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

András Dobó

Supervisor
János Csirik, DSc

Doctoral School of Computer Science
Faculty of Science and Informatics
University of Szeged

Summary of PhD Thesis

Szeged
2019

Introduction

For many natural language processing (NLP) problems, including information retrieval noun compound interpretation (Dobó and Pulman, 2011), spelling correction (Budanitsky and Hirst, 2001) and (Hliaoutakis et al., 2006) among many others, it is crucial to determine the semantic similarity or semantic relatedness of words. While relatedness takes a wide range of relations between words (including similarity) into account, similarity only considers how much the concepts denoted by the words are truly alike. Thus similarity entices relatedness, but not vice versa. For example, the words "bicycle" and "motorbike" are similar, as both denote 2-wheeled vehicles, and thus they are also related. On the other hand, the words "postman" and "mail" are highly related, as usually mails are delivered by postmen, and yet they are not similar, as they denote rather different concepts. Further, the words "furnace" and "voyage" are neither similar nor related.

Motivation

Most models are based on the distributional hypothesis of meaning (Harris, 1954), and thus calculate this similarity or relatedness using distributional data extracted from large corpora. These models can be collectively called as distributional semantic models (DSMs) (Baroni and Lenci, 2010; Baroni et al., 2014). In these models first possible features are identified, usually in the form of context words, and then a weight is assigned for each word-feature pair using complex methods, thus creating feature vectors for all words. The similarity or relatedness of words are then calculated by comparing their feature vectors 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, as also suggested by Levy et al. (2015).

Most papers presenting DSMs focus on only one or two aspects of the problem, and take all the other parameters as granted with some standard setting. For example, the majority of studies simply use cosine as vector similarity measure (e.g. Bruni et al., 2013; Baroni et al., 2014; Speer et al., 2017; Salle et al., 2018) and/or (positive) pointwise mutual information as weighting scheme (e.g. Islam and Inkpen, 2008; Hill et al., 2014; Salle et al., 2018) out of convention. And even in case of the considered parameters, usually only a handful of possible settings are tested for. Further, 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 with multiple possible settings, such as Lapesa and Evert (2014) and Kiela and Clark (2014), but even these are far from truly comprehensive, and do not fully test for the interaction between the different parameters. So, although an extensive analysis of the possible parameters and their combinations would be crucial, as also suggested by (Levy et al., 2015), there has been no research to date that would have evaluated these truly comprehensively. Moreover, despite the fact that the best parameter 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. In this thesis we would like to address these gaps.

Aims and objectives

DSMs have two distinct phases in general. In the first phase statistical information (e.g. raw counts) is extracted from raw data (e.g. a large corpus of raw text), in the form of statistical distributional data. In the second phase, feature vectors are created from the extracted information for each word and these vectors are then compared to each other to calculate the similarity or relatedness of words. 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 between the different parameters.

We have chosen to only study the second phase of the DSMs, as the two phases are relatively distinct and independent from each other, and testing for every single possible combination of the 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 also trying to test for the parameters of the first phase (e.g. source corpus, context type (window-based or dependency-based) and context size) simultaneously would be unreasonable and unmanageable, and is out of scope of this study. Therefore we have omitted the examination of this phase completely, with one exception to this.

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) (Baroni et al., 2014). In order to get a more complete view and due to the huge popularity of predictive models in recent years, in addition to using information extracted from a corpus using a count-vector-based model, we have also done some experiments with information extracted by a predictive model in case of English. Further, later on we also extended our analysis with a model based on semantic vectors constructed from 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 case only a part of the considered parameters could be tested for due to the characteristics of such models. That is part of the reason why we have focused on count-vector-based DSMs more.

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 case of such input 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 (configuration) achieving the highest performance on standard test datasets. We have done our extensive analysis for English, Spanish and Hungarian separately, and then we have compared our findings for the different languages.

