Distributional Representation of Words for Small Corpora using Pre-training Techniques

Pierpaolo Basile, Lucia Siciliani, Gaetano Rossiello, Pasquale Lops
pierpaolo.basile@uniba.it,lucia.siciliani@uniba.it,gaetano.rossiello@uniba.it,pasquale.lops@uniba.it
Department of Computer Science - University of Bari Aldo Moro
Via E. Orabona, 4, 70125 - Bari, ITALY

ABSTRACT

Distributional Semantic Models and word embedding approaches have proved their effectiveness to represent words as mathematical points in a geometric space. Relying on this representation allows computing the relatedness between words according to their distance in the space. This ability is useful for several natural language processing tasks. However, when we have a collection containing a few documents, it is not possible to build an accurate representation of words because we do not have enough information about the co-occurrences of terms. In this paper, we deal with this issue by proposing an approach which relies on Random Indexing and a pre-trained model built on a large and balanced corpus. We perform an evaluation by investigating a real world application scenario in which this approach has been adopted.

CCS CONCEPTS

• **Information systems → Document collection models**; *Information extraction*; Digital libraries and archives; Dictionaries; • **Computing methodologies → Lexical semantics**; *Information extraction*; • **Applied computing → Document analysis**.

KEYWORDS

distributional semantic models, random indexing, distributed representation of words

1 BACKGROUND AND MOTIVATION

Distributional Semantics Models (DSMs) [16] and more recent word embeddings approaches [13] have proved that the distributed representation of words is effective in several natural language processing (NLP) tasks. In particular, by representing words as mathematical points in a geometric space, it is possible to compute word relatedness as the distance in that space: two words are similar if they are close in the geometric space. Both DSM and word embeddings approaches have their roots in the distributional hypothesis [7, 8]: two words are similar if they share similar linguistic contexts. Generally, these approaches exploit words co-occurrences as linguistic contexts. Since words co-occurrences strongly depend on the statistical distribution of words in the corpus, these approaches can be affected by the dimension of the corpus. The domain of the corpus can also affect the semantics captured by the DSM or embeddings. If the target corpus is very specific and focused on a single domain (e.g. sport or politics) the model captures only semantic aspects belonging to that domain.

Moreover, there are other aspects that can affect these approaches, such as the initialisation of the embeddings. Since embeddings are randomly initialised, different results could be obtained by applying several times the same approach on the same corpus. Some DSMs can be affected by the method used to count (weight) the co-occurrences, or by the parameters used to reduce the co-occurrences matrix (e.g. the number of dimensions in the LSA [11] approach). More details about pitfalls in both DSM and word embeddings are discussed in [1, 9, 12].

In this paper, we focus our attention on the corpus dimension issue. In some contexts, we have corpora with few documents, and even in those cases we aim at obtaining a distributed representation of words able to effectively capture their semantics. In such a case, it might be useful to *pre-train* a distributed representation of words on a large balanced corpus and then exploit that representation as starting point for building word vectors on the corpus containing few documents.

Recently, contextual word embeddings, such as ELMo [14], ULMFit [10] and BERT [6], have shown to be effective as transfer learning technique NLP. The main idea is to leverage an unsupervised neural language model trained on a large corpus as a pre-training stage. Then, the resulting pre-trained word embeddings are used to train deep neural networks for supervised NLP tasks [17]. Even if these neural language models can mitigate the problem represented by Out-of-Vocabulary words, i.e. words not seen during the language modelling stage, they require an enormous amount of data and high computational capabilities. For these reasons, dealing with new words in small collections of documents for specific domains still remains an open challenge.

In an attempt to address this limitation, we analyse a specific DSM approach called Random Indexing (RI) [15, 18], an incremental method that makes simple to add new documents to an already existing model. The incremental property of RI is already exploited for discovering implicit connections between terms [4] and for analysing the evolution of language over time [3]. We chose RI because other approaches based on word embeddings are not inherently incremental. It is possible to initialise embeddings with embeddings built on another corpus, but it is not simple to tackle the issue due to out of vocabulary words[1].

