Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

TrentoTeam at SemEval-2017 Task 3: An application of Grice Maxims principles In Ranking Community Question Answers

Mohammed R. H. Qwaider, Abed Alhakim Freihat, Fausto Giunchiglia

August 2017

Technical Report # DISI-17-009

Proceedings of the 11th International Workshop on Semantic Evaluation, 2-3.08.2017 - Vancouver (Canada)

TrentoTeam at SemEval-2017 Task 3: An application of Grice Maxims principles In Ranking Community Question Answers

Mohammed R. H. Qwaider
Fondazione Bruno Kessler
Via Sommarive, 18
Trento, Italy.
qwaider@fbk.eu

Abed Alhakim Freihat
University of Trento
Via Sommarive, 19
Trento, Italy.
abed.freihat@unitn.it

Fausto Giunchiglia
University of Trento
Via Sommarive, 19
Trento, Italy.
fausto@unitn.it

Abstract

In this paper, we present a community answers ranking system which is based on Grice Maxims. In particular, we describe a ranking system which is based on answer relevancy scores, assigned by three main components: Named entity recognition, similarity score, and sentiment analysis.

1 Introduction

This paper describes the Grice Ranking system which participated to SemEval 2017 Task 3 (Nakov et al., 2017) subtask A competition[1]. The SemEval 2017 Task 3 (Nakov et al., 2017) sub-task A focuses on raking a list of answers (10 in our case) as follows. Given a question Q and a list of answers $\langle a_1, ..., a_n \rangle$. Rank these answers according to their relevancy with respect to the question Q. Our participation to the task was mainly motivated by our interest in applying Grice maxims (Grice, 1975) principles to a ranking task to define standards of Grice maxims for ranking tasks. The system follows 3 steps: similarity, entity recognition, and sentiment analysis.

2 Grice maxims principles

Grice main idea is that communication between human beings is logic and rational. Following this idea, any conversation assumes cooperation between the conversation parties. This cooperation supposes in essence four maxims that usually hold in dialogues or conversations. These maxims are:

1. **Quality**: Say only true things.
2. **Quantity**: Be informative.
3. **Relation**: Be relevant in your conversation.
4. **Manner**: Be direct and straightforward.

These maxims have been intensively researched in the domain of linguistics and pragmatics in the last decades, where the researchers focused on how to use Grice theory to explain speaker intention when he says some thing. For example, these maxims explain that the speaker B understands the intention of the speaker A. The same holds for A who understands the indirect Answer of B.

A *What is the time?*

B *The bus left five minutes ago.*

In this work, we use these maxims partially to measure the appropriateness or relevancy of answer(s) of a given question. In this approach, we do not try to understand what the speaker (intentional) means. Instead, we try to understand if the speaker contribution contains (extensional) elements that comply with Grice maxims.

In the following, we explain how we interpret the quantity, relation and manner maxims in our approach. We do not use the quality maxim and it is beyond the scope of our research.

2.1 Quantity

We interpret this maxim as how much an answer is informative by examining whether an answer contains the following informative elements?

1. **Named entities**: A named entity here refers to person, organization, location, or product.
2. **References**: References include web urls, emails, and phone numbers.
3. **Currency**: We consider the presence of currency in an answer as informative element.
4. **Numbers**: In some cases, phone numbers, or currency are not recognized because they are

[1] http://alt.qcri.org/semeval2016/task3/

implicit such as *20000 is a good salary*. For this reason, we consider the presence of numbers (2 digits or more) in answers as an informative element.

Of course, this list is not exhaustive. However, these are the informative elements that we utilize in our approach.

2.2 Relation

We think that defining what is a relevant contribution in the relation maxim is still an open issue (Frederking, 1996). At the same time, we try to discover relevancy indicators and use them in our algorithm. Accordingly, we consider the following as relevancy indicators.

1. **Similarity**: Similarity between the question and the answer.

2. **Imperatives**: Answers that contain imperative verbs such as *try*, *go to*, or *check* indicate that the answerer is explaining a way to solve the problem being discussed.

3. **Expression of politeness**: Expressions of politeness *I would*, *I suggest*, or *I recommend* are usually polite alternatives for imperatives.