For some of the tested parameters a large number of possible settings were tested, more than a thousand in some cases, resulting in trillions of possible combinations altogether. While of course also testing the conventionally used parameter settings, we also proposed numerous new variants in case of some parameters. Further, we have tested a vast number of novel configurations, with some of these new configurations 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 the same 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).

Conclusions

In this thesis 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. Our analysis included novel parameters and novel parameter settings, and tested all parameters simultaneously, thus also taking the possible interaction between the different parameters 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.

With our two-step heuristic approach we were searching for the best configurations for all three languages, and were able to find such novel ones, many of them also incorporating novel parameter settings, that significantly outperformed conventional configurations. Although we have used a heuristic approach for the search due to the vast number of possible configurations, we have been able to verify the validity of this approach and the reliability and soundness of its results. Further, we have also 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 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 different 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. Actually, our best model 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 systems based on vector space models too. So in our view our results could also be useful (with some reservations) outside the scope of DSMs, in case of other NLP and non-NLP problems using vector space models too.

The main results achieved in this thesis are summarized in the next sections, and the papers corresponding to these results are also listed at the end of each section.

Thesis 1: Analysis of the parameters in the creation and comparison of feature vectors in distributional semantic models

During our research we have devised a heuristic approach for the analysis of the parameters in the creation and comparison of feature vectors in distributional semantic models. For this, first we have implemented a bag-of-words information extraction method, which was used to create the raw input data for our heuristic approach. Then we have examined the possible parameters and their potential settings, and done a selection of these for our analysis. Beside experimenting with the standard, conventional settings for the parameters, we have also designed numerous new variants, especially in case of vector similarity measures and weighing schemes. In the thesis, we have given a very detailed overview of both the used heuristic approach, as well as the used parameters and their tested settings, including the detailed formula for many measures. Thus, this work can also serve as a useful reference for future studies that would like to analyse the parameters of distributional semantic models, or just simply use these parameters with some specific settings.

Main results and contributions:

- Devising a heuristic approach that made the analysis of the numberless configurations in the creation and comparison of feature vectors in DSMs possible;
- Construction of numerous novel vector similarity measures;
- Proposal of numerous new weighting schemes;
- Designing several novel settings for the other parameters;
- The corresponding publications are: (Dobó and Csirik, 2012, 2013; Dobó, 2018; Dobó and Csirik, 2019a,b)

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

  András Dobó and János Csirik. Computing semantic similarity using large static corpora. In *SOFSEM 2013: Theory and Practice of Computer Science. LNCS 7741*, pages 491–502, 2013

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

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

András Dobó and János Csirik. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT (in press)*, 2019b

Thesis 2: Semantic similarity of English words

First we have tried our heuristic analysis for the English language, for which we have included our detailed results in the thesis. To prove the validity of our approach, we have done multiple experiments using different count-vector-based, predictive and knowledge-graph-based data as input. As a result of these, we were able to determine the best configuration for all three different types of input. Further, we have done an extensive comparison of our best configurations with conventional and state-of-the are configurations.

Main results and contributions:

- Determining the best configuration for English count-vector-based, predictive and knowledge-graph-based models;
- Demonstration of the validity and reliability of our heuristic approach and the robustness and reliability of the results obtained by it;
- Main conclusions:
  - The settings of the different parameters are dependent on each other, and instead of inspecting the parameters separately they need to be analysed together, also considering the interactions between them;
  - Different parameter settings have to be used in case of count-vector-based, predictive and knowledge-graph-based models;
  - A configuration working well using a given count-vector-based or predictive input also works well using other input of the same type;
  - We could outperform previous state-of-the-art results when using raw counts as input and thus all 10 parameters could be optimized;
  - We were able to find such configurations that perform at least as well, with a slight superiority in a couple of cases, as previous state-of-the-art models, when using semantic vectors as input and thus only 4 out of 10 parameters could be optimized;
  - Our best model, BestSv2UsingSv, based on semantic vectors constructed from a knowledge graph, achieves absolute state-of-the-art results compared to all previous models of any type on the most important test datasets.
- The corresponding publications are: (Dobó and Csirik, 2012, 2013, 2019a,b)

