	311.9	128.0	51	3410
XSum	Train	204017	2202.4	1795.5	15	174045	125.4	30.2	1	399
	Dev	11327	2174.7	1664.9	26	27345	125.5	30.4	5	500
	Test	11333	2214.5	1817.6	14	63937	125.3	30.5	17	410
Samsun	Train	14732	511.1	402.6	0	5492	110.1	60.8	3	300
	Dev	819	499.3	403.3	57	2950	110.1	61.3	17	300
	Test	818	521.5	409.1	49	2793	108.7	58.4	17	300
Gigaword	Train	3803957	181.5	47.3	35	636	51.0	14.7	6	265
	Dev	189651	181.5	47.3	40	520	51.8	14.4	7	158
	Test	1951	170.8	56.4	3	409	52.2	19.3	7	137

Table 4.6: Statistics for all Summarization datasets employed in this work. We report the number of examples for each split and statistics about the length of documents and summaries (in terms of characters, including whitespaces). As a rule of thumb, the number of characters per token is  $\sim 4$  for most Transformer models.

**Samsun** The Samsun dataset is a corpus for conversations Summarization that was introduced by Liu [2019]. The dataset consists of over 11,000 conversations between two participants with the scope of scheduling a meeting. Each conversation contains a set of messages exchanged between the participants, and the goal of the Summarization task is to produce a concise summary of the conversation that captures its most important aspects.

Conversational Summarization poses several challenges that are not present in traditional document Summarization. Conversations are often highly contextualized and contain a large amount of implicit information, such as the participants’ intentions, preferences, and expectations. As a result, summarizing a conversation requires not only identifying important content but also understanding the underlying discourse and social interactions.

In the Samsun dataset, a set of small summary sentences were produced by human annotators. Each summary is intended to capture the key points of the conversation and provide a concise overview of its content.

**Gigaword** Gigaword is a large-scale text corpus for Abstractive Summarization created by the Linguistic Data Consortium (LDC) [Graff et al., 2003]. The dataset contains over one billion words and covers a wide range of topics and domains over more than two decades of events, from 1994 to 2015. The documents are news articles extracted from a variety of sources, such as the North American News, the AQUAINT corpus and the Agence France-Presse. In the Gigaword dataset, every article is paired with a very short summary that captures its key points. These summaries are typically one to three sentences long and were written by human editors.

## 4.3 Benchmark Suites

**GLUE** GLUE (General Language Understanding Evaluation) [Wang et al., 2018] is a collection of datasets for testing systems on a variety of tasks. It includes CoLA, a benchmark for Language Acceptability [Warstadt et al., 2019]; QNLI, MNLI and WNLI, three datasets for Natural Language Inference [Levesque et al., 2012; Wang et al., 2018; Williams et al., 2018]; RTE, to test performance in Textual Entailment [Giampiccolo et al., 2007]; QQP<sup>10</sup> and MRPC

<sup>10</sup><https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>

Dataset	Task type	Classes	Train set		Development set		Test set
			Examples	Examples/Class (%)	Examples	Examples/Class (%)	Examples
CoLA	Classification	2	8551	29.5/70.5	1043	30.9/69.1	1063
MNLI-m	Classification	3	392702	33.3/33.3/33.4	9815	35.5/31.8/32.7	9796
MRPC	Classification	2	3668	32.6/67.4	408	31.6/68.4	1725
QNLI	Classification	2	104743	50.0/50.0	5463	49.5/50.5	5463
QQP	Classification	2	363846	63.1/36.9	40430	63.2/36.8	39095
RTE	Classification	2	2490	50.2/49.8	277	52.7/47.3	3000
SST-2	Classification	2	67349	44.2/55.8	872	49.1/50.9	1821
STS-B	Regression	-	5749	-	1500	-	1379

Table 4.7: Statistics for all datasets inside the GLUE benchmark suite. “%Examples/Class” shows the percentage of examples for each label class. The labels of the test set are hidden, thus we do not show statistics about their class distribution.

for paraphrasing [Dolan and Brockett, 2005]; STS-B for sentence similarity [Cer et al., 2017]; SST-2, to test models in Sentiment Analysis [Socher et al., 2013].

We do not include WNLI in our experiments because it is hard to beat even a trivial majority classifier [Devlin et al., 2019]. As for MNLI, we use the *matched* version, usually indicated with MNLI-m. Statistics for each dataset are reported in Table 4.7. A more fine-grained analysis of each dataset contained in GLUE is given in Appendix B.1. Notice that recently, a new version of this benchmark called SuperGLUE [Wang et al., 2019a] has been released, and we encourage its usage for future research.

## Chapter 5

# Alternative Efficient Objectives

The first part of this work focuses on Transformers pre-training, which is a core part of building Artificial General Intelligence. Transformers pre-training is a task that requires (i) costly and large computational infrastructures and a lot of energy, (ii) high-level skills to manage distributed training and peer-to-peer efficient communications, (iii) collection and cleaning/filtering of large amounts of raw data and (iv) well-tested code implementations to avoid nodes failure [Brown et al., 2020; Strubell et al., 2019].

In this Chapter, we address the first issue mentioned before by proposing new pre-training tasks and objectives that require shorter pre-training to reach the same level of final performance. The other issues are generally addressed by large organizations which develop high-end frameworks to help researchers and developers. We mention: (i) PyTorch [Paszke et al., 2019] and Tensorflow [Abadi et al., 2015], which allow for fast ANN development and training through automatic back-propagation; (ii) DeepSpeed [Rasley et al., 2020] and ColossalAI [Bian et al., 2021] that increment training speed and reduce inference time by sharding model’s weights and through efficient communication protocols between nodes in distributed settings; (iii) Common-Crawl [Wang, 2023] and The Pile [Gao et al., 2020], which are large sources of cleaned text freely available and (iv) PyTorch-Lightning [Falcon and The PyTorch Lightning team, 2019] and HuggingFace [Wolf et al., 2020], which help developers by providing frameworks to abstract from the underlying hardware and repositories with thousands of ready-to-be-used models.

Many recent works focused on improving pre-training efficiency through the implementation of innovative objectives or by improving the model architecture [Lan et al., 2020; Sanh et al., 2020; Turc et al., 2019]. For example, the authors of ELECTRA trained a BERT model as a discriminator instead of a generator [Clark et al., 2020]. Instead of masking a random subset of input tokens, ELECTRA exploits a small Language Model trained with MLM to create challenging replacements. Then, the discriminator is tasked to predict which tokens are original and which are fakes introduced by the smaller generator network.

ELECTRA introduces many innovations, such as more efficient training and the missing discrepancy between pre-training and fine-tuning because no special [MASK] token is used in the discriminator training. However, it is not clear whether the improvements of ELECTRA over BERT derive more from the generator-discriminator architecture or the longer and more costly pre-training.

We provide an extensive description of the works related to this research in Section 3.3.1. Notice that they are orthogonal to our proposed methodology, which emphasizes the efficiency of the pre-training objective and the classification head size. Therefore, they can be combined with our alternative pre-training objectives. Additionally, works related to the models and the tasks we use in this Chapter can be found in Section 3.1 and 3.2, respectively.

We study a batch of new pre-training strategies designed to reduce training costs while

maintaining the same level of final performance. We provide a thoughtful study of both theoretical cost and experimental training time since the two do not always align well due to underlying hardware optimization techniques and different levels of memory usage. Our contributions can be summarized as the following three self-supervised objectives [Di Liello et al., 2022a]:

- **Random Token Substitution (RTS)**: an efficient alternative to ELECTRA that drops the expensive generator network and works by randomly replacing some tokens with others and by tasking the model at recognizing which are originals and which are not;
- **Cluster-based Random Token Substitution (C-RTS)**: a technique that improves over RTS using previous predictions to find more challenging replacements utilizing a simple and efficient statistical approach;
- **Swapped Language Modeling (SLM)**: an alternative to BERT’s MLM which randomly replaces tokens into others and asks the model to predict the original values.

The next Sections will cover a description of our proposed objectives and a comparison of the computational complexity with the SOTA pre-training tasks for Transformers. Then, we provide details about the experiment setting we used to evaluate our claims and results obtained over many different benchmarks. We also report negative findings in Appendix C.2 to help future research.

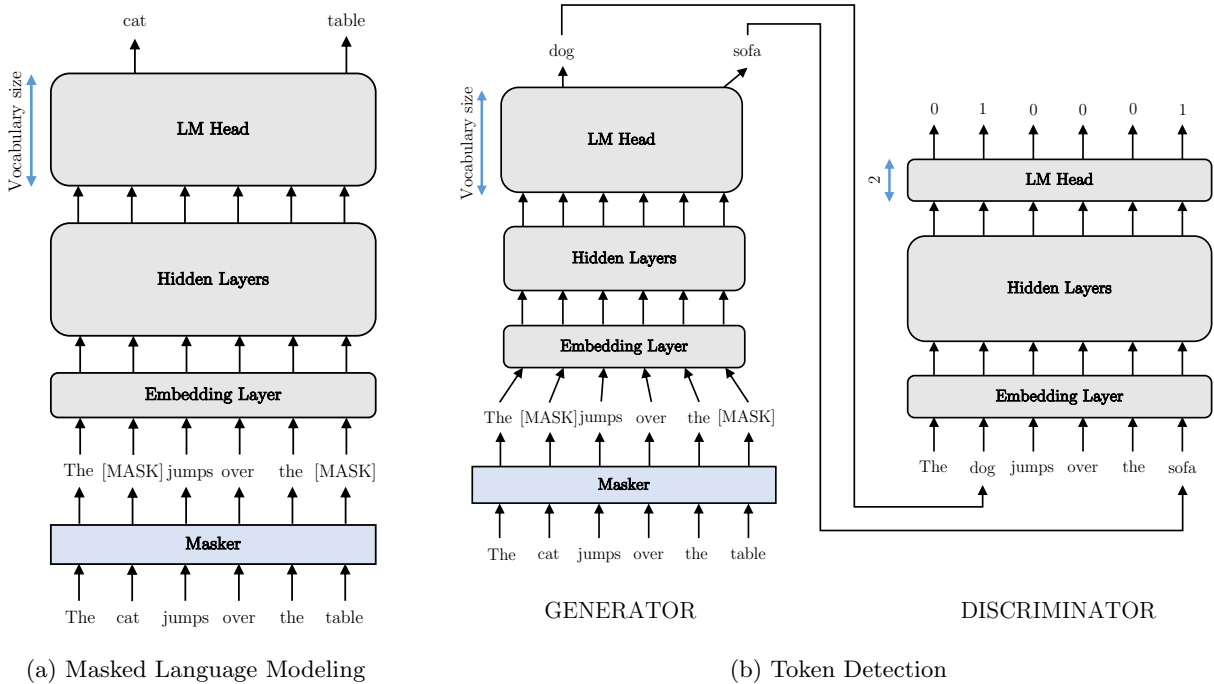


Figure 5.1: BERT’s MLM and ELECTRA’s TD objectives examples.

## 5.1 Alternative Objectives

In this Section, we illustrate our proposed objectives, highlighting the advantages compared to state-of-the-art methods.

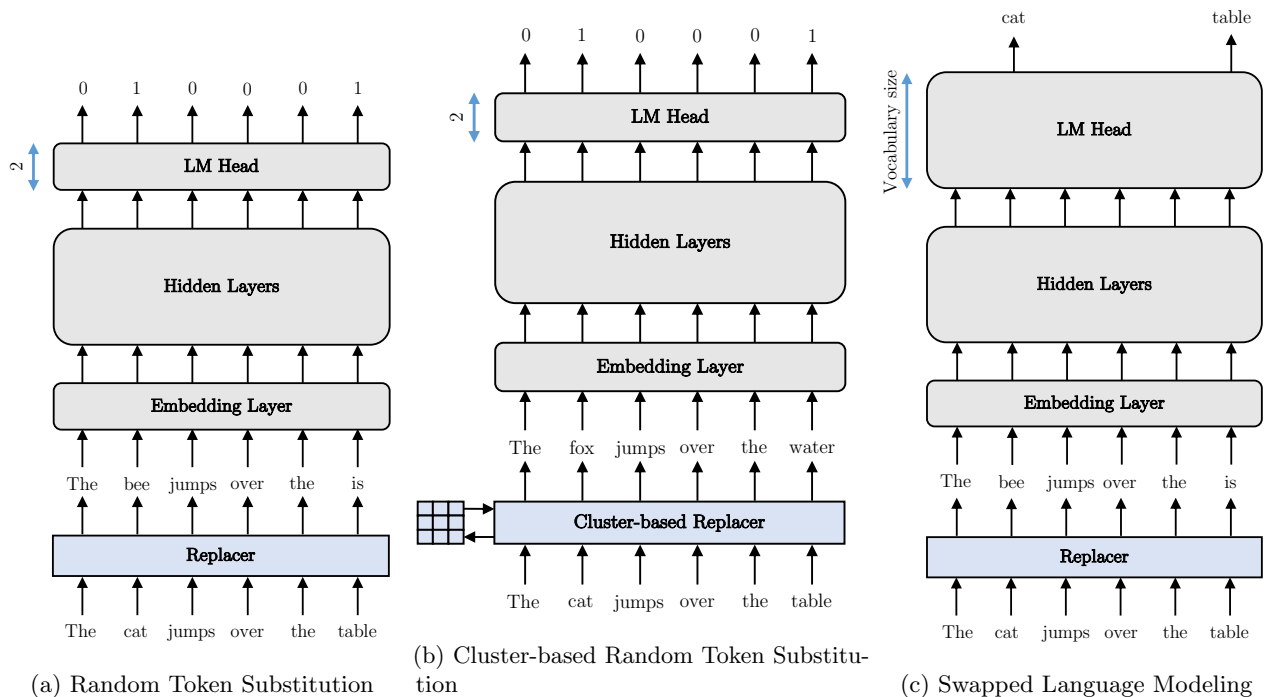


Figure 5.2: Examples for RTS, C-RTS and SLM objectives.

### 5.1.1 Random Token Substitution

RTS is a variant of the ELECTRA architecture in which the expensive generator network is substituted with a straightforward random sampling approach. Examples of how ELECTRA and RTS work are depicted in Figure 5.1b and 5.2a, respectively. RTS uses a very small classification head on top of the Transformer for binary classification, which reduces computational load and memory usage, allowing for larger batch sizes.

In RTS, we randomly select the 15% of the input tokens and we change them into others sampled with uniform distribution from the vocabulary. Then, the model is tasked to recognize which tokens are original and which have been replaced with others.

**Advantages** RTS necessitates a minimal classification head, and predictions are not carried out across the entire vocabulary. Some might argue that removing the MLM head would not result in significant improvements since the embedding layer, which shares weights with the language modeling classification head, still acts as a bottleneck. However, this is not accurate, as the embedding layer functions by indexing the weights matrix to extract only the representation corresponding to each input token. In contrast, the language modeling head computes probabilities for all tokens in the vocabulary, making it computationally much more expensive by several orders of magnitude.

Another important advantage of RTS is that there is no pre-training/fine-tuning discrepancy because the [MASK] token is not used in both pre-training and fine-tuning [Clark et al., 2020].

**Drawbacks** RTS works well with Transformer models containing up to 100M parameters, however, we do not suggest its usage for larger models because it may lead to overfitting, as shown in the successive sections. When the model size increases, the RTS task can be solved with very high accuracies because most replaced tokens are easy to discover, leading to weak error signals and lower model quality.

### 5.1.2 Cluster-based Random Token Substitution

C-RTS is an improvement of RTS that targets the overfitting issue by exploiting the history of previous predictions to find more challenging replacements. The main idea of C-RTS is to model the miss-classification probability of the model for each token independently through a simple static approach.

More formally, given a Transformer model  $\mathcal{M}$  and sequence of tokens  $[t_1, t_2, \dots, t_L]$  in input,  $\mathcal{M}$  outputs a sequence of predictions  $\mathbf{p}$  for the labels  $\mathbf{y}$ , where  $y_i = 0$  means that  $t_i$  is original and  $y_i = 1$  indicates that  $t_i$  was replaced with some other token  $t'_i$  ( $t_i \rightarrow t'_i$ ). Our objective is to maximize:

$$P(p_i = 0 \mid t_i \rightarrow t'_i) \quad (5.1)$$

since we want to find challenging replacements that are hard to be detected by  $\mathcal{M}$ .

Equation 5.1 could be estimated by counting the number of successes and failures of the model  $\mathcal{M}$  at recognizing false input tokens in the previous training steps. An important difference between C-RTS and ELECTRA is that we do statistics over individual tokens, while ELECTRA’s generator uses the whole input context to find challenging alternatives.

**Clustering** Since storing a matrix containing counts for every pair of tokens would lead to a very large memory footprint (vocabulary sizes range from 30K [Devlin et al., 2019] entries to over 120K [He et al., 2020]), we split the vocabulary into  $n$  clusters with similar size. We perform the clustering based on token similarity. However, since we do not have an already-trained embedding layer yet, we train a *word2vec* [Mikolov et al., 2013] model to obtain vectorial representations. We run word2vec on the same data that we use for pre-training, to avoid any performance improvement stemming from the usage of additional text.

In particular, we configure word2vec to use a context size of 2 words on both sides of the target token and we set the embedding size to 300. We use the PyTorch implementation of word2vec [Paszke et al., 2019] and the whole training takes less than 10 minutes on our machine. Given the negligible overhead of this step when compared to the pre-training time of Transformers, we omit those 10 minutes from the final training-time comparisons.

Once word embeddings are computed, we used the K-means algorithm [Lloyd, 1982] to group tokens into clusters. We implemented the script by exploiting the implementation of K-means in the *scikit-learn* [Pedregosa et al., 2011] library. To choose the best number of clusters  $n$ , we first run the clustering algorithm with 20 random restarts for each value of  $n \in \{30, 100, 300, 1000\}$ . This experiment took about 20 minutes on the machine described in Section 1.2. Then, we search for the best number of clusters by running a short pre-training of 5000 steps and by selecting the value that gives the smallest accuracy of C-RTS on the pre-training dev set.

**Statistic-based replacement** Once we partition the tokens of the vocabulary  $\mathbb{V}$  into clusters  $\{\mathbb{C}_1, \dots, \mathbb{C}_n\}$ , we estimate the probability  $P(p_i = 0 \mid t_i \rightarrow t'_i)$  with the following methodology. First, we use a matrix  $\mathbf{F} \in \mathbb{Z}^{n \times n}$  to count the number of successes and failures of the model. At each training step, if  $t_i \in \mathbb{C}_a \rightarrow t'_i \in \mathbb{C}_b$ , then  $\mathbf{F}_{a,b}$  is increased by 1 if  $p_i = 0$  and decreased by 1 otherwise.

As in RTS, we select 15% of the input tokens for replacement. If token  $t_i \in \mathbb{C}_a$  is chosen, we maximize the miss-classification probability by estimating  $P(p_i = 0 \mid t_i \in \mathbb{C}_a \rightarrow t'_i \in \mathbb{C}_b)$  with  $\mathbf{F}$ . Probability of replacing  $t_i$  to  $t'_i$  is computed with:

$$P(t_i \in \mathbb{C}_a \rightarrow t'_i \in \mathbb{C}_b) = P(t_i) P(t'_i \mid \mathbb{C}_b) P(\mathbb{C}_b \mid \mathbb{C}_a) \quad (5.2)$$

where:

- $P(t_i)$  is 15% as in ELECTRA’s TD or BERT’s MLM;
- $P(t'_i | \mathbb{C}_b)$  is equal to  $\frac{1}{|\mathbb{C}_b|}$  because we pick a target token  $t'_i$  with uniform probability from  $\mathbb{C}_b$ ;
- $P(\mathbb{C}_b | \mathbb{C}_a)$  is computed taking advantage of history about previous predictions, which are stored in the matrix  $\mathbf{F}$ .

In particular, the latter is computed by selecting the  $a$ -th row of  $\mathbf{F}$  and by considering it as a multinomial distribution over target clusters. To obtain values from  $\mathbf{F}_a$  interpretable as probabilities, we first apply the *min – max* normalization:

$$\mathbf{F}'_a = \frac{\mathbf{F}_a - \min(\mathbf{F}_a)}{\max(\mathbf{F}_a) - \min(\mathbf{F}_a)} \quad (5.3)$$

to restrict the values into the  $[0, 1]$  interval. Then, we use  $\gamma$ -*softmax* function to smooth the distribution, which allows us to control the temperature:

$$\mathbf{F}''_a = \frac{e^{\mathbf{F}'_{a,b}/\gamma}}{\sum_{k=1}^n e^{\mathbf{F}'_{a,k}/\gamma}} \quad (5.4)$$

The final vector  $\mathbf{F}''_a$  is then used as multinomial distribution to estimate  $P(\mathbb{C}_b | \mathbb{C}_a)$ . An example of C-RTS architecture is depicted in Figure 5.2b.

### 5.1.3 Swapped Language Modeling

The SLM objective is computationally equivalent to MLM, yet it diverges in its approach by omitting the use of a special [MASK] token, resulting in superior performance. In the MLM task, 15% of the input tokens are selected and then treated in three different ways: (i) in 80% of the cases, they are substituted with the special [MASK] token, (ii) 10% of the time they are replaced by other tokens, and (iii) 10% of the time they remain unchanged. SLM can be viewed as a variation of BERT’s MLM, where tokens are consistently transformed into others. The remaining aspects of the model’s architecture and loss functions remain unchanged. Figure 5.2c provides an illustrative example of SLM.

## 5.2 Complexity Analysis

Objective	MLM	RTS	C-RTS	SLM	MLM + TD
LM head complexity	$O( \mathbb{V}  \cdot d)$	$O(d)$	$O(d)$	$O( \mathbb{V}  \cdot d)$	$O( \mathbb{V}  \cdot d) + O(d)$

Table 5.1: Time complexity in number of FLOPS of the classification head employed by different objectives.

In this Section, we analyze the space and time complexity of state-of-the-art objectives for Transformers used in the literature and of our proposals. We focus especially on the size of the language model head because the remaining part of the Transformer is shared across all architectures (ours comprehended). For more details about each objective, refer to Section 2.2.2.1. Table 5.1 summarizes the complexity of the language modeling head for MLM, TD and our proposed objectives.

**Masked Language Modeling & Swapper Language Modeling** MLM and SLM predict the original value or masked or replaced tokens. Figure 5.1a shows an example of input data and targets for MLM while in Figure 5.2c we depict an example of the SLM objective. Both objectives require a classification head that spans the whole vocabulary  $\mathbb{V}$ . Given the final hidden state  $\mathbf{h}_i \in \mathbb{R}^d$  in output from the Transformer stack for each token  $t_i$ , the predictions for each class are performed by multiplying  $\mathbf{h}_i$  with a learnable matrix  $\mathbf{W}_{LM} \in \mathbb{R}^{|\mathbb{V}| \times d}$ :

$$\mathbf{p}_i = \text{softmax}(\mathbf{W}_{LM}\mathbf{h}_i) \quad (5.5)$$

Thus, the computational complexity of the language modeling head for each token is  $O(|\mathbb{V}| \times d)$ . We ignore the complexity of the softmax because it is linear in the hidden size  $d$ .

**Token Detection & Random Token Detection** ELECTRA’s architecture is composed of two blocks: a generator trained with MLM and a discriminator trained with TD. The complexity of MLM is studied in the previous Section. Token Detection is a binary prediction task, thus the language modeling head is composed of a small matrix  $\mathbf{W}_{LM} \in \mathbb{R}^{2 \times d}$  followed by a softmax function. We can derive that the complexity of TD is  $O(d)$ . RTS and C-RTS have the same complexity as ELECTRA’s discriminator since they are similar but the generator is substituted with our algorithms for selecting the token replacements.

## 5.3 Experiments

In this Section, we describe the testing environment we use to measure the accuracy of the reference and proposed objectives. First, we provide details about the pre-training setting, the datasets and the best hyper-parameters we found for each objective. Then, fine-tuning details such as datasets, training parameters and the hardware used to run the experiments are given. Finally, we present results as tables and plots, including comparisons against models released by the authors of BERT when available.

### 5.3.1 Architectures

Every objective has been tested by applying it to the same Transformer architecture, equivalent to that of BERT [Devlin et al., 2019] or RoBERTa [Liu et al., 2019b]. For every model and architecture, we use an uncased vocabulary, which means WordPiece converts text to lowercase before tokenizing. We experiment with models of two different sizes: Base and Small.

**Base** The Base architecture is equivalent to that of BERT<sub>Base</sub>, which has a vocabulary  $\mathbb{V}$  containing 30522 tokens based on WordPiece (see Paragraph A.3.2), a hidden size  $d$  of 768, an intermediate size equal to 3072, 512 positional embeddings, 12 attention heads and 12 layers.

**Small** The Small architecture we employ is similar to that of ELECTRA<sub>Small</sub> [Clark et al., 2020], which in prior experiments resulted to be better than BERT<sub>Small</sub> [Turc et al., 2019]. ELECTRA demonstrated that it is more effective to reduce the *width* (the hidden size) of the Transformer instead of the *height* (the number of layers). We use a vocabulary  $\mathbb{V}$  containing 30522 tokens based again on WordPiece, a hidden size  $d$  of 256, an intermediate size of 1024, 512 positional embeddings 12 attention layers and 4 attention heads.



### 5.3.2 Pre-Training

In order to provide a meaningful and fair comparison between the objectives, we use the same pre-training dataset for all experiments. In more detail, we train all models on the same dump of the English Wikipedia and on the BookCorpus, which are cleaned before training from HTML tags and tables, from lists, indexes and everything not interpretable as raw text. Notice that the original BookCorpus used by Devlin et al. [2019] is not available anymore, and we do not know whether the version we use is of the same quality. For more details about the pre-training corpora, see Section 4.1.

We copied the training set from Devlin et al. [2019], in which the authors train the models for 900K steps with a maximum sequence length of 128 tokens and 100K additional steps with a sequence length of 512. This allows us to save a lot of pre-training time without a significant loss in performance. The increased efficiency derives from the fact that the attention module complexity is quadratic in the sequence length. Thus, for the first 900K steps, the model is theoretically 4 times faster<sup>1</sup> than when training with the full sequence length. However, larger and heavy-trained models such as RoBERTa [Liu et al., 2019b] showed that attention weights and positional embeddings quality may suffer from pre-trainings with shorter sequence lengths due to the lack of computation of long-range dependencies between tokens. Nevertheless, we pre-train with the setting described by Devlin et al. [2019] because the comparison is still fair and because of our limited computational budget.

Regarding the training hyper-parameters, we apply a weight decay of 0.01 to every parameter apart from those in biases and LayerNorms modules and we set  $\epsilon = 10^{-8}$ ,  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  in the optimizer. We use a triangular learning rate scheduler, with a peak of  $1 \cdot 10^{-4}$  (Base) or  $2 \cdot 10^{-4}$  (Small) and 10K warmup steps. The probability in drop-out layers is set to 0.1. Finally, we use a batch size of 256 examples for Base models and 1024 for the Small versions, for optimal utilization of the GPU memory.

We train also a Base and a Small ELECTRA architecture for comparison. Notice that the ELECTRA requires more FLOPS because of the presence of the additional generator network. FLOPS are a general and architecture-independent indicator of model requirements that measure the total number of floating point operations required for training. To address this issue, and train all models with the same compute budget, we reduce the total number of steps of ELECTRA from 900K@128 + 100K@512 to 689K@128 + 77K@512 for both Small and Base models. The generator size is set to  $\frac{1}{3}$  and  $\frac{1}{4}$  for Base and Small models as in the original

<sup>1</sup>Actual speed-up is generally slightly lower because of the embedding layer and the language modeling head whose complexities are linear in the sequence length.

Models	FLOPS	Time	Memory	# Parameters		Objective
				Total	LM Head	
BERT <sub>Base</sub> + MLM	$1.61 \cdot 10^{19}$	3d 14h	13.2GiB	109M	23M	MLM (1.0)
BERT <sub>Base</sub> + RTS	$1.54 \cdot 10^{19}$	2d 22h	9.9GiB	109M	1536	RTS (50.0)
BERT <sub>Base</sub> + C-RTS	$1.54 \cdot 10^{19}$	2d 22h	9.9GiB	109M	1536	C-RTS (50.0)
BERT <sub>Base</sub> + SLM	$1.61 \cdot 10^{19}$	3d 14h	13.2GiB	109M	23M	SLM (1.0)
ELECTRA <sub>Base</sub> + MLM/TD	$1.98 \cdot 10^{19}$	3d 18h	18.0GiB	143M	23M + 1536	MLM (1.0) + TD (50.0)

Table 5.2: FLOPS, time, memory, number of parameters and objectives to pre-train Base models on two A100 GPUs of our machine. Notice that the number of parameters of MLM/SLM and RTS/C-RTS is the same because the weights of the embedding layer and language modeling head are shared. However, the memory usage of MLM/SLM is higher because gradients over the language modeling head are much larger in memory than gradients over the embedding layer. Notice also that FLOPS and training time of C-RTS differ slightly from those reported in Di Liello et al. [2022a] because we offloaded computations over clusters to the CPU. Finally, we report the objectives along with their weight when computing the total loss.

Models	FLOPS	Time	Memory	# Parameters		Objective
				Total	LM Head	
BERT <sub>Small</sub> + MLM	$3.04 \cdot 10^{18}$	1d 7h	12.5GiB	13.5M	3.9M	MLM (1.0)
BERT <sub>Small</sub> + RTS	$2.73 \cdot 10^{18}$	17h	6.3GiB	13.5M	512	RTS (20.0)
BERT <sub>Small</sub> + C-RTS	$2.73 \cdot 10^{18}$	17h	6.3GiB	13.5M	512	C-RTS (20.0)
BERT <sub>Small</sub> + SLM	$3.04 \cdot 10^{18}$	1d 7h	12.5GiB	13.5M	3.9M	SLM (1.0)
ELECTRA <sub>Small</sub> + MLM/TD	$4.87 \cdot 10^{18}$	1d 16h	17.4GiB	18.2M	3.9M + 512	MLM (1.0) + TD (20.0)

Table 5.3: FLOPS, time, memory and number of parameters required to pre-train Small models on the same hardware. For more details, see Table 5.2.

work, respectively<sup>2</sup>. Inspired by Clark et al. [2020], we use FLOPS as a measure of compute requirements. However, actual training time is sometimes misaligned with FLOPS because of hardware acceleration such as NVIDIA Tensor Cores, mixed-precision training, low precision *float32* matrix multiplication and other optimization techniques. Tables 5.2 and 5.3 summarize the training time and FLOPS required by the Base and Small models.

Each experiment is executed on two GPUs of our machine (which is described in Section 1.2), using mixed-precision training.

### 5.3.3 Fine-Tuning

After pre-training every model on the same amount of data and with the same hyper-parameters, we evaluate them on different benchmarks.

**GLUE** First, for a general performance evaluation, we use the GLUE benchmark suite, which contains datasets for many different tasks such as Questions Similarity and Natural Language Inference. For more details about the GLUE dataset, refer to Section 4.3. As in Devlin et al. [2019] and Clark et al. [2020], we omit WNLI from the GLUE datasets because even a trivial majority classifier is better than most Base or Small pre-trained models.

As a starting point, we copy the fine-tuning hyper-parameters of Liu et al. [2019b] for every GLUE task. After that, we evaluate affine configurations to ensure good final accuracy of the models. The final hyper-parameters are described in Table 5.4. Other parameters have been kept constant for all experiments: no weight decay, triangular learning rate with a warmup phase of 10% of the total fine-tuning steps and a maximum sequence length of 128. Regarding the testing, we take the best model based on the evaluation metric of each GLUE dataset. Moreover, we apply early stopping with a patience of 5 and we validate at the end of each training epoch.

Regarding the evaluation metrics, we use those proposed by the original authors of GLUE [Wang et al., 2018] and similar works [Clark et al., 2020; Devlin et al., 2019; Liu et al., 2019b]. In particular, we use accuracy for all tasks apart from CoLA and STS-B, which are evaluated with Matthew and Spearman correlation coefficients, respectively.

<sup>2</sup>Applies to the hidden size, the intermediate size and the number of attention heads.

Hyper-Parameter	CoLA	MNLI-m	MRPC	QNLI	QQP	RTE	SST-2	STS-B
Batch Size	16	32	16	32	32	16	32	16
Learning Rate	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$2 \cdot 10^{-5}$
Max epochs	20	8	10	10	8	40	10	40

Table 5.4: GLUE fine-tuning hyper-parameters.

Finally, we repeat every experiment with 5 random seeds (which influence the classification head initialization and the dataset shuffling), and we send the results of the best-performing checkpoints on the dev set to the GLUE Leaderboard for evaluation on the test set<sup>3</sup>.

**ASNQ** We use ASNQ to evaluate models performance on large Answer Selection datasets. ASNQ contains more than 20M question-answer pairs in the training set, making it the largest dataset for AS2 to date. More details about ASNQ are given in Section 4.2.1. We use a batch size of 2048 examples for each model and a learning rate equal to  $1 \cdot 10^{-5}$ . The values are obtained with a greedy search: we try all the combinations of batch sizes in  $\{128, 256, 512, 1024, 2048, 4096\}$  and learning rates in  $\{2 \cdot 10^{-6}, 5 \cdot 10^{-6}, 1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$ . Since most of the question-answer pairs in the dataset are shorter than 128 tokens, we set the maximum sequence length to this value. We allow models to train up to 10 epochs, and we do early stopping on the development set if MAP does not improve for 4 validations in a row. Other training parameters are the same of the fine-tuning on GLUE. Experiments are repeated 3 times with different initialization seeds to show average and standard deviation in the results.

We measure performance using Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). The first is a soft metric that measures the Average Precision at different levels of recall, while the second is a hard metric that measures only the reciprocal rank of the positive answer with the highest score. For more details about Answer Selection ranking, see Section 2.1.1.2.

**WikiQA, TREC-QA and Transfer-Learning** As a last step, we evaluate the pre-trained models on two small datasets for Answer Sentence Selection: WikiQA and TREC-QA. We first fine-tune the pre-trained models directly on those two datasets. Then, as in TANDA [Garg et al., 2019], we measure performance in Transfer-Learning by taking the checkpoints of the models after a transfer step on ASNQ and fine-tuning them on WikiQA and TREC-QA. We indicate the two transfer steps and subsequent adaptations to the final task with  $\text{ASNQ} \rightarrow \text{WikiQA}$  and  $\text{ASNQ} \rightarrow \text{TREC-QA}$ .

Fine-tuning hyper-parameters are selected again with greedy search. For the batch size, we try values in  $\{32, 64, 128\}$ , while for the learning rate, we search in  $\{2 \cdot 10^{-6}, 5 \cdot 10^{-6}, 1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$ . The maximum sequence length is fixed to 128 tokens while the number of epochs is limited to 40. Again, we use early stopping to spare resources, with a patience of 5.

## 5.4 Results

### 5.4.1 Pre-Training

We start the comparison of the results with a discussion about pre-training times. For each Base model, Table 5.2 shows training time and the total number of FLOPS. RTS and C-RTS use 20% less training time to complete the same number of pre-training steps. The difference in terms of FLOPS is instead smaller, around 5%. Notice that RTS and C-RTS use 45% GPU memory, thus training times could be reduced even further by doubling the batch size. However, we keep the same training parameters for all models to provide a fair comparison.

With Small models, the gains in training time are even larger. RTS and C-RTS allow to reduce pre-training time by 45% and FLOPS by 10%. Gaps are larger than Base models because the embedding layer and the language modeling head are bound to the vocabulary size  $|\mathbb{V}|$ . We do not shrink the vocabulary size because recent works show significant drops in performance

<sup>3</sup><https://gluebenchmark.com/leaderboard/>

Model	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	AVG
BERT <sub>Base</sub> + MLM/NSP [Devlin et al., 2019]	57.6 (1.8)	84.3 (0.4)	82.3 (1.3)	91.0 (0.7)	91.0 (0.2)	68.9 (1.4)	92.6 (0.1)	89.1 (0.3)	82.1
ELECTRA <sub>Base</sub> + MLM/TD	63.4 (1.3)	83.7 (0.2)	87.2 (0.8)	90.4 (0.1)	91.2 (0.1)	74.6 (1.4)	91.4 (0.4)	88.5 (0.2)	83.8
BERT <sub>Base</sub> + MLM	58.1 (1.0)	<b>83.4</b> (0.2)	87.5 (0.5)	<b>90.2</b> (0.3)	90.9 (0.1)	67.4 (1.2)	<b>92.2</b> (0.3)	<b>87.8</b> (0.3)	82.2
<b>BERT<sub>Base</sub> + RTS</b>	58.1 (1.1)	82.7 (0.2)	<b>87.6</b> (1.0)	89.4 (0.3)	90.9 (0.1)	68.5 (1.4)	91.5 (0.3)	86.6 (0.4)	81.9
<b>BERT<sub>Base</sub> + C-RTS</b>	57.4 (0.7)	82.0 (0.3)	84.2 (0.4)	89.6 (0.2)	90.6 (0.1)	66.6 (2.4)	91.5 (0.2)	87.0 (0.2)	81.1
<b>BERT<sub>Base</sub> + SLM</b>	<u>59.6</u> (1.0)	<b>83.4</b> (0.2)	87.5 (0.4)	89.9 (0.2)	<b>91.0</b> (0.1)	<u>69.2</u> (1.2)	92.1 (0.1)	87.6 (0.3)	<b>82.5</b>

Table 5.5: Results on GLUE development set for Base models. We do not employ ensemble models. We measure performance with Accuracy on all tasks apart from CoLA and STS-B, in which we use Matthews correlation coefficient and Spearman’s rank correlation coefficient, respectively. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over BERT<sub>Base</sub> + MLM.

