You and me... in a vector space: modelling individual speakers with distributional semantics

Aurélie Herbelot
Centre for Mind/Brain Sciences
University of Trento
aurelie.herbelot@unitn.it

Behrang QasemiZadeh
DFG Collaborative Research Centre 991
Heinrich-Heine-Universität Düsseldorf
zadeh@phil.hhu.de

Abstract

The linguistic experiences of a person are an important part of their individuality. In this paper, we show that people can be modelled as vectors in a semantic space, using their personal interaction with specific language data. We also demonstrate that these vectors can be taken as representative of 'the kind of person' they are. We build over 4000 speaker-dependent subcorpora using logs of Wikipedia edits, which are then used to build distributional vectors that represent individual speakers. We show that such 'person vectors' are informative to others, and they influence basic patterns of communication like the choice of one's interlocutor in conversation. Tested on an information-seeking scenario, where natural language questions must be answered by addressing the most relevant individuals in a community, our system outperforms a standard information retrieval algorithm by a considerable margin.

1 Introduction

Distributional Semantics (DS) (Turney and Pantel, 2010; Clark, 2012; Erk, 2012) is an approach to computational semantics which has historical roots in the philosophical work of Wittgenstein, and in particular in the claim that 'meaning is use', i.e. words acquire a semantics which is a function of the contexts in which they are used (Wittgenstein, 1953). The technique has been used in psycholinguistics to model various phenomena, from priming to similarity judgements (Lund and Burgess, 1996), and even aspects of language acquisition (Landauer and Dumais, 1997; Kwiatkowski et al., 2012). The general idea is that an individual speaker develops the verbal side of his or her conceptual apparatus from the linguistic experiences he or she is exposed to, together with the perceptual situations surrounding those experiences.

One natural consequence of the distributional claim is that meaning is both speaker-dependent and community-bound. On the one hand, depending on *who* they are, speakers will be exposed to different linguistic and perceptual experiences, and by extension develop separate vocabularies and conceptual representations. For instance, a chef and a fisherman may have different representations of the word *fish* (Wierzbicka, 1984). On the other hand, the vocabularies and conceptual representations of individual people should be close enough that they can successfully communicate: this is ensured by the fact that many linguistic utterances are shared amongst a community.

There is a counterpart to the claim that 'language is speaker-dependent': speakers are language-dependent. That is, the type of person someone is can be correlated with their linguistic experience. For instance, the fact that *fish* and *boil* are often seen in the linguistic environment of an individual may indicate that this individual has much to do with cooking (contrast with high co-occurrences of *fish* and *net*). In some contexts, linguistic data might even be the only source of information we have about a person: in an academic context, we often infer from the papers a person has written and cited which kind of expertise they might have.

This paper offers a model of individuals based on (a subset of) their linguistic experience. That is, we model how, by being associated with particular types of language data, people develop a uniqueness representable as a vector in a semantic space. Further, we evaluate those 'person vectors' along one particular dimension: the type of knowledge we expect them to hold.

The rest of this paper is structured as follows. We first give a short introduction to the topic of modelling linguistic individuality (§2) and we discuss how DS is a suitable tool to represent the associated characteristics for a given person (§3). We describe a model of individuals in a community using 'person vectors' (§4). We then highlight the challenges associated with evaluating such vectors, and propose a prediction task which has for goal to identify someone with a particular expertise, given a certain information need (§5, §6). Concretely, we model a community of over 4000 individuals from their linguistic interaction with Wikipedia (§7). We finally evaluate our model on the suggested task and compare results against a standard information retrieval algorithm.

2 Individuality and how it is seen

A speaker's linguistic experience—what they read, write, say and hear—is individual in all the ways language can be described, from syntax to pragmatics, including stylistics and register. One area of work where linguistic individuality has been extensively studied is author profiling and identification (Zheng et al., 2006; Stamatatos, 2009). It has been shown, in particular, how subtle syntactic and stylistic features (including metalinguistic features such as sentence length) can be a unique signature of a person. This research, often conducted from the point of view of forensic linguistics, has person identification as its main goal and does not delve much into semantics, for the simple reason that the previously mentioned syntactic and structural clues often perform better in evaluation (Baayen et al., 1996).

