Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 127-130 recibido 30-11-2019 revisado 12-01-2020 aceptado 16-01-2020

A comprehensive analysis of the parameters in the creation and comparison of feature vectors in distributional semantic models for multiple languages

Análisis exhaustivo respecto a varios idiomas de los parámetros usados durante la creación y comparación de vectores de propiedades de modelos semánticos distribucionales

András Dobó
Institute of Informatics, University of Szeged
2 Árpád tér, Szeged, 6720 Hungary
dobo@inf.u-szeged.hu

Abstract: PhD thesis written by András Dobó under the supervision of Prof. Dr. János Csirik (University of Szeged). The thesis was defended in Szeged (Hungary) on the 15th of November, 2019. The doctoral committee comprised of Prof. Dr. Márk Jelasity (University of Szeged), Prof. Dr. András Kornai (Budapest University of Technology and Economics), Prof. Dr. Reinhard Köhler (University of Trier), Dr. Rudolf Ferenc (University of Szeged) and Dr. Zsolt Gazdag (University of Szeged). The thesis obtained the grade of *Summa Cum Laude*.
Keywords: distributional semantic models; semantic similarity and relatedness; best combination of parameter settings; English, Spanish and Hungarian

Resumen: Tesis doctoral elaborada por András Dobó con la supervisión del Prof. Dr. János Csirik (Universidad de Szeged). La defensa de la tesis tuvo lugar en Szeged (Hungría), el 15 de noviembre de 2019. Los miembros del comité de doctorado fueron Prof. Dr. Márk Jelasity (Universidad de Szeged), Prof. Dr. András Kornai (Universidad de Tecnología y Economía de Budapest), Prof. Dr. Reinhard Köhler (Universidad de Tréveris), Dr. Rudolf Ferenc (Universidad de Szeged) y Dr. Zsolt Gazdag (Universidad de Szeged). La tesis fue evaluada con la calificación de *Summa Cum Laude*.
Palabras clave: modelos de semántica distribucional; similitud y relación semántica; combinación óptima de configuraciones de parámetros; inglés, español y húngaro

1 Introduction

For many natural language processing problems, including noun compound interpretation (Dobó and Pulman, 2011), it is crucial to determine the semantic similarity or relatedness of words. While relatedness considers a wide range of relations (including similarity), similarity only examines how much the concepts denoted by the words are truly alike.

1.1 Motivation

Most semantic models calculate the similarity or relatedness of words using distributional data extracted from large corpora. These models can be collectively called as distributional semantic models (DSMs). In these models feature vectors are created for each word, usually made up of context words with weights, and the similarity or relatedness of words is then calculated using vector similarity measures. Although DSMs have many possible parameters, a truly comprehensive study of these parameters, also fully considering the dependencies between them, is still missing and would be needed.

Most papers presenting DSMs focus on only one or two aspects of the problem, and take all other parameters as granted with some standard setting. For example, the majority of studies simply use cosine as vector similarity measure and/or (positive) pointwise mutual information as weighting scheme

ISSN 1135-5948. DOI 10.26342/2020-64-16

out of convention. Further, even in case of the considered parameters, usually only a handful of possible settings are tested for. Moreover, there are also such parameters that are completely ignored by most studies and have not been truly studied in the past, not even separately (e.g. smoothing, vector normalization or minimum feature frequency). What's more, as these parameters can influence each other greatly, evaluating them separately, one-by-one, would not even be sufficient, as that would not account for the interaction between them.

There are a couple of studies that consider several parameters for DSMs with multiple possible settings, but even these are far from truly comprehensive, and do not fully test for the interaction between the different parameters. So, although a comprehensive analysis of the possible parameters and their combinations would be crucial, there has been no such research to date. Further, although the best settings for the parameters can differ for different languages, the vast majority of papers consider DSMs for only one language (mostly English), or consider multiple languages but without a real comparison of findings across languages. Our study aims to address these gaps.

1.2 Aims and objectives

DSMs have two distinct phases in general. First, statistical distributional information (e.g. raw counts) is extracted from raw data (e.g. a large corpus). Then, feature vectors are created for words from the extracted information, and the similarity or relatedness of words is calculated by comparing their feature vectors. In our study we take the distributional information extracted in the first phase as already granted, and present a systematic study simultaneously testing all important aspects of the creation and comparison of feature vectors in DSMs, also caring for the interaction of the different parameters.