Model	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	AVG
BERT <sub>Base</sub> + MLM/NSP [Devlin et al., 2019]	53.0	82.7	82.8	89.6	88.2	62.8	91.2	80.6	78.9
ELECTRA <sub>Base</sub> + MLM/TD	60.6	83.6	84.1	90.2	89.0	69.1	92.4	85.5	81.8
BERT <sub>Base</sub> + MLM	53.7	<b>83.6</b>	82.6	<b>89.9</b>	<b>89.1</b>	63.1	<b>92.3</b>	83.6	79.7
<b>BERT<sub>Base</sub> + RTS</b>	<b>57.3</b>	82.6	81.3	89.3	88.9	<b>66.2</b>	91.7	82.2	79.9
<b>BERT<sub>Base</sub> + C-RTS</b>	<b>57.3</b>	82.8	81.9	89.4	88.6	60.9	91.0	82.2	79.3
<b>BERT<sub>Base</sub> + SLM</b>	57.0	83.3	<b>83.1</b>	89.7	88.9	65.0	<b>92.3</b>	<b>83.8</b>	<b>80.4</b>

Table 5.6: Results on the GLUE test set for Base models. For more details about the metrics, see Table 5.5. For the evaluation on the GLUE leaderboard, we select the best models on the development set.

when reducing it [Turc et al., 2019] and because BERT’s vocabulary size is already one of the smallest among recently released pre-trained models. RTS and C-RTS have about  $\frac{1}{3}$  of the memory footprint of MLM or SLM. As with Base models, we could have increased the batch size to speed up the training of RTS and C-RTS even further.

## 5.4.2 Fine-Tuning

In this Section, we present the comparison on different tasks of the models we pre-trained. The statistical significance comparison in the tables is performed compared to BERT<sub>Base</sub> + MLM and BERT<sub>Small</sub> + MLM. Results of the original BERT<sub>Base</sub> + MLM/NSP [Devlin et al., 2019] derive from our fine-tuning on the various tasks in the same setting as the other models, because the original authors do best model selection on the development set.

### 5.4.2.1 Base Models

Results of Base models over the GLUE benchmarks are presented in Table 5.5 for the development set (as in Clark et al. [2020]) and in Table 5.6 for the test set. Results on the development set allow the comparison of the standard deviation in results. This is not possible on the test set because of the limited number of submissions allowed by the GLUE Leaderboard.

On GLUE, the models pre-trained with RTS and C-RTS objectives obtain comparable results with all the other techniques, except for ELECTRA-TD. However, while RTS and C-RTS require about 20% less training time than a similar MLM-based model, ELECTRA-TD requires more time to complete the pre-training and consumes much more memory due to the presence of the additional generator network, thus reducing scalability. In particular, ELECTRA uses 36% more memory than a similar MLM- or SLM-based model and almost doubles the memory required by RTS or C-RTS. The SLM objective we designed outperforms MLM in most tasks while being trained with the same computational budget. In particular, SLM achieves an average GLUE score higher than our MLM-based BERT by 0.7 points and also higher than the original BERT,

Model	ASNQ		WikiQA		TREC-QA	
	MAP	MRR	MAP	MRR	MAP	MRR
BERT <sub>Base</sub> + MLM/NSP [Devlin et al., 2019]	66.8 (0.1)	73.2 (0.2)	82.8 (0.9)	84.2 (1.0)	87.2 (0.9)	92.9 (1.0)
ELECTRA <sub>Base</sub> + MLM/TD	64.9 (0.3)	71.7 (0.4)	81.8 (1.6)	83.2 (1.6)	86.8 (1.4)	92.2 (1.5)
BERT <sub>Base</sub> + MLM	65.5 (0.2)	72.2 (0.3)	79.9 (1.1)	81.2 (1.2)	86.4 (0.7)	91.5 (0.7)
<b>BERT<sub>Base</sub> + RTS</b>	64.6 (0.1)	71.1 (0.2)	78.5 (2.5)	80.1 (2.5)	86.6 (1.5)	91.8 (1.5)
<b>BERT<sub>Base</sub> + C-RTS</b>	64.7 (0.1)	71.5 (0.1)	79.0 (1.9)	80.6 (1.6)	87.0 (0.8)	91.8 (1.1)
<b>BERT<sub>Base</sub> + SLM</b>	<b>65.8 (0.3)</b>	<b>72.7 (0.3)</b>	<b>80.2 (1.5)</b>	<b>81.7 (1.6)</b>	<b>87.3 (1.3)</b>	<b>92.1 (1.5)</b>

Table 5.7: Results on ASNQ, WikiQA and TREC-QA for Base models. The results on ASNQ are on the dev. set. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over BERT<sub>Base</sub> + MLM.

Model	ASNQ → WikiQA		ASNQ → TREC-QA	
	MAP	MRR	MAP	MRR
BERT <sub>Base</sub> + MLM/NSP [Devlin et al., 2019]	87.9 (0.4)	89.3 (0.4)	89.1 (0.4)	93.4 (0.5)
ELECTRA <sub>Base</sub> + MLM/TD	88.4 (0.4)	89.8 (0.4)	88.9 (0.3)	92.0 (0.5)
BERT <sub>Base</sub> + MLM	<b>88.9 (1.0)</b>	<b>90.2 (1.0)</b>	87.6 (0.9)	90.6 (1.1)
<b>BERT<sub>Base</sub> + RTS</b>	87.9 (0.5)	89.3 (0.5)	88.5 (0.5)	<b>93.4 (0.4)</b>
<b>BERT<sub>Base</sub> + C-RTS</b>	87.1 (0.6)	88.3 (0.6)	<b>88.7 (0.4)</b>	<b>92.9 (0.7)</b>
<b>BERT<sub>Base</sub> + SLM</b>	87.7 (0.8)	89.3 (0.7)	87.9 (0.6)	91.0 (0.6)

Table 5.8: Results on WikiQA and TREC-QA for Base models after the transfer step on ASNQ.

which was trained with MLM and NSP (+1.5). SLM shows also that the pre-training/fine-tuning discrepancy is a major issue of MLM-based models such as BERT or RoBERTa.

We compare also with the original BERT model released by Devlin et al. [2019], which uses an additional sentence-level objective to improve performance. Notice that even if Liu et al. [2019b] demonstrated that dropping the NSP objective does not affect final accuracy, we believe this claim is valid only with very long pre-trainings. RoBERTa [Liu et al., 2019b] is trained for a full day on 1024 GPUs, which is about 53 more times the computational requirements of BERT [Devlin et al., 2019]. Thus, with models pre-trained with an academic budget and over tasks where the model has to work on two input sequences, such as AS2, our results show that NSP is still beneficial because it improves the representation of the CLS token, which is used for classification.

We do not compare against the ELECTRA models released by Clark et al. [2020] because they were pre-trained on 20 times the total tokens seen by our models.

Regarding the results in Answer Sentence Selection, which are provided in Table 5.7 and 5.8, RTS and C-RTS show comparable performance with MLM and SLM in most tasks. On WikiQA, results are slightly lower on average, while they are superior on TREC-QA. On ASNQ, RTS and C-RTS are below MLM and SLM by less than 1 point in MAP, while being pre-trained for 20% less time. With the transfer step over ASNQ before the final fine-tuning on WikiQA and TREC-QA, relative differences between objectives do not change. Results are higher on average because the models are accustomed to performing Answer Selection on a large dataset before being fine-tuned. Notice also that the transfer step on ASNQ benefits more WikiQA than TREC-QA. We argue that the reason is that both ASNQ and WikiQA are created from documents sourced from Wikipedia, while TREC-QA is based on human-generated answers. On the other hand, SLM shows his superiority over MLM in 4 tasks out of 5. SLM outperforms a similar MLM-based model by 1.9 points in MAP on WikiQA, by 0.9 on TREC-QA, by 0.3 on ASNQ → WikiQA and by 0.3 on ASNQ, while requiring the same computational budget to pre-train. These results highlight again the advantage of removing the special [MASK] token from the pre-training.

Model	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	AVG
ELECTRA <sub>Small</sub> + MLM/TD	51.2 (0.1)	76.8 (0.2)	85.5 (0.3)	86.9 (0.0)	88.4 (0.0)	63.9 (0.3)	87.0 (0.3)	85.5 (0.1)	78.1
BERT <sub>Small</sub> + MLM	42.2 (0.9)	78.8 (0.2)	80.8 (0.9)	85.7 (0.6)	88.6 (0.1)	58.6 (1.3)	89.4 (0.2)	84.4 (0.2)	76.1
<b>BERT<sub>Small</sub> + RTS</b>	<u>51.2</u> (1.8)	<b>79.9</b> (0.2)	81.5 (0.7)	<b>87.9</b> (0.1)	<b>89.4</b> (0.1)	<u>61.2</u> (0.6)	88.4 (0.4)	<u>85.2</u> (0.2)	78.1
<b>BERT<sub>Small</sub> + C-RTS</b>	<b>51.6</b> (0.9)	<u>79.8</u> (0.1)	81.3 (1.0)	<u>87.0</u> (0.4)	<b>89.4</b> (0.1)	<b>61.6</b> (1.4)	<b>89.5</b> (0.2)	<u>85.3</u> (0.3)	<b>78.2</b>
<b>BERT<sub>Small</sub> + SLM</b>	<u>45.8</u> (0.5)	<u>79.1</u> (0.2)	<b>83.3</b> (0.4)	86.3 (0.3)	<u>89.0</u> (0.1)	<u>60.8</u> (1.4)	88.6 (0.4)	<b>86.0</b> (0.2)	77.4

Table 5.9: Results on GLUE development set for Small models. For more details about the metrics, see Table 5.6. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over BERT<sub>Small</sub> + MLM.

Model	WikiQA		TREC-QA		ASNQ		GLUE
	MAP	MRR	MAP	MRR	MAP	MRR	AVG
ELECTRA <sub>Small</sub> + MLM/TD	75.8 (1.6)	77.5 (1.5)	83.3 (1.0)	88.2 (1.6)	59.2 (0.0)	66.3 (0.0)	75.5
BERT <sub>Small</sub> + MLM	74.5 (1.9)	75.8 (1.9)	85.3 (1.3)	89.9 (1.2)	58.6 (0.3)	66.0 (0.4)	74.1
<b>BERT<sub>Small</sub> + RTS</b>	<b>76.6</b> (0.7)	<b>78.4</b> (0.9)	85.2 (2.4)	90.5 (2.8)	<b>59.8</b> (0.1)	<u>66.9</u> (0.2)	75.4
<b>BERT<sub>Small</sub> + C-RTS</b>	74.0 (1.1)	75.8 (1.3)	86.1 (0.6)	<b>92.5</b> (2.0)	<u>59.6</u> (0.2)	<b>67.1</b> (0.2)	<b>75.7</b>
<b>BERT<sub>Small</sub> + SLM</b>	73.6 (0.8)	75.2 (0.9)	<b>86.3</b> (1.4)	91.7 (1.8)	<u>59.2</u> (0.2)	66.5 (0.3)	<b>75.7</b>

Table 5.10: Results of our Small models pre-trained with RTS, C-RTS and SLM compared with MLM on AS2 datasets and the GLUE test set.

### 5.4.2.2 Small Models

Here we present the results of applying our effective objectives to Small Transformer models. Table 5.9 shows the results over the GLUE development set, while in Table 5.10 we provide the results over 3 AS2 datasets and the average GLUE score on the test set. We do not show performance on Transfer-Learning from ASNQ to either WikiQA or TREC-QA because ASNQ is so large that after the transfer step, the differences in accuracy between the objectives are minimal and not statistically significant.

On the GLUE dev set, Small models trained with RTS and C-RTS outperform a similar MLM-based model on every task. The average GLUE score of RTS and C-RTS is respectively 2.0 and 2.1 points higher than MLM and even higher than SLM on some individual datasets. The only drawback of RTS and C-RTS is that the standard deviation is higher across most tasks, indicating a higher sensitivity to parameters’ initialization. On the GLUE test set, RTS and C-RTS perform accordingly to the dev set, outperforming MLM by 1.3 and 1.6 points. We can conclude that RTS and C-RTS are well suited for Small models, since they require almost half of the pre-training time and deliver superior performance compared to MLM, SLM and ELECTRA’s TD.

The AS2 dataset results exhibit a trend resembling those of GLUE: RTS and C-RTS consistently outperform MLM in all tasks. For instance, WikiQA showcases RTS’s superiority over MLM by 2.1 points in Mean Average Precision, while ASNQ shows a 1.2-point difference in favor of RTS. While results between MLM and RTS on TREC-QA are quite similar, it is noteworthy that BERT with RTS surpasses ELECTRA<sub>Small</sub> trained with TD in all tasks, despite the latter requiring twice the pre-training time and three times the GPU memory.

Regarding C-RTS, its performance, on average, is slightly lower than RTS in AS2 but slightly better on the GLUE test set, with differences in only a few decimal points. However, in the next Section, we will demonstrate that C-RTS exhibits clear advantages over RTS when trained for a longer duration, likely due to the slower convergence of the more challenging pre-training task.

Finally, SLM provides good performance on the GLUE test set, matching the average score of C-RTS and surpassing the expensive ELECTRA<sub>Small</sub> architecture and the MLM-based BERT. On the other Answer Sentence Selection tasks, SLM performance is better than those of BERT

Model	WikiQA	TREC-QA	ASNQ	MRPC	QNLI
<b>BERT<sub>Small</sub> + RTS</b>	76.6 (0.9)	85.5 (1.5)	59.9 (0.2)	81.5 (1.5)	<b>86.9</b> (0.3)
<b>BERT<sub>Small</sub> + C-RTS</b>	<b>77.4</b> (1.0)	<b>86.4</b> (0.9)	<b>60.7</b> (0.1)	<b>84.2</b> (0.1)	<b>86.9</b> (0.1)

Table 5.11: Comparison between RTS and C-RTS. Results for WikiQA and TREC-QA are computed over the test set, while we report the best results over the dev. set for the others. We compare results with MAP for WikiQA, TREC-QA and ASNQ, while we use Accuracy for the others. We underline statistically significant improvements over BERT<sub>Small</sub> + RTS.

with MLM on 2 datasets out of three. Notice that BERT with SLM and MLM share the exactly same architecture, pre-training time and memory footprint.

### 5.4.3 RTS vs C-RTS

The results of the previous Section show that C-RTS achieves accuracies similar to RTS in most benchmarks. To fully demonstrate the potential of a harder training task, we compared RTS and C-RTS when pre-trained for longer on the same data as before. In order to accomplish that, we use a new pre-training configuration in which the maximum sequence length is not reduced for the first steps (e.g. we use always 512 tokens). Then, we set the batch size to 1024 examples and train both models for 200K steps. After every 10K pre-training steps, we evaluate over 5 different benchmarks: ASNQ, WikiQA, TREC-QA, QNLI and MRPC. We selected those datasets because they cover 3 different domains and the sizes range from very few examples (WikiQA) to more than 20M (ASNQ). We exploit Small models for this experiment due to the cost of pre-training Base Transformers.

In Table 5.11 we report the results in the last pre-training step of the comparison described above. C-RTS obtains better results than RTS on 4 tasks out of 5 thanks to the harder training task. On the other hand, both objectives perform comparably from a statistical viewpoint on QNLI. This demonstrates empirically that selecting challenging replacements forces the model to find deeper relations between tokens. We provide the full plots of the comparison in Appendix C.1

## 5.5 Pre-Training Computational Trends

Our goal is to develop models that achieve comparable performance to MLM-based models while requiring less computational resources. Our findings indicate that the performance improvement follows a logarithmic pattern as the size of the pre-training dataset increases and models are trained for longer. Additionally, our computationally lighter models, such as RTS and C-RTS, demonstrate no statistically significant difference compared to more expensive models on most tasks. A comparison in Table 5.12 reveals that the top-performing architectures only outperform

Model	GLUE	FLOPS	Model	GLUE	FLOPS
BERT <sub>Small</sub> + RTS	75.4	× 0.10	BERT <sub>Large</sub> + MLM/NSP [Devlin et al., 2019]	83.3	× 12
BERT <sub>Small</sub> + MLM	74.1	× 0.12	ELECTRA <sub>Base</sub> + MLM/TD [Clark et al., 2020]	85.7	× 21
BERT <sub>Small</sub> + SLM	75.7	× 0.12	RoBERTa <sub>Base</sub> + MLM [Liu et al., 2019b]	86.3	× 53
BERT <sub>Base</sub> + RTS	79.9	× 0.81	ELECTRA <sub>Large</sub> + MLM/TD [Clark et al., 2020]	88.6	× 194
BERT <sub>Base</sub> + MLM	79.7	× 1	RoBERTa <sub>Large</sub> + MLM [Liu et al., 2019b]	88.8	× 200
BERT <sub>Base</sub> + SLM	80.4	× 1	ALBERT <sub>Large</sub> + MLM/SOP [Lan et al., 2020]	90.0	× 1937

Table 5.12: Comparison of computational requirements for state-of-the-art large language models. We report the average GLUE score as well as an estimation of the FLOPS used for the pre-training, normalized on BERT<sub>Base</sub> + MLM.

MLM-based models when trained on significantly larger amounts of data and for much more time. For instance,  $\text{ELECTRA}_{\text{Base}}$  utilizes 21 times more resources than  $\text{BERT}_{\text{Base}}$ , and RoBERTa uses 53 times more. It is worth noting that achieving a score of 90.0 on the GLUE benchmark requires a model to be trained for 2000 times the duration of the original BERT model training, which is quite remarkable. For those reasons, we believe research in finding more efficient architectures, objectives or distillation techniques is extremely important to reduce the carbon footprint of models pre-training.



## Chapter 6

# Self-supervised Objectives for Multi-Sentence Inference

In this Chapter, we propose pre-training objectives designed specifically to improve accuracy in multiple-sentence classification. Many tasks such as AS2, Fact Verification or NLI are solved by training a sentence pair classifier, for example by feeding the Transformer model with a question-answer or a claim-evidence pair. Recent studies [Tymoshenko and Moschitti, 2021; Zhang et al., 2021] showed that modeling multiple candidate sentences helps in learning long-range dependencies and allows the model to compare them, instead of scoring them individually.

A major problem is that off-the-shelf Transformer models struggle at modeling multiple related sentence candidates together. For this reason, we propose a new pre-training technique that exploits weak supervision of pre-training corpora to adapt the model at reasoning over multiple input sentences taken from the same paragraph. We call *Jointwise* the class of models trained with objectives that target relations among multiple input sentences. At the same time, we call *Pairwise* models the class of architectures designed to reason over 1 or 2 input sequences.

Regarding the related works, additional information about the RoBERTa architecture [Liu et al., 2019b] can be found in Section 3.1.2. Additionally, Sections 3.2.1 and 3.2.2 describe the actual state-of-the-art for Answer Sentence Selection and Fact Verification. In Section 3.3.2 instead, a description of works about multi-sentence inference is provided, while in Section 3.3.3 we analyze related research that exploits the weak structure of large corpora for supervision.

To demonstrate the superiority of our approach, we evaluate models pre-trained with our objectives on three datasets for Answer Sentence Selection (ASNQ, WikiQA and TREC-QA) and one dataset for Fact Verification (FEVER). Statistics and details about each dataset are given in Section 4.2.1. We select RoBERTa<sub>Base</sub> [Liu et al., 2019b] as the starting point because it provides good accuracy on a wide range of tasks and does not require expensive hardware to be fine-tuned. We compare against a baseline pre-trained only with common objectives (e.g. MLM) as well as models pre-trained over data formatted for our task but without explicit supervision.

### 6.1 Limits of Pairwise Models

Examples of Pairwise models comprehend Transformers either (i) pre-trained with sentence-level objectives such as NSP [Devlin et al., 2019] or SOP [Lan et al., 2020] or (ii) fine-tuned on Pairwise tasks such as Natural Language Inference, Answer Sentence Selection and Fact Verification.

When applied to downstream tasks, Pairwise models perform effectively in comparing two input sequences, thanks to the Cross-Attention of Transformers. In the AS2 task, they receive a question and a single answer candidate and produce a score indicating the degree of correctness.

However, this approach has limitations as the model cannot efficiently compare and resolve references among multiple answer candidates to select the best answer.

Conversely, in Fact Verification, the current state-of-the-art approach involves providing the model with a claim and a set of evidence sentences. The model then determines whether the claim is supported, neutral, or refuted by the evidence sentences [Liu et al., 2020; Tymoshenko and Moschitti, 2021]. Nevertheless, Transformers are typically pre-trained with token-level and sentence-level objectives, focusing on at most two spans of text. Although models like RoBERTa are trained on very long sequences containing multiple sentences, they are not explicitly trained to identify connections and relations between these sentences. Consequently, they may struggle to learn meaningful dependencies between the candidates during fine-tuning, making them less adaptable to tasks requiring reasoning over multiple inputs.

## 6.2 Multi-Sentence Transformer Model

In this Section, we present the Transformer architecture and the pre-training objective we developed for the multi-sentence training and inference tasks.

### 6.2.1 Joint Encoder Architecture

We created an architecture specifically designed to process and reason over multiple input sequences. We indicate with  $k + 1$  the maximum number of text spans that can be provided to the model. The first input sentence is called *pivot* and is of special importance because we study its relation with the other  $k$  input sequences. For example, in AS2 we place the question in the first position and the model should predict its relation with the other  $k$  answer candidates. Similarly, in Fact Verification we set the claim as the pivot and we predict whether the other  $k$  evidence sentences support or reject the claim.

We experimented with 2 different configurations: the first, referred to as *Fixed* (Figure 6.1), and the second, named *Flexible* (Figure 6.2), both consisting of a RoBERTa Encoder [Liu et al., 2019b]. The main differences between the Fixed and Flexible variants are how inputs are formatted before being fed to the model and how multiple predictions are performed from the last hidden state.

#### 6.2.1.1 Fixed Model

In the Fixed variant of our architecture, each of the  $k + 1$  input text spans has the same Fixed length  $L$ , measured in the number of tokens. We crop or pad on the right all the input sequences whose length is different than  $L$ . Notice that this may lead to pad tokens between different input sentences. This is not an issue because the attention mask will be aware of those special tokens and will disable attention scores in the corresponding positions. We add positional embeddings and sequence id embeddings to the word embeddings. Positional embeddings help the model in computing attention between pairs of tokens in different positions while sequence id embeddings are useful to let the model know to which sequence each token belongs. We are aware that having both positional ids and sequence ids in a Fixed architecture is redundant, however, we keep both embeddings because the overhead is negligible and we can use a single implementation for both Fixed and Flexible models. We also experimented using the same positional ids  $0, \dots, L - 1$  for every sentence, but we didn't measure significant gains. Notice that in this case, token type ids played an important role and couldn't be removed.

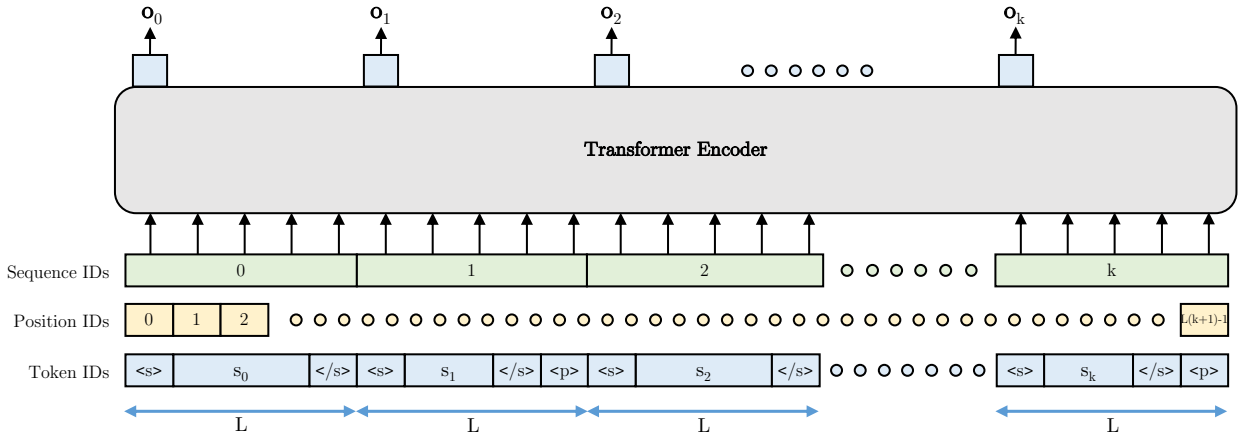


Figure 6.1: Example of Jointwise Fixed architecture.

Jointwise Fixed models perform  $k + 1$  sentence-level predictions for each step, employing a classification head over output embeddings  $[\mathbf{o}_0, \mathbf{o}_1, \dots, \mathbf{o}_k]$ , which correspond to the first token  $\langle s \rangle$  of each input sentence<sup>1</sup>. In particular, given the last hidden state  $\mathbf{H} \in \mathbb{R}^{(k+1)L \times d}$ , the output embeddings  $\mathbf{o}_0, \mathbf{o}_1, \dots, \mathbf{o}_k \in \mathbb{R}^d$  correspond to the rows  $0, L, \dots, kL$  of  $\mathbf{H}$ .

### 6.2.1.2 Flexible Model

The Flexible architecture is similar to the Fixed version but allows input sequences to have a variable length. In Figure 6.2, the length of each input sequence is indicated with  $L_0, \dots, L_k$ . The only constraint in Flexible models is:

$$\sum_{i=0}^k L_i \leq L \tag{6.1}$$

If the input does not match the constraint, we iteratively remove one token at a time from the longest sequence until we do not match the requirement. We use again positional and sequence ids. In particular, preliminary experiments demonstrated that sequence ids are very important

<sup>1</sup> $\langle s \rangle$  is RoBERTa’s “begin of sentence” token, equivalent to BERT’s [CLS].

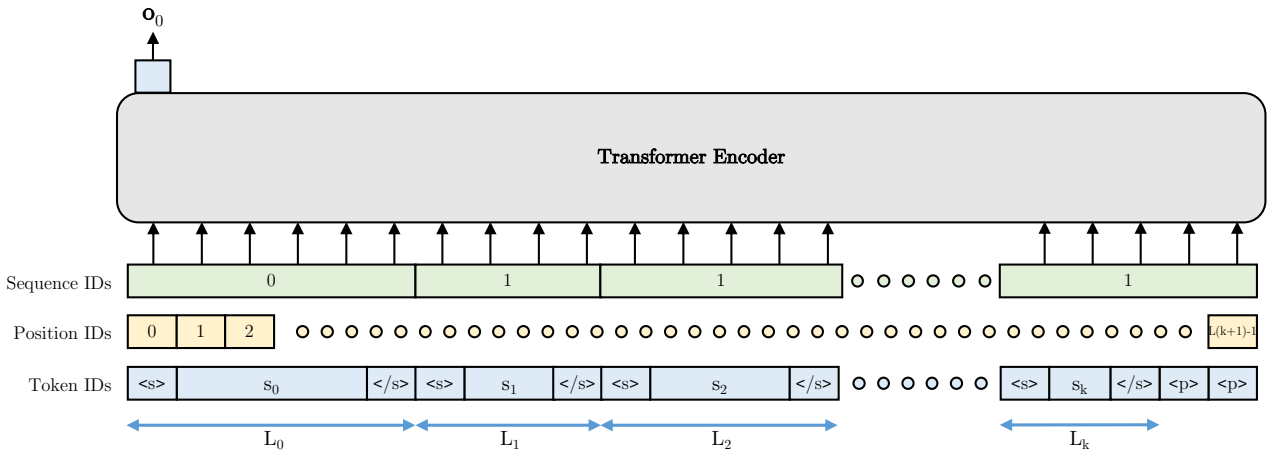


Figure 6.2: Example of Jointwise Flexible architecture.

in this architecture because they help the model to better understand the context for each prediction. In Flexible models, we employ only 2 sequence ids instead of  $k + 1$ . We assign a sequence id equal to 0 to the pivot and 1 to all the other  $k$  sentence candidates. Predictions with Flexible models are simpler because we consider always only the output embedding  $\mathbf{o}_0$  of the first token to perform single or multi-label predictions.

## 6.2.2 Multi-Sentence Inference Tasks

Here, we describe how we adapt AS2 and Fact Verification for the multi-sentence inference task. We denote with  $k$  the maximum number of predictions performed by our Jointwise model.

**AS2** We extend the definition of AS2 given in Section 2.1.1.2 for multi-sentence inference. In particular, we partition the candidates  $\mathbb{C}$  in disjoint subsets  $\mathbb{C}_0, \dots, \mathbb{C}_m$  of size  $k$  with random selection. Then, we feed the model  $\mathcal{M}$  with the question and one subset at a time, obtaining  $k$  predictions over the candidates. If one subset  $\mathbb{C}_i$  is smaller than  $k$ , we add padding on the right until we match the sequence length of the other elements in the batch and we do not predict and back-propagate errors from the corresponding classification heads. Finally, we select the best answer among all subsets as the one having the highest overall score, as in the original task.

Our intuition is that by providing  $\mathcal{M}$  with  $k$  different candidates, it can model interrelated information to find the best answer  $a$  to  $q$  [Zhang et al., 2021].

**Fact Verification** Let’s consider a claim  $c$  and a set of evidence sentences  $\mathbb{E} = \{e_1, \dots, e_n\}$  retrieved using a BERT-based DocIR [Liu et al., 2020]. In the classic definition of the task,  $c$  should be supported/refuted by at least one piece of evidence  $e_i \in \mathbb{E}$  to be labeled accordingly. More details about Fact Verification are given in Section 2.1.2.

As for Joint AS2, we provide the model  $\mathcal{M}$  with the claim  $c$  and the evidence  $\mathbb{E}$ , such that different evidence sentences can be modeled jointly and complement each other [Tymoshenko and Moschitti, 2021].

## 6.2.3 Training and Inference using Joint Encoders

We describe how we exploit output embeddings  $\mathbf{o}_0, \dots, \mathbf{o}_k$  of Fixed and  $\mathbf{o}_0$  of Flexible models to perform predictions. We differentiate between AS2, that requires to predict a label for each candidate, and Fact Verification, which requires to predict a single label for every input example.

### 6.2.3.1 Predict a single label

We experiment with two different strategies to predict a single label from the output embeddings of the model.

**IE<sub>1</sub>** We place a linear layer  $\mathbf{W} \in \mathbb{R}^{2 \times d}$  only on top of  $\mathbf{o}_0$ , that corresponds to the embedding of the first input sentence (similar to BERT [Devlin et al., 2019]). This classification head can be used for both Fixed and Flexible models. We call this method Individual Evidence inference head and we obtain predictions  $p$  with:

$$p = \mathbf{W}\mathbf{o}_0 \tag{6.2}$$

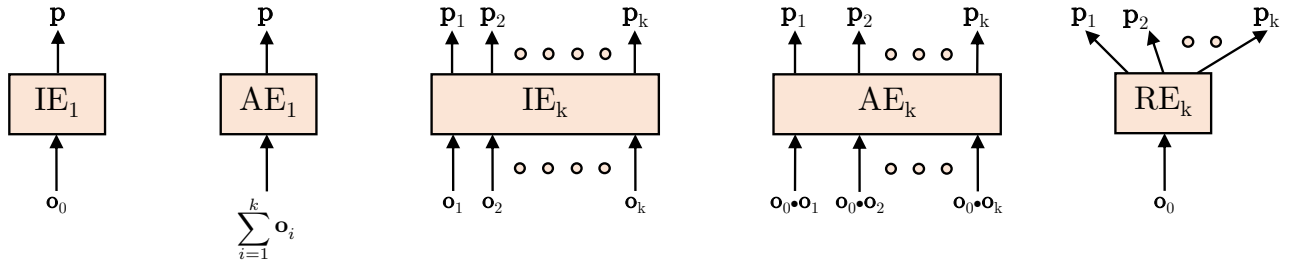


Figure 6.3: Inference heads for Joint Transformer models.

**AE<sub>1</sub>** Here, we apply a single linear layer  $\mathbf{W} \in \mathbb{R}^{2 \times d}$  to the average of all the output embeddings  $\mathbf{o}_1, \dots, \mathbf{o}_k$  to gather information from all input candidates. By definition, this classification head can be used only with Fixed models. We refer to this technique as Aggregated Evidence classification head:

$$p = \mathbf{W} \left( \frac{1}{k} \sum_{i=1}^k \mathbf{o}_i \right) \quad (6.3)$$

### 6.2.3.2 Predict $k$ labels

To predict  $k$  different labels, we experimented with 3 techniques (2 for Fixed and 1 for Flexible models).

**IE<sub>k</sub>** A linear layer with shared weights  $\mathbf{W} \in \mathbb{R}^{2 \times d}$  is applied to each output embedding  $\mathbf{o}_i$  to obtain predictions. By definition, this head works only with Fixed models. We refer to this method as  $k$ -candidate Individual Evidence inference head:

$$p_i = \mathbf{W} \mathbf{o}_i \quad \forall i \in 1, \dots, k \quad (6.4)$$

**AE<sub>k</sub>** In this experiment, let the linear layer be a matrix of shared weights  $\mathbf{W} \in \mathbb{R}^{2 \times 2d}$ . We feed the linear layer with the concatenation of the first output embedding  $\mathbf{o}_0$  with each candidate embedding  $(\mathbf{o}_1, \dots, \mathbf{o}_k)$ . This head is yet again specifically designed only for Fixed models. We call this technique  $k$ -candidate Aggregated Evidence (AE<sub>k</sub>) classification head (“.” means vectors concatenation):

$$p_i = \mathbf{W}(\mathbf{o}_0 \cdot \mathbf{o}_i) \quad \forall i \in 1, \dots, k \quad (6.5)$$

**RE<sub>k</sub>** This last classification head is designed only for Flexible models. Let  $\{\mathbf{W}_i \in \mathbb{R}^{2 \times d}, i \in 1, \dots, k\}$  be a set of  $k$  linear layers. We apply all of them independently on the output embedding  $\mathbf{o}_0$  of the first input token to obtain  $k$  predictions. We name this method  $k$ -candidate Reduced Evidence (RE<sub>k</sub>). Predictions are computed as follows:

$$p_i = \mathbf{W}_i \mathbf{o}_0 \quad \forall i \in 1, \dots, k \quad (6.6)$$

In figure 6.3 we depict every classification head to better catch the differences between them. Furthermore, Table 6.1 provides a summary of the applicable classification heads for each model and task.

Task	Fixed					Flexible				
	IE <sub>1</sub>	IE <sub>k</sub>	AE <sub>1</sub>	AE <sub>k</sub>	RE <sub>k</sub>	IE <sub>1</sub>	IE <sub>k</sub>	AE <sub>1</sub>	AE <sub>k</sub>	RE <sub>k</sub>
MSPP Pre-Training		✓		✓						✓
Answer Sentence Selection		✓		✓						✓
Fact Verification	✓		✓			✓				

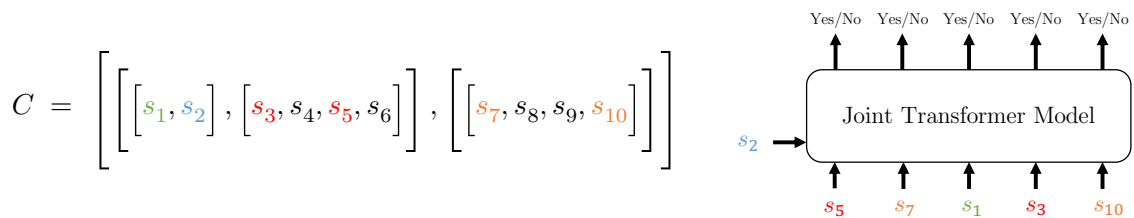
Table 6.1: Compatibility matrix between models, classification heads and downstream tasks.

## 6.2.4 Continuous Pre-Training of Joint Encoders

In long documents, paragraphs are used to explore the document’s topic from various perspectives. Despite this rich source of information, most Transformer pre-training strategies fail to take advantage of it, leaving the unsupervised pre-training phase without any guidance. To improve the ability of Transformer models to understand the relationships between multiple sentences, we introduce a new pre-training task called Multi-Sentences in Paragraph Prediction (MSPP). In this task, the model is given  $k + 1$  sentences  $\{s_0, \dots, s_k\}$  and is asked to predict whether sentences  $\{s_1, \dots, s_k\}$  belong to the same paragraph  $P$  as  $s_0$ , in the document  $\mathbf{D}$ . We use the  $IE_k$  and  $AE_k$  prediction heads for Fixed models and  $RE_k$  with Flexible models to perform the  $k$  predictions. More formally, we create a label  $y_i$  for each sentence  $s_i$  with the following rule:

$$y_i = \begin{cases} 1 & \text{if } s_0, s_i \in P \text{ in } \mathbf{D} \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in \{1, \dots, k\} \quad (6.7)$$

During the pre-training phase, we randomly select a sentence from a paragraph as  $s_0$ . We then select  $k_1$  other sentences from the same paragraph as positive examples,  $k_2$  sentences from other paragraphs in the same document as hard negative examples, and  $k_3$  sentences from documents other than  $\mathbf{D}$  as easy negatives. We ensure  $k_1 + k_2 + k_3 = k$  by sampling more easy negatives ( $k_3$ ) if paragraph  $P$  is not large enough to sample  $k_2$  hard negatives. This method tasks the model to learn the relationship between sentences in a paragraph, improving its ability to understand the underlying structure of long documents. An example of MSPP is given in Figure 6.4.