This paper questions in which way the semantic aspects of someone's linguistic experience contributes to their individuality. One aspect that comes to mind is variations in word usage (as mentioned in the introduction). Unfortunately, this aspect of the problem is also the most difficult to approach computationally, for sheer lack of data: we highlight in §5 some of the reasons why obtaining (enough) speaker-specific language data remains a technical and privacy minefield. Another aspect, which is perhaps more straightforwardly modellable, is the extent to which the type of linguistic material someone is exposed to broadly correlates with *who they are*. It is likely, for instance, that the authors of this paper write and read a lot about linguistics, and this correlates with broad features of theirs, e.g. they are computational linguists and are interested in language. So, as particular stylistic features can predict *who* a person is, a specific semantic experience might give an insight into *what kind* of person they are.

In what follows, we describe how, by selecting a *public subset* of a person's linguistic environment, we can build a representation of that person which encapsulates and summarises a part of their individuality. The term 'public subset' is important here, as the entire linguistic experience of an individual is (at this point in time!) only accessible to them, and the nature of the subset dictates which aspect of the person we can model. For instance, knowing what a particular academic colleague has written, read and cited may let us model their work expertise, while chatting with them at a barbecue party might give us insight into their personal life.

We further contend that what we know about a person conditions the type of interaction we have with them: we are more likely to start a conversation about linguistics with someone we see as a linguist, and to talk about the bad behaviour of our dog with a person we have primarily modelled as a dog trainer. In other words, the model we have of people helps us successfully communicate with them.

3 Some fundamentals of DS

The basis of any DS system is a set of word meaning representations ('distributions') built from large corpora. In their simplest form,[1] distributions are vectors in a so-called *semantic space* where each dimension represents a term from the overall system's vocabulary. The value of a vector along a particular dimension expresses how characteristic the dimension is for the word modelled by the vector (as calculated using, e.g., Pointwise Mutual Information). It will be found, typically, that the vector *cat* has high weight along the dimension *meow* but low weight along *politics*. More complex architectures result in compact representations with reduced dimensionality, which can integrate a range of non-verbal information such as visual and sound features (Feng and Lapata, 2010; Kiela and Clark, 2015).

Word vectors have been linked to conceptual

[1] There are various possible ways to construct distributions, including predictive language models based on neural networks (Mikolov et al., 2013).

representations both theoretically (Erk, 2013) and experimentally, for instance in psycholinguistic and neurolinguistic work (Anderson et al., 2013; Mitchell et al., 2008). The general idea is that a distribution encapsulates information about *what kind of thing* a particular concept might be. Retrieving such information in ways that can be verbalised is often done by looking at the 'nearest neighbours' of a vector. Indeed, a natural consequence of the DS architecture is that similar words cluster in the same area of the semantic space: it has been shown that the distance between DS vectors correlates well with human similarity judgements (Baroni et al., 2014b; Kiela and Clark, 2014). So we can find out what a cat is by inspecting the subspace in which the vector *cat* lives, and finding items such as *animal, dog, pet, scratch* etc.

In what follows, we use this feature of vector spaces to give an interpretable model of an individual, i.e., we can predict that a person might be a linguist by knowing that their vector is the close neighbour of, say, *semantics, reference, model.*

4 A DS model of a community

4.1 People in semantic spaces

Summing up what we have said so far, we follow the claim that we can theoretically talk about the linguistic experience of a speaker in distributional terms. The words that a person has read, written, spoken or heard, are a very individual signature for that person. The sum of those words carries important information about the type of concepts someone may be familiar with, about their social environment (indicated by the registers observed in their linguistic experience) and, broadly speaking, their interests.

We further posit that people's individuality can be modelled as vectors in a semantic space, in a way that the concepts surrounding a person's vector reflect their experience. For instance, a cook might 'live' in a subspace inhabited by other cooks and concepts related to cooking. In that sense, the person can be seen as any other concept inhabiting that space.