  András Dobó and János Csirik. Magyar és angol szavak szemantikai hasonlóságának au-

tomatikus kiszámítása. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 213–224, 2012

András Dobó and János Csirik. Computing semantic similarity using large static corpora. In *SOFSEM 2013: Theory and Practice of Computer Science. LNCS 7741*, pages 491–502, 2013

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

András Dobó and János Csirik. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT (in press)*, 2019b

Thesis 3: Comparison of the finding of our analysis for English, Spanish and Hungarian

After our analysis for English, we have done the same extensive analysis for Spanish and Hungarian too. To be able to evaluate our results for Hungarian, we had to create novel test datasets for the evaluation of semantic models for Hungarian due to the lack of such datasets. After we had all the necessary datasets, we have determined the best configuration for Spanish and Hungarian count-vector-based models using our heuristic approach. Further, we have compared our results both with respect to the first and the second phase of the analysis for the different languages, and have drawn conclusions based on them.

Main results and contributions:

- Creation of new general test datasets for the evaluation of semantic models for Hungarian, published in the Appendix of the thesis;
- Determining the best configuration for Spanish and Hungarian count-vector-based models;
- Main conclusions:
  - Based on the tests and verification steps our heuristic approach was successful for Spanish and Hungarian too, and it gives robust and reliable results;
  - Different configuration produce the best results for the different languages, but all of them are rather language-independent and perform well for the other languages too;
  - As anticipated, there were parameters where the results for Spanish and Hungarian were similar, but different for English;
  - Contrary to our intuition we did not find any parameters that were alike for English and Spanish, but different for Hungarian;
  - To our surprise we also found such a parameter, where the results were similar for English and Hungarian, but different for Spanish.
- The corresponding publications are: (Dobó and Csirik, 2012, 2013, 2019a,b)

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

  András Dobó and János Csirik. Computing semantic similarity using large static corpora.

In *SOFSEM 2013: Theory and Practice of Computer Science. LNCS 7741*, pages 491–502, 2013

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

András Dobó and János Csirik. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT (in press)*, 2019b

References

Marco Baroni and Alessandro Lenci. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd ACL*, pages 238–247, 2014.

Elia Bruni, Nam Khan Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 48:1–47, 2013.

Alexander Budanitsky and Graeme Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources*, page 2, 2001.

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

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

András Dobó and János Csirik. Computing semantic similarity using large static corpora. In *SOFSEM 2013: Theory and Practice of Computer Science. LNCS 7741*, pages 491–502, 2013.

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

András Dobó and János Csirik. Comparison of the best parameter settings in the creation and comparison of feature vectors in distributional semantic models across multiple languages. In *15th International Conference on Artificial Intelligence Applications and Innovations. IFIP AICT (in press)*, 2019b.

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

Zellig S. Harris. Distributional Structure. *WORD*, 10(2-3):146–162, 1954.

Felix Hill, Roi Reichart, and Anna Korhonen. Multi-Modal Models for Concrete and Abstract Concept Meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296, 2014.

Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides G M Petrakis, and Evangelos Milios. Information retrieval by semantic similarity. *International journal on semantic Web and information systems*, 2(3):55–73, 2006.

Aminul Islam and Diana Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2):1–25, 2008.

Douwe Kiela and Stephen Clark. A systematic study of semantic vector space model parameters. In *2nd CVSC at EACL*, pages 21–30, 2014.

Gabriella Lapesa and Stefan Evert. A Large Scale Evaluation of Distributional Semantic Models: Parameters , Interactions and Model Selection. *Transactions of the Association for Computational Linguistics*, 2:531–545, 2014.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the ACL*, 3:211–225, 2015.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. LexVec, 2018. URL `https://github.com/alexandres/lexvec/blob/master/README.md`. [Accessed 01.04.2019].

Robert Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *31st AAAI*, pages 4444–4451, 2017.