The general idea behind our approach is to build a model M_{pre} on a large balanced corpus and then, given a new small collection of documents C_s, build a new model M_s relying on word vectors in M_{pre}. The goal is to deal with the issue of the small dimension of

IIR 2019, September 16–18, 2019, Padova, Italy

[1] Words that occur in the domain corpus but do not occur in the embeddings used for the initialisation.

the corpus C_s, by relying on the information captured during the definition of the model M_{pre}.

Our research question is to prove that in case of a small collection of documents the approach based on pre-training is able to provide better performance with respect to a word representation without pre-training. We provide an evaluation by exploiting a real world application scenario in which given a collection of documents and a set of seed words we want to discover related concepts by exploiting the relatedness computed in the semantic space.

The paper is structured as follow: Section 2 describes the proposed methodology for pre-training word vectors using RI, while Section 3 provides details about the evaluation and reports the results. Finally, Section 4 closes the paper by providing final remarks and future work.

2 METHODOLOGY

Our approach is based on RI. The mathematical insight behind RI is the projection of a high-dimensional space on a lower dimensional one using a random matrix; this kind of projection does not compromise distance metrics[2] [5].

Formally, given a $n \times m$ matrix A and an $m \times k$ matrix R, which contains random vectors, we define a new $n \times k$ matrix B as follows:

$$A^{n,m} \cdot R^{m,k} = B^{n,k} \quad k << m \tag{1}$$

The new matrix B has the property to preserve the distance between points.

Specifically, RI creates the *DSM* in two steps:

(1) A random vector is assigned to each word. This vector is sparse, high-dimensional and ternary, which means that its elements can take values in {-1, 0, 1}. A random vector contains a small number of randomly distributed non-zero elements, and the structure of this vector follows the hypothesis behind the concept of Random Projection;
(2) Random vectors are accumulated by analysing co-occurring words. In particular the semantic vector for any word is computed as the sum of the random vectors for words that co-occur with the analysed word. When computing the sum, we apply some weighting to the random vector. In our case, to reduce the impact of very frequent terms, we use the following weight: $h_i = \sqrt{\frac{th \times C}{\#t_i}}$, where C is the total number of occurrences in the corpus and $\#t_i$ is the occurrences of the term t_i. The idea behind this weighting schema is to penalise most frequent words. The parameter th is generally set to 0.001.

Formally, given a corpus D of n documents, and a vocabulary V of m words extracted form D, we perform two steps: i) we assign a random vector r to each word w in V; ii) we compute a semantic vector sv_i for each word w_i as the sum of all random vectors assigned to words co-occurring with w_i. The context is the set of c words that precede and follow w_i. In our experiment we set c to 5. The second step is defined by the following equation:

$$sv_i = \sum_{d \in D} \sum_{\substack{-c<j<+c \\ j \neq i}} h_j * r_j \tag{2}$$

[2]Only L_2 norm-based distances are preserved.

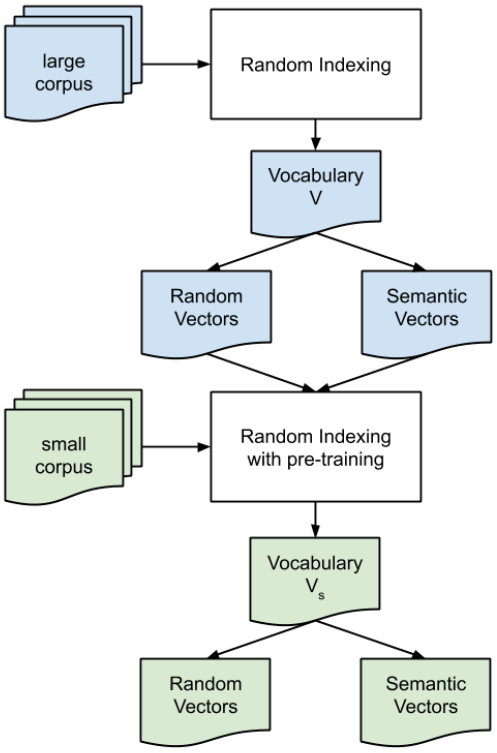

Figure 1: Random Indexing with pre-trained vectors.

where h_j is the weight applied to the context word as previously explained. After these two steps, we obtain a set of semantic vectors assigned to each word in V representing our DSM.