We have chosen to only study the second phase of DSMs, as the two phases are relatively independent from each other, and testing for every possible combination of parameter settings in the second phase is already unfeasible due to the vast number of combinations. So, instead of a full analysis, we already had to use a heuristic approach. Thus, we have omitted the examination of the first phase, as that would have been unreasonable and unmanageable, with one exception.

DSMs relying on information extracted from static corpora have two major categories, based on the type of their first phase: count-vector-based (CVBM) and predictive models (PM; also called word embeddings). In order to get a more complete view and due to the recent popularity of predictive models, in addition to using information extracted by a CVBM, we have also conducted experiments with information extracted by a PM for English. Further, we have also extended our analysis with a model based on a knowledge graph. Our intuition was that there will be a single configuration that achieves the best results in case of all types of models. However, please note that in the latter two cases only a part of the considered parameters could be tested due to the characteristics of such models. We have mainly focused on count-vector-based DSMs partly due to this.

During our research we have identified altogether 10 important parameters for the second phase of count-vector-based DSMs, such as vector similarity measures, weighting schemes, feature transformation functions, smoothing and dimensionality reduction techniques. However, only 4 of these parameters are available when predictive or knowledge-graph-based semantic vectors are used as input, as in those cases the raw counts are not available any more, the weighted vectors are already constructed and their dimensions are usually also reduced.

In the course of our analysis we have simultaneously evaluated each parameter with numerous settings in order to try to find the best possible configuration achieving the highest performance on standard test datasets. We have done our extensive analysis for English, Spanish and Hungarian separately, and then compared our findings for the different languages.

While also testing the conventionally used settings for each parameter, we also proposed numerous new variants in case of some parameters. Therefore, for many parameters a large number of settings (more than a thousand in some cases) were tested, resulting in trillions of possible combinations. All in all, we have considered a vast number of novel configurations, with some of these considerably outperforming the standard configurations that are conventionally used, and thus achieving state-of-the-art results.

First we have done our analysis for English and evaluated the results extensively (Dobó and Csirik, 2019a). Then we have repeated our analysis, with an increased number of settings for several parameters, for English, Spanish and Hungarian, and compared the findings across the different languages (Dobó and Csirik, 2019b).

2 Thesis overview

The thesis (Dobó, 2019) is organized into nine chapters, followed by the Appendices.

Chapter 1 gives an introduction to the topic, and presents our motivations and aims. Then we provide the theoretical background to the topic in Chapter 2, after which Chapter 3 introduces the used data and evaluation methods. Dobó and Csirik (2012) and Dobó and Csirik (2013) describe the inception of our research into the topic, and introduce a part of the used data and evaluation methods.

Chapter 4 is devoted to the detailed description of our analysis (Dobó and Csirik, 2019a), including the presentation of our two-phase heuristic approach and the description of the ten tested parameters. Our novel smoothing technique (Dobó, 2018) tested during our analysis is also presented here.

In Chapter 5 we present the results of our two-phase heuristic analysis for English, and then evaluate these results (Dobó and Csirik, 2019a). This is followed by the detailed comparison of the results for English, Spanish and Hungarian (Dobó and Csirik, 2019b) in Chapter 6.

We draw our conclusions in Chapter 7, which is followed by the English and Hungarian synopsis of the thesis in Chapters 8 and 9, respectively.

In the Appendices, we present the most important vector similarity measures and weighting schemes in detail, together with their formula. We also publish our novel general test datasets for Hungarian here, which, given the lack of previous such datasets, made the evaluation and comparison of our semantic models for Hungarian possible.