In order to compute such person vectors, we expand on a well-known result of compositional distributional semantics (CDS). CDS studies how words combine to form phrases and sentences. While various, more or less complex frameworks have been proposed (Clark et al., 2008; Mitchell and Lapata, 2010; Baroni et al., 2014a), it has repeatedly been found that simple addition of vectors performs well in modelling the meaning of larger constituents (i.e., we express the meaning of *black cat* by simply summing the vectors for *black* and *cat*). To some extent, it is also possible to get the 'gist' of simple sentences by summing their constituent words. The fundamental idea behind simple addition is that, given a coherent set of words (i.e. words which 'belong together and are close in the semantic space), their sum will express the general topic of those words by creating a centroid vector sitting in their midst. This notion of coherence is important: summing two vectors that are far away from each other in the space will result in a vector which is far from both the base terms (this is one of the intuitions used in (Vecchi et al., 2011) to capture semantically anomalous phrases).

We take this idea further by assuming that people are on the whole coherent (see (Herbelot, 2015) for a similar argument about proper names): their experiences reflect who they are. For instance, by virtue of being a chef, or someone interested in cooking, someone will have many interconnected experiences related to food. In particular, a good part of their *linguistic* experiences will involve talking, reading and writing about food. It follows that we can represent a person by summing the vectors corresponding to the words they have been exposed to. When aggregating the vocabulary most salient for a chef, we would hopefully create a vector inhabiting the 'food' section of the space. As we will see in §6, the model we propose is slightly more complex, but the intuition remains the same.

Note that, in spite of being 'coherent', people are not one-sided, and a cook can also be a bungee-jumper in their spare time. So depending on the spread of data we have about a person, our method is not completely immune to creating vectors which sit a little too far away from the topics they encapsulate. This is a limit of our approach which could be solved by attributing a set of vectors, rather than a single representation, to each person. In this work, however, we do not consider this option and assume that the model is still discriminative enough to distinguish people.

4.2 From person vectors to interacting agents

In what sense are person vectors useful representations? We have said that, as any distribution in a semantic space, they give information about *the type of thing/person* modelled by the vector. We also mentioned in §2 that knowing who someone is (just like knowing *what* something is) influences our interaction with them. So we would like to model in which ways our people representations help us successfully communicate with them.

For the purpose of this paper, we choose an information retrieval task as our testbed, described in §5. The task, which involves identifying a relevant knowledge holder for a particular question, requires us to embed our person vectors into simple agent-like entities, with a number of linguistic, knowledge-processing and communicative capabilities. A general illustration of the structure of each agent is shown in Fig. 1. An agent stores (and dynamically updates) a) a person vector; b) a memory which, for the purpose of our evaluation (§5), is a store of linguistic experiences (some data the person has read or written, e.g. information on Venezuelan cocoa beans). The memory acts as a knowledge base which can be queried, i.e. relevant parts can be 'remembered' (e.g. the person remember reading about some Valrhona cocoa, with a spicy flavour). Further, the agent has some awareness of others: it holds a model of its community consisting of other people's vectors (e.g., the agent knows Bob, who is a chef, and Alice, who is a linguist). When acted by a particular communication need, the agent can direct its attention to the appropriate people in its community and engage with them.

5 Evaluating person vectors

5.1 The task

To evaluate our person vectors, we choose a task which relies on having a correct representation of the expertise of an individual.

Let's imagine a person with a particular *information* need, for instance, getting sightseeing tips for a holiday destination. Let's also say that we are in a pre-Internet era, where information is typically sought from other actors in one's real-world community. The communication process associated with satisfying this information need takes two steps: a) identifying the actors most likely to hold relevant knowledge (perhaps a friend who has done the trip before, or a local travel agent); b) asking them to share relevant knowledge.

In the following, we replicate this situation using a set of agents, created as described in §4.