We apply the classical RI approach to the balanced large corpus as described above. We obtain two spaces: i) the set of random vectors assigned to each word w in V and ii) the set of semantic vectors SV built by accumulating random vectors. The set SV contains a semantic vectors for each word in V.

Given a small collection of documents S we want to apply RI by relying on vectors built on the large corpus. Since RI is an incremental approach we can reuse the vectors built on the large collection. In particular:

(1) we extract the vocabulary V_s from S. V_s can contain words that already occur in V. For these words $w_j \in V \cap V_s$ we reuse both the random vector and the semantic vector assigned to w_j. For words $w_k \in V_s \setminus V$ we build new random vectors and initialise the semantic vectors coordinates to zero.
(2) we perform the accumulation of random vectors by analysing word co-occurrences as describe above for the classical RI approach.

The output of this process is composed of two new sets of random vectors and semantic vectors as reported in Figure 1.

3 EVALUATION

The goal of the evaluation is to prove that in case of a small collection of documents the approach based on RI with pre-training is

able to provide better performance with respect to a word representation based on RI without pre-training. However, the datasets usually used to evaluate word similarity performance are based on common concepts or common entities. This kind of datasets is not suitable for evaluating specific domain collections as in our case. For that reason we design an in-vivo evaluation by integrating our approach in an already existing system called Semantic Framework [2]. The Semantic Framework provides a set of tools and services for analysing, indexing and searching a collection of documents for the Public Administration. Moreover, the framework also provides services for discovering related words and concepts starting from both a collection of documents and the description of two concepts. In particular, given the description of two concepts given as a set of words, the tool is able to provide a ranked list of other words that are somehow related to both the initial concepts. This is the specific scenario we have chosen for the empirical evaluation.

3.1 Extraction of Related Words

Given two concepts c_1 and c_2, along with their descriptions given as set of words d_1 and d_2, and given a collection of documents, the goal is to extract a ranked list of words related to both c_1 and c_2. The method relies on both the distributed representation of words and the similarity between words in the geometric space. In particular, given a collection of documents, we build a DSM where each word is represented as a vector. For each concept, e.g. c_1 and c_2, we build a vector representation by computing the centroid of the vectors of the keywords occurring in the concept description (e.g. d_1 and d_2).

The second step is to compute the list of related words: given the two vectors describing the concepts c_1 and c_2, we retrieve the neighbourhood for each concept by using cosine similarity. In the last step we normalised each list using the z-norm normalisation and we create the final list by averaging the score of words occurring in both the lists (intersection). The final list is ranked and the top-N words are returned to the user. The whole process is sketched in Figure 2.

The GUI provided by the Semantic Framework for building the list of related words is shown in Figure 3. The tool allows: i) to build a matrix with a specific number of rows and columns; ii) to define each concept on the rows and columns by associating a description which is then adopted to build the corresponding vector representation. It is worth to notice that the cell in the matrix reports multi-word expressions, instead of single words, since the collection of documents has been indexed by exploiting the Semantic Framework services, able to automatically extract phrases from documents [2] .

3.2 Evaluation Setup

We evaluate our method on several collections of documents in both English and Italian. Four English collections are taken from the TALIA European project and two Italian collections are provided by the Apulia Region. In particular, English collections concern deliverable of European projects related to the Interreg-Mediterranean program, while the two Italian collections contain project proposals of two research programs funded by the Apulia Region. Table 1 shows the statistics about the collections.

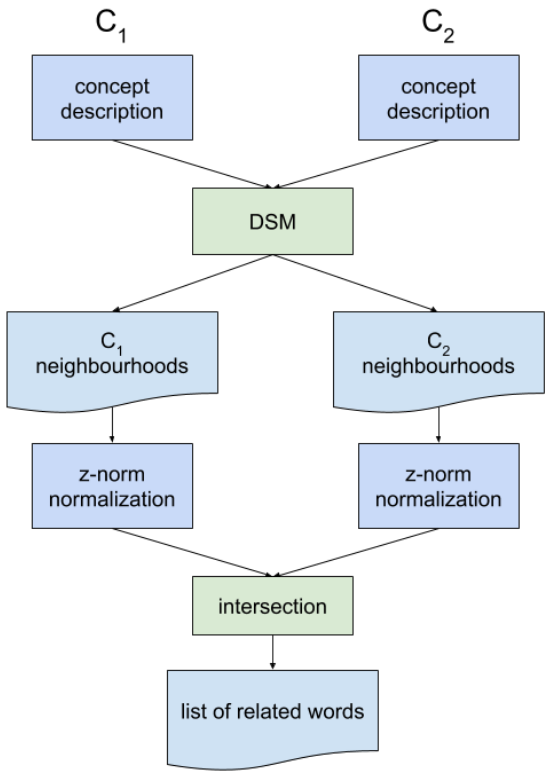

Figure 2: Extraction of related words given two concepts.