4. **Factoid answer particles**: For factoid questions *is/are* or *does/do*, the answer particles *yes/no* indicate the relevancy of the answer.

5. **Domain specific terms**: Domain specific terms indicate relevancy. For example, terms such as *CV*, *NOC*, *torrent*,... are domain specific terms . Using such terms indicates also that the answerer is trying to explain how to solve the problem being discussed.

Again, this list is not exhaustive and it would be much better for our approach if we could use more concrete criteria that indicate the relation maxim.

2.3 Manner

Grice summarizes this maxim as (a speaker contribution is expected to be clear) and he gives four criteria that indicate not violating this maxim: *Avoid obscurity of expressions, avoid ambiguity, be brief*, and *be orderly*. We did not use these criteria for the difficulty of applying them. Instead, we give the following criteria that can be used to judge that a speaker contribution complies with or violates the manner maxim.

1. **Be positive**: By this criterion, we mean that the speaker contribution is expected to be tolerant and permissive.

2. **Avoid frustrating utterances**: Answers that contain such expressions are usually not useful in the conversation.

3. **Avoid ironic and humbling expressions**: We mean here that the answer tends to be formal and professional and that the answerer is aiming to give a direct useful contribution.

4. **Avoid insulting and degrading expressions**: Answers that contain such expressions are not expected to be be useful.

We may also consider the grammatical and orthographic correctness as a criterion. We did not consider it because many of the members of Qatar Living are not native speakers of English.

3 Implementation

In the following, we present the ranking algorithm , where we start with explaining the used resources. Then, we illustrate some experiments that we have conducted in the framework of our approach, and finally we describe the Grice Ranking system.

3.1 Resources

We used the following resources in our algorithm.
Quality: No resources and this maxim was not used in the implementation.
Quantity: We adopted pre-trained openNLP[2] name finder model for named entity recognition (NER) to our domain data. we needed this model because of the low performance of the state of the art NER systems on the training data. We have trained the openNLP NER system on an self annotated subset of the training data set. The generated model reached 87% F1-measure. Both the annotated corpus and the model are downloadable online[3].
Relation: We have used the following resources:

a) *Similarity*: We used Word2Vec[4](Turian et al., 2010) and Brown and clark (Agerri and Rigau, 2016) embeddings.

[2]https://opennlp.apache.org/
[3]https://www.researchgate.net/project/Named-Entity-Recognizer-For-Qatar
[4]https://github.com/ragerri/cluster-preprocessing/

b) *Imperatives and Expression of politeness*: We used an OpenNLP POS-tagger to detect these expressions. We reward answers that contains such expressions.

c) *Domain specific terms*: Using the training data, we have compiled a small dictionary that contains domain specif terms such as *router, CV, NOC, torrent,....* The terms in the dictionary are not classified and of course they are not exhaustive. Answers that contain such expressions are also rewarded.

Manner: We used two sentiment polarity lists[5], one positive sentiment list and other negative sentiment list.

a) *Be positive*: We used the positive sentiment list to reward answers that contain positive expressions.

b) *Avoid frustrating expressions*: We used the negative sentiment list to penalize answers that contain such expressions.

c) *Avoid ironic and humbling expressions*: The negative sentiment list includes some of the ironic and humbling expressions. We used the training data to extend the list with new expressions that we found in the training data. Answers that contain such expressions are penalized.

d) *Avoid insulting and degrading expressions*: The negative sentiment list includes some of these expressions. We have extended the list with new expressions that we found in the training data. We penalize answers that contain such expressions.

3.2 Experiments

In the following, we describe some of the experiments that we conducted to evaluate the proposed algorithm which is described in next section. We evaluate the systems using the test set taken from SemEval 2016 (Nakov et al., 2016), where we used MAP (mean average precision) as performance measure.

Experiment 1 (similarity run): Rank the answers of a question using TF-IDF as a similarity function from the most similar answer to less relevant one.

[5]https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107

Experiment 2 (clusters/word representation 1): We experimented mixing different combinations of word embeddings and similarity measure to rank the answers. We used Brown embedding with N-grams level, with a weight of 0.5 to embedding similarity and 0.5 to string similarity.