Los Angeles, often referred to by its initials L.A., is a major city in the U.S. state of California. With a 2020 population of 3,898,747, it is the largest city in the state, as well as the second-largest city in the United States following New York City.

The area that became Los Angeles was claimed by Spain in 1542. It became a part of Mexico in 1821 following the Mexican War of Independence. At the end of the Mexican-American War, Los Angeles was purchased as part of the Treaty of Guadalupe Hidalgo and became part of the United States. Los Angeles was incorporated as a municipality in 1850.

New York, often called New York City or NYC, is the most populous city in the United States. With a 2020 population of 8,804,190 distributed over 300.46 square miles (778.2 km<sup>2</sup>), New York City is the most densely populated major city in the United States. It is more than twice as populous as Los Angeles, the nation’s second-largest city. The city also has a population that is larger than that of 38 individual U.S. states.

Figure 6.4: Example of MSPP objective.

## 6.3 Experiments

To assess the performance of our joint Transformers, we conduct experiments on three AS2 datasets and one Fact Verification dataset. We do not use commonly used language model benchmarks for our evaluation, such as GLUE, because they only involve Pairwise classification tasks, which is not sufficient for our study.

### 6.3.1 Continuous Pre-Training

In our experiments, we want to avoid improvements stemming from the usage of additional data. For this reason, we use the same pre-training corpora of RoBERTa: English Wikipedia, BookCorpus, OpenWebText and CC-News. For more details about the datasets, see Section 4.1. RoBERTa was also trained on the STORIES dataset, however, we omit it because it is not publicly available anymore<sup>2</sup>.

In all of our experiments, we set  $k = 5$  as in Tymoshenko and Moschitti [2021]; Zhang et al. [2021] because it aligns well with the number of candidates/evidence sentences in the datasets and because it does still allow an efficient training with a reasonable input sequence length. We set also  $L = 64$  in the experiments. Thus, Fixed models will have at most 6 input sequences of length 64 while Flexible models will also accept up to 6 input sequences whose total length is at most 384 tokens. However, since Flexible models do not apply internal padding, we found that setting the maximum sequence length  $L$  to 256 does not affect performance but increases training speed.

The pre-training dataset is created with the MSPP strategy by setting  $k_1 = 1, k_2 = 2, k_3 = 2$ . With these settings, we have a good balance of soft and hard negative examples. The resulting dataset contains about 180M examples. We do continuous pre-training starting from a RoBERTa<sub>Base</sub> checkpoint because training from scratch would require a very large computational budget<sup>3</sup>. The original checkpoint is only augmented with sequence-level embeddings and a small classification head for binary predictions. While doing continuous pre-training, we found it beneficial to keep the original MLM objective along with MSPP. We train the model over the dataset described above for 100K steps with a batch size of 4096. We used a triangular learning rate with a peak value of  $5 \cdot 10^{-5}$  and 10K warmup steps. We set the optimizer with  $\beta_1 = 0.9, \beta_2 = 0.999$  and  $\epsilon = 10^{-8}$  to update the weights of the model. We also apply a weight decay of 0.01 and dropout with probability 0.1 to regularize model training. Since we trained in mixed-precision (*fp16*), we did not use gradient clipping because gradients are already scaled by the NVIDIA Apex algorithm automatically. For the pre-training, we experimented with 2 Fixed and 1 Flexible model: the first trained with  $IE_k$ , the second with  $AE_k$  and the Flexible with  $RE_k$ . As in BERT and ALBERT [Devlin et al., 2019; Lan et al., 2020], we use an equal weight for the MLM and MSPP to compute the final loss value. The training takes about 2.2 days (Fixed) and 1.7 days (Flexible) on our machine, which is described in Section 1.2, using mixed precision and the parameters described above. In terms of efficiency, our continuous pre-trainings account for about a 7% (Fixed) and a 5% (Flexible) increase in computation compared to the original RoBERTa pre-training.

<sup>2</sup>[https://github.com/tensorflow/models/tree/archive/research/lm\\_commonsense#1-download-data-files](https://github.com/tensorflow/models/tree/archive/research/lm_commonsense#1-download-data-files)

<sup>3</sup>RoBERTa<sub>Base</sub> was trained on 1024 NVIDIA V100 GPUs for 1 day, which corresponds to an expense of more than 100,000\$ on Amazon Web Services or similar cloud computing providers.

### 6.3.2 Fine-Tuning

We evaluate the 3 continuously pre-trained checkpoints on ASNQ, WikiQA, TREC-QA and FEVER, using the correct classification heads, i.e.  $IE_k$ ,  $AE_k$  or  $RE_k$  for AS2 and  $IE_1$  or  $AE_1$  for Fact Verification.

We compare against many baselines from other works. On AS2, we compare with (i) RoBERTa<sub>Base</sub>, used as a Pairwise Cross-Encoder, which is the standard practice [Garg et al., 2019] and (ii) RoBERTa<sub>Base</sub> with our classification heads but without MSPP pre-training. The latter allows us to evaluate how RoBERTa<sub>Base</sub> without the specialized pre-training objective adapts at processing  $k + 1$  input sequences.

On FEVER instead, we compare with GEAR [Zhou et al., 2019], KGAT [Liu et al., 2020] and Transformer-XH [Zhao et al., 2020]. We also include 3 models from Tymoshenko and Moschitti [2021]: (i) a RoBERTa<sub>Base</sub> for multi-sentence classification with  $IE_1$  prediction head, (ii) two Pairwise Cross-Encoder RoBERTa<sub>Base</sub> with heads that do max-pooling and weighted-sum to gather the context for the classification.

We optimize our models with a maximum sequence length of 384 tokens for all experiments, which is 64 multiplied by the number of sentences  $k + 1$ . Regarding ASNQ, we train the model for a maximum of 10 epochs using a learning rate of  $10^{-5}$  and a batch size of 512 examples, with the FuseAdam optimizer. We warmup for only 5000 steps, and we use early stopping based on the Mean Average Precision (MAP) on the development set. For WikiQA and TREC-QA, we train the model using a learning rate of  $2 \cdot 10^{-6}$  and a batch size of 32, with 1000 warmup steps. We train for up to 40 epochs and use early stopping based on the MAP metric on the development set. Finally, for the FEVER dataset, we use a batches of 64 example, a learning rate of  $10^{-5}$ , and 1000 warmup steps. We use early stopping based on the Accuracy metric on the development set.

We evaluate models’ performance on Answer Sentence Selection with Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Precision@1 (P@1), as in similar works [Garg et al., 2019]. For the FEVER dataset, we measure performance with Label Accuracy, which is a standard metric in Fact Verification [Liu et al., 2020; Tymoshenko and Moschitti, 2021; Zarharan et al., 2021]. It evaluates the accuracy in predicting whether a claim is supported/refuted/neutral to the set of retrieved evidence sentences.

## 6.4 Results

In this Section, we report results obtained by fine-tuning the baselines and our MSPP pre-trained models on AS2 and Fact Verification benchmarks. We report results from original papers when available. On AS2, we analyze the results after a standard fine-tuning as well as the results of chaining our Jointwise models with a common Pairwise re-ranker. The statistical significance in AS2 is compared to RoBERTa<sub>Base</sub> [Liu et al., 2019b].

### 6.4.1 Answer Sentence Selection

Results on all the AS2 datasets are reported in Table 6.2. We experiment using all multi-label heads defined in Section 6.2.3, i.e.  $IE_k$ ,  $AE_k$  and  $RE_k$ . Results are reported against a Pairwise baseline which uses the same Transformer encoder and has the same parameters as our Joint models (which in some cases have only a slightly larger classification head, with negligible effects on fine-tuning/inference time). The Table shows two main discoveries: (i) models fine-tuned directly on the downstream tasks in a multi-sentence prediction setting do not achieve high



Model	ASNQ			WikiQA			TREC-QA		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
RoBERTa <sub>Base</sub> Fixed ( $IE_k$ )	62.1 (0.8)	65.6 (1.2)	57.4 (1.3)	39.4 (1.6)	40.3 (1.8)	19.7 (1.9)	41.9 (2.4)	50.8 (3.9)	30.9 (5.4)
RoBERTa <sub>Base</sub> Fixed ( $AE_k$ )	62.3 (0.7)	66.0 (1.0)	57.7 (1.1)	39.0 (2.8)	39.7 (2.9)	18.7 (3.9)	42.3 (3.2)	49.2 (5.0)	29.7 (6.9)
RoBERTa <sub>Base</sub> Flexible ( $RE_k$ )	63.7 (0.3)	70.2 (0.4)	58.9 (0.6)	77.8 (2.3)	79.6 (2.4)	69.5 (3.4)	84.8 (0.7)	91.3 (1.7)	84.8 (3.1)
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	66.9 (0.1)	73.1 (0.1)	61.8 (0.2)	85.3 (0.9)	86.5 (1.0)	77.1 (2.1)	89.3 (0.9)	93.1 (1.0)	87.9 (2.2)
<b>RoBERTa<sub>Base</sub> Fixed (<math>IE_k</math>) + MSPP</b>	<b>67.2</b> (0.2)	<b>73.7</b> (0.2)	<b>63.0</b> (0.3)	<b>88.5</b> (1.5)	<b>89.0</b> (1.5)	<b>82.7</b> (2.2)	<b>91.1</b> (0.5)	<b>95.2</b> (1.3)	<b>91.7</b> (2.2)
<b>RoBERTa<sub>Base</sub> Fixed (<math>AE_k</math>) + MSPP</b>	<b>67.3</b> (0.2)	<b>73.7</b> (0.2)	<b>63.0</b> (0.3)	<b>87.9</b> (1.4)	<b>89.0</b> (1.5)	<b>81.9</b> (2.6)	90.1 (1.0)	93.6 (0.6)	88.7 (0.8)
<b>RoBERTa<sub>Base</sub> Flexible (<math>RE_k</math>) + MSPP</b>	<b>69.4</b> (0.0)	<b>75.7</b> (0.0)	<b>65.3</b> (0.1)	<b>88.7</b> (0.8)	<b>90.0</b> (0.8)	<b>83.5</b> (1.2)	<b>91.0</b> (0.7)	<b>95.0</b> (0.8)	<b>91.2</b> (1.2)
TANDA RoBERTa <sub>Base</sub> [Garg et al., 2019]	-	-	-	88.9	90.1	-	91.4	95.2	-

Table 6.2: Results on ASNQ, WikiQA and TREC-QA.  $IE_k$ ,  $AE_k$  and  $RE_k$  are the prediction heads we use both for MSPP pre-training and fine-tuning. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over RoBERTa<sub>Base</sub>.

re-ranking scores; (ii) MSPP pre-training is fundamental to accustom the model to reason over multiple sentences.

Regarding the first point, the low performance of models fine-tuned directly on the multi-sentence tasks can be explained by considering that attention scores are not pre-trained to compare many sentences. Most pre-training objectives focus on token-level dependencies (e.g. MLM, TD, SLM) while sentence-level objectives such as NSP or SOP consider at most 2 input sequences. This is further confirmed by the results on ASNQ, which is a large dataset, and those on WikiQA and TREC-QA, which contain only a few thousand examples. Results on ASNQ are higher in the standard fine-tuning setting because the model can leverage the large quantity of annotated examples (about 4M, each containing 1 question and  $k = 5$  answer candidates) to learn multi-sentence relations.

Regarding the Fixed models, both checkpoints pre-trained with MSPP and fine-tuned with either  $IE_k$  and  $AE_k$  improve over the Pairwise baseline by a large margin on WikiQA and TREC-QA. Specifically, the Joint Fixed model with  $IE_k$  classification head improves by 5.6 points in Precision@1 on WikiQA and by 3.8 points on TREC-QA. Similarly, the Joint model with  $AE_k$  head outperforms the baseline by 4.8 in Precision@1 on WikiQA and by 0.8 in TREC-QA. On ASNQ, improvements are modest because the dataset is larger and reduces the differences seen in pre-training, but are still consistent. In particular, the Joint model with  $IE_k$  and  $AE_k$  show a gain in MAP of 0.3 and 0.4 points respectively. When compared with Precision@1 instead, improvements are larger and equal to 1.2 points for both models.

The Flexible model we present reaches the best results on ASNQ and WikiQA, with a gap of 3.5 points in P@1 over the baseline on ASNQ and 6.4 points over the WikiQA baseline. On TREC-QA, MAP and P@1 are aligned with the other Joint models, with gains over the Pairwise baseline of 1.7 points in MAP and 3.3 in Precision@1. To the best of our knowledge, results on ASNQ are the actual state-of-the-art when compared with other models with a similar number of parameters<sup>4</sup>. This shows that using a Flexible architecture is overall the best choice to design a multi-sentence inference model for Answer Sentence Selection.

Finally, we compare with TANDA, which is the actual state-of-the-art in AS2 for WikiQA and TREC-QA. Our methods demonstrate similar performance on both datasets, with a maximum difference of 0.3 points when measured with MAP and MRR. Notice that TANDA exploits a large labeled dataset for AS2 (ASNQ) to perform a pre-fine-tuning step and transfer knowledge into smaller datasets. We use instead only raw data from large corpora without manual annotation.

Model	ASNQ			WikiQA			TREC-QA		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	66.9 (0.1)	73.1 (0.1)	61.8 (0.2)	85.3 (0.9)	86.5 (1.0)	77.1 (2.1)	89.3 (0.9)	93.1 (1.0)	87.9 (2.2)
<b>RoBERTa<sub>Base</sub> Fixed (IE<sub>k</sub>) + MSPP</b>	<u>71.3</u> (0.4)	73.1 (0.5)	<u>63.9</u> (0.8)	<b>88.5</b> (1.4)	<u>89.0</u> (1.5)	<u>82.7</u> (3.0)	<b>93.5</b> (1.2)	<b>95.4</b> (1.4)	<b>92.2</b> (1.8)
<b>RoBERTa<sub>Base</sub> Fixed (AE<sub>k</sub>) + MSPP</b>	<b>71.5</b> (0.3)	<b>73.4</b> (0.3)	<b>64.3</b> (1.1)	87.9 (0.8)	88.7 (0.7)	82.1 (1.1)	<b>93.5</b> (1.1)	<u>94.9</u> (1.2)	<u>91.2</u> (1.4)
<b>RoBERTa<sub>Base</sub> Flexible (RE<sub>k</sub>) + MSPP</b>	<u>70.7</u> (0.5)	72.5 (0.6)	<u>64.1</u> (1.0)	<u>88.1</u> (1.1)	<b>89.1</b> (1.2)	<b>82.8</b> (1.9)	90.7 (0.7)	92.2 (0.8)	89.7 (1.2)
Superior limit	86.5			97.9			100.0		

Table 6.3: Results on ASNQ, WikiQA and TREC-QA for Joint Fixed and Flexible models when chained after the RoBERTa<sub>Base</sub> Pairwise re-ranker. In the last row, we show the maximum achievable performance based on the Hit-Rate at  $k = 5$  (HR@5) of the Pairwise re-ranker. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over RoBERTa<sub>Base</sub>.

### 6.4.1.1 Re-Rankers Chain

In this Section, we study whether Jointwise models can be chained with a Pairwise Cross-Encoder to improve the performance of the latter. More specifically, we use a RoBERTa<sub>Base</sub> Cross-Encoder to find the top  $k$  more relevant answers for each question. After that, we apply the Jointwise model to better re-rank the top  $k$  answers retrieved by the baseline Cross-Encoder.

Results are reported in Table 6.3. The last row of the Table shows the upper limit to Jointwise model performance. The limit is calculated as the Hit-Rate@ $k$  of the Pairwise re-ranker. This comparison is needed because if the Pairwise baseline is not able to rank at least one positive candidate in the top  $k$ , Jointwise encoders will process only negative examples and will receive a bad score by the evaluation metrics.

In this experiment, Fixed models provide better overall performance, especially on the TREC-QA dataset. On the other hand, the Flexible model provides similar performance to Fixed models on ASNQ and WikiQA, but falls short behind on TREC-QA, still providing a better re-ranking of the first  $k$  candidates than the baseline Pairwise Cross-Encoder. On average, performance is lower than applying Jointwise models directly to the whole datasets: we suppose that our models work better when applied to all data because more candidates can be compared with each other. Moreover, the accuracy of Jointwise models when applied after a Pairwise re-ranker is limited by the latter’s performance. This is further confirmed by checking the results on TREC-QA, in which most Jointwise models perform better than the previous setting because the Pairwise re-ranker was able to score always at least a positive candidate in the top  $k$  (HR@ $k = 100.0\%$ ).

## 6.4.2 Fact Verification

We present results on the Fact Verification task by comparing our models against many baselines from the literature. Table 6.4 reports the Label Accuracy on the development and test sets of the FEVER dataset.

Flexible models without our ad-hoc pre-training task reach remarkable performance on both the development and test set, obtaining a Label Accuracy higher than some previous state-of-the-art systems. On the other hand, when our models are further trained with the MSPP objective, we obtain state-of-the-art results among models based on a Base architecture (about 120M of parameters). In particular, by testing models on the Fever Shared Task Leaderboard, we obtain a significant gain of 1.29 points when using our Jointwise Flexible model with MSPP. On the development set, all of our Joint models outperform the baselines by a margin of at least 1.20 points in Label Accuracy.

Finally, notice that our Joint Flexible model achieves a score slightly below the actual

<sup>4</sup>In Chapter 7 we will show that large models can achieve even higher scores.

Model	FEVER	
	L.A. Dev	L.A. Test
GEAR [Zhou et al., 2019]	70.69	71.60
DOMLIN [Stammach and Neumann, 2019]	71.54	-
KGAT with RoBERTa <sub>Base</sub> [Liu et al., 2020]	78.29	74.07
Transformer-XH [Zhao et al., 2020]	78.05	72.39
RoBERTa <sub>Base</sub> Pairwise + MaxPool [Tymoshenko et al., 2021]	79.82	-
RoBERTa <sub>Base</sub> Pairwise + WgtSum [Tymoshenko et al., 2021]	80.01	-
DOMLIN++ [Stammach and Ash, 2020-10]	77.48	76.60
DREAM [Zhong et al., 2020]	79.16	<b>76.85</b>
RoBERTa <sub>Base</sub> Fixed (IE <sub>1</sub> )	78.92 (0.46)	73.01
RoBERTa <sub>Base</sub> Fixed (AE <sub>1</sub> )	78.85 (0.36)	72.94
RoBERTa <sub>Base</sub> Flexible (IE <sub>1</sub> )	79.25 (0.27)	73.56
<b>RoBERTa<sub>Base</sub> Fixed (IE<sub>1</sub>) + MSPP (IE<sub>k</sub>)</b>	81.21 (0.24)	74.39
<b>RoBERTa<sub>Base</sub> Fixed (AE<sub>1</sub>) + MSPP (AE<sub>k</sub>)</b>	81.21 (0.16)	74.13
<b>RoBERTa<sub>Base</sub> Flexible (IE<sub>1</sub>) + MSPP (RE<sub>k</sub>)</b>	<b>82.10</b> (0.10)	75.36

Table 6.4: Results on FEVER dev and test sets. We use the prediction heads IE<sub>1</sub> and AE<sub>1</sub> for fine-tuning and either IE<sub>k</sub>, AE<sub>k</sub> or RE<sub>k</sub> for the MSPP pre-training. ‘-’ means results are not reported in the original paper. We use bold formatting for our proposals and the best results.

state-of-the-art on FEVER: DREAM and DOMLIN++. However, those models use a better retrieval system (we are limited by the performance of the BERT-based Doc-IR) and larger models for the Fact Verification task. In particular, DREAM Fact Verification systems is based on an XLNet<sub>Large</sub> [Yang et al., 2019] model while DOMLIN++ uses a RoBERTa<sub>Large</sub> [Liu et al., 2019b] architecture. Both models contain about 360M parameters, which is much more than the 124M parameters of our Jointwise RoBERTa<sub>Base</sub>.

## 6.5 Latency Analysis

Although our Joint model allows for longer input sequences in the Transformer, it also decreases the number of forward passes required by an equivalent Pairwise Cross-Encoder. To provide a simplified analysis of the latency in AS2, let  $L$  be the length of the longest sentence. In the case of the Pairwise Cross-Encoder,  $k$  forward passes (one for each candidate  $c_i$ ) of the Transformer are needed with a sequence length of  $2L$ . However, our Joint model only requires a single forward pass over the Transformer with an input length of  $L(k + 1)$  (one pass with the query and  $k$  answer candidates).

Since the Transformer’s Self-Attention has a quadratic relationship with the input sequence length, we can expect the inference time of our Joint model to be approximately  $O((k + 1)^2 L^2)$ , and  $O(4L^2 k)$  for the Pairwise Cross-Encoder. Thus, the ratio between the inference time of our Joint model and the Pairwise Cross-Encoder is:

$$r = \frac{O((k + 1)^2 L^2)}{O(4L^2 k)} = O(k) \quad (6.8)$$

However,  $k$  and  $L$  are usually small numbers and the Transformer contains also many affine transformations and an embedding layer that scale linearly in the input sequence length. We observe empirically that when fine-tuning on the AS2 datasets with a batch size of 32, there is only a slight increase in the average inference latency equal to 14.1% for Fixed and 9% for Flexible models. Flexible models are faster because they allow to shrink batches dynamically when the input examples are all shorter than the allowed maximum sequence length.



## Chapter 7

# Self-Supervised Objectives for AS2

In this Chapter, we study pre-training sentence-level objectives specifically designed for Answer Sentence Selection. Annotating large datasets is time-consuming and expensive (because of the large set of possible answer candidates retrieved for each question). For this reason, we aim at pre-training Transformer models that could be adapted to various AS2 datasets even if they contain a reduced number of training examples.

We propose three novel self-supervised sentence-level objectives that could be used in pre-training along with MLM to gather paragraph-level semantics from single or multiple documents into the model. The objectives we propose do not need additional expensive labeled datasets (such as in [Garg et al. \[2019\]](#)), but leverage semi-structured knowledge from large pre-training corpora such as Wikipedia or the CommonCrawl. In order to align with the sentence-pair structure of the Answer Sentence Selection task, our pre-training objectives are specifically designed to function with a pair of input text sequences.

The first objective can be summarized as predicting whether two random spans of text are extracted from the same paragraph of a document. The second objective instead provides the model with a paragraph and a sentence and should predict whether the sentence belongs to that paragraph in the original document. Finally, the third objective tasks the model to predict whether two paragraphs are extracted from the same original document.

We provide a detailed overview of the Transformer models used in this Chapter in [Section 3.1](#). Moreover, works related to the task we address here can be found in [Section 3.2.1](#), while for similar techniques that exploit the weak supervision of raw documents, the reader should refer to [Section 3.3.3](#).

By doing continuous pre-training starting from different pre-trained checkpoints such as RoBERTa [[Liu et al., 2019b](#)] and ELECTRA [[Clark et al., 2020](#)], we show that raw data can be exploited to increase final accuracy on different AS2 datasets. The next Section examines the reason why common pre-training tasks are not effective at structuring information from large pre-training corpora. Next, we describe our alternative objectives in detail. Then, we present the experimental setting and the results of our objectives when compared with common Transformer baselines. To help future research, we report a set of negative results in [Appendix D.1](#). Finally, we show how we achieve new state-of-the-art results on some datasets by continuously pre-training larger models with our newly proposed tasks.

## 7.1 The issue with common Objectives

Several Transformer models, including BERT, RoBERTa, ELECTRA, ALBERT, DeBERTa, and others, are trained using token-level objectives, sometimes supplemented with sentence-level

tasks. BERT, RoBERTa, ALBERT, and DeBERTa primarily employ the Masked Language Modeling (MLM) objective as their main training strategy, while ELECTRA focuses on the Token Detection task. BERT also incorporates the Next Sentence Prediction (NSP) objective to enhance performance, particularly in classification tasks involving two sentences, such as Natural Language Inferencing, AS2, or Textual Entailment. However, Liu et al. [2019b] demonstrated that by extending the training duration, NSP can be omitted without compromising final accuracy. Additionally, Lan et al. [2020] indicated that NSP is a trivial task with easily attainable high accuracies, resulting in weak error signals for the model. Consequently, they proposed Sentence Order Prediction (SOP) as a more challenging alternative to NSP.

In this Chapter, we demonstrate that token-level objectives are good at capturing short-range dependencies in the input text, but are not optimal for reasoning over inputs composed of separated spans of text, e.g., a question and an answer. Most masked tokens in MLM or replaced tokens in TD can be correctly predicted by looking at the surrounding context [Beltagy et al., 2020; Zaheer et al., 2020]. Moreover, we show that NSP and SOP are not challenging enough to train the model at reasoning over two input spans of text and do not align well with the final downstream AS2 task.

## 7.2 Alternative Objectives for AS2

Large sources of data such as Wikipedia, the CommonCrawl or the Pile are collections of documents, and each document is divided into a list of paragraphs by humans. The paragraphs of a single document generally describe the same argument from different perspectives. The objectives we propose in this Chapter exploit this weak supervision to create 3 self-supervised tasks that do not need manually annotated data.

We define pre-training corpus  $\mathbb{C}$  in the following way:

- the corpus is a set containing millions of documents:  $\mathbb{C} = \{\mathbf{D}_1, \dots, \mathbf{D}_k\}$ ;
- each document is a sequence of paragraphs:  $\mathbf{D}_i = [P_1, \dots, P_n]$ , where  $n$  is the number of paragraphs in  $\mathbf{D}_i$ ;
- paragraphs are lists of sentences:  $P_j = [s_1, \dots, s_m]$ , where  $m$  is the number of sentences in  $P_j$ .

We use the notation  $s_{ab}$  to indicate the subsequence of a paragraph  $P_j$  that starts from sentence  $a$  and ends with sentence  $b$ ,  $s_{ab} = s_a \cdot \dots \cdot s_b$  (we use “ $\cdot$ ” to denote text concatenation). Then, we define our objectives as follows.

### 7.2.1 Spans in the Same Paragraph (SSP)

The first task consists in detecting whether two spans of text  $s_{ab}$  and  $s_{cd}$  are extracted from the same paragraph  $P_j \in \mathbf{D}_i$ . More formally, given two subsequences  $s_{ab}, s_{cd}$  such that  $s_{ab} \cap s_{cd} = \emptyset$ , the task is to predict whether  $s_{ab}, s_{cd} \in P_j$ . We create positive example pairs by sampling two disjoint sequences  $s_{ab}$  and  $s_{cd}$  from the same paragraph  $P_j$  of some document  $\mathbf{D}_i$ . Then, we create  $k_1$  hard negatives by sampling  $s_{ab} \in P_j$  and  $s_{cd} \in P_l$  from two different paragraphs in the same document, i.e.  $P_j, P_l \in \mathbf{D}_i$  and  $j \neq l$ . This forces the model at recognizing that the general topic of the text spans is the same but described from different viewpoints. Finally, we create  $k_2$  easy negatives by sampling  $s_{ab} \in P_j$  and  $s_{cd} \in P_l$  such that  $P_j \in \mathbf{D}_i, P_l \in \mathbf{D}_f$  and  $i \neq f$ .

The task is designed to mimic the input of Transformer models for the Answer Sentence Selection task, which is composed of a question and a candidate answer. The question corresponds



History of Los Angeles. The area that became Los Angeles was claimed by Spain in 1542. It became a part of Mexico in 1821 following the Mexican War of Independence. At the end of the Mexican-American War, Los Angeles was purchased as part of the Treaty of Guadalupe Hidalgo and became part of the United States. Los Angeles was incorporated as a municipality in 1850. The discovery of oil in the 1890s brought rapid growth to the city. The city was further expanded with the completion of the Los Angeles Aqueduct in 1913.

Figure 7.1: Example of SSP objective.

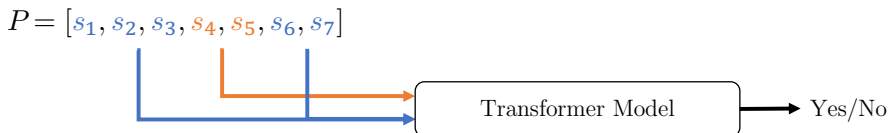
to  $s_{ab}$  while the answer is replaced by  $s_{cd}$ . Since  $s_{ab}$  and  $s_{cd}$  are two sequences of sentences in some paragraphs, we describe the sampling method we use to define their length. Questions in most cases are composed of a single sentence, with rare peaks of up to three sentences. Thus, we sample the length of  $s_{ab}$  in the interval  $[1, 3]$ , with weighted probabilities of 70%, 20% and 10%, respectively. Then, since most answers in the AS2 datasets are composed of one to five sentences, we randomly sample the length of  $s_{cd}$  in  $[1, 5]$  with probabilities of 14% for 1 and 5, and 24% for the other lengths.

Regarding the number of hard and easy negatives sampled for each positive pair, we try to obtain  $k_1 = 2$  hard negatives from within the same paragraph and  $k_2 = 2$  easy negatives from paragraphs in other documents. When there are not enough sentences in the paragraph to create  $k_1$  hard negatives, we force  $k_1 + k_2 = 4$  by sampling more easy negatives. An example of the SSP objective is depicted in Figure 7.1.

## 7.2.2 Spans in Paragraph (SP)

SP is a task that challenges the model at recognizing whether a span of text  $s_{ab}$  is extracted from some paragraph  $P_j \in \mathbf{D}_i$ . More specifically, we divide every paragraph  $P_j$  in the corpus into two subsets  $s_{ab}$  and  $P_j \setminus s_{ab}$ . Then, we create positive example pairs by sampling  $s_{ab}$  and  $P_j \setminus s_{ab}$ , with  $s_{ab} \in P_j$ . For the hard negatives, we use the same  $s_{ab} \in P_j$  but paired with other reduced paragraphs  $P_l \setminus s_{cd}$  in the same document, such that  $P_j, P_l \in \mathbf{D}_i$  and  $j \neq l$ . This should force the model at recognizing that the topic is the same but addressed from different perspectives. Finally, we create easy negatives by pairing  $s_{ab} \in P_j$  with reduced paragraphs  $P_l \setminus s_{cd}$  sampled from other documents, i.e.  $P_j \in \mathbf{D}_i, P_l \in \mathbf{D}_f$  such that  $i \neq f$ .

As in SSP, we define a sampling strategy for the length of  $s_{ab}$ . Since  $s_{ab}$  represents the question when fine-tuning on the downstream task, we apply the same rule of SSP, by sampling the length of  $s_{ab}$  in the interval  $[1, 3]$ , with weighted probabilities of 70%, 20% and 10%. The second part is always the text that remains in paragraph  $P_i$  after the split, and its length may



History of Los Angeles. The area that became Los Angeles was claimed by Spain in 1542. It became a part of Mexico in 1821 following the Mexican War of Independence. At the end of the Mexican-American War, Los Angeles was purchased as part of the Treaty of Guadalupe Hidalgo and became part of the United States. Los Angeles was incorporated as a municipality in 1850. The discovery of oil in the 1890s brought rapid growth to the city. The city was further expanded with the completion of the Los Angeles Aqueduct in 1913.

Figure 7.2: Example of SP objective.

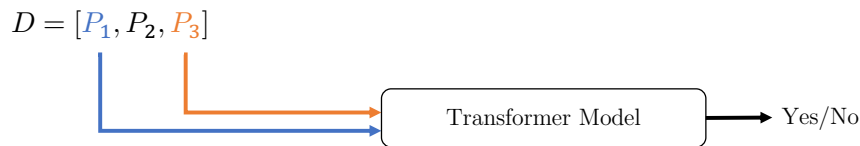
vary considerably from one corpus to another. For example, Wikipedia has longer paragraphs than CC-News and OpenWebText on average, leading to longer pairs. We do not set any constraint on the length of the right part of the pair because the tokenizer will take care of truncating sequences that exceed the maximum sequence length allowed.

As in SSP, we sample up to  $k_1 = 2$  hard negatives from within the same document and then we sample  $k_2$  easy negatives until  $k_1 + k_2 = 4$ . An example of the SP objective is depicted in Figure 7.2.

### 7.2.3 Paragraphs in the Same Document (PSD)

The last new pre-training task we propose asks the model to classify whether two entire paragraphs are sampled from the same document. Formally, positive pairs are created by sampling two paragraphs  $P_j$  and  $P_l$  such that  $P_j, P_l \in \mathbf{D}_i$  and  $j \neq l$ . In PSD there is no distinction between easy and hard negatives because there is no way of creating negatives from within the same document. We create negative pairs by sampling  $P_j$  and  $P_l$  from different documents, i.e.  $P_j \in \mathbf{D}_i$  and  $P_l \in \mathbf{D}_f$  given  $i \neq f$ .

We do not crop paragraphs to reduce their length because this task will be performed by the tokenizer, by iteratively removing one token from the longest sequence in the pair until the total length is within the limit. For each positive pair, we create  $k = 4$  additional negative examples. An example of PSD is given in Figure 7.3.



Los Angeles, often referred to by its initials L.A., is a major city in the U.S. state of California. With a 2020 population of 3,898,747, it is the largest city in the state, as well as the second-largest city in the United States following New York City.

The city of Los Angeles lies in a basin in Southern California. Adjacent to the Pacific Ocean extending through the Santa Monica Mountains and into the San Fernando Valley, it covers about 469 square miles (1,210 km<sup>2</sup>).

Los Angeles has a diverse economy with a broad range of industries. It has one of the busiest container ports in the Americas. The Los Angeles metropolitan area had a gross metropolitan product of over \$1.0 trillion in 2018.

Figure 7.3: Example of PSD objective.

## 7.3 Experiments

In this Section, we describe the experimental setting used to evaluate the effectiveness of our pre-training objectives.

### 7.3.1 Continuous Pre-Training

We do continuous pre-training starting from RoBERTa, ELECTRA, DeBERTa and DeBERTaV3 checkpoints. We apply all of our objectives to RoBERTa and ELECTRA, while for DeBERTa and DeBERTaV3, since they are more expensive to be continuously pre-trained, we apply only the objectives that show more consistent performance across all tasks, even though it may not always be the best. We also experiment by mixing the objectives and challenging the model



Model	Steps	LR	BS	MSL	# Tokens	Effort	Objectives
RoBERTa <sub>Base</sub> + SSP	400K	$1 \cdot 10^{-4}$	4096	128	210B	+2.5%	MLM (1.0) + SSP (1.0)
RoBERTa <sub>Base</sub> + SP	200K	$1 \cdot 10^{-4}$	4096	256	210B	+5.0%	MLM (1.0) + SP (1.0)
RoBERTa <sub>Base</sub> + PSD	200K	$1 \cdot 10^{-4}$	4096	256	210B	+5.0%	MLM (1.0) + PSD (1.0)
RoBERTa <sub>Base</sub> + ALL	400K	$1 \cdot 10^{-4}$	4096	128	210B	+2.5%	MLM (1.0) + ALL (1.0)
ELECTRA <sub>Base</sub> + SSP	400K	$1 \cdot 10^{-4}$	4096	128	210B	+5.0%	MLM (1.0) + TD (50.0) + SSP (1.0)
ELECTRA <sub>Base</sub> + SP	200K	$1 \cdot 10^{-4}$	4096	256	210B	+10.0%	MLM (1.0) + TD (50.0) + SP (1.0)
ELECTRA <sub>Base</sub> + PSD	200K	$1 \cdot 10^{-4}$	4096	256	210B	+10.0%	MLM (1.0) + TD (50.0) + PSD (1.0)
ELECTRA <sub>Base</sub> + ALL	400K	$1 \cdot 10^{-4}$	4096	128	210B	+5.0%	MLM (1.0) + TD (50.0) + ALL (1.0)
DeBERTa <sub>Base</sub> + ALL	100K	$1 \cdot 10^{-4}$	4096	128	52B	+1.25%	MLM (1.0) + SP (1.0)
DeBERTa <sub>Large</sub> + ALL	100K	$1 \cdot 10^{-4}$	4096	128	52B	+1.25%	MLM (1.0) + ALL (1.0)
DeBERTaV3 <sub>Base</sub> + ALL	100K	$1 \cdot 10^{-4}$	4096	128	52B	+0.625%	MLM (1.0) + TD (50.0) + ALL (1.0)
DeBERTaV3 <sub>Large</sub> + ALL	100K	$1 \cdot 10^{-4}$	4096	128	52B	+0.625%	MLM (1.0) + TD (50.0) + ALL (1.0)

Table 7.1: Hyper-parameters used for the continuous pre-training of several Transformer models with our specialized objectives. The “Effort” column indicates the amount of additional FLOPS required compared to the original pre-training. “MSL” is the maximum allowed input length. “LR” is the learning rate while “BS” refers to the training batch size. The last column describes the used training tasks along with their weight (in round brackets).

to automatically identify the task and predict the correct label. We mix the datasets for the continuous pre-training at the example level, thus a batch can contain samples from both SSP, SP or PSD, with equal probability. The other pre-training parameters are the same as SSP. Those experiments are indicated with “ALL” in the following Sections.

To avoid improvements deriving from the usage of more data, we exploit the same corpora used for the original pre-training of the models. In particular, we do continuous pre-training with SSP, SP and PSD on English Wikipedia, BookCorpus, OpenWebText and CC-News. For ELECTRA, the BookCorpus is not included in the original pre-training, however, it accounts for less than 0.5% of the training examples. After the pre-processing of the corpora with the techniques described above, we obtain about 420M examples for SP, 260M for SP and 200M for PSD.