3 Main contributions

We have presented a very detailed and systematic analysis of the possible parameters used during the creation and comparison of feature vectors in distributional semantic models, for English, Spanish and Hungarian, filling a serious research gap. We have identified 10 important parameters of count-vector-based models and 4 relevant ones in case of using semantic vectors as input, and tested numerous settings for all of them. The main contributions of our work are as follows:

- Our analysis included novel parameters and novel parameter settings, and tested all parameters simultaneously, thus also taking their interaction into account. To our best knowledge, we are the first to do such a detailed analysis for these parameters, and also to do such an extensive comparison of them across multiple languages.
- Our novel two-step heuristic approach made the search for the best configurations among the numberless possibilities feasible for all three languages, and thus we were able to find such novel ones, many of them also incorporating novel parameter settings, that significantly outperformed conventional configurations.
- Although we had to use a heuristic approach due to the vast size of the search space, we have been able to verify the validity of this approach and the reliability and soundness of its results.
- While we have found that different configurations are best in case of models with count-vector-based, predictive and knowledge-graph-based semantic vectors as input, we have verified that a configuration performing well on given input data also works well on other input data of the same type.
- In accordance with our intuition, there were several parameters that worked very similarly in case of all three languages. We also found such parameters that were alike for Spanish and Hungarian, and different for English, which we also anticipated. However, it was interesting to see that there was such a parameter that worked similarly for English and Hungarian, but not for Spanish, and that we did not find any parameters that worked similarly for the two Indo-European languages, but differently for Hungarian.
- Although we have found that the very best results are produced by differ-

ent configurations for the different languages, our cross-language tests showed that all of them work rather well for all languages. Based on this we think that we could find such configurations that are rather language-independent, and give robust and reliable results.

- To be able to compare our results with the previous state-of-the-art, we have run such tests where the same data was used as input for both the previous state-of-the-art configurations and our configurations. In case of using raw counts as input and thus being able to optimize all 10 of our examined parameters, our best configurations contained novel parameter settings and clearly outperformed previous state-of-the-art configurations, with a considerable margin in most cases. When using semantic vectors as input and thus only being able to optimize 4 out of 10 parameters, our best configurations, also incorporating novel parameter settings, performed at least as well as the previous state-of-the-art, with a slight superiority in a couple of cases. All in all, our best model actually achieved absolute state-of-the-art results compared to all previous models of any type on the most important test datasets. Based on these results we think that our analysis was successful, and we were able to present such new parameter settings and new configurations that are superior to the previous state-of-the-art.

As it could be seen, the size of the input corpus, as well as the used information extraction method greatly influences the results. Therefore we think that doing an analysis similar to our current one for the information extraction phase of DSMs would be a principal direction for future research. Further, in our opinion it would be important to test our proposed new configurations using corpora magnitudes larger than that we could use. It would be even better if our whole heuristic analysis could also be repeated on these huge corpora. Further, although our results seem rather robust and reliable for Spanish and Hungarian too, it would be interesting to redo our analysis on larger and more reliable Spanish and Hungarian datasets, when such datasets will become available in the future.

We think that with this study we significantly contributed to the better understanding of the working and properties of DSMs. Although fully reliable conclusions from our results can only be drawn with respect to DSMs, we think that similar conclusions would hold for other NLP and non-NLP systems based on vector space models too.

For reproducibility and transparency, we have made our code and our most important resources publicly available at: https://github.com/doboandras/dsm-parameter-analysis/.

References

Dobó, A. 2018. Multi-D Kneser-Ney Smoothing Preserving the Original Marginal Distributions. *Research in Computing Science*, 147(6):11–25.

Dobó, A. 2019. *A comprehensive analysis of the parameters in the creation and comparison of feature vectors in distributional semantic models for multiple languages.* Ph.D. thesis, University of Szeged.

Dobó, A. and J. Csirik. 2012. Magyar és angol szavak szemantikai hasonlóságának automatikus kiszámítása. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 213–224, Szeged, Hungary.

Dobó, A. and J. Csirik. 2013. Computing semantic similarity using large static corpora. In *39th International Conference on Current Trends in Theory and Practice of Computer Science*, pages 491–502, Špindlerův Mlýn, Czech Republic.

Dobó, A. and J. Csirik. 2019a. A comprehensive study of the parameters in the creation and comparison of feature vectors in distributional semantic models. *Journal of Quantitative Linguistics.*

Dobó, A. and J. Csirik. 2019b. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 487–499, Hersonissos, Greece.

Dobó, A. and S. G. Pulman. 2011. Interpreting noun compounds using paraphrases. *Procesamiento del Lenguaje Natural*, 46:59–66.