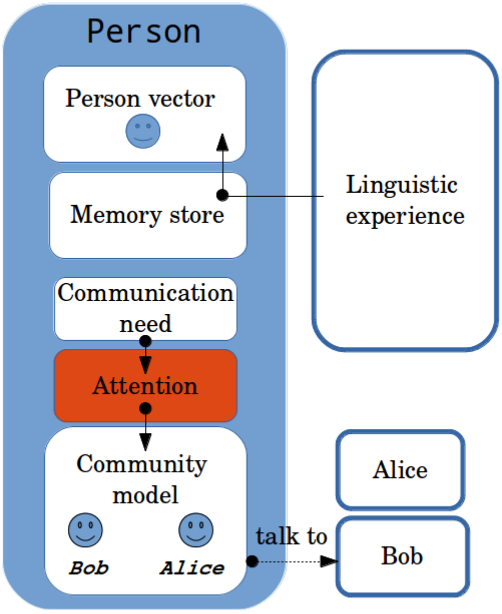

Figure 1: A person is exposed to a set of linguistic experiences. Computationally, each experience is represented as a vector in a memory store. The sum of those experiences make up the individual's 'person vector'. The person also has a model of their community in the form of other individuals' person vectors. In response to a particular communication need, the person can direct their attention to the relevant actors in that community.

We assume that those agents are fully connected and aware of each other, in a way that they can direct specific questions to the individuals most likely to answer them. Our evaluation procedure tests whether, for a given information need, expressed in natural language by one agent (e.g. *What is Venezuelan chocolate like?*), the community is modelled in a way that an answer can be successfully obtained (i.e. an agent with relevant expertise has been found, and 'remembers' some information that satisfies the querier's need). Note that we are not simulating any real communication between agents, which would require that the information holder generates a natural language answer to the question. Rather, the contacted agent simply returns the information in its memory store which seems most relevant to the query at hand. We believe this is enough to confirm that the person vector was useful in acquiring the information: if the querying agent contacts the 'wrong' person, the system has failed in successfully fulfulling the information need.

5.2 Comparative evaluation

We note that the task we propose can be seen as an information retrieval (IR) problem over a distributed network: a query is matched to some relevant knowledge unit, with all available knowledge being split across a number of 'peers' (the individuals in our community). So in order to know how well the system does at retrieving relevant information, we can use as benchmark standard IR software.

We compare the performance of our system with a classic, centralised IR algorithm, as implemented in the *Apache Lucene* search engine. Lucene is an open source library for implementing (unstructured) document retrieval systems, which has been employed in many full-text search engine systems (for an overview of the library, see (Bialecki et al., 2012)). We use the out-of-the-box 'standard' indexing solution provided by Lucene,[2] which roughly implements a term-by-document Vector Space Model, in which terms are lemmatised and associated to documents using their *tf-idf* scores (Spärck-Jones, 1972) computed from the input Wikipedia corpus of our evaluation. Similarly, queries are parsed using Lucene's standard query parser and then searched and ranked by the computed 'default' similarities.[3]

Our hypothesis is that, if our system can match the performance of a well-known IR system, we can also conclude that the person vectors were a good summary of the information held by a particular agent.

5.3 Data challenges

Finding data to set up the evaluation of our system is an extremely challenging task. It involves finding a) personalised linguistic data which can be split into coherent 'linguistic experiences'; b) realistic natural language queries; c) a gold standard matching queries and relevant experiences. There is very little openly available data on people's personal linguistic experience. What is available comes mostly from the Web science and user personalisation communities and such data is either not annotated for IR evaluation purposes (e.g. (von der Weth and Hauswirth, 2013)), or proprietary and not easily accessible or re-distributable (e.g. (Collins-Thompson et al., 2011)). Conversely, standard IR datasets do not give any information about users' personal experiences. We attempt to solve this conundrum by using information freely available on Wikipedia. We combine a Wikipedia-based Question Answering (QA) dataset with contributor logs from the online encyclopedia.

We use the freely available 'WikiQA' dataset of (Yang et al., 2015).[4] This dataset contains 3047 questions sampled from the Bing search engine's data. Each question is associated with a Wikipedia page which received user clicks at query time. The dataset is further annotated with the particular sentence in the Wikipedia article which answers the query – if it exists. Many pages that were chosen by the Bing users do not actually hold the answer to their questions, reducing the data to 1242 queries and the 1194 corresponding pages which can be considered relevant for those queries (41% of all questions). We use this subset for our experiments, regarding each document in the dataset as a 'linguistic experience', which can be stored in the memory of the agent exposed to it.