In Table 7.1, we report the hyper-parameters used for each continuous pre-training experiment. For every model, we combine our objectives with the original tasks, i.e. MLM for RoBERTa and DeBERTa and TD for ELECTRA and DeBERTaV3. We found this beneficial since our long continuous pre-training may reduce models’ generalization by only performing binary classification on the output embedding of the first token.

Regarding the length of the continuous pre-training, we also report the total number of tokens provided to each model. For reference, RoBERTa<sub>Base</sub> and DeBERTaV3<sub>Base</sub> were pre-trained over about 2000B tokens, ELECTRA<sub>Base</sub> over more than 500B tokens and DeBERTa<sub>Base</sub> over about 1000B tokens. DeBERTa<sub>Large</sub> and DeBERTaV3<sub>Large</sub> were trained over the same amount of tokens of the corresponding Base versions. For more details about the original pre-trained models, refer to Section 3.1. Notice that for DeBERTaV3 models, we couldn’t exploit the original generators for Token Detection, as they were not released. We used DeBERTa<sub>Small</sub> and DeBERTaV3<sub>Base</sub> instead as the generators for DeBERTaV3<sub>Base</sub> and DeBERTaV3<sub>Large</sub>, respectively. We don’t think this issue has noticeable effects on final accuracy. The column “Effort” of the table shows the effective amount of computational FLOPS required for our continuous pre-training compared to the original model training. This ratio is computed by taking into account the number of total training steps, the batch size and the maximum sequence length. Notice that the Attention Mechanism complexity is quadratic in the input length. Thus, by training with a limit of 128 or 256, we reduce the computational complexity, and thus training time, by  $\frac{1}{16}$  and  $\frac{1}{4}$  respectively (original models were all trained with a sequence length of 512). To further demonstrate that our achievements do not stem from this additional short pre-training, we

Model	Accuracy	F1
RoBERTa <sub>Base</sub> + SSP	91.8	83.1
ELECTRA <sub>Base</sub> + SSP	90.4	79.9
RoBERTa <sub>Base</sub> + SP	91.3	83.3
ELECTRA <sub>Base</sub> + SP	89.9	80.1
RoBERTa <sub>Base</sub> + PSD	83.5	61.4
ELECTRA <sub>Base</sub> + PSD	82.3	57.1
BERT <sub>Base</sub> [Devlin et al., 2019] (NSP)	96.9	97.1
ALBERT <sub>Base</sub> [Lan et al., 2020] (SOP)	93.7	94.7
StructBERT [Wang et al., 2019b] (SSO)	92.3	93.4

Table 7.2: Accuracy and F1-score of several objectives on the pre-training validation set.

compare models continuously pre-trained with and without our objectives in Section 7.4.3.

Regarding the optimization and the hardware, we set  $\beta_1 = 0.9$  and  $\beta_2 = 0.99$ , the weight decay to 0.01 and we use a triangular learning rate with 10% warmup steps. We trained all models on our machine (see Section 1.2 for more details), and each experiment required about 2.5 days of continuous pre-training.

### 7.3.1.1 Comparison of objectives hardness

Here we compare common sentence-level objectives with our proposals in terms of final accuracy. As explained before, we state that NSP [Devlin et al., 2019], SOP [Lan et al., 2020], SSO [Wang et al., 2019b] and other sentence-level tasks are trivial and do not provide meaningful signals to the model. In Table 7.2 we report the accuracy on the development set after the continuous pre-training with our proposed objectives (SSP, SP and PSD) and the performance of other sentence-level tasks from the literature. Since we don’t have access to the pre-training development sets of BERT, ALBERT and StructBERT, the performance of NSP, SOP and SSO are measured over a small random shard of OpenWebText and CC-News, which were not used in the pre-training of those two models. Thus, we expect slightly higher performance on the original data because here they are tested out of domain.

The results indicate that performing sentence-level predictions is more challenging for SSP and SP compared to NSP, SOP and SSO. Surprisingly, solving PSD proves to be even more difficult, despite having access to two large paragraphs for predictions, unlike SSP and SP, which only rely on two small text spans. Notably, we find it intriguing that the accuracy of SSP and SP is similar. This similarity can be attributed to the fact that the two tasks are not significantly different in practice, considering that the average paragraph length in the pre-training corpora is 1.4 sentences. Consequently, after filtering the paragraphs that are too small, it frequently occurs that the remaining portion of the paragraph in SP, after removing the text span, is merely another short text span, similar to the case in SSP.

## 7.3.2 Fine-Tuning

To demonstrate the superiority of our approach, we fine-tune all the models listed in Table 7.1 on 6 different datasets for Answer Sentence Selection. Specifically, we use ASNQ, WikiQA, TREC-QA, WQA, NewsAS2 and TriviaAS2, which are all described in Section 4. We use the same hyper-parameters applied in the continuous pre-training apart from the learning rate, the batch size, the number of warmup steps of the triangular learning rate and the total number of epochs. In Table 7.3 we provide a summary of the search space of all hyper-parameters divided by dataset. For every experiment, the number of warmup steps is computed to match

Parameter	ASNQ	WikiQA & TREC-QA	WQA	NewsAS2 & TriviaAS2
Batch Size	{1024, 2048}	{16, 32, 64, 128}	{64, 128, 256, 512}	{128, 256, 512}
Learning Rate	$\{1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{5 \cdot 10^{-6}, 1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{5 \cdot 10^{-6}, 1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$
Max epochs	6	40	20	10

Table 7.3: Hyper-parameters search space for the fine-tuning over several AS2 datasets.

the number of training steps in the first epoch. We optimize the hyper-parameters using the development set and using only the baselines, thus the best hyper-parameters for the models continuously pre-trained with our objectives may vary slightly. We do not conduct an extensive grid search because it is too expensive.

Performance of each model is measured with common metrics for AS2: Mean Average Precision, Mean Reciprocal Rank and Precision@1, as in related works [Garg et al., 2019; Lauriola and Moschitti, 2021]. In particular, in the description of the results we focus on Precision@1 because this is the rate at which a correct answer is returned to the user.

## 7.4 Results

We divide results across the different models we used to improve readability. We begin by describing results on Base models and then we show the accuracy of Large architectures on the same tasks. Then, we demonstrate empirically that the improvements shown in this Chapter do not stem from the additional pre-training but rather from our harder tasks. Lastly, we offer insights into the reasons behind the varying performance of different objectives on different datasets. In the tables, the statistical significance is computed always with respect to the vanilla model of each block. We report results from the original papers when available, otherwise we fine-tuned the models on the different tasks.

### 7.4.1 Base Models

The results over Base models are reported in Tables 7.4 and 7.5 for ASNQ, WikiQA and TREC-QA, and in Tables 7.6 and 7.7 for WQA, NewsAS2 and TriviaAS2. Results are split across several tables for better readability. This Section proceeds by describing the results for each objective we propose.

**SSP** Our Spans in the Same Paragraph objective improves models’ accuracy on all tasks compared to the baselines, apart from the combination with RoBERTa and ELECTRA on TREC-QA. The advantages are more marked on small datasets, which confirms that our SSP objective is particularly effective when fine-tuning data is scarce. More on this in Section 7.4.4. For example, SSP with either RoBERTa or ELECTRA improves the Precision@1 by 2.9 points on ASNQ and by 4.6 points on WikiQA. In the second table, SSP shows minor improvements in Precision@1 on WQA, NewsAS2 and TriviaAS2 of 0.6, 0.2 and 1.9 points respectively.

**SP** The Spans in Paragraph objective provides significant improvements across all 6 datasets and combined with every architecture. For example, when applied to RoBERTa or ELECTRA, it improves the Precision@1 by up to 3.6 points on ASNQ, 4.7 on WikiQA and 1.4 on TriviaAS2. As before, the general pattern shows discrete improvements on large datasets and very high gains on small datasets such as WikiQA. We believe this is the most flexible objective, and we

Model	ASNQ			WikiQA			TREC-QA		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
TANDA RoBERTa <sub>Base</sub> ♣	-	-	-	88.5 (0.8)	89.9 (0.8)	83.0 (1.3)	90.1 (0.4)	94.1 (0.6)	89.7 (0.7)
CT RoBERTa <sub>Base</sub> [Soldaini et al., 2020]	66.3	65.4	53.2	89.9	91.0	-	92.4	96.7	-
TANDA RoBERTa <sub>Base</sub> [Garg et al., 2019]	-	-	-	88.9	90.1	-	91.4	95.2	-
RLAS-BIABC [Gharagozlou et al., 2022]	-	-	-	88.8 (3.6)	89.1 (1.7)	-	89.8 (1.5)	90.6 (9.2)	-
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	66.9 (0.1)	73.1 (0.1)	61.8 (0.1)	85.8 (1.3)	87.2 (1.3)	78.3 (2.8)	89.7 (0.7)	94.4 (1.1)	90.0 (1.9)
<b>RoBERTa<sub>Base</sub> + SSP</b>	<b>68.1</b> (0.2)	<b>74.5</b> (0.3)	<b>64.1</b> (0.3)	<b>88.7</b> (0.3)	<b>89.9</b> (0.4)	<b>82.9</b> (0.7)	89.3 (0.7)	93.6 (0.6)	88.5 (1.2)
<b>RoBERTa<sub>Base</sub> + SP</b>	<b>68.3</b> (0.1)	<b>74.5</b> (0.2)	<b>64.1</b> (0.2)	<b>87.7</b> (0.3)	<b>88.9</b> (0.4)	81.0 (0.8)	90.1 (0.8)	94.7 (1.3)	<b>90.9</b> (2.6)
<b>RoBERTa<sub>Base</sub> + PSD</b>	<b>67.7</b> (0.2)	<b>73.7</b> (0.3)	<b>62.6</b> (0.4)	86.4 (1.1)	88.0 (1.0)	80.5 (1.6)	<b>90.3</b> (0.5)	<b>95.1</b> (0.7)	90.3 (1.3)
<b>RoBERTa<sub>Base</sub> + ALL</b>	<b>68.0</b> (0.1)	<b>74.1</b> (0.2)	<b>63.9</b> (0.4)	<b>88.2</b> (0.4)	<b>89.5</b> (0.4)	<b>82.5</b> (0.9)	89.3 (0.7)	93.4 (0.6)	87.9 (1.2)
TANDA ELECTRA <sub>Base</sub> ♣	-	-	-	90.2 (0.8)	91.4 (0.7)	85.6 (1.1)	91.6 (0.7)	95.5 (0.7)	92.6 (1.5)
ELECTRA <sub>Base</sub> [Clark et al., 2020]	67.5 (0.2)	73.6 (0.2)	62.4 (0.4)	85.0 (2.6)	86.5 (2.7)	77.1 (4.0)	89.9 (0.4)	94.0 (0.9)	90.3 (1.7)
<b>ELECTRA<sub>Base</sub> + SSP</b>	<b>69.7</b> (0.2)	<b>75.7</b> (0.2)	<b>65.3</b> (0.3)	<b>88.6</b> (1.4)	<b>90.0</b> (1.4)	<b>82.5</b> (2.0)	89.6 (0.7)	93.5 (0.9)	88.5 (1.9)
<b>ELECTRA<sub>Base</sub> + SP</b>	<b>69.0</b> (0.1)	<b>75.1</b> (0.1)	<b>65.0</b> (0.2)	<b>88.1</b> (1.5)	89.5 (1.5)	81.8 (2.3)	90.3 (0.7)	94.6 (0.7)	91.2 (1.5)
<b>ELECTRA<sub>Base</sub> + PSD</b>	<b>68.9</b> (0.3)	<b>75.1</b> (0.3)	<b>65.3</b> (0.4)	85.6 (0.7)	87.3 (0.6)	78.6 (0.7)	87.9 (1.1)	92.2 (1.1)	85.9 (2.2)
<b>ELECTRA<sub>Base</sub> + ALL</b>	<b>69.3</b> (0.2)	<b>75.2</b> (0.2)	<b>65.0</b> (0.2)	<b>87.3</b> (1.2)	88.7 (1.1)	80.8 (1.9)	<b>90.4</b> (0.4)	<b>95.5</b> (1.0)	<b>92.6</b> (1.8)

Table 7.4: Results of our RoBERTa<sub>Base</sub> and ELECTRA<sub>Base</sub> models on ASNQ, WikiQA and TREC-QA. We report several baselines such as TANDA, Cascade Transformer (CT) and RLAS-BIABC. ♣: our versions of TANDA. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either RoBERTa<sub>Base</sub> or ELECTRA<sub>Base</sub>.

Model	ASNQ			WikiQA			TREC-QA		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Base</sub> [He et al., 2020]	68.2 (0.1)	74.4 (0.1)	64.4 (0.3)	88.0 (2.1)	<b>89.3</b> (1.9)	81.6 (3.1)	89.9 (1.2)	94.1 (1.2)	89.7 (2.5)
<b>DeBERTa<sub>Base</sub> + ALL</b>	<b>69.9</b> (0.2)	<b>76.3</b> (0.2)	<b>66.9</b> (0.3)	<b>88.2</b> (1.0)	<b>89.3</b> (0.9)	<b>82.4</b> (1.7)	<b>91.1</b> (0.9)	<b>94.6</b> (1.1)	<b>90.7</b> (2.2)
DeBERTaV3 <sub>Base</sub> [He et al., 2021]	70.5 (0.1)	77.1 (0.1)	67.8 (0.2)	88.9 (2.3)	90.1 (2.1)	83.0 (3.6)	90.0 (0.3)	93.2 (0.6)	89.1 (0.8)
<b>DeBERTaV3<sub>Base</sub> + ALL</b>	<b>71.0</b> (0.2)	<b>77.2</b> (0.2)	<b>68.2</b> (0.3)	<b>89.5</b> (1.4)	<b>90.7</b> (1.5)	<b>84.3</b> (2.2)	<b>91.3</b> (0.9)	<b>94.1</b> (1.2)	<b>91.5</b> (1.2)

Table 7.5: Results of our DeBERTa<sub>Base</sub> and DeBERTaV3<sub>Base</sub> models on ASNQ, WikiQA and TREC-QA. We underline statistically significant improvements over the corresponding baselines.

empirically confirm this claim, because it is composed of short spans of text that mimic the question and paragraphs gathered from multiple documents that mimic the answer.

**PSD** Models trained with Paragraph is the Same Document outperform the baselines on all datasets apart from the combination with ELECTRA on TREC-QA<sup>1</sup>. The improvements are consistent compared to the baselines, however, the average gains are smaller when compared with SSP or SP. We suspect that this is the consequence of PSD working at a higher level, by performing predictions over entire paragraphs, which may not align optimally with the AS2 task. Among the most significant results, we cite the 3.1 points of improvements in Precision@1 on ASNQ and the gains of 0.9 and 1.3 on WQA and TriviaAS2 respectively.

**ALL** In this experiment, we show the results of combining all the proposed objectives for a single continuous pre-training. When applied to RoBERTa, ELECTRA and DeBERTa, the results show improvements across all datasets and more agnostic results to the target tasks, which means that this objective should be the choice when testing on new datasets. Regarding the two DeBERTa architectures, both models augmented with the combination of SSP, SP and PSD outperform the baselines by a significant margin on all datasets. In particular, we report the gains in Precision@1 of 2.5 points on ASNQ, 2.4 on TREC-QA and 2.3 points on TriviaAS2.

<sup>1</sup>The results of ELECTRA+PSD on TREC-QA are significantly lower than the average, we suspect there was an issue with the model initialization for this experiment.

Model	WQA			NewsAS2			TriviaAS2		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
TANDA RoBERTa <sub>Base</sub> ♣	69.1 (0.1)	78.0 (0.2)	66.2 (0.2)	82.3 (0.2)	85.1 (0.3)	76.1 (0.5)	77.3 (0.5)	83.0 (0.6)	75.5 (0.6)
CT RoBERTa <sub>Base</sub> [Soldaini et al., 2020]	57.8	76.9	67.5	-	-	-	-	-	-
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	68.4 (0.1)	77.3 (0.2)	65.4 (0.2)	82.4 (0.2)	85.2 (0.3)	76.4 (0.6)	76.9 (0.6)	82.2 (0.5)	73.1 (0.5)
<b>RoBERTa<sub>Base</sub> + SSP</b>	<u>69.2</u> (0.1)	<u>77.7</u> (0.1)	<u>65.7</u> (0.2)	82.3 (0.3)	85.3 (0.3)	76.5 (0.4)	<b>77.5</b> (0.3)	<b>82.7</b> (0.4)	<b>74.3</b> (0.6)
<b>RoBERTa<sub>Base</sub> + SP</b>	<u>69.4</u> (0.1)	<u>78.0</u> (0.2)	<u>66.0</u> (0.4)	<b>82.7</b> (0.4)	<b>85.6</b> (0.4)	<b>77.2</b> (0.7)	77.1 (0.2)	<b>82.7</b> (0.3)	<u>74.0</u> (0.5)
<b>RoBERTa<sub>Base</sub> + PSD</b>	<u>69.4</u> (0.1)	<u>78.0</u> (0.1)	<u>66.0</u> (0.2)	82.4 (0.2)	85.0 (0.2)	76.3 (0.3)	76.0 (0.1)	81.5 (0.1)	72.2 (0.1)
<b>RoBERTa<sub>Base</sub> + ALL</b>	<b>69.5</b> (0.2)	<b>78.2</b> (0.1)	<b>66.2</b> (0.3)	82.2 (0.4)	85.1 (0.3)	76.2 (0.7)	77.0 (0.3)	82.4 (0.5)	73.6 (0.6)
TANDA ELECTRA <sub>Base</sub> ♣	70.7 (0.1)	79.5 (0.1)	68.2 (0.1)	82.3 (0.4)	85.1 (0.5)	76.4 (0.9)	74.2 (0.2)	80.1 (0.1)	70.6 (0.2)
ELECTRA <sub>Base</sub> [Clark et al., 2020]	70.1 (0.2)	78.6 (0.2)	66.9 (0.3)	82.0 (0.2)	84.8 (0.2)	76.0 (0.2)	73.3 (0.7)	79.1 (1.1)	68.9 (1.3)
<b>ELECTRA<sub>Base</sub> + SSP</b>	<u>70.6</u> (0.0)	<u>79.2</u> (0.1)	<u>67.5</u> (0.2)	82.2 (0.2)	85.0 (0.2)	76.2 (0.4)	<u>74.7</u> (0.5)	<u>80.4</u> (0.2)	<u>70.8</u> (0.3)
<b>ELECTRA<sub>Base</sub> + SP</b>	<u>70.6</u> (0.1)	<u>79.2</u> (0.2)	<u>67.5</u> (0.3)	<u>82.4</u> (0.1)	<u>85.2</u> (0.2)	76.5 (0.3)	<u>74.5</u> (0.3)	80.0 (0.4)	70.3 (0.8)
<b>ELECTRA<sub>Base</sub> + PSD</b>	<u>70.7</u> (0.1)	<u>79.2</u> (0.2)	<u>67.8</u> (0.4)	<b>82.5</b> (0.2)	<u>85.4</u> (0.2)	<u>76.8</u> (0.3)	<u>74.1</u> (0.3)	79.8 (0.3)	70.2 (0.3)
<b>ELECTRA<sub>Base</sub> + ALL</b>	<b>70.7</b> (0.0)	<b>79.4</b> (0.0)	<u>67.6</u> (0.1)	<b>82.5</b> (0.3)	<b>85.5</b> (0.3)	<b>76.9</b> (0.5)	<b>74.8</b> (0.5)	<b>80.7</b> (0.6)	<b>71.4</b> (0.9)

Table 7.6: Results of our RoBERTa<sub>Base</sub> and ELECTRA<sub>Base</sub> models on WQA, NewsAS2 and TriviaAS2. ♣: our versions of TANDA. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either RoBERTa<sub>Base</sub> or ELECTRA<sub>Base</sub>.

Model	WQA			NewsAS2			TriviaAS2		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Base</sub> [He et al., 2020]	69.5 (0.0)	78.5 (0.2)	66.6 (0.3)	<b>82.9</b> (0.5)	85.5 (0.5)	77.0 (0.8)	77.8 (0.2)	83.3 (0.4)	75.1 (0.4)
<b>DeBERTa<sub>Base</sub> + ALL</b>	<b>69.7</b> (0.1)	<b>78.6</b> (0.1)	<b>68.0</b> (0.2)	<b>82.9</b> (0.3)	<b>85.7</b> (0.4)	<b>77.3</b> (0.6)	<b>79.5</b> (0.3)	<b>84.2</b> (0.5)	<b>76.5</b> (0.8)
DeBERTaV3 <sub>Base</sub> [He et al., 2021]	70.7 (0.2)	78.9 (0.1)	67.8 (0.1)	82.8 (0.4)	85.6 (0.3)	77.2 (0.5)	80.4 (0.2)	85.1 (0.4)	77.2 (0.5)
<b>DeBERTaV3<sub>Base</sub> + ALL</b>	<b>71.1</b> (0.1)	<b>79.4</b> (0.1)	<b>68.4</b> (0.1)	<b>83.1</b> (0.2)	<b>85.8</b> (0.2)	<b>77.8</b> (0.6)	<b>81.1</b> (0.3)	<b>86.0</b> (0.6)	<b>79.5</b> (0.8)

Table 7.7: Results of our DeBERTa<sub>Base</sub> and DeBERTaV3<sub>Base</sub> models on WQA, NewsAS2 and TriviaAS2. We underline statistically significant improvements over the corresponding baselines.

Finally, notice that we match the performance of TANDA on several datasets such as WikiQA, WQA and NewsAS2, even outperforming it in some scenarios. TANDA uses ASNQ to transfer knowledge about the Answer Sentence Selection task before fine-tuning on smaller datasets, leading to higher final accuracy. We match the performance of TANDA in many experiments while using only unlabeled data that do not require expensive annotation processes. We compare our models also with the Cascade Transformer [Soldaini and Moschitti, 2020]. Notice that this technique performs an additional training step over ASNQ before fine-tuning over the target datasets, as TANDA does. We surpass the Cascade Transformer by a large margin on ASNQ with all of our continuously pre-trained models. Moreover, we surpass the performance of RLAS-BIABC with our DeBERTaV3<sub>Base</sub> + ALL model on both WikiQA and TREC-QA by a significant margin of 0.7 and 1.5 MAP points, respectively.

## 7.4.2 Large Models

In Tables 7.8 and 7.9 we report the results of combining larger models with our specialized objectives. On ASNQ, DeBERTa<sub>Large</sub> and DeBERTaV3<sub>Large</sub> provide an improvement of 0.7 and 1.7 points in Precision@1, respectively. To the best of our knowledge, DeBERTaV3<sub>Large</sub> + ALL is the actual state-of-the-art on ASNQ. On the two small datasets (WikiQA and TREC-QA), the improvements are smaller than with the Base models, because the baselines are already very capable of adapting to the AS2 task using less data, thanks to their larger size. Nevertheless, on TREC-QA and WikiQA we achieve the highest results to date among methods that do not exploit additional data, and the second place behind a TANDA model based on RoBERTa<sub>Large</sub> otherwise. As a last experiment, we combined our continuously pre-trained checkpoints with TANDA, by performing a transfer step over ASNQ with a batch size of 2048 and learning rate

Model	ASNQ			WikiQA			TREC-QA		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
TANDA BERT <sub>Large</sub> [Garg et al., 2019]	-	-	-	90.4	91.2	-	91.2	96.7	-
TANDA RoBERTa <sub>Large</sub> [Garg et al., 2019]	-	-	-	92.0	93.3	-	94.3	97.4	-
DeBERTa <sub>Large</sub> [He et al., 2020]	72.6 (0.2)	78.5 (0.2)	68.8 (0.3)	89.5 (0.3)	90.8 (0.4)	84.0 (0.7)	91.9 (0.2)	94.7 (1.1)	90.7 (1.7)
<b>DeBERTa<sub>Large</sub> + ALL</b>	<b>73.4</b> (0.1)	<b>78.9</b> (0.1)	<b>69.5</b> (0.1)	<u>90.1</u> (0.1)	91.2 (0.2)	<u>85.1</u> (0.5)	91.8 (1.1)	94.6 (0.7)	90.7 (1.7)
<b>TANDA DeBERTa<sub>Large</sub> + ALL</b>	-	-	-	<u>91.4</u> (1.0)	<u>92.4</u> (0.8)	<u>86.8</u> (1.6)	<b>92.3</b> (0.6)	<b>94.9</b> (1.6)	<b>93.1</b> (1.7)
DeBERTaV3 <sub>Large</sub> [He et al., 2021]	73.8 (0.1)	79.0 (0.1)	69.5 (0.1)	90.9 (1.2)	92.0 (1.1)	86.2 (2.0)	91.7 (0.7)	94.7 (0.5)	91.8 (0.8)
<b>DeBERTaV3<sub>Large</sub> + ALL</b>	<b>74.3</b> (0.2)	<b>80.0</b> (0.2)	<b>71.2</b> (0.4)	91.6 (1.2)	92.6 (1.4)	87.3 (1.9)	92.3 (0.3)	94.6 (0.7)	93.8 (1.6)
<b>TANDA DeBERTaV3<sub>Large</sub> + ALL</b>	-	-	-	<b>92.7</b> (0.7)	<b>93.9</b> (0.8)	<b>89.5</b> (1.5)	<b>95.4</b> (0.7)	<b>98.4</b> (0.9)	<b>97.1</b> (1.5)

Table 7.8: Results of our Large models on ASNQ, WikiQA and TREC-QA. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either DeBERTa<sub>Large</sub> or DeBERTaV3<sub>Large</sub>.

Model	WQA			NewsAS2			TriviaAS2		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Large</sub> [He et al., 2020]	72.6 (0.0)	80.8 (0.1)	69.7 (0.1)	<b>84.1</b> (0.4)	<b>86.7</b> (0.3)	78.6 (0.5)	82.7 (0.3)	86.8 (0.5)	80.2 (0.9)
<b>DeBERTa<sub>Large</sub> + ALL</b>	<b>73.2</b> (0.0)	<b>81.3</b> (0.1)	<b>70.3</b> (0.1)	<b>84.1</b> (0.1)	<b>86.7</b> (0.2)	<b>78.9</b> (0.4)	<b>83.2</b> (0.3)	<b>87.5</b> (0.3)	<b>81.4</b> (0.5)
DeBERTaV3 <sub>Large</sub> [He et al., 2021]	72.4 (0.2)	80.3 (0.2)	69.5 (0.3)	<b>85.0</b> (0.1)	87.5 (0.1)	80.2 (0.3)	82.5 (0.4)	87.0 (0.7)	80.5 (1.1)
<b>DeBERTaV3<sub>Large</sub> + ALL</b>	<b>73.5</b> (0.2)	<b>81.3</b> (0.3)	<b>70.9</b> (0.5)	<b>85.0</b> (0.3)	<b>87.6</b> (0.3)	<b>80.5</b> (0.4)	<b>84.3</b> (0.4)	<b>88.4</b> (0.3)	<b>82.7</b> (0.6)

Table 7.9: Results of our Large models on WQA, NewsAS2 and TriviaAS2. We underline statistically significant improvements over the corresponding baselines.

of  $10^{-5}$  (other hyper-parameters are the same of Table 7.3), followed by the fine-tuning over WikiQA and TREC-QA. Our TANDA DeBERTaV3<sub>Large</sub> + ALL achieves state-of-the-art accuracy on both WikiQA and TREC-QA, with an impressive MAP of 92.7 and 95.4, respectively.

Regarding WQA, NewsAS2 and TriviaAS2, we show competitive performance on all benchmarks, outperforming the baselines in the best scenario by 1.4 points of P@1 on ASNQ, 0.3 on NewsAS2 and 2.2 points on TriviaAS2.

### 7.4.3 Only MLM Continuous Pre-Training

Model	ASNQ		WikiQA		TREC-QA		WQA	
	MAP	P@1	MAP	P@1	MAP	P@1	MAP	P@1
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	66.9 (0.1)	61.8 (0.1)	85.8 (1.3)	78.3 (2.8)	89.7 (0.7)	90.0 (1.9)	68.4 (0.1)	65.4 (0.2)
RoBERTa <sub>Base</sub> + SSP Data only	<u>67.1</u> (0.1)	<u>63.4</u> (0.3)	84.5 (0.7)	76.7 (0.9)	88.8 (0.6)	87.4 (1.3)	68.1 (0.2)	64.6 (0.3)
RoBERTa <sub>Base</sub> + SP Data only	<u>67.2</u> (0.2)	<u>62.8</u> (0.2)	84.7 (0.8)	76.8 (1.6)	89.8 (0.3)	88.8 (1.3)	67.8 (0.1)	64.0 (0.2)
RoBERTa <sub>Base</sub> + PSD Data only	<u>67.3</u> (0.2)	<b>64.1</b> (0.3)	85.6 (1.4)	79.1 (1.6)	89.6 (1.0)	87.1 (2.8)	67.9 (0.2)	63.5 (0.2)
<b>RoBERTa<sub>Base</sub> + SSP</b>	<u>68.1</u> (0.2)	<b>64.1</b> (0.3)	<b>88.7</b> (0.3)	<b>82.9</b> (0.7)	89.3 (0.7)	88.5 (1.2)	<u>69.2</u> (0.1)	<u>65.7</u> (0.2)
<b>RoBERTa<sub>Base</sub> + SP</b>	<b>68.3</b> (0.1)	<b>64.1</b> (0.2)	<u>87.7</u> (0.3)	81.0 (0.8)	90.1 (0.8)	<b>90.9</b> (2.6)	<b>69.4</b> (0.1)	<b>66.0</b> (0.4)
<b>RoBERTa<sub>Base</sub> + PSD</b>	<u>67.7</u> (0.2)	<u>62.6</u> (0.4)	<u>86.4</u> (1.1)	80.5 (1.6)	<b>90.3</b> (0.5)	90.3 (1.3)	<b>69.4</b> (0.1)	<b>66.0</b> (0.2)

Table 7.10: Comparison between pre-training with our specialized objectives and only over the data used for the continuous pre-training, but without supervision. We omit the MRR metric for better readability. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either RoBERTa<sub>Base</sub>.

In this Section, we compare models continuously pre-trained using only the MLM objective and with the combination of MLM and our proposed tasks. We show that SSP, SP and PSD are of fundamental importance to reach the highest accuracy on Answer Selection datasets and that improvements should not be attributed to the additional data. Table 7.10 summarizes the results on ASNQ, WikiQA, TREC-QA and WQA of our continuously pre-trained models compared with similar models trained over the same data but using only the MLM objective.

Our proposals consistently outperform models trained solely with the MLM objective across all tasks, even outperforming the baseline on smaller datasets. This can be attributed to the negative impact of resetting optimizer and scheduler states when training resumes. Additionally, adapting partitioned schedulers and optimizer states to new training configurations with a different number of training devices would have been challenging, even if the original complete checkpoints were available.

#### 7.4.4 Few-Shot Setting

This Section analyzes the performance of our models in a few-shot scenario, demonstrating how our specialized continuous pre-training using unlabeled data can help when target data is scarce.

In the Tables 7.11, 7.12, 7.13 and 7.14 we show the performance of our continuously pre-trained models in the few-shot setting. In this experiment, we reduce the size of the training sets by 1000 times for ASNQ and by 20 and 50 times for WikiQA and TREC-QA. Regarding WQA, NewsAS2 and TriviaAS2, the reduction factor is set to 150 for the former and 1000 for the last two. This leaves us with about 20K examples for ASNQ, 1000 for WikiQA, TREC-QA and WQA and 1800 for TriviaAS2 and NewsAS2. In order to ensure model convergence, we keep a higher number of examples for ASNQ. This adjustment is necessary due to the significantly small ratio of positive to negative examples, which is approximately 1 in 350. We leave the validation and test sets unchanged in this experiment to better compare with previous results.

The results show that on average, our models outperform the baselines by a large margin featuring also a much lower standard deviation when the training data is scarce, thus confirming the benefits of our continuous pretraining. For example, on ASNQ and using Base models, our objectives increase the P@1 over the baselines from 6 to 24 points. On DeBERTa<sub>Large</sub> and DeBERTaV3<sub>Large</sub> the improvements are smaller but consistent, with gains of about 4 points in P@1.

WikiQA and TREC-QA have been reduced fewer times and thus improvements are more subtle. Nevertheless, every model we train with our specialized objective outperforms the corresponding baseline by a significant margin, with gains of up to 17 points in P@1. Notice that ELECTRA is the weaker model across all tasks. We believe that this is due to the shorted pre-training and thus it struggles at adapting to new tasks with less supervision. As a comparison, ELECTRA’s number of pre-training tokens is half of those seen by DeBERTa and  $\frac{1}{4}$  compared to RoBERTa and DeBERTaV3.

On WQA and NewsAS2, the results follow a similar trend. In the few-shot setting over WQA, which is the harder task, our continuously pre-trained models outperform the baselines from 2 to 10 points in Precision@1. For example, RoBERTa<sub>Base</sub> + PSD achieves 6.2 more points in P@1 when compared with the original RoBERTa<sub>Base</sub>. On NewsAS2, improvements range from 2 to 16 points, highlighting again how our specialized continuous pre-training is essential when target data is scarce.

Finally, we notice how our objectives combined with RoBERTa and ELECTRA struggle at improving performance on TriviaAS2. We suspect there is a misalignment between the structure of our task and the way data have been collected in the TriviaAS2 corpora. Since our objective show significant improvements when trained on the full dataset, we think that the reduced dataset is not large enough to correct the misalignment mentioned before.

Model	ASNQ			WikiQA			TREC-QA		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	31.0 (1.7)	36.8 (1.9)	23.1 (1.6)	71.8 (1.3)	73.0 (1.4)	57.8 (1.9)	77.6 (0.1)	86.1 (1.3)	77.5 (2.2)
<b>RoBERTa<sub>Base</sub> + SSP</b>	<b>37.9</b> (0.3)	<b>44.4</b> (0.3)	<b>31.1</b> (0.3)	<u>75.6</u> (1.1)	<u>76.8</u> (1.0)	<u>63.6</u> (1.6)	<u>79.2</u> (0.2)	86.5 (0.1)	79.4 (0.0)
<b>RoBERTa<sub>Base</sub> + SP</b>	36.8 (0.1)	43.8 (0.2)	29.8 (0.6)	72.8 (1.1)	74.0 (1.2)	60.5 (1.4)	<b>82.3</b> (0.7)	88.7 (0.5)	80.4 (0.8)
<b>RoBERTa<sub>Base</sub> + PSD</b>	<u>34.7</u> (0.5)	<u>41.3</u> (0.3)	<u>27.2</u> (0.2)	68.1 (0.6)	70.2 (0.6)	54.6 (0.5)	<u>81.1</u> (1.9)	<u>88.7</u> (1.8)	<u>81.5</u> (3.1)
<b>RoBERTa<sub>Base</sub> + ALL</b>	<u>37.7</u> (0.1)	<u>44.3</u> (0.1)	<u>30.2</u> (0.1)	<b>76.7</b> (0.2)	<b>78.1</b> (0.2)	<b>66.1</b> (0.2)	<u>82.2</u> (0.2)	<b>90.8</b> (0.5)	<b>85.8</b> (0.8)
ELECTRA <sub>Base</sub> [Clark et al., 2020]	17.3 (1.7)	21.2 (2.0)	10.3 (1.7)	63.8 (0.3)	65.1 (0.3)	49.1 (1.7)	72.7 (1.3)	85.0 (1.6)	75.0 (2.9)
<b>ELECTRA<sub>Base</sub> + SSP</b>	<u>40.6</u> (0.5)	<u>47.8</u> (0.5)	<u>33.6</u> (0.7)	<u>75.7</u> (0.1)	<u>76.6</u> (0.1)	<u>64.0</u> (0.5)	<u>78.4</u> (0.4)	<u>87.1</u> (0.5)	77.5 (0.8)
<b>ELECTRA<sub>Base</sub> + SP</b>	<u>39.4</u> (0.4)	<u>45.8</u> (0.4)	<u>31.2</u> (0.4)	<u>76.6</u> (0.7)	<b>78.4</b> (0.7)	<b>66.4</b> (1.2)	<u>82.0</u> (0.5)	<u>88.4</u> (0.7)	<u>80.9</u> (1.5)
<b>ELECTRA<sub>Base</sub> + PSD</b>	<u>35.8</u> (0.9)	<u>42.9</u> (0.8)	<u>29.2</u> (0.8)	<u>74.9</u> (0.2)	<u>75.9</u> (0.2)	<u>62.2</u> (0.2)	<b>82.6</b> (0.3)	<b>88.5</b> (0.7)	<b>81.6</b> (1.5)
<b>ELECTRA<sub>Base</sub> + ALL</b>	<u>41.5</u> (0.4)	<u>48.3</u> (0.5)	<u>34.4</u> (0.7)	<b>76.7</b> (0.6)	<b>77.8</b> (0.6)	<b>63.9</b> (1.1)	<u>82.0</u> (0.2)	<u>88.3</u> (0.6)	<u>81.4</u> (0.8)

Table 7.11: **Few-shot** setting results of RoBERTa<sub>Base</sub> and ELECTRA<sub>Base</sub> models over ASNQ, WikiQA and TREC-QA. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either RoBERTa<sub>Base</sub> or ELECTRA<sub>Base</sub>.