Collection	Language	#documents	#occurrences
T1	EN	200	1,460,713
T2	EN	221	1,802,770
T3	EN	158	1,360,393
T4	EN	579	4,623,876
A1	IT	55	623,575
A2	IT	87	876,654

Table 1: Statistics about collections used during the evaluation.

As corpus for the pre-training step, we adopt the British National Corpus (BNC)[3] for the English language and the Paisà[4] corpus for the Italian.

BNC is a 100 million word collection of samples of written and spoken language from several sources, designed to represent a wide cross-section of British English both spoken and written.

The Paisà corpus is a large collection of Italian web texts in which documents were selected in two different ways. A part of the corpus was constructed by querying the Web exploiting 50,000 word pairs combining terms from an Italian basic vocabulary list. The remaining documents come from the Italian versions of various Wikimedia Foundation projects, namely: Wikipedia, Wikinews,

[3]http://www.natcorp.ox.ac.uk/
[4]https://www.corpusitaliano.it/

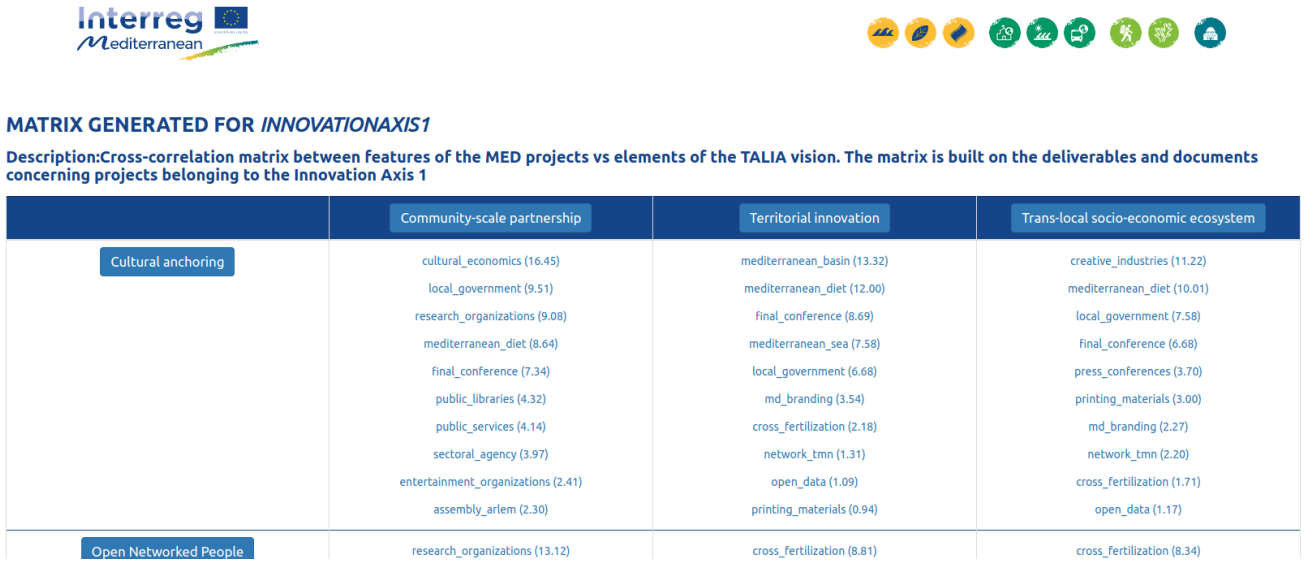

Figure 3: GUI for the related word extraction.

Wikisource, Wikibooks, Wikiversity, Wikivoyage. The corpus contains approximately 380,000 documents coming from about 1,000 different websites, for a total of about 250 million words.