To model individuals, we download a log of Wikipedia contributions (March 2015). This log is described as a 'log events to all pages and users'. We found that it does not, in fact, contain all possible edits (presumably because of storage issues). Of the 1194 pages in our WikiQA subset, only 625 are logged. We record the usernames of all contributors to those 625 documents, weeding out contributors whose usernames contain the string *bot* and have more than 10,000 edits (under the assumption that those are, indeed, bots). Finally, for each user, we download and clean all articles they have contributed to.

In summary, we have a dataset which consists of a) 662 WikiQA queries linked to 625 documents relevant for those queries; b) a community of 4379 individuals/agents, with just over 1M documents spread across the memories of all agents.

6 Implementation

Our community is modelled as a distributed network of 4379 agents $\{a_1, \ldots, a_{4379}\}$. Each agent a_k has two components: a) a personal profile component, which fills the agent's memory with information from the person's linguistic experience (i.e., documents she/he reads or edits) and calculates the corresponding person vector; b) an 'attention' component which gets activated when

[2] Ver. 5.4.1, obtained from http://apache.lauf-forum.at/lucene/java/5.4.1.

[3] For an explanation of query matching and similarity computation see http://lucene.apache.org/core/5_4_1/core/org/apache/lucene/search/similarities/Similarity.html.

[4] http://aka.ms/WikiQA

a communication need is felt. All agents share a common semantic space $\mathcal{S}$ which gives background vectorial representations for words in the system's vocabulary. In our current implementation, $\mathcal{S}$ is given by the CBOW semantic space of (Baroni et al., 2014b), a 400-dimension vector space of 300,000 items built using the neural network language model of (Mikolov et al., 2013). This space shows high correlation with human similarity judgements (i.e., $\rho = 0.80$) over the 3000 pairs of the *MEN dataset* (Bruni et al., 2012). Note that using a standard space means the we assume shared meaning presentations across the community (i.e., at this stage, we don't model inter-speaker differences at the lexical item level).

Person vectors: A person vector is the normalised sum of that person's linguistic experiences:

$$\vec{p} = \sum_{1..k..n} \vec{e_k}. \tag{1}$$

As mentioned previously, in our current setup, linguistic experiences correspond to documents.

Document/experience vectors: we posit that the (rough) meaning of a document can be expressed as an additive function acting over (some of) the words of that document. Specifically, we sum the 10 words that are most characteristic for the document. While this may seem to miss out on much of the document's content, it is important to remember that the background DS representations used in the summation are already rich in content: the vector for *Italy*, for instance, will typically sit next to *Rome*, *country* and *pasta* in the semantic space. The summation roughly captures the document's content in a way equivalent to a human describing a text as being *about so and so*.

We need to individually build document vectors for potentially sparse individual profiles, without necessitating access to the overall document collection of the system (because a_k is not necessarily aware of a_m's experiences). Thus, standard measures such as *tf-idf* are not suitable to calculate the importance of a word for a document. We alleviate this issue by using a static list of word entropies (calculated over the ukWaC 2 billion words corpus, (Baroni et al., 2009)) and the following weighting measure:

$$w_t = \frac{freq(t)}{log(H(t)+1)}, \tag{2}$$

where $freq(t)$ is the frequency of term t in the document and $H(t)$ is its entropy, as calculated over a larger corpus. The representation of the document is then the weighted sum of the 10 terms[5] with highest importance for that text:

$$\vec{e} = \sum_{t \in t_1 ... t_{10}} w_t * \vec{t}. \tag{3}$$

Note that both vectors $\vec{t}$ and $\vec{e}$ are normalised to unit length.

For efficiency reasons, we compute weights only over the first 20 lines of documents, also following the observation that the beginning of a document is often more informative as to its topic than the rest (Manning et al., 2008).