Model	ASNQ			WikiQA			TREC-QA		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Base</sub> [He et al., 2020]	34.6 (1.1)	40.8 (1.0)	26.2 (0.9)	72.3 (2.2)	73.5 (2.2)	57.8 (3.0)	80.1 (1.0)	88.3 (0.9)	79.4 (1.5)
<b>DeBERTa<sub>Base</sub> + ALL</b>	<b>40.0</b> (0.6)	<b>47.0</b> (0.8)	<b>32.9</b> (1.2)	<b>77.6</b> (0.1)	<b>78.9</b> (0.1)	<b>67.2</b> (0.2)	<b>83.4</b> (0.0)	<b>89.9</b> (0.1)	<b>82.4</b> (0.2)
DeBERTa <sub>Large</sub> [He et al., 2020]	43.9 (0.5)	50.7 (0.6)	36.4 (0.7)	75.7 (1.0)	<b>77.0</b> (1.0)	63.9 (1.5)	82.4 (3.0)	90.8 (0.7)	83.3 (0.8)
<b>DeBERTa<sub>Large</sub> + ALL</b>	<u>45.9</u> (0.1)	<b>53.5</b> (0.2)	<u>40.1</u> (0.5)	<b>76.1</b> (1.4)	<b>77.0</b> (1.4)	<b>64.6</b> (1.9)	<b>86.6</b> (0.1)	<b>92.3</b> (0.5)	<b>85.8</b> (0.8)
DeBERTaV3 <sub>Base</sub> [He et al., 2021]	21.7 (7.8)	26.4 (9.2)	14.8 (7.7)	65.3 (2.3)	67.1 (2.3)	51.8 (3.3)	69.3 (1.7)	79.4 (3.0)	66.7 (4.7)
<b>DeBERTaV3<sub>Base</sub> + ALL</b>	<b>37.8</b> (0.7)	<b>45.0</b> (0.9)	<b>31.5</b> (0.8)	<u>74.4</u> (0.2)	<u>75.4</u> (0.2)	<b>61.0</b> (0.2)	<b>82.3</b> (0.3)	<b>88.5</b> (0.2)	<b>80.9</b> (0.0)
DeBERTaV3 <sub>Large</sub> [He et al., 2021]	43.5 (5.0)	50.5 (5.7)	36.1 (6.0)	69.5 (1.3)	71.0 (1.2)	57.4 (2.5)	83.0 (2.3)	89.5 (1.5)	81.4 (2.2)
<b>DeBERTaV3<sub>Large</sub> + ALL</b>	<b>45.7</b> (0.6)	<b>53.0</b> (0.7)	<b>40.5</b> (0.9)	<b>78.8</b> (0.9)	<b>79.6</b> (0.8)	<b>68.1</b> (1.0)	<b>87.7</b> (0.2)	<b>92.5</b> (0.4)	<b>87.7</b> (0.8)

Table 7.12: **Few-shot** setting results of DeBERTa and DeBERTaV3 models over ASNQ, WikiQA and TREC-QA. We underline statistically significant improvements over the corresponding baselines.

Model	WQA			NewsAS2			TriviaAS2		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	58.4 (0.3)	66.8 (0.2)	51.1 (0.2)	59.3 (1.7)	63.8 (2.0)	46.0 (3.5)	<b>44.2</b> (1.0)	<b>56.3</b> (1.5)	<b>40.0</b> (2.1)
<b>RoBERTa<sub>Base</sub> + SSP</b>	<u>60.3</u> (0.0)	<u>68.9</u> (0.2)	<u>54.0</u> (0.3)	<u>65.5</u> (0.1)	<u>70.1</u> (0.1)	<u>55.9</u> (0.2)	40.3 (0.1)	52.8 (0.4)	37.3 (0.5)
<b>RoBERTa<sub>Base</sub> + SP</b>	60.0 (0.2)	68.4 (0.2)	53.0 (0.3)	<b>67.5</b> (0.0)	<b>72.2</b> (0.0)	<b>58.1</b> (0.1)	42.4 (0.2)	54.0 (0.4)	37.9 (0.4)
<b>RoBERTa<sub>Base</sub> + PSD</b>	<b>62.6</b> (0.1)	<b>71.5</b> (0.2)	<b>57.3</b> (0.3)	<u>62.1</u> (0.2)	<u>66.6</u> (0.2)	<u>51.6</u> (0.1)	38.4 (0.3)	49.4 (0.1)	32.6 (0.3)
<b>RoBERTa<sub>Base</sub> + ALL</b>	<u>59.6</u> (0.2)	<u>68.0</u> (0.3)	<u>52.6</u> (0.4)	<u>66.5</u> (0.0)	<u>71.2</u> (0.1)	<u>56.7</u> (0.0)	41.2 (0.1)	52.8 (0.3)	37.0 (0.6)
ELECTRA <sub>Base</sub> [Clark et al., 2020]	56.1 (1.5)	64.9 (1.3)	49.1 (1.3)	53.9 (3.0)	58.4 (3.2)	39.3 (3.2)	41.7 (0.6)	53.4 (0.7)	37.3 (1.2)
<b>ELECTRA<sub>Base</sub> + SSP</b>	<u>61.6</u> (0.1)	<u>70.5</u> (0.2)	<u>56.1</u> (0.3)	<u>65.3</u> (0.1)	<u>69.7</u> (0.2)	<u>54.8</u> (0.4)	41.4 (0.1)	54.1 (0.1)	37.8 (0.0)
<b>ELECTRA<sub>Base</sub> + SP</b>	<u>61.6</u> (0.1)	<u>69.4</u> (0.1)	<u>54.1</u> (0.2)	<b>65.7</b> (0.1)	<b>70.3</b> (0.2)	<b>55.2</b> (0.2)	<u>42.6</u> (0.2)	52.7 (0.4)	35.6 (0.6)
<b>ELECTRA<sub>Base</sub> + PSD</b>	<b>64.3</b> (0.0)	<b>72.9</b> (0.0)	<b>59.1</b> (0.1)	<u>62.0</u> (0.1)	<u>67.0</u> (0.1)	<u>51.3</u> (0.3)	38.8 (0.1)	49.6 (0.0)	33.5 (0.3)
<b>ELECTRA<sub>Base</sub> + ALL</b>	<u>63.5</u> (0.1)	<u>71.6</u> (0.1)	<u>56.8</u> (0.3)	<u>65.5</u> (0.3)	<u>69.8</u> (0.3)	<u>54.7</u> (0.6)	<b>43.3</b> (0.2)	<b>55.1</b> (0.4)	<b>38.6</b> (0.6)

Table 7.13: **Few-shot** setting results of RoBERTa<sub>Base</sub> and ELECTRA<sub>Base</sub> models over WQA, NewsAS2 and TriviaAS2. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either RoBERTa<sub>Base</sub> or ELECTRA<sub>Base</sub>.

Model	WQA			NewsAS2			TriviaAS2		
	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Base</sub> [He et al., 2020]	61.4 (0.9)	70.4 (0.8)	55.1 (1.1)	64.5 (1.6)	69.3 (1.6)	54.4 (2.1)	42.4 (0.3)	54.1 (0.3)	38.4 (0.4)
<b>DeBERTa<sub>Base</sub> + ALL</b>	<b>61.9</b> (0.1)	<b>70.8</b> (0.1)	<b>56.4</b> (0.3)	<b>66.4</b> (0.1)	<b>71.0</b> (0.1)	<b>56.2</b> (0.2)	<b>43.9</b> (0.2)	<b>55.1</b> (0.2)	<b>39.2</b> (0.3)
DeBERTa <sub>Large</sub> [He et al., 2020]	<b>63.9</b> (0.2)	<b>72.9</b> (0.3)	<b>59.5</b> (0.5)	66.5 (3.3)	<b>71.5</b> (3.2)	56.4 (3.0)	41.9 (0.5)	52.8 (0.6)	36.7 (0.7)
<b>DeBERTa<sub>Large</sub> + ALL</b>	63.2 (0.3)	72.1 (0.3)	58.2 (0.5)	<b>66.8</b> (0.3)	71.4 (0.3)	<b>57.7</b> (0.4)	<b>45.0</b> (2.3)	<b>57.5</b> (2.1)	<b>42.8</b> (2.5)
DeBERTaV3 <sub>Base</sub> [He et al., 2021]	59.1 (3.2)	68.0 (3.5)	53.1 (4.5)	56.4 (5.9)	60.5 (6.3)	43.4 (7.9)	43.6 (2.6)	56.2 (2.8)	41.2 (2.5)
<b>DeBERTaV3<sub>Base</sub> + ALL</b>	<b>64.8</b> (0.1)	<b>73.3</b> (0.2)	<b>59.6</b> (0.2)	<b>66.5</b> (0.2)	<b>71.3</b> (0.2)	<b>57.4</b> (0.2)	<b>45.6</b> (0.2)	<b>57.9</b> (0.2)	<b>41.8</b> (0.3)
DeBERTaV3 <sub>Large</sub> [He et al., 2021]	62.4 (1.7)	71.2 (1.8)	57.1 (2.6)	64.3 (4.4)	68.9 (4.5)	54.8 (5.3)	43.1 (0.8)	55.8 (1.8)	41.0 (2.5)
<b>DeBERTaV3<sub>Large</sub> + ALL</b>	<b>64.3</b> (0.2)	<b>73.1</b> (0.3)	<b>59.7</b> (0.4)	<b>67.5</b> (0.1)	<b>72.3</b> (0.1)	<b>58.8</b> (0.1)	<b>46.1</b> (0.5)	<b>58.9</b> (0.3)	<b>43.4</b> (0.3)

Table 7.14: **Few-shot** setting results of DeBERTa and DeBERTaV3 models over WQA, NewsAS2 and TriviaAS2. We underline statistically significant improvements over the baselines.



### 7.4.5 General Considerations

In this section, we provide an overview of the general performance of our models across various datasets and elucidate the factors contributing to the slight improvements observed in some scenarios.

**Objective-dataset alignment** Our objectives provide significant advantages on some datasets and more limited gains on others. First, consider that among the 6 datasets we use for the evaluation, 4 of them gather answer candidates from within the same document (ASNQ, WikiQA, NewsQAS2, TriviaAS2) while the other two contain answer candidates extracted from several documents across the web. We suspect this difference influences the performance of our models when trained with SSP, SP and PSD. The reason is that while SSP and SP work more at the paragraph/sub-paragraph level, PSD is trained only over paragraphs across different documents. For this reason, PSD aligns generally better than the other techniques on WQA and TREC-QA, while SSP and SP feature the greatest advantages on the other datasets. As a rule of thumb, we suggest exploiting PSD when the target task contains answer candidates gathered from multiple documents and when the question or the answer is longer than a pair of sentences. Otherwise, we suggest using SSP or SP. If the target dataset is very heterogeneous or not known in advance, we suggest choosing the models trained with the combination of our objectives (“ALL”) instead, as they provide more stable results across all datasets.

**Limited improvements** The gains on WQA and TREC-QA are limited when using our techniques or even state-of-the-art methods, such as TANDA. For TREC-QA, this can be explained by noticing that we are close to a perfect re-ranking, and thus improving further is difficult. In [Soldaini and Moschitti \[2020\]](#), the authors manually investigated the TREC-QA dataset and discovered that a few answer candidates are not annotated correctly. Considering also that the number of questions in the test set is only 68, this explains why it is hard to obtain significant gains on this dataset.

Regarding WQA, this dataset contains answer candidates extracted from billions of documents across the web. The answer candidates contain many typos, are sometimes malformed and do not always have correct annotations. For those reasons, our methods and other state-of-the-art methods struggle at improving the output ranking.



## Chapter 8

# Self-supervised objectives for Contextual AS2

In this research, we study the effects of providing Answer Selection architectures with additional context to improve accuracy in candidates’ evaluation.

Previous studies [Han et al., 2021; Lauriola and Moschitti, 2021] have exploited pre-trained Transformer encoders directly for Contextual AS2 tasks by fine-tuning them on inputs containing multiple sentences with different roles: the question, the answer candidate, and the surrounding context (previous and following sentences). However, this structured input poses challenges during fine-tuning as standard pre-training approaches do not align well with the downstream Contextual AS2 task. The language model lacks knowledge of the role of each sentence, requiring the learning of extended sentence-level embeddings during fine-tuning, which has shown empirical underperformance. This issue becomes more pronounced with limited fine-tuning data, indicating difficulties in leveraging context effectively.

To address these challenges, we propose three pre-training objectives that align structurally with the final Contextual AS2 task, which is described in Section 2.1.1.2. These objectives leverage the mild form of supervision contained in large datasets of raw text to pre-train the context slots in the Transformer text input. Specifically, we exploit the human subdivision of large documents in smaller paragraphs, which address the same general topic from different perspectives.

We report a description of the models used for the experiments in Section 3.1, while for works related to the Answer Sentence Selection and the Contextual AS2 tasks, refer to Section 3.2.1. Additionally, we provide an overview of related research that exploits the document structure in Section 3.3.3.

The effectiveness of our strategies is evaluated on 7 datasets using several popular pre-trained Transformers. The experimental results demonstrate that our approaches, which incorporate structural pre-training, effectively adapt Transformers to process contextualized input. Compared to the baselines, our methods achieve up to an 8% improvement in accuracy on certain downstream datasets.

## 8.1 Contextual Models

Contextual models process 3 inputs at a time: the question, the answer candidate and the context. Contrary to the Jointwise models discussed in Chapter 6, Contextual models focus on a single answer candidate and leverage the additional context to enhance the ranking accuracy.

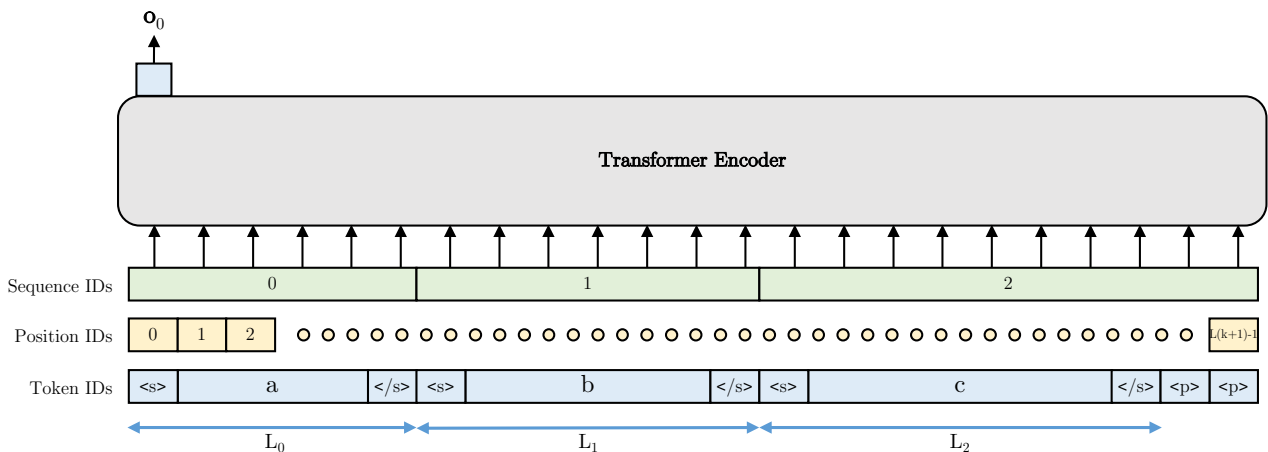


Figure 8.1: Example of Contextual model.

### 8.1.1 Injecting Context into Transformers

Common pre-trained Transformers such as BERT [Devlin et al., 2019] and RoBERTa [Liu et al., 2019b] struggle at reasoning over more than 2 spans of text. The reason is that they were trained either only with token-level objectives (e.g. MLM, TD) or with a combination of token-level and sentence-level objectives over two text spans, such as NSP or SOP. We invite the reader to refer to Section 2.2.2 and 7.1 for further information about the common pre-training objectives and their issues, respectively.

We extend the original Transformer and tokenizer inputs by incorporating a third slot for context. An illustration of an extended model for Contextual training is shown in Figure 8.1. As depicted in the picture, there exists a token type id for each distinct input, and these embeddings require training either during the pre-training or directly in fine-tuning. Additionally, the pre-trained model checkpoints available from recent works do not encompass information about the contextual input. To address these concerns, in the next Section we propose several pre-training strategies for Contextual models pre-training.

## 8.2 Context-aware Pre-training Objectives

Contextual models are architectures designed to process more than 2 inputs efficiently. This poses practical challenges because common Transformer models are pre-trained on 1 or 2 spans of text at a time [Clark et al., 2020; Devlin et al., 2019; Liu et al., 2019b]. Thus, it is difficult to accustom those architectures to reason over many text spans only while fine-tuning. With the term “span”, we mean a sequence of 1 or more sentences within the same document. Notice that even if most Transformer models are trained over long spans of text, there is no strong supervision over the relations between the different parts of the input. For example, the NSP and SOP objective of BERT and ALBERT tasks the model only to predict the order of 2 long text spans.

For those reasons, we propose 3 alternative pre-training objectives that are designed to adapt the model at exploiting additional context to better re-rank answer candidates in fine-tuning. The proposed objectives are extensions of SSP, which is described in Chapter 7. We explain the intuition behind SSP with an example.

Consider a passage from Wikipedia that consists of three sentences:

- $s_1$ : Harry lives in a cupboard under the stairs in the house of the Dursleys, his aunt, uncle, and cousin, Dudley.
- $s_2$ : At the age of 11, Harry starts attending Hogwarts School of Witchcraft and Wizardry, where he takes classes in Charms, Transfiguration, and others.
- $s_3$ : He also participates in Potions lessons, where he encounters Professor Severus Snape, who immediately displays a dislike for him.

When presented with a question like “Which classes did Harry take in Hogwarts?”, a conventional language model can easily identify answers that follow the format “ $X$  takes classes in  $Y$ ”. It achieves this by matching the subject of the question with the object of the answer, focusing on the shared predicate of taking classes. However, the same language model would struggle to select answers of the form “ $X$  participates in  $Y$ ”, as it requires understanding the entire predicate’s argument structure of taking vs participating. By pre-training a language model using the SSP task, it can learn the implicit connections between concepts mentioned in  $s_3$ , such as “participating in Potions lessons”, and the concept of “taking classes in Charms, Transfiguration, and others” mentioned in  $s_2$ . These sentences belong to the same paragraph, enabling the model to reason about concepts and establish relationships between entities.

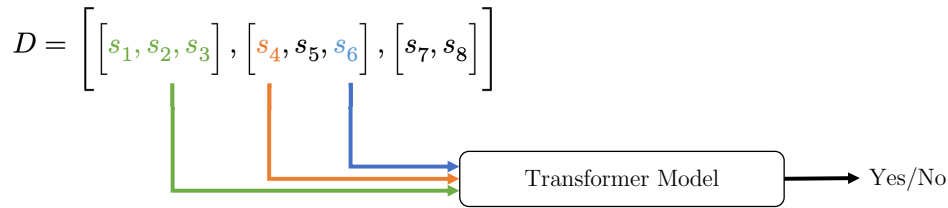
The learned semantics, which involve connecting sentences within the same paragraph, prove to be valuable in downstream tasks. The model can leverage previously acquired relations between entities and concepts and apply them when matching questions with potential answers. By utilizing relations from one sentence, it can generate questions that can be answered by information retrieved from another sentence. This tendency is particularly common among sentences within the same paragraph since every paragraph provides different perspectives on the same general topic.

We study three different methods to augment SSP with suitable contextual information, denoted as  $c$ . Below, we provide a detailed explanation of how we sample the context  $c$  from the documents for the contextual pre-training of the models. Regarding  $a$  and  $b$ , they are sampled in the same way described in Section 7.2.1. For each positive pair, we sample up to  $k_1 = 2$  hard negatives from within the same document and  $k_2$  easy negatives from other documents until  $k_1 + k_2 = 4$ , as in SSP.

### 8.2.1 Static Document-level Context

In this initial experiment, we established the context  $c$  to be the first paragraph  $P_1$  of a document  $\mathbf{D} = [P_1, \dots, P_n]$  from which the answer  $b$  is selected. The rationale behind this choice is that the first paragraph of most documents provides a general overview of the document’s topic [Chang et al., 2020]. By doing this, the context  $c$  assists in predicting whether  $a$  is extracted from the same paragraph as  $b$ . An example of the SDC objective is provided in Figure 8.2.

We refer to this as static document-level context because the contextual information  $c$  remains constant for any answer  $b$  obtained from the same document  $\mathbf{D}$ . Specifically, we generate positive examples by randomly sampling  $a$  and  $b$  from a single paragraph  $P_i \in \mathbf{D}$ , where  $i > 1$ . For the selected  $a$ , we create challenging negative examples by randomly selecting a sentence  $b$  from different paragraphs  $P_j \in \mathbf{D}$ , where  $j \neq i \wedge j > 1$ . In these negative examples, we still set  $c = P_1$  since  $b$  belongs to the same document  $\mathbf{D}$ . We generate easy negatives for a given  $a$  by sampling  $b$  from a random paragraph  $P'_i$  in another document  $\mathbf{D}' \neq \mathbf{D}$ . In this scenario, we choose  $c$  as the first paragraph  $P'_1$  of  $\mathbf{D}'$  because the context in the downstream AS2 task relates to the answer candidate rather than the question.



Los Angeles, often referred to by its initials L.A., is a major city in the U.S. state of California. With a 2020 population of 3,898,747, it is the largest city in the state, as well as the second-largest city in the United States following New York City. Los Angeles is known for its Mediterranean climate, ethnic and cultural diversity, being the home of the Hollywood film industry, and its sprawling metropolitan area.

The majority of the city proper lies in a basin in Southern California adjacent to the Pacific Ocean. In the west, it extends partly through the Santa Monica Mountains and north into the San Fernando Valley. It covers about 469 square miles (1,210 km<sup>2</sup>), and is the county seat of Los Angeles County, which is the most populous county in the United States with an estimated 9.86 million residents as of 2022.

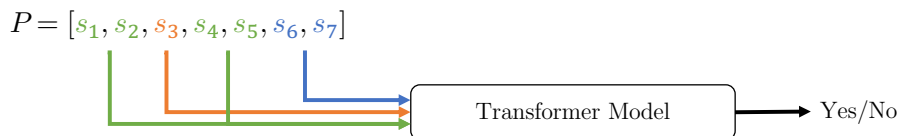
Los Angeles has a diverse economy with a broad range of industries. It has one of the busiest container ports in the Americas. The Los Angeles metropolitan area had a gross metropolitan product of over \$1.0 trillion in 2018.

Figure 8.2: Example of SDC objective.  $a$  is in orange,  $b$  in blue and the context  $c$  in green.

## 8.2.2 Dynamic Paragraph-level Context

In Dynamic Paragraph-level Context (DPC), we select the context  $c$  as the remaining text of the paragraph  $P_i$  from which  $b$  is extracted. Positive examples are created by sampling both  $a$  and  $b$  from the same paragraph  $P_i \in \mathbf{D}$ . For the positives, the context  $c$  is equal to  $c = P_i \setminus \{a, b\}$ . Notice that removing  $a$  and  $b$  from  $P_i$  is of fundamental importance not to make the task trivial, otherwise, the model could exploit  $P_i$  to decide whether  $a$  and  $b$  originated from the same paragraph. Hard negatives are created by randomly selecting  $b$  from another paragraph  $P_j \in \mathbf{D}, j \neq i$ . In this case, the context  $c$  is set to whatever remains in  $P_j$  after the removal of  $b$ :  $c = P_j \setminus \{b\}$ . Finally, we create easy negatives by sampling  $b$  from a paragraph  $P'_i$  in another document  $\mathbf{D}' \neq \mathbf{D}$ . Here, we set the context to  $c = P'_i \setminus \{b\}$ .

Notice that the context is always extracted from the same document or paragraph of  $b$ . The reason is that in Question Answering, the context is tied to the answer, and should provide more details on it rather than on the question. An example of DPC objective is depicted in Figure 8.3.



History of Los Angeles. The area that became Los Angeles was claimed by Spain in 1542. It became a part of Mexico in 1821 following the Mexican War of Independence. At the end of the Mexican-American War, Los Angeles was purchased as part of the Treaty of Guadalupe Hidalgo and became part of the United States. Los Angeles was incorporated as a municipality in 1850. The discovery of oil in the 1890s brought rapid growth to the city. The city was further expanded with the completion of the Los Angeles Aqueduct in 1913.

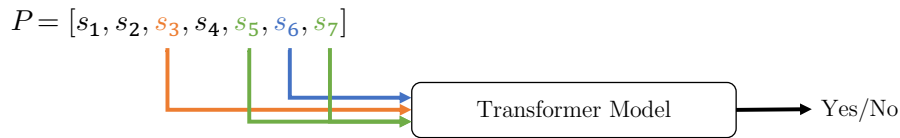
Figure 8.3: Example of DPC objective.  $a$  is in orange,  $b$  in blue and the context  $c$  in green.

### 8.2.3 Dynamic Sentence-level Local Context

In this approach, we designate the context  $c$  as the surrounding text encompassing the sentence  $b$ , which includes the sentences preceding and succeeding  $b$  within the paragraph  $P_i$ . To address exceptional cases, we require that either the previous or the next sentence of  $b$  must exist (e.g., if  $b$  is the final sentence in the paragraph  $P$ , the next sentence may not be present). We refer to this as Dynamic Sentence-level Local Context (DSLCLC), as the contextual information  $c$  is defined at the sentence level and varies accordingly for each sentence  $b$  extracted from the paragraphs.

Similar to the SDC and DPC, we create positive pairs by randomly sampling  $a$  and  $b$  from the same paragraph  $P_i \in \mathbf{D}$ , with  $c$  representing the local context surrounding  $b$  within  $P_i$  (ensuring  $a$  is not part of  $c$ ). We automatically exclude paragraphs that are insufficient in length to guarantee the formation of a positive example. For hard negatives, we sample  $b$  from another paragraph  $P_j \in \mathbf{D}$ , where  $j \neq i$ . As for easy negatives, we sample  $b$  from a paragraph  $P'_i \in \mathbf{D}'$ , where  $\mathbf{D}' \neq \mathbf{D}$  (in both cases,  $c$  is defined as the sentences around  $b$ ).

An example of DSLCLC objective is given in Figure 8.4.



History of Los Angeles. The area that became Los Angeles was claimed by Spain in 1542. It became a part of Mexico in 1821 following the Mexican War of Independence. At the end of the Mexican-American War, Los Angeles was purchased as part of the Treaty of Guadalupe Hidalgo and became part of the United States. Los Angeles was incorporated as a municipality in 1850. The discovery of oil in the 1890s brought rapid growth to the city. The city was further expanded with the completion of the Los Angeles Aqueduct in 1913.

Figure 8.4: Example of DSLCLC objective.  $a$  is in orange,  $b$  in blue and the context  $c$  in green.

## 8.3 Experiments

In this Section, we describe the setting in which we pre-trained and fine-tuned the models. We highlight the datasets we use, the hyper-parameters we use in training and the amount of compute effort for each execution.

### 8.3.1 Continuous Pre-Training

We perform continuous pre-training starting from checkpoints of several state-of-the-art models. In particular, we test all of our objectives over RoBERTa and ELECTRA and then, based on which objective provides the better performance overall in preliminary experiments, we train also two DeBERTaV3 models. We also experiment by mixing all objectives, thus forcing the model to gather information from three different types of context to better understand entities and relations in the answers. We indicate those experiments with “ALL”, as in the previous Chapter.

To avoid improvements derived from the usage of more data, we exploit the training corpora used to pre-train the original checkpoints of RoBERTa, ELECTRA and DeBERTaV3. We do continuous pre-training over the English Wikipedia, the BookCorpus, OpenWebText and CC-News by using a combination of our objectives (SDC, DPC and DSLCLC) with the original pre-training tasks, i.e. MLM for RoBERTa and TD for ELECTRA and DeBERTaV3. As before,

Model	Steps	LR	BS	MSL	# Tokens	Effort	Objectives
RoBERTa <sub>Base</sub> + SSP (SDC/DPC/DSLCL/ALL)	400K	$1 \cdot 10^{-4}$	4096	128	210B	+2.5%	MLM (1.0) + SSP (1.0)
ELECTRA <sub>Base</sub> + SSP (SDC/DPC/DSLCL/ALL)	400K	$1 \cdot 10^{-4}$	4096	128	210B	+5.0%	MLM (1.0) + TD (50.0) + SSP (1.0)
DeBERTa <sub>Base</sub> + SPP (ALL)	100K	$1 \cdot 10^{-4}$	4096	128	52B	+1.25%	MLM (1.0) + SSP (1.0)
DeBERTa <sub>Large</sub> + SPP (ALL)	100K	$1 \cdot 10^{-4}$	4096	128	52B	+1.25%	MLM (1.0) + SSP (1.0)
DeBERTaV3 <sub>Base</sub> + SPP (ALL)	100K	$1 \cdot 10^{-4}$	4096	128	52B	+0.625%	MLM (1.0) + TD (50.0) + SSP (1.0)
DeBERTaV3 <sub>Large</sub> + SPP (ALL)	100K	$1 \cdot 10^{-4}$	4096	128	52B	+0.625%	MLM (1.0) + TD (50.0) + SSP (1.0)

Table 8.1: Hyper-parameters used for the continuous pre-training of several Transformer models with our specialized objectives. The “Effort” column indicates the amount of additional FLOPS required compared to the original pre-training. “MSL” is the maximum allowed input length. “LR” is the learning rate while “BS” refers to the training batch size. The last column describes the used training tasks along with their weight (in round brackets).

we found that keeping the original training tasks allows us to reach higher levels of accuracy. Moreover, notice that for DeBERTaV3<sub>Large</sub> models, we can not exploit the original generator for Token Detection, as it was not released. Thus, we use a DeBERTaV3<sub>Base</sub> instead as the generator for DeBERTaV3<sub>Large</sub>. We pre-process the datasets described above according to the definition of our objectives, obtaining about 330M of samples for SDC and 200M for DPC and DSLC.

In Table 8.1 we show the hyper-parameters we use in each experiment, highlighting also the additional computational effort required over the original pre-training. One of the advantages of our approaches is that once continuously pre-trained, our checkpoints can be applied to different AS2 settings with minimal effort. The amount of tokens seen in pre-training by each initial checkpoint is described in Section 3.1. Notice that as in SSP, we keep a constant maximum sequence length of 128 tokens because in preliminary experiments we didn’t find particular advantages with higher values. This reduces the computational complexity by a factor of 16 compared to the original sequence length of 512 tokens.

Regarding the optimization hyper-parameters and the hardware used, we set  $\beta_1 = 0.9$  and  $\beta_2 = 0.99$  in the FuseAdam optimizer, a weight decay of 0.01 and a triangular learning rate with 10% warmup steps. All models are trained on our machine (see Section 1.2) and require about 3.0 days of training each with *fp16* and DeepSpeed [Rasley et al., 2020]. There may be slight variations compared with the results published in Di Liello et al. [2023] because we fixed an important bug regarding the initialization of the token types embedding layer.

### 8.3.2 Fine-Tuning

We evaluate the models described in the previous Section on several datasets for Contextual Answer Sentence Selection. More specifically, we fine-tune over ASNQ, WikiQA, TREC-QA, IQAD, NewsAS2 and TriviaAS2. We do not test on WQA as in the previous Chapter because it provides no way of retrieving the original documents, but we use IQAD instead, which follows a similar data distribution and provides context. In all datasets, the context is always composed of the sentences before and after the candidate answers in the original document. The statistics for the datasets can be retrieved in Section 4.2.1. For TREC-QA, we do not have context because answers are not extracted from documents (or the documents have not been publicly released). In this case, we exploit a large generative language model to generate additional context for each answer candidate. Specifically, we utilize Falcon [Almazrouei et al., 2023], a recently released LLM with 40B parameters. We use the version fine-tuned on various datasets from Baize [Xu et al., 2023], which contains instruction-based conversations between real users about several topics<sup>1</sup>. We provide the model with the following prompt: “Provide additional

<sup>1</sup>The model we use can be found here: <https://huggingface.co/tiiuae/falcon-40b-instruct>



Parameter	ASNQ	WikiQA & TREC-QA	IQAD	NewsAS2 & TriviaAS2
Batch Size	{1024, 2048}	{16, 32, 64, 128}	{256, 512, 1024}	{128, 256, 512}
Learning Rate	$\{1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{5 \cdot 10^{-6}, 1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$
Max epochs	6	40	20	10

Table 8.2: Hyper-parameters search space used for the fine-tuning over several contextual AS2 datasets.

context to the text ‘{answer}’ in two sentences: ”, by dynamically substituting {answer} with the answer candidate in the dataset. For the generation, we use a temperature of 0.7, a  $\text{top}_k$  of 8 and sampling. Moreover, we force the output length to be between 16 and 64 tokens.

For the fine-tuning, we use the same hyper-parameters used in the pre-training but for the batch size, the learning rate, the number of warmup steps and the number of training epochs. Summaries of the hyper-parameters search spaces for each downstream dataset are reported in Table 8.2. The number of warmup steps in the scheduler is set to match the number of steps in the first epoch. We also increase the maximum sequence length to 256 tokens in every experiment over ASNQ, NewsAS2, TriviaAS2 and IQAD, since the concatenation of the question, the answer candidate and the context of those datasets often exceeds 128 tokens. As in previous Chapters, we use the development set to find the best batch size and learning rate for each combination of baseline models and target datasets. We do not conduct a grid search over our continuously pre-trained models because it is too expensive and they adapt well to the best hyper-parameters chosen for the baselines.

We measure performance again with Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Precision@1 (P@1) as in related works [Garg et al., 2019; Lauriola and Moschitti, 2021].

## 8.4 Results

In this section, we present the outcomes obtained by refining our pre-trained models over various AS2 datasets. To facilitate comprehension of the enhancements, we divide the results into base and large models. We report results from the original papers when available, otherwise, we fine-tune the models on the different tasks. Statistical significance is computed compared to models from Lauriola et al., 2021, which are fine-tuned by us for a fair comparisons because the authors perform an additional transfer step on ASNQ and because they report results on the development set.

### 8.4.1 Base Models

Results for this experiment are reported in Table 8.3 and 8.4 for ASNQ, WikiQA and TREC-QA, in Table 8.5 for both IQAD test splits and in Table 8.6 and 8.7 for NewsAS2 and TriviaAS2. This Section continues by describing the results of our Spans in the Same Paragraph (SSP) objective augmented with the different context extraction techniques we developed. We compare against the original pre-trained checkpoints of each model as well as against models fine-tuned with context, as in Lauriola et al., 2021.

**SDC** Our continuously pre-trained models with SSP and Static Document-level Context improve over the baselines fine-tuned with either only question-answer pairs or question-answer-context tuples. On ASNQ, our  $\text{RoBERTa}_{\text{Base}} + \text{SSP}$  (SDC) increases the P@1 by 5.2 and 1.1 points compared to the two baselines, respectively. Similarly,  $\text{ELECTRA}_{\text{Base}} + \text{SSP}$  (SDC) improves the P@1 by 4.6 and 1.5 points over the corresponding baselines.

Model	CTX	ASNQ			WikiQA			TREC-QA		
		MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	✗	68.2 (0.5)	74.2 (0.3)	63.5 (0.5)	85.1 (1.9)	86.5 (1.9)	77.2 (3.1)	89.9 (1.7)	93.7 (1.6)	89.2 (3.1)
RoBERTa <sub>Base</sub> [Lauriola et al., 2021]	✓	71.6 (0.6)	77.3 (0.5)	67.6 (0.6)	84.4 (1.5)	86.0 (1.3)	77.0 (2.1)	88.2 (1.7)	93.5 (2.0)	88.2 (2.9)
RoBERTa <sub>Base</sub> + SSP (SDC)	✓	<u>73.1</u> (0.5)	<u>78.4</u> (0.6)	<u>68.7</u> (0.8)	<u>87.8</u> (0.6)	<u>89.2</u> (0.5)	<u>81.8</u> (0.9)	88.2 (0.9)	92.3 (2.2)	86.3 (3.7)
RoBERTa <sub>Base</sub> + SSP (DPC)	✓	<b>73.2</b> (0.4)	<b>78.5</b> (0.3)	<b>69.2</b> (0.5)	<b>89.9</b> (0.2)	<b>91.1</b> (0.2)	<b>85.2</b> (0.4)	<b>90.5</b> (0.2)	95.3 (0.9)	91.2 (1.5)
RoBERTa <sub>Base</sub> + SSP (DSLCL)	✓	<u>72.9</u> (0.4)	<u>78.4</u> (0.2)	<u>69.0</u> (0.3)	<u>87.8</u> (0.9)	<u>89.2</u> (0.8)	<u>81.6</u> (1.3)	<b>90.5</b> (0.3)	<b>95.7</b> (0.4)	<b>92.2</b> (0.8)
RoBERTa <sub>Base</sub> + SSP (ALL)	✓	<u>72.9</u> (0.6)	77.9 (0.6)	68.2 (0.8)	<u>88.2</u> (0.9)	<u>89.6</u> (0.8)	<u>82.4</u> (1.7)	90.3 (1.4)	94.7 (1.2)	90.7 (1.7)
ELECTRA <sub>Base</sub> [Clark et al., 2020]	✗	69.3 (0.0)	75.1 (0.1)	65.0 (0.2)	85.7 (0.9)	87.1 (0.9)	78.5 (1.6)	89.9 (0.4)	94.0 (0.9)	90.3 (1.7)
ELECTRA <sub>Base</sub> [Lauriola et al., 2021]	✓	72.3 (0.6)	77.9 (0.8)	68.1 (0.8)	83.1 (1.3)	84.4 (1.4)	73.8 (2.1)	88.1 (1.1)	92.9 (1.4)	87.3 (3.1)
ELECTRA <sub>Base</sub> + SSP (SDC)	✓	<u>74.7</u> (0.5)	<u>79.5</u> (0.3)	<u>69.6</u> (0.3)	<b>88.7</b> (0.1)	<b>90.0</b> (0.2)	<b>82.9</b> (0.2)	<b>90.1</b> (0.5)	94.0 (0.7)	89.7 (1.5)
ELECTRA <sub>Base</sub> + SSP (DPC)	✓	<u>74.4</u> (0.2)	<b>79.5</b> (0.2)	<b>70.5</b> (0.2)	88.0 (0.6)	89.2 (0.6)	<u>81.3</u> (0.6)	<u>89.4</u> (0.0)	<b>95.5</b> (0.8)	<b>92.6</b> (1.5)
ELECTRA <sub>Base</sub> + SSP (DSLCL)	✓	<u>74.3</u> (0.3)	<u>79.4</u> (0.5)	<u>70.0</u> (0.8)	<u>87.0</u> (0.9)	<u>88.5</u> (0.9)	<u>79.7</u> (1.4)	89.2 (0.8)	<u>95.0</u> (0.4)	<u>90.7</u> (0.8)
ELECTRA <sub>Base</sub> + SSP (ALL)	✓	<u>73.8</u> (0.4)	78.7 (0.3)	68.8 (0.4)	<u>87.5</u> (0.5)	<u>89.1</u> (0.5)	<u>81.5</u> (0.7)	<u>89.6</u> (0.2)	93.5 (0.4)	88.7 (0.8)

Table 8.3: Results of RoBERTa<sub>Base</sub> and ELECTRA<sub>Base</sub> contextual models over ASNQ, WikiQA and TREC-QA. “CTX” tells whether the context was used in fine-tuning. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either RoBERTa<sub>Base</sub> or ELECTRA<sub>Base</sub> with context.