Experiment 3 (clusters/word representation 2): Using Brown and Clark with weight of 0.3 to string similarity and 0.7 to cluster similarity.

Experiment 4 (clusters/word representation 3): Including word2vec to Brown and clark, with a low-level features, like word shape with the same weight of 0.3 to string similarity and 0.7 to cluster similarity.

Experiment 5 (similarity rule based): In this experiment, we run the system in two phases:

1. Rank the comments depending on their token-based similarity score.

2. Re-rank them based on background rules such as downgrading the answers of the same person which we considered as duplicates.

3.3 Grice Maxims Based Ranking Algorithm

In the following, we present the ranking algorithm.[6].

Input: Q: $\langle p, qText\rangle$, L: $\langle a_1, ..., a_n\rangle$
p: The person who is asking
$qText$: The question text.
$a_i = \langle p_i, aText_i, score_i\rangle$.
p_i: The person who answered a_i
$qText_i$: The answer text of a_i
$score_i$: The relevancy of a_i.
Output: L: The input list after sorting.
algorithm `GriceMaximsRanking(`Q`,` L`)`
begin
  foreach `answer` a_i **in** L
    if $p_i = p$ **then** $score_i = i * -100$
    else
    $score_i = |SM_{qi}| + |NE_i| + |RE_i| + |CN_i| + |IM_i| + |DT_i| + |PS_i|$;
    $score_i -= |NS_i| + |IR_i| + |ID_i|$;
  `sort` L ;
  return L;
end

[6] SM: Similarity between question and answer. NE: Named entities. RE: Reference expressions. CN: Currency and numbers. IM: Imperative and polite expressions. DT: Domain specific terms. PS: Positive sentiment words. NS: Negative sentiment words. IR: Ironic and humbling words. ID: Insulting and degrading words. p: Refers to the person who is asking. $qText$: Refers to the question text

The algorithm works in four steps as follows.

1. The algorithm checks whether the answerer is the same person who asked the question. The answers made by person who asked the question are downgraded such that they become the last answers in the list.

2. For the rest of the answers, we compute the similarity between the question Q and the answer a_i, where $0 \leq SM_{qi} \leq n$ $(n = |L|)$.

3. Then, based on Grice maxims, the answers are rewarded or penalized as follows.

   a The answer a_i is rewarded according to the number of entities, reference expressions, currency and numbers, imperatives, domain specific terms, and positive sentiment words.

   b On the other hand, a_i is penalized according to the number of negative sentiment, ironic, and insulting words.

4. After rewarding and penalizing all answers, we then sort the list of answers according to their achieved scores in a descending order. Best answer is the first answer in the list and so on.

Evaluating the algorithm on the same test set in the previous experiments, we get MAP=0.7151. The best system in SemEval 2016 (Filice et al., 2016) achieved MAP=79.19 as shown in Table 1.

System	MAP
Baseline	0.5280
Experiment 3	0.5596
Experiment 1	0.5839
Experiment 2	0.6089
SemEval-2016 Worst System	0.6224
Experiment 5	0.6403
Experiment 4	0.6422
Our System	0.7151
SemEval-2016 Best System	0.7919

Table 1: Results of some community Question Answer Ranking approaches in SemEval 2016.

4 Results

Our system obtained a rank[7] of 12 out of 13 participated systems and a MAP of 78.56. It beat the IR baseline by 6 points, and the last system LaSIGE-primary by 15 points, with a difference of 10 points from the best system.

[7] http://alt.qcri.org/semeval2017/task3/

References

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence* 238:63 – 82.

Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1116–1123. http://www.aclweb.org/anthology/S16-1172.

R. Frederking. 1996. Grice's maxims: do the right thing. *Proc. of AAAI SpringSymp. on Compl. Implicature: Computational Approaches to Interpreting and Generating Conversational Implicature* .

H. P. Grice. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, Academic Press, San Diego, CA, pages 41–58.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Vancouver, Canada, SemEval '17.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. Semeval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, SemEval '16, pages 525–545. http://www.aclweb.org/anthology/S16-1083.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, pages 384–394. http://www.aclweb.org/anthology/P10-1040.