Attention: The 'attention' module directs the agent to the person most relevant for its current information need. In this paper, it is operationalised as cosine similarity between vectors. The module takes a query q and translates it into a vector $\vec{q}$ by summing the words in the query, as in Eq. 3. It then goes through a 2-stage process: 1) find potentially helpful people by calculating the cosine distance between $\vec{q}$ and all person vectors $\vec{p_1}...\vec{p_n}$; 2) query the m most relevant people, who will calculate the distance between $\vec{q}$ and all documents in their memory, $D_k = \{d_1...d_t\}$. Receive the documents corresponding to the highest scores, ranked in descending order.

7 Describing the community

7.1 Qualitative checks

As a sanity check, it is possible to inspect where each experience/document vector sits in the semantic space, by looking at its 'nearest neighbours' (i.e., the m words closest to it in the space). We show below two documents with their nearest neighbours, as output by our system:

```
Artificial_intelligence:
ai artificial intelligence intelligent
computational research researchers
computing cognitive computer

Anatoly_Karpov:
chess ussr moscow tournament ukraine
russia soviet russian champion opponent
```

We also consider whether each user inhabits a seemingly coherent area of the semantic space. The following shows a user profile, as output by our system, which corresponds to a person with an interest in American history:

[5] We experimented with a range of values, not reported here for space reasons.

# agents	# docs
2939	1-100
944	100-500
226	500-1000
145	1000-2000
82	2000-5000
15	10000-200000

Table 1: Distribution of documents across people. For example, 2939 agents contain 1–100 documents.

```
name = [...]
topics = confederate indians american
americans mexican mexico states army
soldiers navy
coherence = 0.452686176513
p_vector:0.004526 0.021659 [...] 0.029680
```

The profile includes a username and the 10 nearest neighbours to the user's p_k vector (which give a human-readable representation of the broad expertise of the user), the corresponding coherence figure (see next section for information about coherence) and the actual person vector for that agent.

7.2 Quantitative description

Distribution of documents across agents: An investigation of the resulting community indicates that the distribution of documents across people is highly skewed: 12% of all agents only contain one document, 31% contain less than 10 documents. Table 1 shows the overall distribution.

Topic coherence: We compute the 'topic coherence' of each person vector, that is, the extent to which it focuses on related topics. We expect that it will be easier to identify a document answering a query on e.g. baking if it is held by an agent which contains a large proportion of other cooking-related information. Following the intuition of (Newman et al., 2010), we define the coherence of a set of documents $d_1, \cdots, d_n$ as the mean of their pairwise similarities:

$$\begin{aligned} Coherence(d_{1 \ldots n}) = \; & mean\{Sim(d_i, d_j), \\ & ij \in 1 \ldots n, i < j\} \end{aligned} \quad (4)$$

where Sim is the cosine similarity between two documents.

The mean coherence over the 4379 person vectors is 0.40 with a variance of 0.06. The high variance is due to the number of agents containing one document only (which have coherence 1.0). When only considering the agents with at least two documents, the mean coherence is 0.32, with variance 0.01. So despite a high disparity in memory sizes, the coherence is roughly stable. For reference, a cosine similarity of 0.32 in our semantic space corresponds to a fair level of relatedness: for instance, some words related to *school* at the 0.30 level are *studied, lessons, attend, district, church.*

# relevant docs	# agents containing doc
176	1
169	2-4
100	5-9
64	10-19
45	20-49
49	50-99
19	100-199
3	200-399

Table 2: Redundancy of relevant documents across people. For example, 176 documents are found in one agent; 169 documents are found in 2–4 agents, etc.

Information redundancy: we investigate the redundancy of the created network with respect to our documents of interest: given a document D which answers one or more query in the dataset, we ask how many memory stores contain D. This information is given in Table 2. We observe that 176 documents are contained in only one agent out of 4379. Overall, around 70% of the documents that answer a query in the dataset are to be found in less than 10 agents. So as far as our pages of interest are concerned, the knowledge base of our community is minimally redundant, making the task all the more challenging.