Model	CTX	ASNQ			WikiQA			TREC-QA		
		MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Base</sub> [He et al., 2020]	✗	69.1 (0.3)	75.3 (0.5)	65.3 (0.6)	<b>87.1</b> (0.8)	<b>88.6</b> (0.8)	<b>81.0</b> (1.5)	89.7 (0.7)	94.2 (1.2)	90.2 (1.7)
DeBERTa <sub>Base</sub> [Lauriola et al., 2021]	✓	71.1 (0.2)	76.2 (0.2)	66.2 (0.3)	84.8 (0.9)	86.4 (1.2)	77.1 (2.4)	89.6 (0.8)	94.1 (0.4)	89.7 (1.5)
DeBERTa <sub>Base</sub> + SSP (ALL)	✓	<b>72.2</b> (0.1)	<b>77.9</b> (0.2)	<b>67.7</b> (0.2)	83.8 (1.1)	85.2 (1.0)	75.5 (1.8)	<b>90.2</b> (1.2)	<b>95.0</b> (1.4)	<b>90.7</b> (2.2)
DeBERTaV3 <sub>Base</sub> [He et al., 2021]	✗	70.8 (0.2)	77.0 (0.2)	67.5 (0.4)	86.7 (1.8)	88.0 (1.7)	79.3 (2.8)	<b>90.6</b> (1.2)	93.4 (0.7)	90.6 (1.7)
DeBERTaV3 <sub>Base</sub> [Lauriola et al., 2021]	✓	74.4 (1.1)	79.8 (1.1)	70.3 (1.4)	77.9 (3.7)	79.3 (3.6)	66.7 (5.7)	89.8 (1.2)	93.3 (1.3)	88.5 (2.6)
DeBERTaV3 <sub>Base</sub> + SSP (ALL)	✓	<b>75.1</b> (0.5)	<b>80.5</b> (0.4)	<b>71.3</b> (0.3)	<b>87.2</b> (0.5)	<b>88.6</b> (0.4)	<b>80.9</b> (0.5)	90.2 (1.1)	<b>94.3</b> (0.7)	<b>91.5</b> (1.2)

Table 8.4: Results of DeBERTa<sub>Base</sub> and DeBERTaV3<sub>Base</sub> contextual models over ASNQ, WikiQA and TREC-QA. We underline statistically significant improvements over the baselines augmented with context.

On smaller datasets such as WikiQA, improvements are larger because the baselines struggle at reasoning over 3 input text spans since fine-tuning data is scarce and pre-training was performed on at most 2 input text spans. For these reasons, our models augmented with SDC improve the P@1 by more than 4 points on WikiQA for both RoBERTa and ELECTRA. On TREC-QA, SDC struggles to outperform both baselines, probably because the first paragraph of a document used in SDC does not align well with the type of context generated by the large LLM we employ.

In the industrial test splits (IQAD Bench 1&2), our models feature significant gains over both baselines. For example, on IQAD Bench 1, RoBERTa<sub>Base</sub> + SSP (SDC) improves the P@1 by 2.8%. Results are shown relative to the corresponding baseline because those datasets are subjected to Amazon’s internal data license. Notice that the baselines augmented with context struggle to outperform the original Pairwise baseline because IQAD does not contain enough data to accustom the model to effectively exploit the contextual information.

Finally, we analyze the improvements we achieve on NewsAS2 and TriviaAS2. On these datasets, RoBERTa and ELECTRA combined with SSP (SDC) outperform the baselines augmented with context by 1 about point in P@1 for every combination. A particular case is represented by ELECTRA over TriviaAS2, in which our model outperforms the corresponding baseline by 3.5 points in P@1.

**DPC** In DPC, the context is created by retrieving the remainder of the paragraph after two spans of text are extracted for SSP. On ASNQ, the checkpoints of RoBERTa and ELECTRA continuously pre-trained with SSP (DPC) improve the P@1 by up to 1.6 points over the baselines augmented with context, demonstrating that specialized objectives in pre-training are very important to accustom the model at the final task.

Model	CTX	IQAD Bench 1			IQAD Bench 2		
		MAP	MRR	P@1	MAP	MRR	P@1
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	✗		Baseline		Baseline		
RoBERTa <sub>Base</sub> [Lauriola et al., 2021]	✓	-0.1%	-0.1%	-0.1%	+0.4%	+0.0%	+0.4%
<b>RoBERTa<sub>Base</sub> + SSP (SDC)</b>	✓	<b>+1.1%</b>	<b>+1.3%</b>	<b>+2.8%</b>	+0.2%	+0.0%	+0.4%
<b>RoBERTa<sub>Base</sub> + SSP (DPC)</b>	✓	+0.4%	+0.4%	+1.3%	+0.4%	+0.2%	<b>+0.9%</b>
<b>RoBERTa<sub>Base</sub> + SSP (DSLCL)</b>	✓	+0.1%	+0.5%	+1.9%	+0.1%	+0.0%	+0.3%
<b>RoBERTa<sub>Base</sub> + SSP (ALL)</b>	✓	+0.6%	+1.0%	+2.2%	<b>+0.6%</b>	<b>+0.3%</b>	<b>+0.9%</b>
ELECTRA <sub>Base</sub> [Clark et al., 2020]	✗		Baseline		Baseline		
ELECTRA <sub>Base</sub> [Lauriola et al., 2021]	✓	-0.6%	-0.6%	-1.0%	-0.4%	-0.4%	-0.9%
<b>ELECTRA<sub>Base</sub> + SSP (SDC)</b>	✓	<b>+1.2%</b>	<b>+0.6%</b>	<b>+0.6%</b>	<b>+0.9%</b>	<b>+0.9%</b>	<b>+1.4%</b>
<b>ELECTRA<sub>Base</sub> + SSP (DPC)</b>	✓	+0.4%	-0.3%	-0.6%	+0.4%	+0.2%	+0.1%
<b>ELECTRA<sub>Base</sub> + SSP (DSLCL)</b>	✓	+1.0%	+0.5%	+0.6%	+0.2%	+0.2%	+0.0%
<b>ELECTRA<sub>Base</sub> + SSP (ALL)</b>	✓	+0.7%	<b>+0.6%</b>	+0.4%	+0.5%	+0.5%	+0.3%

Table 8.5: Results of RoBERTa<sub>Base</sub> and ELECTRA<sub>Base</sub> contextual models over both IQAD test sets. We use bold formatting for our proposals and the best results.

Model	CTX	NewsAS2			TriviaAS2		
		MAP	MRR	P@1	MAP	MRR	P@1
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	✗	82.1 (0.2)	84.8 (0.2)	76.0 (0.1)	77.3 (0.7)	82.6 (0.7)	74.4 (0.9)
RoBERTa <sub>Base</sub> [Lauriola et al., 2021]	✓	83.2 (0.3)	85.5 (0.4)	77.1 (0.6)	77.7 (0.5)	83.2 (0.3)	75.3 (0.3)
<b>RoBERTa<sub>Base</sub> + SSP (SDC)</b>	✓	<u>83.8</u> (0.1)	<u>86.4</u> (0.1)	<u>78.0</u> (0.1)	<b>78.7</b> (0.4)	<b>84.2</b> (0.4)	<b>76.6</b> (0.8)
<b>RoBERTa<sub>Base</sub> + SSP (DPC)</b>	✓	83.3 (0.3)	<u>86.0</u> (0.2)	76.9 (0.5)	77.7 (0.5)	83.0 (0.4)	74.6 (0.8)
<b>RoBERTa<sub>Base</sub> + SSP (DSLCL)</b>	✓	83.6 (0.3)	<u>86.2</u> (0.4)	77.9 (0.6)	77.9 (0.3)	83.3 (0.5)	75.0 (0.9)
<b>RoBERTa<sub>Base</sub> + SSP (ALL)</b>	✓	<b>84.0</b> (0.2)	<b>86.6</b> (0.3)	<b>78.3</b> (0.5)	<u>78.5</u> (0.3)	<b>83.8</b> (0.2)	75.9 (0.5)
ELECTRA <sub>Base</sub> [Clark et al., 2020]	✗	82.2 (0.2)	84.9 (0.4)	76.1 (0.6)	73.4 (0.4)	79.2 (0.5)	69.3 (0.8)
ELECTRA <sub>Base</sub> [Lauriola et al., 2021]	✓	83.0 (0.4)	85.6 (0.4)	77.1 (0.7)	73.7 (0.2)	79.1 (0.7)	69.3 (1.4)
<b>ELECTRA<sub>Base</sub> + SSP (SDC)</b>	✓	<u>83.7</u> (0.5)	<u>86.3</u> (0.4)	<u>78.2</u> (0.7)	<u>75.5</u> (0.4)	<u>81.6</u> (0.6)	<u>72.8</u> (0.9)
<b>ELECTRA<sub>Base</sub> + SSP (DPC)</b>	✓	<u>83.7</u> (0.1)	<u>86.4</u> (0.3)	<u>78.2</u> (0.5)	<u>75.4</u> (0.1)	<u>81.3</u> (0.1)	<u>72.6</u> (0.3)
<b>ELECTRA<sub>Base</sub> + SSP (DSLCL)</b>	✓	<u>83.5</u> (0.1)	<u>86.3</u> (0.1)	<u>78.0</u> (0.2)	<u>75.6</u> (0.8)	<u>81.2</u> (0.4)	<u>72.1</u> (0.3)
<b>ELECTRA<sub>Base</sub> + SSP (ALL)</b>	✓	<b>84.0</b> (0.2)	<b>86.6</b> (0.1)	<b>78.7</b> (0.2)	<b>75.9</b> (0.4)	<b>81.7</b> (0.3)	<b>73.0</b> (0.5)

Table 8.6: Results of RoBERTa<sub>Base</sub> and ELECTRA<sub>Base</sub> contextual models over NewsAS2 and TriviaAS2. We underline statistically significant improvements over either RoBERTa<sub>Base</sub> or ELECTRA<sub>Base</sub> with context.

On the two small datasets (WikiQA and TREC-QA), both RoBERTa and ELECTRA with SSP (DPC) outperform the two baselines by a relevant margin. For example, ELECTRA<sub>Base</sub> + SSP (DPC) improves the P@1 by 2.3 and 5.3 points on TREC-QA compared to the Pairwise and the contextual baseline, respectively.

On the two IQAD internal benchmarks, the results of exploiting SSP (DPC) are mixed. Our objective combined with RoBERTa shows consistent gains on both the test splits, with improvements up to 1.3% in P@1. With ELECTRA, our objective provides consistent improvements over the contextual baseline but struggles at competing with the Pairwise model. This further demonstrates that context is a precious source of information that is difficult to be exploited properly, especially on IQAD, which contains real user questions that are noisy, contain typos and may not be semantically correct.

Finally, on NewsAS2 and TriviaAS2, our ELECTRA<sub>Base</sub> augmented with SSP (DPC) provides improvements over both datasets, with gains in the range of 1-3 P@1 points over the baselines. At the same time, RoBERTa<sub>Base</sub> + SSP (DPC) outperforms the Pairwise baseline but falls short behind the contextual baseline.

**DSLCL** The last context type we analyze is DSLCL, in which we exploit the sentences around the second span of SSP as the context. On ASNQ, RoBERTa and ELECTRA augmented with SSP (DSLCL) outperform the stronger baseline by 1.4 and 1.9 points in P@1 respectively. Similarly, on WikiQA and TREC-QA, every combination of our models with SSP (DSLCL) shows

Model	CTX	NewsAS2			TriviaAS2		
		MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Base</sub> [He et al., 2020]	✗	82.8 (0.2)	85.5 (0.1)	76.9 (0.1)	79.2 (0.3)	83.9 (0.5)	76.0 (0.7)
DeBERTa <sub>Base</sub> [Lauriola et al., 2021]	✓	84.1 (0.1)	86.6 (0.1)	78.5 (0.3)	79.9 (0.0)	85.0 (0.2)	<b>78.2</b> (0.2)
<b>DeBERTa<sub>Base</sub> + SSP (ALL)</b>	✓	<b>84.3</b> (0.2)	<b>86.9</b> (0.3)	<b>79.0</b> (0.5)	<b>80.5</b> (0.6)	<b>85.5</b> (0.8)	<b>78.2</b> (1.1)
DeBERTaV3 <sub>Base</sub> [He et al., 2021]	✗	83.0 (0.4)	85.8 (0.4)	77.2 (0.5)	<b>81.0</b> (0.5)	85.6 (0.5)	78.8 (0.5)
DeBERTaV3 <sub>Base</sub> [Lauriola et al., 2021]	✓	83.7 (0.2)	86.3 (0.2)	78.0 (0.5)	80.8 (0.3)	85.7 (0.5)	78.7 (0.4)
<b>DeBERTaV3<sub>Base</sub> + SSP (ALL)</b>	✓	<b>84.4</b> (0.3)	<b>87.0</b> (0.3)	<b>78.9</b> (0.4)	<b>81.0</b> (0.2)	<b>86.0</b> (0.3)	<b>79.6</b> (0.3)

Table 8.7: Results of DeBERTa<sub>Base</sub> and DeBERTaV3<sub>Base</sub> contextual models over NewsAS2 and TriviaAS2. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either DeBERTa<sub>Base</sub> or DeBERTaV3<sub>Base</sub> with context.

improvements over the baselines, with gains between 0.4 and 4.4 points of P@1.

On Amazon’s internal datasets, RoBERTa with SSP (DSL) outperforms both the Pairwise and the contextual baseline by a significant margin. For example, on IQAD Bench 1, our model improves the MRR and P@1 by 0.5 and 1.9 points respectively. When combining SSP (DSL) with ELECTRA and testing over IQAD Bench 1, our model features significant gains of 1.0 MAP and 0.6 P@1 points. On IQAD Bench 2 instead results of our models are significantly better when compared with the contextual baseline and also slightly better than the Pairwise RoBERTa<sub>Base</sub>.

On NewsAS2 and TriviaAS2, the gains of our models are similar to those on ASNQ. In particular, RoBERTa<sub>Base</sub> + SSP (DSL) outperforms the Pairwise and contextual baselines on NewsAS2 by a margin of 1.9 and 0.8 P@1 points respectively. When tested on TriviaAS2 instead, the results are in line with the contextual baselines. Regarding ELECTRA combined with SSP (DSL), it outperforms the strongest baseline by 0.9 and 2.8 P@1 points on NewsAS2 and TriviaAS2 respectively.

**ALL** In this experiment, we combine all the proposed context-gathering methodologies for a single continuous pre-training, indicated with SSP (ALL). When applied to RoBERTa, ELECTRA and DeBERTa, the trend is clear: our pre-trained models outperform the baseline on almost all benchmarks. The only exception is represented by DeBERTa<sub>Base</sub> + SSP (ALL) when tested over WikiQA, in which the additional context is enough to hurt performance, while SSP (ALL) leads to even lower accuracies. We suspect there is an interaction between the DeBERTa<sub>Base</sub> pre-training setting and the context of the WikiQA dataset. Since the results of SSP (ALL) are more agnostic compared to the target dataset, we suggest the usage of this objective when testing on new or unknown data.

Among the best results obtained among Base models, we cite DeBERTaV3<sub>Base</sub> + SSP (ALL), which reaches a P@1 of 71.3 on ASNQ. Moreover, DeBERTa<sub>Base</sub> + SSP (ALL) evaluated on NewsAS2 achieves the highest P@1 of 79.0 among all the models in this work built over a Base architecture.

## 8.4.2 Large Models

In Tables 8.8 and 8.9 we report the results of combining larger models with our SSP objective with context gathered in three different ways (ALL).

On ASNQ, our DeBERTa<sub>Large</sub> and DeBERTaV3<sub>Large</sub> + SSP (ALL) reach remarkable performance on all metrics (second and first place respectively in the global ranking). When combining DeBERTa<sub>Large</sub> with SSP (ALL) and testing over WikiQA, there is a degradation of the performance, as with DeBERTa<sub>Base</sub> (more details about this issue are given in the

Model	CTX	ASNQ			WikiQA			TREC-QA		
		MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Large</sub> [He et al., 2020]	✗	76.6 (0.1)	81.7 (0.1)	73.1 (0.2)	89.9 (0.8)	91.1 (0.7)	84.5 (1.5)	91.5 (0.3)	94.2 (0.1)	90.2 (2.2)
DeBERTa <sub>Large</sub> [Lauriola et al., 2021]	✓	76.4 (0.3)	81.3 (0.2)	72.0 (0.3)	88.8 (1.0)	90.1 (0.8)	82.7 (1.8)	90.7 (0.8)	94.0 (0.6)	90.7 (2.2)
<b>DeBERTa<sub>Large</sub> + SSP (ALL)</b>	✓	<b>77.5</b> (0.1)	<b>82.4</b> (0.1)	<b>73.7</b> (0.1)	87.3 (0.9)	88.6 (1.1)	81.6 (2.1)	<u>91.6</u> (0.1)	94.1 (0.8)	92.2 (2.2)
<b>TANDA DeBERTa<sub>Large</sub> + SSP (ALL)</b>	✓	-	-	-	<b>91.7</b> (0.7)	<b>93.0</b> (0.6)	<b>88.2</b> (1.3)	<b>93.1</b> (0.7)	<b>95.6</b> (0.2)	<b>93.4</b> (1.0)
DeBERTaV3 <sub>Large</sub> [He et al., 2021]	✗	74.5 (0.2)	79.8 (0.2)	70.3 (0.2)	86.2 (1.7)	87.5 (1.7)	79.9 (2.7)	91.2 (0.3)	93.7 (0.8)	91.2 (1.5)
DeBERTaV3 <sub>Large</sub> [Lauriola et al., 2021]	✓	78.6 (0.2)	82.9 (0.2)	74.0 (0.3)	87.3 (0.8)	88.6 (0.7)	80.8 (1.2)	91.5 (0.5)	94.4 (0.8)	92.9 (1.2)
<b>DeBERTaV3<sub>Large</sub> + SSP (ALL)</b>	✓	<b>78.8</b> (0.3)	<b>83.5</b> (0.3)	<b>75.1</b> (0.5)	<u>89.5</u> (1.1)	<u>90.8</u> (1.2)	<u>84.7</u> (1.7)	91.9 (1.0)	94.5 (1.1)	<b>93.5</b> (1.8)
<b>TANDA DeBERTaV3<sub>Large</sub> + SSP (ALL)</b>	✓	-	-	-	<b>91.8</b> (0.5)	<b>93.0</b> (0.7)	<b>88.2</b> (0.7)	<b>93.1</b> (0.4)	<b>96.0</b> (0.7)	92.6 (0.5)

Table 8.8: Results of DeBERTa<sub>Large</sub> and DeBERTaV3<sub>Large</sub> contextual models over ASNQ, WikiQA and TREC-QA. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over either DeBERTa<sub>Large</sub> or DeBERTaV3<sub>Large</sub> with context.

Model	CTX	NewsAS2			TriviaAS2		
		MAP	MRR	P@1	MAP	MRR	P@1
DeBERTa <sub>Large</sub> [He et al., 2020]	✗	84.8 (0.0)	87.3 (0.1)	79.7 (0.2)	82.9 (0.2)	87.2 (0.2)	81.2 (0.4)
DeBERTa <sub>Large</sub> [Lauriola et al., 2021]	✓	85.0 (0.1)	87.4 (0.2)	79.7 (0.2)	83.5 (0.2)	87.5 (0.3)	81.7 (0.3)
<b>DeBERTa<sub>Large</sub> + SSP (ALL)</b>	✓	<b>86.2</b> (0.3)	<b>88.4</b> (0.2)	<b>81.3</b> (0.3)	<b>84.4</b> (0.1)	<b>88.9</b> (0.2)	<b>83.4</b> (0.2)
DeBERTaV3 <sub>Large</sub> [He et al., 2021]	✗	85.0 (0.2)	87.5 (0.2)	79.8 (0.3)	84.5 (0.2)	88.5 (0.2)	82.9 (0.4)
DeBERTaV3 <sub>Large</sub> [Lauriola et al., 2021]	✓	85.8 (0.1)	88.1 (0.2)	80.5 (0.3)	83.5 (0.1)	88.4 (0.1)	82.8 (0.1)
<b>DeBERTaV3<sub>Large</sub> + SSP (ALL)</b>	✓	<b>86.4</b> (0.2)	<b>88.9</b> (0.2)	<b>81.8</b> (0.4)	<b>85.5</b> (0.3)	<b>89.4</b> (0.2)	<b>84.2</b> (0.3)

Table 8.9: Results of DeBERTa<sub>Large</sub> and DeBERTaV3<sub>Large</sub> contextual models over NewsAS2 and TriviaAS2. We underline statistically significant improvements over either DeBERTa<sub>Large</sub> or DeBERTaV3<sub>Large</sub> with context.

previous Section). On the other hand, DeBERTaV3<sub>Large</sub> + SSP (ALL) on WikiQA shows improvements of up to 3.9 P@1 points. On TREC-QA, DeBERTa<sub>Large</sub> and DeBERTaV3<sub>Large</sub> combined with our specialized continuous pre-training lead to improvements of 1.5 and 0.6 points in P@1, respectively. We experiment also the combination of our techniques with TANDA, which results in our models obtaining the third and fourth place in the global ranking of WikiQA. Notice that the transfer step over ASNQ is able to override the misalignment between DeBERTa<sub>Large</sub> + SSP (ALL) and WikiQA. On TREC-QA, we obtain the third place on the global ranking with TANDA DeBERTaV3<sub>Large</sub> + SSP (ALL), just a few decimal MAP points behind our TANDA DeBERTaV3<sub>Large</sub> + ALL from Chapter 7 and TANDA RoBERTa<sub>Large</sub> from Garg et al. [2019].

On NewsAS2 and TriviaAS2, the combination of Large DeBERTa models with our SSP objective provides increases in accuracy on both datasets, with gains up to 1.6 and 1.7 respectively, when measuring performance with P@1.

### 8.4.3 General Considerations

This Section analyzes the general pattern of performance gains and provides suggestions to choose between the objectives for different use cases.

**Objective-dataset alignment** As in Section 7.4.5, we describe the alignment between the different objectives and the target datasets. First, notice that our datasets can be divided into two categories based on how answer candidates are extracted from relevant documents. In ASNQ, WikiQA, NewsAS2 and TriviaAS2, all the answer candidates for a given question are extracted from a single document (Wikipedia for ASNQ and WikiQA, CNN/DailyMail for NewsAS2 and quiz-league websites for TriviaAS2). On the other hand, IQAD and TREC-QA contain answer candidates extracted from multiple documents sourced from the web. This

results in a more homogeneous context for the former datasets because all answer candidates belong to the same document and a more heterogeneous context for the latter.

Our DPC (Document Paragraph-level Context) and DSLC (Document Sentence-level Local Context) pre-training methods share similarities in terms of the context used to assist SSP. While DPC employs the remaining of paragraph  $P$  as context after excluding  $a$  and  $b$ , DSLC utilizes the preceding and subsequent sentences to  $b$  within  $P$ . Through empirical observation, we find that often the contexts employed by DPC and DSLC partially overlap because of the limited number of sentences per paragraph in the original corpora. This explains why models trained using both approaches yield comparable performance.

Regarding the IQAD dataset, we observe that the SDC (Static Document-level Context) approach surpasses the performance of DPC and DSLC. In SDC, the context  $c$  can significantly differ from  $a$  and  $b$ , representing the first paragraph of the document. This distinction proves beneficial for leveraging information and effectively ranking answer candidates extracted from multiple documents, potentially from diverse domains, as required in IQAD. Consequently, we recommend employing DPC and DSLC when answer candidates originate from the same document, while SDC is preferable when candidates are sourced from multiple documents.

Finally, we expected TREC-QA to benefit more from SDC than DPC and DSLC, given the nature of the dataset, in which answer candidates are extracted from multiple documents. However, we suspect that the additional context generated by the Large Language Model we employ relates strictly to every answer candidate, thus leveraging better the continuous pre-training with DPC and DSLC.

**Limited improvements** Similar to the results over WQA in Chapter 7, the improvements in IQAD are comparatively smaller than those observed in the other datasets. There are several reasons for this. Firstly, the questions in IQAD are genuine queries from users, and they are frequently ambiguous and inadequately structured. Additionally, IQAD incorporates answer candidates obtained from a vast collection of web documents extracted with Elasticsearch. These documents have not been reviewed by humans and may contain errors in syntax, poorly constructed sentences, and inaccurate information. Moreover, the annotation process relies on user feedback, which is not always reliable in terms of accuracy.

## Chapter 9

# Self-supervised objectives for Summarization

In this last Chapter, we show that specialized pre-training objectives can improve accuracy in tasks other than Answer Sentence Selection or Fact Verification. For this experiment, we set the goal of developing ad-hoc objectives for Summarization using Auto-Regressive models.

Summarization involves condensing a given long text or document into a shorter version while preserving the key information and main points. It aims to provide a concise and coherent summary that captures the essential meaning and context of the original text. In this experiment, we focus on Abstractive Summarization, which differs from Extractive Summarization in many ways. Abstractive Summarization involves generating new sentences that capture the essence of the source text, while Extractive Summarization involves selecting and combining existing sentences. More information about the Summarization tasks is given in Section 2.1.3. For a detailed description of the models exploited in this Chapter, see Section 3.1. Moreover, we provide works related to Summarization in Section 3.2.3, while regarding the exploitation of large corpora to create self-supervised tasks, refer to Section 3.3.3.

Abstractive Summarization requires models able to process long documents and re-write them in a more concise form. Thus, in this experiment, we employ two state-of-the-art Generative Transformer-based models: BART [Lewis et al., 2020] and T5 [Raffel et al., 2020]. Both models are built over an Encoder-Decoder architecture, in which the encoder is fed with the original document and the decoder generates the summary, one token at a time.

## 9.1 Summarization-oriented Pre-training Objectives

High-quality Summarizations datasets are scarce for several reasons:

- **Human Effort:** Creating high-quality Summarization datasets requires significant human effort. Skilled annotators or domain experts need to read and understand the source texts and then generate appropriate summaries. This process can be time-consuming and costly, especially for large-scale datasets;
- **Subjectivity and Complexity:** Summarization is a subjective task, and different annotators may generate different summaries for the same source text. This subjectivity makes it challenging to achieve a consensus on what constitutes a “good” summary. Additionally, summarizing complex or technical content often requires domain expertise, further complicating the annotation process;

- Lengthy Source Texts: Summarization datasets often consist of long documents or articles that need to be condensed into shorter summaries. This adds to the complexity and time required for annotation;

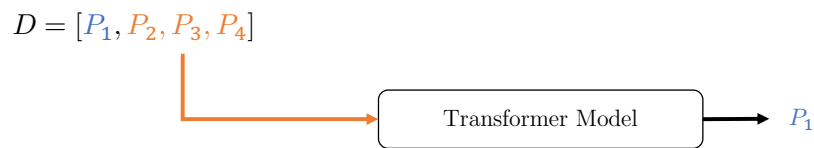
For example, all 4 datasets we use for the evaluation of our approaches have important weaknesses:

- CNN DailyMail [Hermann et al., 2015]: Summaries in the CNN DailyMail dataset are often longer than human-written summaries, which can hinder the production of concise and coherent summaries by models trained on this dataset;
- Samsun [Liu, 2019]: The small size of the Samsun dataset restricts topic diversity and generalizability for models trained on this dataset;
- XSum [Narayan et al., 2018]: The XSum dataset consists of extremely short summaries, typically a single sentence. While this simplifies the Summarization task, it limits the expressiveness and depth of the summaries, as they may not capture the full context of the source articles;
- Gigaword [Graff et al., 2003]: The Gigaword dataset is generated using a rule-based approach, resulting in extractive summaries that may not fully capture the key ideas of the source articles. This limitation restricts the dataset’s utility for training models that aim to perform abstractive Summarization.

For the reasons mentioned above, we develop a specialized training objective that adapts the models to perform Summarization already while pre-training, such that afterward, the model is more independent of the fine-tuning dataset quality.

### 9.1.1 Static Document-level Summary

Static Document-Level Summary (SDS) is an objective built over the ideas developed in the previous Chapters. We exploit the raw subdivision of large documents in paragraphs as a light form of supervision. We task the model at predicting the first paragraph given all the successive. The reason is that the first paragraph acts often as an introduction or summary of the whole document’s content. More formally, given a document  $\mathbf{D} = [P_1, \dots, P_m]$ , the model is fed with  $[P_2, \dots, P_m]$  and should predict  $P_1$ . The training is performed with Teacher-Forcing, and we



Los Angeles, often referred to by its initials L.A., is a major city in the U.S. state of California. With a 2020 population of 3,898,747, it is the largest city in the state, as well as the second-largest city in the United States following New York City.

The city of Los Angeles lies in a basin in Southern California. Adjacent to the Pacific Ocean extending through the Santa Monica Mountains and into the San Fernando Valley, it covers about 469 square miles (1,210 km<sup>2</sup>).

Los Angeles has a diverse economy with a broad range of industries. It has one of the busiest container ports in the Americas. The Los Angeles metropolitan area had a gross metropolitan product of over \$1.0 trillion in 2018.

The area that became Los Angeles was claimed by Spain in 1542. It became a part of Mexico in 1821 following the Mexican War of Independence. At the end of the Mexican-American War, Los Angeles was purchased as part of the Treaty of Guadalupe Hidalgo and became part of the United States. Los Angeles was incorporated as a municipality in 1850.

Figure 9.1: Example of SDS objective. The input is in orange and the gold labels in blue.



utilize several generative algorithms, such as Beam Search [Freitag and Al-Onaizan, 2017], while evaluating. An example of the objective is depicted in Figure 9.1.

## 9.2 Experiments

We describe the experimental setting for the pre-training and the evaluation of SDS.

### 9.2.1 Continuous Pre-Training

We do continuous pre-training starting from  $\text{BART}_{\text{Base}}$  and  $\text{T5}_{\text{Base}}$  publicly available checkpoints. Training those models from scratch would have required a huge amount of resources, which are out of the scope of this Thesis. Moreover, by showing that a small amount of continuous pre-training is enough to improve performance in Summarization, we can state that our techniques can be easily adapted to other architectures and corresponding pre-trained models, such as GPT [Radford and Narasimhan, 2018].

We do continuous pre-training on the same datasets used to pre-train the original checkpoints of  $\text{BART}_{\text{Base}}$  and  $\text{T5}_{\text{Base}}$ . More specifically, we exploit Wikipedia, the BookCorpus, CC-News and OpenWebText for BART and the clean split of the C4 dataset for T5. The datasets have been processed for the SDS task described above and we filter documents with first paragraphs shorter than 2 sentences or with less than 50 characters. Finally, we reduced the resulting datasets to 100M training examples with random sampling.

In Table 9.1 we report the main hyper-parameters used while doing continuous pre-training. We fix the maximum input sequence length of the encoder to 512 tokens, while the limit over the decoder input, which is the gold summary, is set to 256 tokens. We use the Teacher-Forcing (TF) objective for the training and generative algorithms such as Beam Search for fine-tuning. Teacher-Forcing is an effective technique for training generative models, wherein the decoder receives the correct gold summaries as input, and gold labels are generated by shifting the summary tokens to the left by one position.

Regarding the optimization, we employ FuseAdam with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$  and a weight decay of 0.01 as in the other experiments, and we use a triangular learning rate scheduler which warms up to the first 10% of the training steps. We run the two pre-trainings on our machine (see Section 1.2 for more details) with *fp16* and DeepSpeed [Rasley et al., 2020].  $\text{BART}_{\text{Base}}$  requires about 3.5 days before completing the 100K training steps while T5 needs 4 days even if it is fed with half of the tokens of the former. This can be explained by the architectures:  $\text{BART}_{\text{Base}}$  is a Transformer model with 6 layers in the encoder and 6 in the decoder.  $\text{T5}_{\text{Base}}$  instead uses 12 layers in both the encoder and the decoder, thus being a much larger model.

Model	Steps	LR	BS	MSL	# Tokens	Effort	Objectives
$\text{BART}_{\text{Base}}$ + SDS	100K	$1 \cdot 10^{-4}$	2048	512/256	104B/52B	+5.0%	TF (1.0)
$\text{T5}_{\text{Base}}$ + SDS	100K	$1 \cdot 10^{-4}$	1024	512/256	52B/26B	+5.0%	TF (1.0)

Table 9.1: Hyper-parameters used for the continuous pre-training with the SDS objective. “MSL” is the maximum allowed input length on both the encoder and the decoder. “# Tokens” shows the tokens seen by the encoder and the decoder respectively. “Effort” is the amount of FLOPS used in this experiment compared to the original pre-training of the models. “TF” is the Teacher-Forcing objective used for the training.

Parameter	CNN-DailyMail	XSum	Samsum	Gigaword
Batch Size	{64, 128, 256}	{64, 128, 256}	{16, 32, 64}	{128, 256, 512}
Learning Rate	$\{5 \cdot 10^{-6}, 1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{5 \cdot 10^{-6}, 1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$	$\{1 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$
Max epochs	20	20	40	8
Max document length	1024	768	256	128
Max summary length	256	128	128	64

Table 9.2: Hyper-parameters search space for the fine-tuning of our continuously pre-trained models.

Parameter	CNN-DailyMail	XSum	Samsum	Gigaword
Min new tokens	56	14	8	4
Max new tokens	142	96	72	40
Num beams	2	2	4	4
Length penalty	2.0	1.5	1.5	1.0
Temperature	0.7	0.7	0.7	0.7
Top- $k$	8	8	8	8

Table 9.3: Generation hyper-parameters used in the evaluation of our models. They are compatible with HuggingFace [Wolf et al., 2020] generation pipelines.

## 9.2.2 Fine-Tuning

We extensively evaluate our approaches and the baselines on 4 datasets for Summarization: CNN-DailyMail, XSum, Samsum and Gigaword. They cover a wide range of Summarization tasks branches: (i) CNN-DailyMail consists of news articles accompanied by verbose human-generated summaries; (ii) XSum contains short news articles and single-sentence summaries; (iii) Samsum is composed of dialogues between fictional characters along with multi-sentence summaries and (iv) Gigaword is a large-scale dataset consisting of news articles and their machine-extracted headline summaries. More information about each Summarization dataset is given in Section 4.2.3.

We report the summary of the hyper-parameters search space in Table 9.2. We adapt the document and summary maximum length based on the datasets’ statistics to reduce training time and energy consumption. For the generation, we report the main hyper-parameters in Table 9.3. Notice that results may vary slightly from those reported in the original works because (i) we do not perform best model selection and (ii) we use a different decoding setting. Regarding the optimization, we exploit the same optimizer we use for the continuous pre-training, and we dynamically set the learning scheduler warmup to match the number of steps in the first epoch.

## 9.3 Evaluation

Evaluating generative tasks is not trivial because natural language is ambiguous and there are many ways of expressing the same concept. Thus, to measure the performance of our models and perform an accurate comparison, we employ several different metrics: ROUGE- $\{1, 2, L, LSum\}$  [Lin, 2004], Bleu [Papineni et al., 2002] and BLEURT [Sellam et al., 2020].

BLEU calculates the overlap of  $n$ -grams between the generated summary and the gold summaries. It focuses on precision, measuring how well the generated text matches the references. ROUGE, on the other hand, is a comprehensive evaluation package with multiple metrics. One of its metrics, ROUGE- $n$  (with  $n \in \{1, 2\}$  in our test setting), is similar to BLEU as it counts common  $n$ -grams. However, ROUGE- $n$  is recall-oriented and emphasizes the ability to retrieve important information rather than exact matching. ROUGE-L evaluates the Longest Common Subsequence (LCS) between the generated summary and the references.

It captures similarities beyond simple word matches. Finally, ROUGE-S measures similarity using skip-bigram co-occurrence statistics between the generated summary and the set of target references.