We pre-train vectors by using Random Indexing with a vector dimension equals to 300 with 10 non-zero elements in the random vector. We limit the vocabulary dimension to 100,000 by taking into account the most frequent words.

The code for building Random Indexing with pre-training is freely available on GitHub[5].

3.3 Results

One expert of the TALIA project and one expert of the Apulia Region provided a set of concepts pairs for which the list of related words is extracted as described in Section 3.1. In particular, experts provide 24 pairs for the English collections, and 12 pairs for the Italian ones. Each list has been evaluated by two experts. In particular, two lists are provided to each expert, one built by using pre-trained RI and another one using only RI. The expert does not know the method used to build the list. Given the pair of concepts (with their descriptions) and the two lists of related words the expert must judge which list provides more significant words. Finally, we compute the percentage of times that both the experts prefer the list built through the pre-trained RI.

Analysing the experts' judgements we observe that they agree on the higher significance of the list created with the pre-trained RI the 84% of times for the English, while for the Italian the agreement is 75%. This first in-vivo evaluation provides encouraging results and suggests that the pre-training is fundamental when the collection contains few documents. We plan to design an in-vitro evaluation by developing a specific dataset.

[5]We will release the URL in case of acceptance.

4 CONCLUSIONS

In this paper, we propose a pre-training strategy for building a distributional semantics model when a small collection of documents is involved. In particular, we extend an existing DSM approach called Random Indexing by introducing a pre-training step that relies on a large and balanced corpus. We have integrated our method in a tool for the semantic analysis of documents and we designed an in-vivo evaluation that involves two languages (English and Italian) and six collections of documents.

Results prove that the approach based on pre-training provides better results. This suggests that in case of a small collection of documents the additional information provided by a large corpus might help to improve the quality of the distributional model.

As future work, we plan to develop a pre-training strategy for approaches based on word embeddings and design an in-vitro evaluation by building a specific dataset.

ACKNOWLEDGMENTS

This work is partially funded by the "TALIA - Territorial Appropriation of Leading-edge Innovation Action" project, Interreg-Mediterranean program, priority axis 1: Promoting Mediterranean innovation capacities to develop smart and sustainable growth, Programme specific objective 1.1 to increase transnational activity of innovative clusters and networks of key sectors of the MED area (2018-2019).

REFERENCES

[1] Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association of Computational Linguistics* 6 (2018), 107–119.

[2] Pierpaolo Basile, Annalina Caputo, Marco Di Ciano, Gaetano Grasso, Gaetano Rossiello, and Giovanni Semeraro. 2017. SEPIR: a semantic and personalised information retrieval tool for the public administration based on distributional semantics. *International Journal of Electronic Governance* 9, 1-2 (2017), 132–155.

[3] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In *First Italian Conference on Computational Linguistics CLiC-it.*

[4] Trevor Cohen, Roger Schvaneveldt, and Dominic Widdows. 2010. Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics* 43, 2 (2010), 240 – 256. https://doi.org/10.1016/j.jbi.2009.09.003

[5] Sanjoy Dasgupta and Anupam Gupta. 1999. *An elementary proof of the Johnson-Lindenstrauss lemma.* Technical Report. Technical Report TR-99-006, International Computer Science Institute, Berkeley, California, USA.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[7] John Rupert Firth. 1957. *A synopsis of linguistic theory.* Studies in linguistic analysis.

[8] Zellig S. Harris. 1968. *Mathematical Structures of Language.* New York: Interscience.

[9] Johannes Hellrich and Udo Hahn. 2016. Bad companyâĂŤneighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2785–2796.

[10] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *ACL (1)*. Association for Computational Linguistics, 328–339.

[11] 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* 104, 2 (1997), 211–240.

[12] Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3 (2015), 211–225.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

[14] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*. Association for Computational Linguistics, 2227–2237.

[15] Magnus Sahlgren. 2005. An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, Vol. 5.

[16] Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Ph.D. Dissertation. Stockholm: Stockholm University, Faculty of Humanities, Department of Linguistics.

[17] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *BlackboxNLP@EMNLP*. Association for Computational Linguistics, 353–355.

[18] Dominic Widdows and Kathleen Ferraro. 2008. Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.