8 Evaluation

The WikiQA dataset gives us information about the document d_{gold} that was clicked on by users after issuing a particular query q. This indicates that d_{gold} was relevant for q, but does not give us information about which other documents might have also be deemed relevant by the user. In this respect, the dataset differs from fully annotated IR collections like the TREC data (Harman, 1993). In what follows, we report *Mean Reciprocal Rank (MRR)*, which takes into account that only one document per query is considered relevant in our dataset:

$$MRR = \sum_{q \in Q} P(q), \quad (5)$$

where Q is the set of all queries, and $P(q)$ is the precision of the system for query q. $P(q)$ itself is

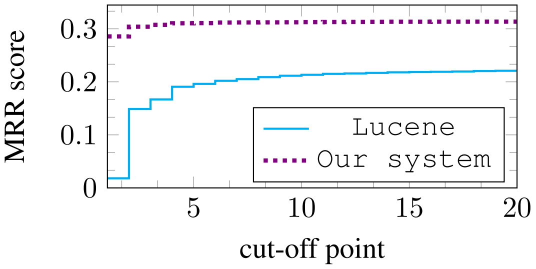

Figure 2: MRR for Lucene and our system (best 5 person vectors).

given by:

$$P(q) = \begin{cases} \frac{1}{r_q} & \text{if } r_q < \text{cutoff} \\ 0 & \text{otherwise} \end{cases},$$

where r_q is the rank at which the correct document is returned for query q, and the cutoff is a predefined number of considered results (e.g., top 20 documents).

The MRR scores for Lucene and our system are shown in Fig. 2. The x-axis shows different cut-off points (e.g., cut-off point 10 means that we are only considering the top 10 documents returned by the system). The graph gives results for the case where the agent contacts the $p = 5$ people potentially most relevant for the query. We also tried $m = \{10, 20, 50\}$ and found that end results are fairly stable, despite the fact that the chance of retrieving at least one 'useful' agent increases. This is due to the fact that, as people are added to the first phase of querying, confusion increases (more documents are inspected) and the system is more likely to return the correct page at a slightly lower rank (e.g., as witnessed by the performance of Lucene's centralised indexing mechanism).

Our hypothesis was that matching the performance of an IR algorithm would validate our model as a useful representation of a community. We find, in fact, that our method considerably outperforms Lucene, reaching $MRR = 0.31$ for $m = 5$ against $MRR = 0.22$. This is a very interesting result, as it suggests that retaining the natural relationship between information and knowledge holders increases the ability of the system to retrieve it, and this, despite the intrinsic difficulty of searching in a distributed setting. This is especially promising, as the implementation presented here is given in its purest form, without heavy pre-processing or parameter setting. Aside from a short list of common stopwords, the agent only uses simple linear algebra operations over raw, non-lemmatised data.

MRR figures are not necessarily very intuitive, so we inspect how many times an agent is found who *can* answer the query (i.e. its memory store contains the document that was marked as holding the answer to the query in WikiQA). We find that the system finds a helpful hand 39% of the time for $m = 5$ and 52% at $m = 50$. These relatively modest figures demonstrate the difficulty of our task and dataset. We must however also acknowledge that finding appropriate helpers amongst a community of 4000 individuals is highly non-trivial.

Overall, the system is very precise once a good agent has been identified (i.e., it is likely to return the correct document in the first few results). This is shown by the fact that the MRR only increases slightly between cut-off point 1 and 20, from 0.29 to 0.31 (compare with Lucene, which achieves $MRR = 0.02$ at rank 1). This behaviour can be explained by the fact that the agent overwhelmingly prefers 'small' memory sizes: 78% of the agents selected in the first phase of the querying process contain less than 100 documents. This is an important aspect which should guide further modelling. We hypothesise that people with larger memory stores are perhaps less attractive to the querying agent because their profiles are less topically defined (i.e., as the number of documents browsed by a user increases, it is more likely that they cover a wider range of topics). As pointed out in §4, we suggest that our person representations may need more structure, perhaps in the form of several coherent 'topic vectors'. It makes intuitive sense to assume that a) the interests of a person are not necessarily close to each other (e.g. someone may be a linguist and a hobby gardener); b) when a person with an information need selects 'who can help' amongst their acquaintances, they only consider the relevant aspects of an individual (e.g., the hobby gardener is a good match for a query on gardening, irrespectively of their other persona as a linguist).