Among the parametric metrics, we utilize BLEURT, which stands out from BLEU or ROUGE by focusing specifically on comparing the encodings of the generated text and the target references. The advantage of using encodings is that concepts expressed differently but with similar meanings should have comparable embedding vectors. BLEURT employs a supervised learning approach, using human judgments to create a training dataset. The dataset comprises pairs of system-generated responses and reference responses, along with quality scores indicating the text’s quality. These pairs train a regression model that maps the generated response’s features to a quality score. BLEURT then outputs a similarity score, reflecting the degree of similarity between the generated and reference summaries.

We do not measure performance with BERT-Score [Zhang et al., 2019] because preliminary evaluations showed a similarity between generated and target summaries of  $\sim 96\%$  for every experiment, even using a RoBERTa<sub>Large</sub> model as sentences encoder. Moreover, BERT-Score is not fair and includes Social Biases [Sun et al., 2022].

A key limitation of these metrics is that the evaluation becomes more accurate as the number of target summaries increases. Since every dataset we consider has only a single gold summary for each input document, in Section 9.4.1 we show a possible solution to this issue.

## 9.4 Results

Model	CNN/DailyMail						XSum					
	R-1	R-2	R-L	R-LSum	BLEU	BLEURT	R-1	R-2	R-L	R-LSum	BLEU	BLEURT
BART <sub>Base</sub> [Lewis et al., 2020]	42.4 (0.1)	19.3 (0.1)	29.3 (0.0)	39.6 (0.0)	16.0 (0.0)	36.4 (0.1)	41.7 (0.0)	18.6 (0.0)	33.3 (0.0)	33.4 (0.2)	13.6 (0.1)	50.1 (0.1)
<b>BART<sub>Base</sub> + SDS</b>	<b>42.7</b> (0.3)	<b>19.4</b> (0.2)	<b>29.5</b> (0.2)	<b>39.7</b> (0.3)	<b>16.2</b> (0.2)	<b>36.8</b> (0.1)	<b>42.4</b> (0.0)	<b>19.3</b> (0.0)	<b>34.0</b> (0.0)	<b>34.0</b> (0.0)	<b>14.1</b> (0.0)	<b>50.9</b> (0.0)
T5 <sub>Base</sub> [Raffel et al., 2020]	40.7 (0.0)	18.4 (0.1)	27.4 (0.0)	37.6 (0.0)	29.3 (0.1)	34.9 (0.1)	37.8 (0.2)	14.6 (0.2)	28.8 (0.3)	28.8 (0.3)	18.0 (0.6)	46.4 (0.3)
<b>T5<sub>Base</sub> + SDS</b>	<b>41.6</b> (0.1)	<b>19.2</b> (0.1)	<b>28.3</b> (0.1)	<b>38.5</b> (0.1)	<b>30.3</b> (0.1)	<b>35.2</b> (0.1)	<b>38.2</b> (0.1)	<b>15.0</b> (0.1)	<b>29.3</b> (0.0)	<b>29.3</b> (0.0)	<b>18.9</b> (0.6)	<b>47.1</b> (0.0)

Table 9.4: Results on CNN/DailyMail and Xsum. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over the baselines.

Model	Samsun						Gigaword					
	R-1	R-2	R-L	R-LSum	BLEU	BLEURT	R-1	R-2	R-L	R-LSum	BLEU	BLEURT
BART <sub>Base</sub> [Lewis et al., 2020] *	47.8 (0.3)	22.8 (0.4)	38.0 (0.5)	44.0 (0.4)	16.6 (0.5)	54.3 (0.1)	38.7 (0.1)	19.4 (0.1)	35.6 (0.1)	35.6 (0.1)	14.6 (0.1)	38.9 (0.1)
<b>BART<sub>Base</sub> + SDS</b>	<b>48.9</b> (0.1)	<b>24.1</b> (0.1)	<b>39.3</b> (0.1)	<b>45.1</b> (0.1)	<b>17.7</b> (0.1)	<b>54.7</b> (0.1)	<b>38.8</b> (0.0)	<b>19.4</b> (0.0)	<b>35.6</b> (0.1)	<b>35.6</b> (0.1)	<b>14.7</b> (0.1)	<b>39.0</b> (0.1)
T5 <sub>Base</sub> [Raffel et al., 2020] *	46.0 (0.3)	22.3 (0.3)	36.0 (0.3)	42.1 (0.3)	25.0 (0.1)	52.5 (0.2)	35.5 (0.1)	16.8 (0.1)	32.6 (0.2)	32.7 (0.2)	13.6 (0.0)	36.3 (0.0)
<b>T5<sub>Base</sub> + SDS</b>	<b>47.6</b> (0.2)	<b>22.9</b> (0.2)	<b>37.6</b> (0.2)	<b>43.1</b> (0.2)	24.6 (0.2)	<b>53.8</b> (0.2)	<b>36.1</b> (0.1)	<b>17.2</b> (0.1)	<b>33.1</b> (0.1)	<b>33.1</b> (0.1)	<b>13.8</b> (0.0)	<b>36.5</b> (0.0)

Table 9.5: Results on Samsun and Gigaword. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over the baselines.

In this Section, we analyze the effects of the continuous pre-training on the final Summarization tasks. In Table 9.4 and 9.5 we report the results achieved on CNN-DailyMail, Xsum, Samsun and Gigaword.

The general pattern shows that models augmented with our SDS objective outperform the baselines on all tasks. For example, T5<sub>Base</sub> + SDS improves ROUGE-1 and ROUGE-2 F1-Scores by 0.9 and 0.8 points on CNN/DailyMail. BART<sub>Base</sub> + SDS also features gains over the baselines of 0.3 points of ROUGE-1 and 0.4 points when summaries are compared with BLEURT. Regarding Xsum, both T5<sub>Base</sub> and BART<sub>Base</sub> outperform the corresponding baselines by 0.7 and 0.4 ROUGE-1 points, respectively. On BLEU and BLEURT, the gap increases up to

0.8 points. Notice that almost all gains over the baselines are statistically significant thanks to the very small standard deviation of the evaluations.

We state again that a perfect model does not score 100.0 in metrics such as ROUGE and BLEU because the same concepts can be expressed in different ways that are equally correct. We consider a ROUGE-1 of about 50.0 as a very good result, given that the actual state-of-the-art in Abstractive Summarization, MoCa [Zhang et al., 2022b], achieves 48.9 on CNN/DailyMail. We do not compare with this technique because (i) the experimental setting is different, (ii) they use a different architecture and (iii) the total number of parameters of the models is different.

On the Samsun benchmark, the improvements are even larger. BART<sub>Base</sub> + SDS outperform its baseline by 1.3 points when performance is measured with ROUGE-2 F1-Score while T5<sub>Base</sub> augmented with SDS provides gains of 1.6 ROUGE-1 and 1.3 BLEURT points.

Finally, on Gigaword our methodologies provide mixed improvements based on the model’s architecture. On BART<sub>Base</sub>, our model is on par with the baseline while T5<sub>Base</sub> + SDS outperforms the original model by a small but statistically significant gap. This demonstrates empirically that benchmarks such as Gigaword that do not align well with our objectives are not significantly penalized by the additional continuous pre-training. The reason Gigaword does not align well is that its summaries are very short and automatically generated, while we use the entire first paragraph of documents from various sources as a form of self-supervision. We leave the study of exploiting the titles of large collections of documents as summaries for future work.

### 9.4.1 Evaluation with Silver Summaries

Model	CNN/DailyMail						XSum					
	R-1	R-2	R-L	R-LSum	BLEU	BLEURT	R-1	R-2	R-L	R-LSum	BLEU	BLEURT
BART <sub>Base</sub> [Lewis et al., 2020] *	55.5 (0.1)	36.7 (0.1)	41.7 (0.1)	52.4 (0.1)	51.6 (0.1)	50.1 (0.1)	46.4 (0.1)	22.1 (0.0)	36.1 (0.0)	37.4 (0.1)	29.5 (0.2)	53.8 (0.1)
<b>BART<sub>Base</sub> + SDS</b>	<b>55.9</b> (0.1)	<b>37.6</b> (0.2)	<b>42.5</b> (0.2)	<b>53.0</b> (0.1)	<b>52.2</b> (0.3)	<b>50.1</b> (0.1)	<b>47.0</b> (0.0)	<b>22.8</b> (0.0)	<b>36.7</b> (0.1)	<b>38.0</b> (0.1)	<b>29.9</b> (0.1)	<b>54.4</b> (0.0)
T5 <sub>Base</sub> [Raffel et al., 2020] *	53.5 (0.0)	34.6 (0.1)	39.6 (0.1)	50.4 (0.0)	50.5 (0.2)	49.8 (0.1)	44.5 (0.0)	20.2 (0.0)	33.4 (0.1)	35.1 (0.1)	28.7 (0.1)	51.4 (0.0)
<b>T5<sub>Base</sub> + SDS</b>	<b>53.8</b> (0.1)	<b>34.8</b> (0.1)	<b>40.0</b> (0.1)	<b>50.7</b> (0.0)	<b>51.3</b> (0.2)	<b>50.0</b> (0.2)	<b>45.1</b> (0.1)	<b>21.0</b> (0.1)	<b>34.1</b> (0.1)	<b>35.9</b> (0.1)	<b>30.7</b> (0.2)	<b>51.8</b> (0.1)

Table 9.6: Results on CNN/DailyMail and XSum in which the test set is augmented with 8 additional machine-generated summaries. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over the baselines.

Model	Samsun						Gigaword					
	R-1	R-2	R-L	R-LSum	BLEU	BLEURT	R-1	R-2	R-L	R-LSum	BLEU	BLEURT
BART <sub>Base</sub> [Lewis et al., 2020] *	55.2 (0.1)	29.4 (0.3)	42.9 (0.5)	51.0 (0.2)	44.9 (0.8)	58.8 (0.1)	38.9 (0.1)	19.6 (0.1)	35.8 (0.1)	35.8 (0.1)	15.0 (0.1)	40.9 (0.1)
<b>BART<sub>Base</sub> + SDS</b>	<b>56.4</b> (0.1)	<b>30.6</b> (0.1)	<b>44.3</b> (0.1)	<b>52.3</b> (0.1)	<b>46.4</b> (0.0)	<b>59.6</b> (0.1)	<b>39.2</b> (0.0)	<b>19.8</b> (0.1)	<b>36.0</b> (0.1)	<b>36.1</b> (0.1)	<b>15.2</b> (0.2)	<b>41.1</b> (0.1)
T5 <sub>Base</sub> [Raffel et al., 2020] *	54.6 (0.2)	29.0 (0.1)	42.3 (0.3)	49.9 (0.2)	42.9 (0.4)	58.9 (0.1)	36.1 (0.1)	17.1 (0.1)	32.8 (0.2)	32.8 (0.2)	13.1 (0.1)	39.9 (0.1)
<b>T5<sub>Base</sub> + SDS</b>	<b>54.8</b> (0.2)	<b>29.3</b> (0.3)	42.0 (0.3)	<b>49.9</b> (0.2)	<b>43.8</b> (0.6)	58.7 (0.2)	<b>36.5</b> (0.0)	<b>17.4</b> (0.1)	<b>33.2</b> (0.1)	<b>33.2</b> (0.1)	<b>13.4</b> (0.2)	<b>40.3</b> (0.2)

Table 9.7: Results on Samsun and Gigaword in which the test set is augmented with 8 additional machine-generated summaries. We use bold formatting for our proposals and the best results. We underline statistically significant improvements over the baselines.

As we mentioned before, the evaluation of generated summaries is optimal when there is a large number of references for each document. We try to solve this issue by generating  $n$  additional silver summaries for each document exploiting a very large language model: Falcon [Almazrouei et al., 2023]. Specifically, we use the version fine-tuned on instruction-based conversations between users, called “falcon-40B-instruct”. We provide the model with the following prompt: “Document {document}\n\nSummary:”, which is what is suggested by the authors. We generate 8 additional summaries for each document in the datasets test sets and we exploit the best checkpoints of the previous Section as a starting point. Since the additional summaries have

not been reviewed by humans, the results may vary based not only on the performance of our models and the baselines but also on the capacity of Falcon in creating high-quality and diverse summaries. Results are shown in Table 9.6 and 9.7.

Generally, the results are higher than the previous Section because we report the highest similarity between the generated summaries and all the references. On CNN/DailyMail, our methods outperform the baselines on ROUGE and BLEU metrics, with increments of up to 0.9 points. On XSum, BART and T5 combined with our SDS objective improve the BLUERT by 0.6 and 0.4 points respectively.

The performance on the Samsun dataset, which contains summaries of dialogues, is improved by  $\text{BART}_{\text{Base}} + \text{SDS}$  by 1.2 ROUGE-1/2 points and by more than 1.3 on ROUGE-L and ROUGE-LSum, compared to the baseline. At the same time,  $\text{T5}_{\text{Base}} + \text{SDS}$  features accuracies similar to the baseline, probably because the additionally generated summaries align better with the baseline than our model.

Finally, we observe notable and statistically significant enhancements in our models when employing SDS on Gigaword, as compared to the baselines. It is worth noting that improvements in this dataset typically fall within the range of decimal points to a few integer points [Zhang et al., 2022b]. This demonstrates again that specialized pre-training can help models at adapting to the final task already while pre-training.

## 9.4.2 General Considerations

In this Section, we provided an overview of the improvements achievable by performing a continuous pre-training on Auto-Regressive models such as  $\text{BART}_{\text{Base}}$  and  $\text{T5}_{\text{Base}}$ , by exploiting only unlabeled documents that can be crawled freely from the web. Notice that our methodologies are orthogonal to related works that apply custom decoding strategies, specialized fine-tuning and other techniques. For example, in MoCa [Zhang et al., 2022b] or BRIO [Liu et al., 2022], our models could easily replace BART and T5 as starting checkpoints before the application of the ad-hoc fine-tuning for Summarization, which is described in the referenced works. We leave this research direction as a future work.



# Chapter 10

## Conclusions

In this Chapter, we summarize the discoveries and the best results achieved in this Thesis. We study several innovative objectives for the efficient pre-training and the specialized continuous pre-training of Transformer-based models. We perform experiments on 4 different pre-training datasets, namely Wikipedia, BookCorpus, OpenWebText and CC-News and we evaluate our models on 21 benchmarks, covering a wide range of tasks: Fact Verification, Question Answering (Answer Sentence Selection), Summarization, Linguistic Acceptability, Sentiment Analysis, Paraphrasing, Natural Language Inference and Textual Entailment. We also release 2 new datasets for Answer Sentence Selection, which we call NewsAS2 and TriviaAS2 (derived from NewsQA and TriviaQA).

### 10.1 Alternative Efficient Objectives

Regarding the efficient pre-training, we develop 3 alternative objectives to BERT’s MLM and ELECTRA’s TD, which we name RTS, C-RTS and SLM. Those results are presented in Chapter 5.

We discover through empirical evaluation that RTS and C-RTS can pre-train a language model to the level of performance of MLM while using fewer computational resources. Specifically, RTS and C-RTS save 20% (base models) and 45% (small models) training time to match or even outperform the performance of MLM on several tasks. For example, BERT<sub>Base</sub> + RTS outperforms BERT<sub>Base</sub> + MLM by 0.2 points on the GLUE benchmark and by 0.2 MAP points on TREC-QA. C-RTS instead outperforms a similar BERT<sub>Small</sub> model trained with MLM by 1.6 points on GLUE and by 0.8 and 1.0 MAP points on TREC-QA and ASNQ, respectively.

We also propose SLM, which focuses instead on providing a higher level of performance while requiring the same computational budget of MLM for the pre-training. SLM applied to BERT<sub>Base</sub> outperforms MLM by 0.7 points on the GLUE benchmark and has superior performance on 4 out of 5 Answer Sentence Selection datasets.

Future works in this research direction include the application of our objectives to more efficient models such as ALBERT [Lan et al., 2020] or DistilBERT [Sanh et al., 2020]. We noticed that with our objectives, the smaller the number of parameters, the larger the improvements. Moreover, we plan to train a larger language model with SLM, which was shown to outperform MLM in most tasks while requiring the same computational budget for pre-training.

## 10.2 Specialized Objectives

We design several training functions that align structurally with the final downstream task. The main advantage of our specialized training tasks is that they do not require manually annotated data, because they exploit the weak structure of large corpora like Wikipedia as a form of self-supervision. Moreover, they are orthogonal to related works, because our checkpoints can be easily exploited as a starting point for more advanced techniques since they share the same architectures of the original models.

### 10.2.1 Multi-Sentence Inference

In this research, we develop a self-supervised pre-training objective we call MSPP, that accustoms a Transformer model to efficiently perform inference over many input text spans. Multi-Sentence Inference can provide better accuracies than common Pairwise classifiers in different tasks. For example, in Answer Sentence Selection it can compare different sentence candidates to find the best answer to a question. In Fact Verification instead, it can leverage several retrieved evidence sentences to predict whether the claim is supported.

We perform continuous pre-training on raw data starting from a RoBERTa<sub>Base</sub> checkpoint and we evaluate our models over several datasets for AS2 and for Fact Verification. We show that checkpoints continuously pre-trained with MSPP can outperform Pairwise baselines by a large margin. For example, we improved the P@1 on ASNQ by 3.5 points, on WikiQA by 8.2 points and on TREC-QA by 3.8 points. When tested in Fact Verification on FEVER, we reach state-of-the-art Label Accuracy on the dev set and results close to other methods using larger models on the test set. Results for these experiments are described in Chapter 6.

In future work, we plan to apply our models to evidence retrieved with a more powerful IR pipeline, since we are currently limited by the performance of a BERT-based DocIR. Moreover, we would like to exploit larger language models to achieve state-of-the-art accuracy on the FEVER test set and even better re-rankings on the AS2 datasets.

### 10.2.2 Answer Sentence Selection

In this research direction, we develop ad-hoc pre-training objectives specifically for the Answer Sentence Selection task. We exploit sources of raw text such as CC-News to create 3 self-supervised objectives over two text spans that align structurally with AS2, called SSP, SP and PSD. The experiments are provided in Chapter 7.

After a continuous pre-training step over large unlabeled corpora, we show that our objectives can help Transformers at predicting whether an answer is correct for a given question. For example, our RoBERTa<sub>Base</sub> + SSP outperforms the corresponding baseline by 2.3, 4.6 and 1.2 P@1 points on ASNQ, WikiQA and TriviaAS2, respectively. We also reach very high scores on ASNQ when combining our objectives with our DeBERTaV3<sub>Large</sub>, surpassing the 70.0 points in P@1 for the first time. For NewsAS2 and TriviaAS2, since they are freshly released within this work, we set very strong baselines for future research. By performing a transfer step over ASNQ as in TANDA before fine-tuning and starting from our DeBERTaV3<sub>Large</sub> + ALL checkpoint, we reach also the new state-of-the-art on WikiQA and TREC-QA, with MAPs of 92.7 and 95.4 points, respectively.

Finally, we show that our objectives can help especially when data is scarce. We study the consequences of reducing all the training datasets to contain only a few thousand examples. The results show improvements in P@1 up to 16 points depending on the objective and the target dataset.



As a future research direction, we plan to combine our pre-trained checkpoints with more refined techniques, such as RLAS-BIABC or TANDA. Moreover, if the computational resources will allow it, we plan to apply our specialized objectives to larger models, such as DeBERTaV2-XXLarge [He et al., 2020].

### 10.2.3 Contextual Answer Sentence Selection

In Chapter 8, we continue the development of pre-training tasks by addressing Contextual AS2. In this downstream task, the system is provided with additional text that could be exploited for resolving entities in answer candidates, and thus performing a better re-ranking. We extended the SSP objective described before with additional context extracted from source documents with 3 different techniques: SDC, DPC and DSLC.

After a continuous pre-training over several millions of documents using only self-supervision, we show that our objectives outperform similar baselines augmented with context by a large margin. The reason is that even though the baselines receive the context as input, they don't know how to effectively take advantage of it for a better re-ranking. On the contrary, after continuous pre-training our models have learned to fully exploit this additional information.

We achieve the state-of-the-art on ASNQ with DeBERTaV3<sub>Large</sub> + SSP (ALL), with a MAP of 78.8 and a P@1 of 75.1 points, and we also set very strong baselines for contextual NewsAS2 and TriviaAS2. Other examples of performance improvements are shown on WikiQA and TREC-QA, where our specialized training improves performance by up to 8.0 points in P@1.

In the future, we plan to exploit Large Language Models (LLMs) to generate better context, with architectures like Falcon [Almazrouei et al., 2023], as we did for TREC-QA. Moreover, we want to perform a better mixed continuous pre-training using all types of context (SDC, DPC, DSLC) together by challenging the model at predicting both the binary label for SSP and the type of context provided.

### 10.2.4 Summarization

The last task we address in this work is Summarization, in Chapter 9. We develop a special training objective (SDS) that tasks the model to predict the first paragraph of a model given all the others. We perform a continuous pre-training with our objective and Auto-Regressive models such as BART and T5.

By extensively evaluating our continuously pre-trained models on 4 datasets for Summarization, we show statistically significant gains compared to vanilla models. For example, our T5<sub>Base</sub> + SDC improves ROUGE-1 by 0.9 and 0.4 points on CNN/DailyMail and XSum, respectively. Additionally, our BART<sub>Base</sub> + SDS outperforms its corresponding baseline by 1.6 and 0.6 ROUGE-1 points on Samsun and Gigaword, which is remarkable considering that typically the improvements in Summarization range between a few decimal points and a few units.

Notice that the performance on Summarization is difficult to evaluate, because each document may have several correct summaries. To address this issue, we designed an experiment in which additional silver summaries for the test sets are generated with a Large LM, e.g. Falcon. By testing on the augmented test sets, we show significant improvements in all the metrics used for the evaluation. For example, our BART<sub>Base</sub> + SDS model improves the BLEURT score by 0.8 and 0.6 points on Samsun and XSum, respectively. Similarly, T5<sub>Base</sub> + SDS increases the BLUE score by 0.8 points on CNN/DailyMail and by 2.0 points on XSum.

Future research directions include the application of our techniques to large models, such as BART<sub>Large</sub> and T5<sub>Large</sub>, as well as exploiting LLMs like BLOOMZ [Scao et al., 2022] to generate additional high-quality summaries both for training and evaluation.



# Bibliography

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.
- Qingyao Ai, Keping Bi, Jiafeng Guo, and W. Bruce Croft. Learning a deep listwise context model for ranking refinement. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, page 135–144, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3209985. URL <https://doi.org/10.1145/3209978.3209985>.
- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1409.0473>.
- Gianni Barlacchi, Ivano Lauriola, Alessandro Moschitti, Marco Del Tredici, Xiaoyu Shen, Thuy Vu, Bill Byrne, and Adrià de Gispert. FocusQA: Open-domain question answering with a context in focus. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5195–5208, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.381. URL <https://aclanthology.org/2022.findings-emnlp.381>.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL <https://arxiv.org/abs/2004.05150>.
- Weijie Bian, Si Li, Zhao Yang, Guang Chen, and Zhiqing Lin. A compare-aggregate model with dynamic-clip attention for answer selection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17*, page 1987–1990, New York, NY, USA, 2017a. Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3133089. URL <https://doi.org/10.1145/3132847.3133089>.

- Weijie Bian, Si Li, Zhao Yang, Guang Chen, and Zhiqing Lin. A compare-aggregate model with dynamic-clip attention for answer selection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17*, page 1987–1990, New York, NY, USA, 2017b. Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3133089. URL <https://doi.org/10.1145/3132847.3133089>.
- Zhengda Bian, Hongxin Liu, Boxiang Wang, Haichen Huang, Yongbin Li, Chuanrui Wang, Fan Cui, and Yang You. Colossal-ai: A unified deep learning system for large-scale parallel training. *arXiv preprint arXiv:2110.14883*, 2021.
- Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.
- Daniele Bonadiman and Alessandro Moschitti. A study on efficiency, accuracy and document structure for answer sentence selection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5211–5222, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.457. URL <https://aclanthology.org/2020.coling-main.457>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- Stefano Campese, Ivano Lauriola, and Alessandro Moschitti. Quadro: Dataset and models for question-answer database retrieval. *ArXiv*, abs/2304.01003, 2023.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL <https://aclanthology.org/S17-2001>.
- Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. Pre-training tasks for embedding-based large-scale retrieval. *ArXiv*, abs/2002.03932, 2020.
- Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.427. URL <https://aclanthology.org/2022.acl-long.427>.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL <https://aclanthology.org/W14-4012>.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker

Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL <https://aclanthology.org/P19-1285>.

Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. Torchmetrics - measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70):4101, 2022. doi: 10.21105/joss.04101. URL <https://doi.org/10.21105/joss.04101>.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.

Luca Di Liello, Daniele Bonadiman, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Alessandro Moschitti. Cross-language transformer adaptation for frequently asked questions, sep 2020.

Luca Di Liello, Matteo Gabburo, and Alessandro Moschitti. Effective pretraining objectives for transformer-based autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5533–5547, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.405. URL <https://aclanthology.org/2022.findings-emnlp.405>.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. Paragraph-based transformer pre-training for multi-sentence inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2521–2531, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.181. URL <https://aclanthology.org/2022.naacl-main.181>.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. Pre-training transformer models with sentence-level objectives for answer sentence selection. In *Proceedings of the 2022*

- Conference on Empirical Methods in Natural Language Processing*, pages 11806–11816, Abu Dhabi, United Arab Emirates, December 2022c. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.810. URL <https://aclanthology.org/2022.emnlp-main.810>.
- Luca Di Liello, Siddhant Garg, and Alessandro Moschitti. Context-aware transformer pre-training for answer sentence selection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–468, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.40. URL <https://aclanthology.org/2023.acl-short.40>.
- William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL <https://aclanthology.org/I05-5002>.
- Cícero Nogueira dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks. *CoRR*, abs/1602.03609, 2016. URL <http://arxiv.org/abs/1602.03609>.
- Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. ISSN 0364-0213. doi: [https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E). URL <https://www.sciencedirect.com/science/article/pii/036402139090002E>.
- William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019. URL <https://github.com/Lightning-AI/lightning>.
- Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5801. URL <https://aclanthology.org/D19-5801>.
- Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3207. URL <https://aclanthology.org/W17-3207>.
- Matteo Gabburo, Rik Koncel-Kedziorski, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. Knowledge transfer from answer ranking to answer generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9481–9495, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.645. URL <https://aclanthology.org/2022.emnlp-main.645>.
- Philip Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020.
- Siddhant Garg, Thuy Vu, and Alessandro Moschitti. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection, 2019.

- Hamid Gharagozlou, Javad Mohammadzadeh, Azam Bastanfard, Saeed Shiry Ghidary, and Ripon Chakraborty. Rlas-biabc: A reinforcement learning-based answer selection using the bert model boosted by an improved abc algorithm. *Intell. Neuroscience*, 2022, jan 2022. ISSN 1687-5265. doi: 10.1155/2022/7839840. URL <https://doi.org/10.1155/2022/7839840>.
- Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. Contextual lstm (clstm) models for large scale nlp tasks. *ArXiv*, abs/1602.06291, 2016.
- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/W07-1401>.
- John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.72. URL <https://aclanthology.org/2021.acl-long.72>.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34, 2003.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML'20*. JMLR.org, 2020.
- Rujun Han, Luca Soldaini, and Alessandro Moschitti. Modeling context in answer sentence selection systems on a latency budget. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3005–3010, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.261. URL <https://aclanthology.org/2021.eacl-main.261>.
- Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520. URL <https://doi.org/10.1080/00437956.1954.11659520>.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020. URL <https://arxiv.org/abs/2006.03654>.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. Deberv3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543, 2021. URL <https://arxiv.org/abs/2111.09543>.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015. URL <http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend>.
- Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural computation*, 9: 1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

- Chao-Chun Hsu, Eric Lind, Luca Soldaini, and Alessandro Moschitti. Answer generation for retrieval-based question answering systems. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4276–4282, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.374. URL <https://aclanthology.org/2021.findings-acl.374>.
- Peter Izsak, Moshe Berchansky, and Omer Levy. How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.831. URL <https://aclanthology.org/2021.emnlp-main.831>.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL <https://aclanthology.org/P17-1147>.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. doi: 10.1162/tacl.a.00300. URL <https://aclanthology.org/2020.tacl-1.5>.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL <https://aclanthology.org/2020.emnlp-main.550>.
- Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’20*, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401075. URL <https://doi.org/10.1145/3397271.3401075>.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- John F. Kolen and Stefan C. Kremer. *Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies*, pages 237–243. 2001. doi: 10.1109/9780470544037.ch14.
- Kimmo Koskenniemi. A general computational model for word-form recognition and production. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 178–181, Stanford, California, USA, July 1984. Association for Computational Linguistics. doi: 10.3115/980491.980529. URL <https://aclanthology.org/P84-1038>.
- Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia,



- July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL <https://aclanthology.org/P18-1007>.
- Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL <https://aclanthology.org/D18-2012>.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl.a.00276. URL <https://aclanthology.org/Q19-1026>.
- Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.
- Ivano Lauriola and Alessandro Moschitti. Answer sentence selection using local and global context in transformer models. In *ECIR 2021*, 2021. URL <https://www.amazon.science/publications/answer-sentence-selection-using-local-and-global-context-in-transformer-models>.
- Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.
- Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86*, page 24–26, New York, NY, USA, 1986. Association for Computing Machinery. ISBN 0897912241. doi: 10.1145/318723.318728. URL <https://doi.org/10.1145/318723.318728>.
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'12*, page 552–561. AAAI Press, 2012. ISBN 9781577355601.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL <https://aclanthology.org/2020.acl-main.703>.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021*

- Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.21. URL <https://aclanthology.org/2021.emnlp-demo.21>.
- Luca Di Liello, Pierfrancesco Ardino, Jacopo Gobbi, Paolo Morettin, and Stefano Teso. Efficient generation of structured objects with constrained adversarial networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- Luca Di Liello, Olga Uryupina, and Alessandro Moschitti. Language transfer for identifying diagnostic paragraphs in clinical notes. In *Italian Conference on Computational Linguistics*, 2021.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- Peter J. Liu, Yu-An Chung, and Jie Ren. Summae: Zero-shot abstractive text summarization using length-agnostic auto-encoders, 2019a.
- Yang Liu. Fine-tune bert for extractive summarization, 2019. URL <https://arxiv.org/abs/1903.10318>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019b.
- Yixin Liu and Pengfei Liu. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.135. URL <https://aclanthology.org/2021.acl-short.135>.
- Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.207. URL <https://aclanthology.org/2022.acl-long.207>.
- Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.655. URL <https://aclanthology.org/2020.acl-main.655>.
- Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.
- Alexandra Sasha Luccioni, Sylvain Viguiier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *ArXiv*, abs/2211.02001, 2022.

- Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL <https://aclanthology.org/D15-1166>.
- Agnes Lydia and Sagayaraj Francis. Adagrad - an optimizer for stochastic gradient descent. Volume 6:566–568, 05 2019.
- Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.
- Benjamin Muller, Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind, and Alessandro Moschitti. Cross-lingual open-domain question answering with answer sentence generation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 337–353, Online only, November 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.aacl-main.27>.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL <https://aclanthology.org/D18-1206>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL <https://aclanthology.org/N18-1202>.
- Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446, 2021.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.
- Anand Rajaraman and Jeffrey David Ullman. *Data Mining*, page 1–17. Cambridge University Press, 2011. doi: 10.1017/CBO9781139058452.002.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’20, page 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406703. URL <https://doi.org/10.1145/3394486.3406703>.
- Mathieu Ravaut, Shafiq Joty, and Nancy Chen. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.309. URL <https://aclanthology.org/2022.acl-long.309>.
- Radim Rehurek and Petr Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL <https://arxiv.org/abs/1908.10084>.
- Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL <https://doi.org/10.1561/1500000019>.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Francois Yvon, Matthias Gall'e, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klammer, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla A. Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heizerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Francois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruo Chen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenek Kasner, Alice Rueda,

- Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Undreaaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Olusola Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emily Baylor, Ezinwanne Ozoani, Fatim T Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Livia Macedo Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, M. K. K. Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, A. Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, R. Eisenberg, Robert Martin, Rodrigo L. Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aaroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, T. A. Laud, Th'eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yun chao Xu, Zhee Xiao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100, 2022.
- Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012.
- Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12):54–63, nov 2020. ISSN 0001-0782. doi: 10.1145/3381831. URL <https://doi.org/10.1145/3381831>.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL <https://aclanthology.org/2020.acl-main.704>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL <https://aclanthology.org/P16-1162>.
- Aliaksei Severyn and Alessandro Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15*, page 373–382, New

- York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336215. doi: 10.1145/2766462.2767738. URL <https://doi.org/10.1145/2766462.2767738>.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2074. URL <https://aclanthology.org/N18-2074>.
- Gehui Shen, Yunlun Yang, and Zhi-Hong Deng. Inter-weighted alignment network for sentence pair modeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1179–1189, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1122. URL <https://aclanthology.org/D17-1122>.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL <https://aclanthology.org/D13-1170>.
- Luca Soldaini and Alessandro Moschitti. The cascade transformer: an application for efficient answer sentence selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5697–5708, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.504. URL <https://aclanthology.org/2020.acl-main.504>.
- Dominik Stammach and Elliott Ash. e-fever: Explanations and summaries for automated fact checking. In Emiliano De Cristofaro and Preslav Nakov, editors, *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32 – 43, Arlington, VA, 2020-10. Hacks Hackers. ISBN 978-1-7359904-0-8. doi: 10.3929/ethz-b-000453826. Conference for Truth and Trust Online (TTO 2020) (virtual); Conference Location: online; Conference Date: October 16-17, 2020; Due to the Coronavirus (COVID-19) the conference was conducted virtually.
- Dominik Stammach and Guenter Neumann. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6616. URL <https://aclanthology.org/D19-6616>.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL <https://aclanthology.org/P19-1355>.
- Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.245. URL <https://aclanthology.org/2022.emnlp-main.245>.

- Chuanqi Tan, Furu Wei, Qingyu Zhou, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. Context-aware answer selection with hierarchical gated recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26:540–549, 2018.
- Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey, 2020.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL <https://aclanthology.org/N18-1074>.
- Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2): 26–31, 2012.
- Juan-Manuel Torres-Moreno. *Automatic Text Summarization*. 10 2014. ISBN 978-1-84821-668-6. doi: 10.1002/9781119004752.ch1.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- Quan Hung Tran, Tuan Lai, Gholamreza Haffari, Ingrid Zukerman, Trung Bui, and Hung Bui. The context-dependent additive recurrent neural net. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1274–1283, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1115. URL <https://aclanthology.org/N18-1115>.
- Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL <https://aclanthology.org/W17-2623>.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019.
- Kateryna Tymoshenko and Alessandro Moschitti. Strong and light baseline models for fact-checking joint inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4824–4830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.426. URL <https://aclanthology.org/2021.findings-acl.426>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.



- Alexei Vinokourov, Nello Cristianini, and John Shawe-Taylor. Inferring a semantic representation of text via cross-language correlation analysis. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL <https://proceedings.neurips.cc/paper/2002/file/d5e2fbef30a4eb668a203060ec8e5eef-Paper.pdf>.
- Ellen M. Voorhees and Dawn M. Tice. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May 2000. European Language Resources Association (ELRA). URL <http://www.lrec-conf.org/proceedings/lrec2000/pdf/26.pdf>.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL <https://aclanthology.org/W18-5446>.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537, 2019a. URL <http://arxiv.org/abs/1905.00537>.
- Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. What is the Jeopardy model? a quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 22–32, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/D07-1003>.
- Shuohang Wang and Jing Jiang. A compare-aggregate model for matching text sequences. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=HJTzHtqee>.
- Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. Structbert: Incorporating language structures into pre-training for deep language understanding, 2019b.
- Xinyue Wang. Large web archive collection infrastructure and services, 2023. URL <http://hdl.handle.net/10919/113345>.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl.a.00290. URL <https://aclanthology.org/Q19-1040>.
- Jeremy West, Dan Ventura, and Sean Warnick. A theoretical foundation for inductive transfer. Talk, 2007. Department of Computer Science, Faculty Advisor: Dan Ventura.
- Michael Widenius, Davis Axmark, and Paul DuBois. *MySQL Reference Manual*. O’Reilly & Associates, Inc., USA, 1st edition, 2002. ISBN 0596002653.
- Wikipedia contributors. Natural language processing — Wikipedia, the free encyclopedia, 2004. URL [https://en.wikipedia.org/wiki/Natural\\_language\\_processing](https://en.wikipedia.org/wiki/Natural_language_processing). [Online; accessed 22-July-2004].

- Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL <https://aclanthology.org/N18-1101>.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6>.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data, 2023.
- Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. A survey on green deep learning. *ArXiv*, abs/2111.05193, 2021.
- Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1237. URL <https://aclanthology.org/D15-1237>.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf).
- Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. A compare-aggregate model with latent clustering for answer selection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*, page 2093–2096, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3358148. URL <https://doi.org/10.1145/3357384.3358148>.
- Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062, 2020. URL <https://arxiv.org/abs/2007.14062>.
- Majid Zarharan, Mahsa Ghaderan, Amin Pourdabiri, Zahra Sayedi, Behrouz Minaei-Bidgoli, Sauleh Eetemadi, and Mohammad Taher Pilehvar. ParsFEVER: a dataset for Farsi fact extraction and verification. In *Proceedings of \*SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 99–104, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.starsem-1.9. URL <https://aclanthology.org/2021.starsem-1.9>.

- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020a.
- Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avi Sil, and Todd Ward. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5461–5468, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.440. URL <https://aclanthology.org/2020.emnlp-main.440>.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022a.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2019. URL <https://arxiv.org/abs/1904.09675>.
- Xingxing Zhang, Yiran Liu, Xun Wang, Pengcheng He, Yang Yu, Si-Qing Chen, Wayne Xiong, and Furu Wei. Momentum calibration for text generation, 2022b.
- Zeyu Zhang, Thuy Vu, and Alessandro Moschitti. Joint models for answer verification in question answering systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3252–3262, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.252. URL <https://aclanthology.org/2021.acl-long.252>.
- Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *The Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020. URL <https://www.microsoft.com/en-us/research/publication/transformer-xh-multi-evidence-reasoning-with-extra-hop-attention/>.
- Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.549. URL <https://aclanthology.org/2020.acl-main.549>.
- Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1085. URL <https://aclanthology.org/P19-1085>.
- Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015.



# Appendix A

## Background & Related Work

### A.1 Data Organization

**Structured data** Data are structured when they are defined as a set of entities/concepts and relations between them. Some of the oldest and most common techniques to store structured data are tables and relational databases, such as MySQL [Widenius et al., 2002] or MariaDB<sup>1</sup>. Those databases are usually queried in a specific language (e.g. SQL) to extract interesting entities and relations.

Other relevant techniques to store structured data are formal ontologies and knowledge graphs. Ontologies are representations of a domain through the usage of categories, entities and relations. The goal of an ontology is to represent a domain with all its characteristics and properties. Those structures are usually represented with graphs or databases.

On the other hand, knowledge graphs are similar to ontologies but more data-driven, and the representation of data may change based on the scope. When building a knowledge graph, the objective is to allow users to find target concepts and related entities as quickly as possible. Those structures are usually backed by a graph database (GDB).

**Semi-structured data** Data are semi-structured when they are organized similarly to structured data but attributes are variable and there is no general schema that defines concepts and entities. Moreover, entities and relations may contain attributes that relate to unstructured data. An example of a database to store semi-structured data is MongoDB<sup>2</sup>, in which the user is allowed to store custom dictionaries of key-value objects. Another example of semi-structured data is Wikipedia: the entities are the different arguments described by every page and relations are the links between different pages.

The growth of the World Wide Web is increasing exponentially the amount of semi-structured data available for Information Retrieval. Most of the data on the WWW is semi-structured because each blog, newspaper or website uses a different organization schema, or doesn't use one at all. Today, most of the algorithms based on neural networks for NLP are first pre-trained on huge amounts of semi-structured data before being specialized on downstream applications. For example, in our experiments, we create self-supervised training functions exploiting the weak structure of large corpora, which only define a paragraph-level subdivision.

**Unstructured data** Unstructured data in NLP are represented by raw text, without any kind of supervision. Unstructured text can be seen as a sequence of words, numbers, dates and

---

<sup>1</sup><https://mariadb.org>

<sup>2</sup><https://www.mongodb.com>

other forms of free text. Recent advances in NLP allow the conversion from unstructured to semi-structured data through methods that search for structures inside the text, such as entity extraction algorithms (NER) and pattern recognition systems.

## A.2 Language Models

### A.2.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a specialized type of neural network suited for sequential data processing, including natural language, time series, and audio. While RNNs share their origin with Feed-Forward neural networks, they belong to a distinct category due to the potential presence of cycles in the information flow graph.

When applied to text, RNNs create a sentence-level representation by traversing the input text from left to right. RNNs consist of recurrent units, where each unit's output becomes the input for the same unit at the next time step. This design enables the network to maintain an internal state (memory) that can capture dependencies between elements in the input sequence. In processing a sequence of elements, an RNN typically passes the input elements through the network one at a time, using each output as the input for the next time step. The network's internal state is updated based on both the current input and the previous internal state.

Training RNNs involves employing backpropagation through time, a variant of the backpropagation algorithm used for standard Feed-Forward neural networks. The network is unrolled over time to treat it as a standard Feed-Forward neural network, allowing for gradient computation and weight updates through gradient descent.

However, RNNs suffer from certain drawbacks, including inefficiency due to their inability to process tokens in parallel and the vanishing information problem for early tokens in long sequences [Kolen and Kremer, 2001].

### A.2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent a type of Artificial Neural Networks (ANNs) initially designed for Computer Vision tasks, including image classification and object recognition, but they have also been successfully adapted for NLP tasks.

In CNNs, convolution operates by extracting features from input data using a set of learnable filters. These filters slide over the input data, performing a dot product between the filter entries and the input data, generating a feature map.

An important characteristic of CNNs is weight sharing, where the same filters are utilized to scan the entire input. This feature reduces the number of parameters the network needs to learn and improves efficiency. CNNs also incorporate pooling layers and activation functions, which respectively downsample the feature maps and introduce non-linearity into the model.

For NLP tasks, CNNs treat the input text as a 2D matrix, with each word represented as a row and each word embedding represented as a column. Similar to image recognition tasks, the filters in CNNs for NLP slide over the input text, performing dot products between the filter entries and the input text, resulting in feature maps.

A notable distinction between CNNs for NLP and image recognition lies in the use of pooling. While pooling is employed in image recognition to downsample feature maps, it is typically avoided in NLP tasks to prevent the loss of vital information. Instead, max pooling is often employed to extract the most salient features from the feature map.

One significant advantage of CNNs over RNNs is their efficiency in processing input text in parallel, enabling them to take advantage of hardware acceleration technologies for faster multiplication of large matrices.

## A.3 Tokenization methodologies

Here we provide an in-depth analysis of the most common tokenization techniques.

### A.3.1 Byte-Pair Encoding

BPE is a tokenization algorithm that follows the philosophy of representing frequent words with single tokens while splitting uncommon words in multiple tokens. It is derived from the compression algorithm [Gage, 1994] that iteratively replaces the most frequent pair of bytes with a new, unseen, byte. In the BPE tokenizer, the vocabulary  $\mathbb{V}$  is built by iteratively merging the token pair that appears more frequently in the training data and adding it to  $\mathbb{V}$ . Algorithm 1 provides more details about the vocabulary creation procedure. This approach is bottom-up because it starts from a small vocabulary  $\mathbb{V}$  containing only single characters and builds up new tokens through the concatenation of frequent pairs. BPE’s encoding and decoding are

---

**Algorithm 1** Example of vocabulary creation with BPE tokenizer.

---

**Require:** Training documents  $\mathbb{D}$ , vocabulary size  $k$

```

1: function BPE( $\mathbb{D}$ ,  $k$ )
2:    $\mathbb{V} \leftarrow \text{characters}(\mathbb{D})$  ▷ Initialize  $\mathbb{V}$  with all the single characters in the training data
3:   while  $|\mathbb{V}| < k$  do
4:      $\{(t_i, t_{i+1}), c_i\}_{i \in \mathbb{I}} \leftarrow \text{get\_bigrams\_count}(\mathbb{D}, \mathbb{V})$  ▷ Get bigrams counts splitting over  $\mathbb{V}$ 
5:      $j \leftarrow \arg \max_{i \in \mathbb{I}} c_i$  ▷ Select most frequent token pair
6:      $t \leftarrow t_j \cdot t_{j+1}$  ▷ “.” is the concatenation operator
7:      $\mathbb{V} \leftarrow \mathbb{V} \cup \{t\}$ 
8:      $\mathbb{D} \leftarrow \text{replace\_bigram\_with\_unigram}(\mathbb{D}, [t_i, t_{i+1}], t)$ 
9:   end while
10:  return  $\mathbb{V}$ 
11: end function

```

---

performed with a greedy algorithm. The encoding procedure searches iteratively for the longest token  $t \in \mathbb{V}$  that matches the start of the given input sequence. Usually, before the vocabulary creation input sentences are split over whitespaces and punctuation to ensure no token crosses word boundaries. Moreover, special characters are prepended or appended to every input word to indicate the start or the end. Regarding the decoding, it is a simple concatenation of every token. The division in words by inserting whitespaces is ensured to be correct by taking advantage of the additional special characters inserted before the tokenization. The following is an example of how the BPE tokenizer of RoBERTa [Liu et al., 2019b], which is a very prominent language model, splits text into tokens:

“Harry Potter and the Philosopher’s Stone”

↓

[“Harry”, “ĠPotter”, “Ġand”, “Ġthe”, “ĠPhilos”, “opher”, “s”, “ĠStone”]

Notice how the word “Philosopher’s” will be decoded correctly thanks to the special character “Ġ” use by RoBERTa to indicate the start of a new word in the original text.

Among the advantages of BPE there is the ability to tokenize words that were not seen at training time by only using the available tokens. In the worst case, the new word will be split in single characters, which were part of the initial vocabulary. BPE is used by many Transformer models such as GPT-2 [Radford et al., 2019], GPT-3 [Brown et al., 2020] and RoBERTa [Liu et al., 2019b].

### A.3.2 WordPiece

WordPiece is a tokenization algorithm derived from BPE in which the selection of the token pair to merge is performed with a more sophisticated technique. In particular, WordPiece does not select the symbol pairs that are most commonly used, but rather the ones that will increase the likelihood of the training data when added to the vocabulary.

This means that it calculates the likelihood of each symbol pair by dividing the joint probability of the pair by the probability of the individual symbols. The symbol pair with the highest ratio is then selected. This approach is different from BPE, as WordPiece takes into account the potential loss of information when merging two symbols to determine if it is a worthwhile decision.

The algorithm to create the vocabulary  $\mathbb{V}$  is the same as BPE (Algorithm 1), after changing line 5 into:

$$j \leftarrow \arg \max_{i \in \mathbb{I}} \frac{P(t_i, t_{i+1})}{P(t_i)P(t_{i+1})} \quad (\text{A.1})$$

One of the benefits of using WordPiece is that it has been found to be less prone to over-segmenting words, and it has been shown to be effective when used in neural Machine Translation models. WordPiece is used in well known Transformer models such as BERT [Devlin et al., 2019], DistilBERT [Sanh et al., 2020] and ELECTRA [Clark et al., 2020].

### A.3.3 UnigramLM

UnigramLM is a top-down tokenization approach which iteratively shrinks the vocabulary size by removing tokens. UnigramLM usually initializes the vocabulary with all the words and their substrings in training data. Then, given some loss function such as the log-likelihood, at each iteration the algorithm trains a simple unigram language model over the input data  $\mathbb{D}$  and selects a subset of tokens that, if removed, would increase the loss over the data the lowest. In other words, at each iteration the subset of tokens that influence the loss the less are removed. Common implementations remove the about the 10% of the tokens at each iteration until the vocabulary reaches the desired size. It is also a good practice to include in the vocabulary all the single characters of a language to ensure every word can be tokenized and without recurring to the OOV token. Algorithm 2 provides an example of UnigramLM vocabulary creation. Since UnigramLM is based on vocabulary reduction and not on merging rules like BPE and WordPiece, there may be several ways of tokenizing the same string. This may hurt language models performance because different tokenization of the same string may be provided to the model. For this reason, an UnigramLM tokenizer saves also the unigram language model over the final vocabulary, see line 17 of Algorithm 2. The final language model  $M$  is then used to compute the likelihood of every possible tokenization. In practice the most likely tokenization is chosen to have a deterministic split of text in subwords.

UnigramLM alone is not used by any relevant language model. However, its vocabulary creation technique is often implemented along with Sentence-Piece tokenization.



---

**Algorithm 2** Example of vocabulary creation with UnigramLM tokenizer.

---

**Require:** Training documents  $\mathbb{D}$ , vocabulary size  $k$ , fraction  $\alpha$  of tokens to remove at each iteration

```

1: function UNIGRAMLM( $\mathbb{D}$ ,  $k$ )
2:    $\mathbb{V} \leftarrow \text{words\_and\_subwords}(\mathbb{D})$  ▷ Initialize  $\mathbb{V}$  with words and subwords in the training data
3:   while  $|\mathbb{V}| > k$  do
4:      $M \leftarrow \text{train\_unigram\_language\_model}(\mathbb{D}, \mathbb{V})$  ▷ Train Unigram LM over  $\mathbb{D}$ 
5:      $\mathbb{L} = \{\}$  ▷ Will contain loss decrease for each token
6:     for  $t \in \mathbb{V}$  do
7:        $M' \leftarrow \text{train\_unigram\_language\_model}(\mathbb{D}, \mathbb{V} \setminus \{t\})$ 
8:        $\mathbb{L}_t \leftarrow P_M(\mathbb{D}) - P_{M'}(\mathbb{D})$  ▷ Loss difference by removing token  $t$  from  $\mathbb{V}$ 
9:     end for
10:    for  $(|\mathbb{V}| \cdot \alpha)$  times do ▷ Remove the  $\alpha$  fraction of less influent tokens
11:       $t = \arg \min_i \mathbb{L}_i$ 
12:       $\mathbb{V} \leftarrow \mathbb{V} \setminus \{t\}$  ▷ Trim vocabulary
13:       $\mathbb{L} \leftarrow \mathbb{L} \cup \{\mathbb{L}_t\}$  ▷ Avoid selecting again  $\mathbb{L}_t$  from  $\mathbb{L}$ 
14:    end for
15:  end while
16:   $M \leftarrow \text{train\_unigram\_language\_model}(\mathbb{D}, \mathbb{V})$  ▷ Train final Unigram LM over  $\mathbb{D}$ 
17:  return  $\mathbb{V}, M$ 
18: end function

```

---

### A.3.4 Sentence-Piece

Sentence-Piece is not a new tokenization technique by itself, but rather a variant of the previously discussed algorithms in which the input text is not split over whitespaces to separate words. By treating the whitespaces between words as all the other characters, Sentence-Piece can be trained on streams of raw text in several languages without the need of ad-hoc adaptations. For example, it can successfully tokenize Chinese or Japanese text, which does not contain whitespaces, as well as English or German.

The vocabulary creation of a Sentence-Piece tokenizer can be based on either Byte-Pair Encoding or UnigramLM. Thus, it can tokenize almost any text without producing OOV tokens. Sentence-Piece is used in many start-of-the-art language models such as ALBERT [Lan et al., 2020], XLNet [Yang et al., 2019], DeBERTaV2 [He et al., 2020], DeBERTaV3 [He et al., 2021] and Google T5 [Raffel et al., 2020].

### A.3.5 Loss functions, Back-Propagation and Optimization

During the pre-training phase of language models, the primary objective is typically framed as a classification task. With token-level objectives, the model is trained to predict the original input by reconstructing it from various forms of corruption applied during training. To achieve this, a classification layer, such as a linear projection, is employed to map the output embeddings of the model to different categorical labels, enabling it to effectively classify and reconstruct the original input. In sentence-level tasks and fine-tuning, the output embeddings of the model are exploited to predict a categorical label that represents a feature of the whole input, such as the opinion of a user in Sentiment Analysis, or the relation between question and answer in Question Answering.

**Loss function** The loss function is a metric that measures the accuracy of the predictions of a model compared to the gold labels. Transformer models commonly use the Cross-Entropy Loss function for both pre-training and fine-tuning stages. The Cross-Entropy Loss measures the average number of bits required to identify an event drawn from the distribution of the

true labels  $\mathcal{L}$  when using a scheme optimized for the predicted probabilities distribution  $\mathcal{P}$ . It provides a measure of the dissimilarity between these two distributions, indicating how well the predicted probabilities align with the gold labels. In the context of classification, it calculates the loss by comparing the predicted class probabilities with the categorical labels. Given a label  $y \in \mathbb{C}$ , where the latter is the set of possible classes, and a vector of predictions over the classes  $[p_1, \dots, p_{|\mathbb{C}|}]$ , the Cross-Entropy loss can be computed as follows:

$$L_{CE}(\mathbf{p}, y) = - \sum_{i=1}^{|\mathbb{C}|} \mathbb{1}_{\{y=i\}} \log \frac{e^{p_i}}{\sum_{i=0}^{|\mathbb{C}|} e^{p_i}} \quad (\text{A.2})$$

Moreover, the Cross-Entropy definition can be simplified when the task is binary classification and the predictions are already softmaxed:

$$L_{CE}(p, y) = -(y \log p + (1 - y) \log(1 - p)) \quad (\text{A.3})$$

Once the loss over the predictions and the labels is computed, the back-propagation algorithm is used to compute the gradients for the parameters of the model.

**Back-propagation** Given the error in the predictions, the back-propagation algorithm computes the gradients of all the parameters, enabling the model to update its weights and improve its performance through gradient descent.

Mathematically, let's consider a model  $\mathcal{M}$  that is composed of a sequence of parametric functions  $f_1, \dots, f_n$  which are applied in order to an input  $\mathbf{x}$ :

$$\mathbf{p} = f_n \circ f_{n-1} \circ \dots \circ f_1(\mathbf{x}) \quad (\text{A.4})$$

We call  $\mathbf{h}_i$  the output of the layer  $f_i$ . Given some loss function, such as the Cross-Entropy defined before, and a target label  $y$ , we can compute the error  $L$  with some loss function, such as Cross-Entropy:

$$L = L_{CE}(\mathbf{p}, y) \quad (\text{A.5})$$

Then, by applying the Chain Rule [Goodfellow et al., 2016], we can compute the partial derivatives for each layer  $f_i$  in the model  $\mathcal{M}$ . This rule allows us to propagate gradients through multiple layers of a neural network, by unrolling the gradients computed over the whole model to the gradients of each layer. This is performed as follows:

$$\nabla \mathcal{M} = \frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L}{\partial f_n} \cdot \frac{\partial f_n}{\partial f_{n-1}} \cdot \dots \cdot \frac{\partial f_2}{\partial f_1} \cdot \frac{\partial f_1}{\partial \mathbf{x}} \quad (\text{A.6})$$

Then, for each layer  $f_i$ , we can exploit the output of the previous layer  $\mathbf{h}_{i-1}$ , the output of the current layer  $\mathbf{h}_i$  and the derivatives computed for the next layer  $\frac{\partial f_{i+1}}{\partial f_i}$  to compute the gradients of the current layer parameters.

**Optimization** By efficiently applying the Chain Rule, we can propagate the gradients backward through the network, enabling the calculation of parameter updates and optimization through gradient descent. This process allows the network to learn and improve its performance by adjusting the weights based on the computed gradients. Suppose you have a layer  $i$  which is a parametric function  $f_i$  with weights  $\mathbf{W}_i$  and gradients  $\nabla \mathbf{W}_i$ , previously computed with back-propagation. The common approach to update  $\mathbf{W}_i$  is gradient descent:

$$\mathbf{W}'_i = \mathbf{W}_i - \eta \nabla \mathbf{W}_i \quad (\text{A.7})$$

where  $\eta$  is called the learning rate and determines the speed at which the model converges. The original Gradient Descent algorithm is no longer used due to its sensitivity and instability while searching for the local minima. Common optimizer algorithms such as Adam [Kingma and Ba, 2017], RMSProp [Tieleman et al., 2012] and Adagrad [Lydia and Francis, 2019] improve over Gradient Descent by storing additional data for each weight in the model. They allow to maintain a more general direction toward the local minima while optimizing, thus increasing the training stability and the model generalization capabilities.

## A.4 Machine Reading Comprehension Branches

### A.4.1 Extractive Question Answering

Extractive Question Answering involves finding specific answers to questions within a given text or set of texts. This is different from Generative (or Abstractive) QA, which involves generating new text as an answer to a question, rather than extracting an answer from existing text.

In Extractive QA, the system is given a question and a set of documents, and is expected to identify the specific piece of text within the documents that provides the answer to the question. This typically involves identifying the relevant parts of the text, such as named entities or specific phrases, and extracting them as the answer.

More formally, in Extractive QA a model  $\mathbf{C}$  is given a question  $q$  and a set of documents  $\mathbb{D}$ , which may contain answer spans  $a_i$  in some  $D_i \in \mathbb{D}$ . The goal of the model is to extract the answer spans  $a_i$  from  $\mathbb{D}$  given the question  $q$ . Typically, the performance are measured by checking the degree of overlap between the gold answer spans and those predicted by the model. The prediction of the answer spans by the model  $\mathbf{C}$  can be performed with different techniques. For example, the model may output the start and end positions of the correct answer or a binary classifier over each token of the documents in  $\mathbb{D}$  could be used to predict the belonging to the answer.

### A.4.2 Generative Question Answering

Generative Question Answering regards the generation of new text as an answer to a question, rather than extracting an answer from existing documents. This involves using Machine Learning techniques to learn how to generate coherent and appropriate responses to questions, based on a large dataset of examples. The goal is to generate correct answers that should be indistinguishable from human-generated text. The input set of retrieved documents could be used as a source of knowledge to help in the answer generation. However, some recent models have such a large number of internal parameters that are able to memorize very large quantities of information already while pre-training, and thus can answer most questions without external knowledge [Brown et al., 2020; Raffel et al., 2020].

In general, measuring the performance of generative models is a hard task, and evaluating generated answers is no exception. The main problem is that the model may output text with the same meaning of the gold answer but with a different formulation and using different terms. Thus, the best evaluation of generative model's outputs is manual annotation by well-trained crowd-workers. However, this method it is very expensive and the quality of the annotation is subject to variations based on annotator's salary, educational background and geographical location.

As an alternative, automatic evaluation techniques can be applied to evaluate generated answers. The most common metrics for automatic evaluation are BLEU [Papineni et al., 2002]

and ROUGE [Lin, 2004], which were developed for Machine Translation and Summarization respectively. BLEU works by counting the ratio of  $n$ -grams that overlap between the generated answer and the target answer(s). On the other hand, ROUGE is an evaluation package containing different metrics. ROUGE-N for example, is similar to BLEU and works by counting common  $n$ -grams, however it is recall oriented and not precision-based as BLEU. ROUGE-L searches instead for the Longest Common Subsequences (LCS) between the generated answer and the gold references. Finally, we mention ROUGE-S, which uses skip-bigram co-occurrence statistics to measure the similarity between the generated string and the set of target references. Since generated answers may contain the same meaning of the target reference but expressed in a different form, the larger the number of gold references for each question, the better the evaluation.

### A.4.3 Abstractive Question Answering

Abstractive Question Answering, is a type of generative QA that involves generating answers by summarizing or paraphrasing the information in a given document. This requires the system to understand and reason about the text, and to identify the most important or relevant information to include in the answer. Abstractive QA is strictly related to applications such as Summarization, where the goal is to generate a concise and coherent summary of a document or set of documents.

Since Abstractive QA is a generative task such as Generative QA, the metrics used for the evaluation are mostly the same: human evaluation when possible or automatic evaluation with metrics like BLEU, ROUGE and BERT-Score otherwise.

# Appendix B

## Datasets

### B.1 GLUE

The GLUE (General Language Understanding Evaluation) benchmark is a collection of diverse datasets designed to evaluate the performance of Natural Language Understanding (NLU) models. The benchmark consists of nine individual datasets, each targeting a different aspect of language understanding. We omit the WNLI dataset in our evaluation as similar works [Clark et al., 2020; Devlin et al., 2019] because even a trivial majority classifier performs better than most of the models considered in this work. Here are detailed descriptions of each dataset:

- **CoLA (Corpus of Linguistic Acceptability)**: This dataset focuses on grammatical correctness. It contains sentences from a variety of sources, including books, articles, and websites. The task is to determine whether a given sentence is grammatically correct or not.
- **SST-2 (Stanford Sentiment Treebank)**: The SST-2 dataset is derived from movie reviews, where each sentence is labeled with its sentiment (positive or negative). The task is to classify the sentiment of each sentence accurately.
- **MRPC (Microsoft Research Paraphrase Corpus)**: This dataset comprises sentence pairs extracted from various sources, such as online forums and news. The goal is to determine whether the two sentences in each pair are paraphrases or not.
- **QQP (Quora Question Pairs)**: The QQP dataset is created from pairs of questions asked on Quora, a question-and-answer platform. The task is to predict whether two questions are semantically equivalent or not.
- **STS-B (Semantic Textual Similarity Benchmark)**: STS-B contains sentence pairs along with their similarity scores. The dataset covers diverse domains such as news, captions, and forum threads. The objective is to predict the degree of semantic similarity between the two sentences.
- **MNLI (Multi-Genre Natural Language Inference)**: MNLI consists of sentence pairs from various genres, such as fiction, government reports, and telephone conversations. The goal is to determine the relationship between the two sentences: entailment, contradiction, or neutral.
- **QNLI (Question-answering Natural Language Inference)**: QNLI is derived from the Stanford Question Answering Dataset (SQuAD). It involves sentence pairs where one sentence is a question and the other is a snippet from a Wikipedia article. The task is

to determine whether the given question can be answered correctly using the provided snippet.

- **RTE (Recognizing Textual Entailment)**: The RTE dataset consists of sentence pairs collected from news articles. The objective is to determine if the premise sentence entails, contradicts, or remains neutral compared to the hypothesis sentence.
- **WNLI (Winograd Natural Language Inference)**: WNLI focuses on resolving pronoun references using contextual understanding. The dataset contains sentence pairs with ambiguous pronouns, and the task is to determine if the pronoun in the second sentence refers to the same entity as in the first sentence.

These datasets cover a wide range of Natural Language Understanding tasks, including syntactic and semantic analysis, sentiment classification, paraphrase identification, textual entailment, and pronoun resolution. They serve as a standardized benchmark to evaluate the performance of different language models and techniques in a variety of linguistic tasks.

## Appendix C

# Alternative Efficient Objectives

### C.1 RTS vs C-RTS

In this Section, we provide more details about the comparison between RTS and C-RTS on a longer pre-training, which should better highlight the differences between the two algorithms.

In Figure C.1 we show the results of the comparison between RTS and C-RTS over the 4 datasets. We do not show the plot for ASNQ because fine-tuning the models on this dataset every 10K steps is very expensive. We repeated every fine-tuning on the downstream tasks 5 times with different initialization seeds. On 3 benchmarks out of 4, C-RTS is superior thanks to the harder training task. Choosing more challenging replacements instead of random ones forces the model to find subtler relations between tokens. On QNLI, both objectives perform equally from a statistical viewpoint. We argue that Small models are more influenceable than Base counterparts because they have less parameters. Thus, larger datasets with thousands of examples such as QNLI reduce the differences seen while pre-training.

### C.2 Negative Results

In scientific research, negative results hold valuable lessons for future investigations. This Section presents experiments and observations that diverged from initial hypotheses, offering important insights to guide future researchers.

#### C.2.1 C-RTS sampling always from the same Cluster

We conduct an experiment where we simplify C-RTS by replacing tokens only with tokens within the same cluster. However, we discover that it is essential to allow sampling from other clusters as well because the model demonstrates the ability to learn how tokens are clustered after a significant number of training steps.

#### C.2.2 Position-based techniques

In this experiment, we explore altering the positions of tokens by masking certain positional encodings, similar to the MLM approach. The objective is to determine the original position of masked tokens in the original text. Although the classification head for this technique is smaller than MLM (slightly larger than RTS due to BERT’s 512 possible positions), the results on the GLUE benchmark are about 3.0 points worse.

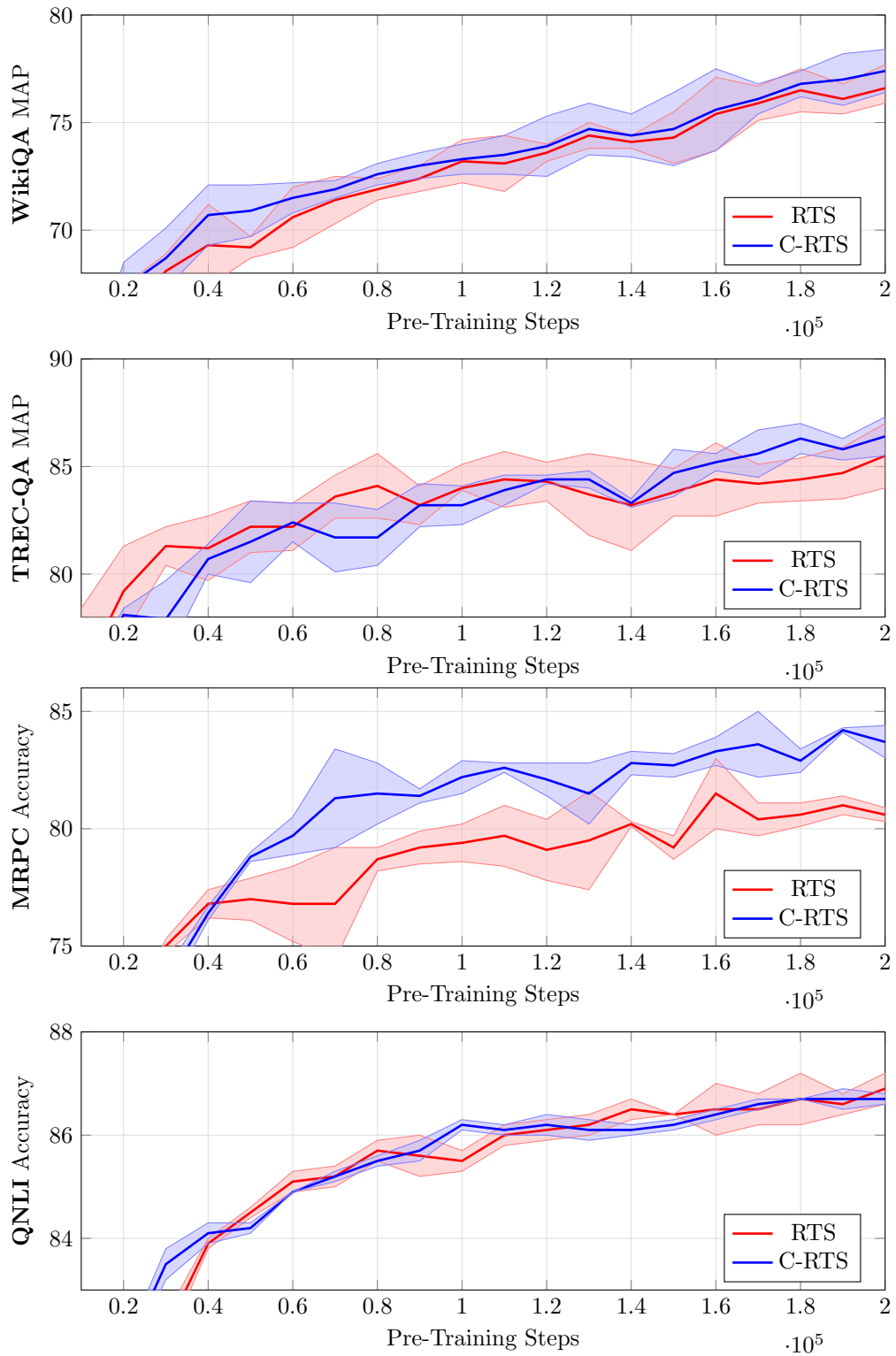


Figure C.1: Performance comparison of RTS and C-RTS on WikiQA, TREC-QA, MRPC and QNLI. We show both average and standard deviation after different fine-tunings with 5 different random seeds.



We also define another objective where we (i) shuffle the positions of some input tokens and (ii) predict their original positions. We randomly select 15% of the tokens and permute them. Although this approach performs 1.5 points below MLM on the GLUE average, it is almost as fast as RTS. Combining position-based objectives with token-level objectives could be a potential future research direction.

### C.2.3 LM head-on ELECTRA’s discriminator

Based on the successful performance of SLM, we have implemented and tested a version of SLM for ELECTRA. However, applying SLM directly to the ELECTRA discriminator is not feasible because predictions are only made for altered tokens, which undermines the task of detecting fake tokens. To address this, we propose an alternative objective called SLM-All. In SLM-All, instead of predicting only on tampered tokens, the model is tasked with predicting the original values of all tokens while also reproducing the unchanged inputs in the output.

We evaluated this approach on GLUE, WikiQA, ASNQ, and TREC-QA. Although it required more time to be pre-trained (1.42 times the time required by a similarly sized MLM-based model), the results were generally inferior compared to other approaches.

The significant efficiency gap between SLM-All and other models is due to the fact that the latter predicts MLM-like tokens for every output embedding of each token, not just the 15% as in SLM or MLM. Even with a 25% reduction in the number of training steps for the ELECTRA model (to balance the presence of the additional generator with size 1/3), it still utilizes slightly more FLOPS than BERT-MLM. Specifically, it achieved a GLUE average score of 80.02 on the test leaderboard and MAP scores of 78.7, 86.7, 64.9 in WikiQA, TREC-QA, and ASNQ, respectively.



# Appendix D

## Self-Supervised Objectives for AS2

### D.1 Negative Results

In this Section, we briefly describe the experiments that didn’t lead to accuracy improvements of our language models. We believe negative results are important to help other authors at selecting future research directions.

#### D.1.1 Additional positives from Large Language Model

Model	ASNQ		ASNQ + 1 Positive		WikiQA		WikiQA + 1 Positive	
	MAP	P@1	MAP	P@1	MAP	P@1	MAP	P@1
RoBERTa <sub>Base</sub> [Liu et al., 2019b]	66.9 (0.1)	61.8 (0.1)	67.1 (0.2)	62.4 (0.2)	85.8 (1.3)	78.3 (2.8)	85.7 (1.2)	77.9 (1.9)
<b>RoBERTa<sub>Base</sub> + SSP</b>	68.1 (0.2)	64.1 (0.3)	68.6 (0.2)	64.3 (0.1)	88.7 (0.3)	82.9 (0.7)	87.6 (1.1)	80.9 (2.1)
<b>RoBERTa<sub>Base</sub> + SP</b>	68.3 (0.1)	64.1 (0.2)	68.2 (0.3)	63.6 (0.3)	87.7 (0.3)	81.0 (0.8)	86.9 (0.2)	79.2 (0.9)
<b>RoBERTa<sub>Base</sub> + PSD</b>	67.7 (0.2)	62.6 (0.4)	67.1 (0.2)	61.9 (0.3)	86.4 (1.1)	80.5 (1.6)	86.8 (1.2)	80.2 (2.3)
<b>RoBERTa<sub>Base</sub> + ALL</b>	68.0 (0.1)	63.9 (0.4)	67.7 (0.1)	63.8 (0.2)	88.2 (0.4)	82.5 (0.9)	87.5 (0.9)	81.1 (1.9)
ELECTRA <sub>Base</sub> [Clark et al., 2020]	67.5 (0.2)	62.4 (0.4)	68.1 (0.3)	63.8 (0.4)	85.0 (2.6)	77.1 (4.0)	82.5 (1.5)	73.1 (1.7)
<b>ELECTRA<sub>Base</sub> + SSP</b>	69.7 (0.2)	65.3 (0.3)	67.6 (0.2)	63.2 (0.2)	88.6 (1.4)	82.5 (2.0)	87.1 (1.3)	80.0 (2.6)
<b>ELECTRA<sub>Base</sub> + SP</b>	69.0 (0.1)	65.0 (0.2)	68.7 (0.3)	64.0 (0.4)	88.1 (1.5)	81.8 (2.3)	85.9 (0.6)	77.8 (1.1)
<b>ELECTRA<sub>Base</sub> + PSD</b>	68.9 (0.3)	65.3 (0.4)	68.0 (0.2)	62.3 (0.2)	85.6 (0.7)	78.6 (0.7)	85.1 (1.2)	77.5 (1.3)
<b>ELECTRA<sub>Base</sub> + ALL</b>	69.3 (0.2)	65.0 (0.2)	68.8 (0.3)	64.3 (0.5)	87.3 (1.2)	80.8 (1.9)	86.2 (1.4)	78.5 (2.3)

Table D.1: Results in the **data augmentation** setting on ASNQ and WikiQA.

Most of the datasets for AS2 have a limited number of positive question-answer pairs. For example, the training set of ASNQ has about 1.1 positive answers every 356 candidates. The other datasets have instead a slightly higher ratio between positive answer and the total number of candidates sentences for each query:  $\frac{0.5}{9.6}$  in WikiQA,  $\frac{5.4}{46}$  in TREC-QA,  $\frac{0.23}{15}$  in WQA and about  $\frac{1.7}{28}$  in NewsAS2 and TriviaAS2. However, the gap between positives and negatives examples is still very large.

For those reasons, we try to increment the number of positive answer candidate by exploiting Falcon-40B-instruct [Almazrouei et al., 2023], a Large Language Model with 40 billion parameters fine-tuned for instruction-based conversations. Specifically, we generate a single additional positive answer for each query by providing Falcon with the question and possibly by adding a question mark at the end if it was missing.

Notice that we do not add positive answer candidates to the development and test sets to perform a fair comparison with the results provided in the previous Sections. We test the vanilla RoBERTa<sub>Base</sub> and ELECTRA<sub>Base</sub> checkpoints as well as the models further pre-trained with

our specialized objectives on ASNQ and WikiQA. The latter should provide a meaningful test suite being a very large dataset with very few positives and a small dataset with a reasonable number of positive answers, respectively.

We report the results of the experiment in Table D.1. On ASNQ, there are rare combinations of models and objectives that benefits from the additional positives. For example, the two baselines  $\text{RoBERTa}_{\text{Base}}$  and  $\text{ELECTRA}_{\text{Base}}$  show an improvements of the P@1 by 0.6 points and 1.2 points respectively. Our continuously pre-trained models generally perform lower by 1.0 P@1 points when fine-tuned with the additional positives, apart from  $\text{RoBERTa}_{\text{Base}} + \text{SSP}$ , which gains 0.2 points in P@1.

On WikiQA, no model benefits from the additional positives. In particular, there is a decrease between 0.3 and 4.0 points in P@1, depending on the combination between models architectures and specialized our pre-training objectives.