Finally, we note that all figures reported here are below their true value (including those pertaining to *Lucene*). This is because we attempt to retrieve the page labelled as containing the answer to the query in the WikiQA dataset. Pages which are relevant but not contained in WikiQA are incorrectly given a score of 0. For instance, the query *what classes are considered humanities* returns *Outline* *of the humanities* as the first answer, but the chosen document in WikiQA is *Humanities*.

9 Conclusion

We have investigated the notion of ‘person vector’, built from a set of linguistic experiences associated with a real individual. These ‘person vectors’ live in the same semantic space as concepts and, as any semantic vector, give information about the kind of entity they describe, i.e. what kind of person someone is. We modelled a community of speakers from 1M ‘experiences’ (documents read or edited by Wikipedians), shared across over 4000 individuals. We tested the representations obtained for each individual by engaging them into an information-seeking task necessitating some understanding of the community for successful communication. We showed that our system outperforms a standard IR algorithm, as implemented by the Lucene engine. We hope to improve our modelling by constructing structured sets of person vectors that explicitly distinguish the various areas of expertise of an individual.

One limit of our approach is that we assumed person vectors to be unique across the community, i.e. that there is some kind of ground truth about the representation of a person. This is of course unrealistic, and the picture that Bob has of Alice should be different from the picture that Kim has of her, and again different from the picture that Alice has of herself. Modelling these fine distinctions, and finding an evaluation strategy for such modelling, is reserved for future work.

A more in-depth analysis of our model would also need to consider more sophisticated composition methods. We chose addition in this paper for its ease of implementation and efficiency, but other techniques are known to perform better for representing sentences and documents (Le and Mikolov, 2014)).

We believe that person vectors, aside from being interesting theoretical objects, are also useful constructs for a range of application, especially in the social media area. As a demonstration of this, we have made our system available at `https://github.com/PeARSearch` in the form of a distributed information retrieval engine. The code for the specific experiments presented in this paper is at `https://github.com/PeARSearch/PeARS-evaluation`.

Acknowledgements

We thank Germán Kruszewski, Angeliki Lazaridou and Ann Copestake for interesting discussions about this work. The first author is funded through ERC Starting Grant COMPOSES (283554).

References

Andrew J Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In *EMNLP*, pages 1960–1970.

Harald Baayen, Hans Van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014a. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014b. Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247.

A. Bialecki, R. Muir, and G. Ingersoll. 2012. Apache Lucene 4. pages 17–24.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of ACL*, pages 136–145.

Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. 2008. A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 133–140.

Stephen Clark. 2012. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics – second edition*. Wiley-Blackwell.

Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian de la Chica, and David Sontag. 2011. Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412. ACM.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Katrin Erk. 2013. Towards a semantics for distributional representations. In *Proceedings of the Tenth International Conference on Computational Semantics (IWCS2013)*.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *NAACL-HLT2010*, pages 91–99, Los Angeles, California, June.

Donna K. Harman. 1993. The first text retrieval conference (TREC-1). *Information Processing & Management*, 29(4):411–414.

Aurélie Herbelot. 2015. Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 151–161.

Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL*, pages 21–30.

Douwe Kiela and Stephen Clark. 2015. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *EMNLP*.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *EACL*, pages 234–244, Avignon, France.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments, & Computers*, 28:203–208, June.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press, Cambridge, UK.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429, November.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *NAACL*, pages 100–108.

Karen Spärck-Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (Linear) maps of the impossible: capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9. Association for Computational Linguistics.

Christian von der Weth and Manfred Hauswirth. 2013. Dobbs: Towards a comprehensive dataset to study the browsing behavior of online users. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013*, volume 1, pages 51–56. IEEE.

Anna Wierzbicka. 1984. Cups and mugs: Lexicography and conceptual analysis. *Australian Journal of Linguistics*, 4(2):205–255.

Ludwig Wittgenstein. 1953. *Philosophical investigations*. Wiley-Blackwell (reprint 2010).

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WIKIQA: A Challenge Dataset for Open-Domain Question Answering. In *EMNLP*.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393.