Corpus specificity in LSA and word2vec: the role of out-of-domain documents

Edgar Altszyler
UBA, FCEyN, DC.
ICC, UBA-CONICET
ealtszyler@dc.uba.ar

Mariano Sigman
U. Torcuato Di Tella - CONICET.
msigman@utdt.edu

Diego Fernández Slezak
UBA, FCEyN, DC,
ICC, UBA-CONICET
dfslezak@dc.uba.ar

Abstract

Despite the popularity of word embeddings, the precise way by which they acquire semantic relations between words remain unclear. In the present article, we investigate whether LSA and word2vec capacity to identify relevant semantic relations increases with corpus size. One intuitive hypothesis is that the capacity to identify relevant associations should increase as the amount of data increases. However, if corpus size grows in topics which are not specific to the domain of interest, signal to noise ratio may weaken. Here we investigate the effect of corpus specificity and size in word-embeddings, and for this, we study two ways for progressive elimination of documents: the elimination of random documents vs. the elimination of documents unrelated to a specific task. We show that word2vec can take advantage of all the documents, obtaining its best performance when it is trained with the whole corpus. On the contrary, the specialization (removal of out-of-domain documents) of the training corpus, accompanied by a decrease of dimensionality, can increase LSA word-representation quality while speeding up the processing time. From a cognitive-modeling point of view, we point out that LSA's word-knowledge acquisitions may not be efficiently exploiting higher-order co-occurrences and global relations, whereas word2vec does.

1 Introduction

The main idea behind corpus-based semantic representation is that words with similar meanings tend to occur in similar contexts (Harris, 1954). This proposition is called *distributional hypothesis* and provides a practical framework to understand and compute the semantic relationship between words. Based in the *distributional hypothesis*, Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer and Dumais, 1997; Hu et al., 2007) and word2vec (Mikolov et al., 2013a,b), are one of the most important methods for word meaning representation, which describes each word in a vectorial space, where words with similar meanings are located close to each other.

Word embeddings have been applied in a wide variety of areas such as information retrieval (Deerwester et al., 1990), psychiatry (Altszyler et al., 2018; Carrillo et al., 2018), treatment optimization(Corcoran et al., 2018), literature (Diuk et al., 2012) and cognitive sciences (Landauer and Dumais, 1997; Denhière and Lemaire, 2004; Lemaire and Denhi, 2004; Diuk et al., 2012).

LSA takes as input a training Corpus formed by a collection of documents. Then a word by document co-occurrence matrix is constructed, which contains the distribution of occurrence of the different words along the documents. Then, usually, a mathematical transformation is applied to reduce the weight of uninformative high-frequency words in the words-documents matrix (Dumais, 1991). Finally, a linear dimensionality reduction is implemented by a truncated *Singular Value Decomposition*, SVD, which projects every word in a subspace of a predefined number of dimensions, k. The success of LSA in capturing the latent meaning of words comes from this low-dimensional mapping. This representation improvement can be explained as a consequence of the elimination of the noisiest dimensions (Turney and Pantel, 2010).

Word2vec consists of two neural network models, Continuous Bag of Words (CBOW) and Skip-gram. To train the models, a sliding window is

Proceedings of the 3rd Workshop on Representation Learning for NLP, pages 1–10
Melbourne, Australia, July 20, 2018. ©2018 Association for Computational Linguistics

moved along the corpus. In the CBOW scheme, in each step, the neural network is trained to predict the center word (the word in the center of the window based) given the context words (the other words in the window). While in the skip-gram scheme, the model is trained to predict the context words based on the central word. In the present paper, we use the skip-gram, which has produced better performance in Mikolov et al. (2013b).

Despite the development of new word representation methods, LSA is still intensively used and has been shown that produce better performances than word2vec methods in small to medium size training corpus (Altszyler et al., 2017).

1.1 Training Corpus Size and Specificity in Word-embeddings

Over the last years, great effort has been devoted to understanding how to choose the right parameter settings for different tasks (Quesada, 2011; Dumais, 2003; Landauer and Dumais, 1997; Lapesa and Evert, 2014; Bradford, 2008; Nakov et al., 2003; Baroni et al., 2014). However, considerably lesser attention has been given to study how different corpus used as input for training may affect the performance. Here we ask a simple question on the property of the corpus: is there a monotonic relation between corpus size and the performance? More precisely, what happens if the topic of additional documents differs from the topics in the specific task? Previous studies have surprisingly shown some contradictory results on this simple question.

On the one hand, in the foundational work, Landauer and Dumais (1997) compare the word-knowledge acquisition between LSA and that of children's. This acquisition process may be produced by 1) direct learning, enhancing the incorporation of new words by reading texts that explicitly contain them; or 2) indirect learning, enhancing the incorporation of new words by reading texts that do not contain them. To do that, they evaluate LSA semantic representation trained with different size corpus in multiple-choice synonym questions extracted from the TOEFL exam. This test consists of 80 multiple-choice questions, in which its requested to identify the synonym of a word between 4 options. In order to train the LSA, Landauer and Dumais used the TASA corpus (Zeno et al., 1995).

Landauer and Dumais (1997) randomly replaced exam-words in the corpus with non-sense words and varied the number of corpus' documents selecting nested sub-samples of the total corpus. They concluded that LSA improves its performance on the exam both when training with documents with exam-words and without them. However, as could be expected, they observed a greater effect when training with exam-words. It is worth mentioning that the replacement of exam-words with non-sense words may create incorrect documents, thus, making the algorithm acquire word-knowledge from documents which should have an exam-word but do not. In the Results section, we will study this indirect word acquisition in the TOEFL test without using non-sense words.

Along the same line, Lemaire and Denhiere (2006) studied the effect of high-order co-occurrences in LSA semantic similarity, which goes further in the study of Landauer's indirect word acquisition.

In their work, Lemaire and Denhiere (2006) measure how the similarity between 28 pairs of words (such as bee/honey and buy/shop) changes when a 400-dimensions LSA is trained with a growing number of paragraphs. Furthermore, they identify for this task the marginal contribution of the first, second and third order of co-occurrence as the number of paragraphs is increased. In this experiment, they found that not only does the first order of co-occurrence contribute to the semantic closeness of the word pairs, but also the second and the third order promote an increment on pairs similarity. It is worth noting that Landauer's indirect word acquisition can be understood in terms of paragraphs without either of the words in a pair, and containing a third or more order co-occurrence link.

So, the conclusion from Lemaire and Denhiere (2006) and Landauer and Dumais (1997) studies suggest that increasing corpus size results in a gain, even if this increase is in topics which are unrelated for the relevant semantic directions which are pertinent for the task.

However, a different conclusion seems to result from other sets of studies. Stone et al. (2006) have studied the effect of Corpus size and specificity in a document similarity rating task. They found that training LSA with smaller subcorpus selected for the specific task domain maintains or even improves LSA performance. This corresponds to the intuition of noise filtering, when removing infor-

mation from irrelevant dimensions results in improvements of performance.

In addition, Olde et al. (2002) have studied the effect of selecting specific subcorpus in an automatic exam evaluation task. They created several subcorpus from a Physics corpus, progressively discarding documents unrelated to the specific questions. Their results showed small differences in the performance between the LSA trained with original corpus and the LSA trained with the more specific subcorpus.

It is well known that the number of LSA dimensions (k) is a key parameter to be duly adjusted in order to eliminate the noisiest dimensions (Landauer and Dumais, 1997; Turney and Pantel, 2010). Excessively high k values may not eliminate enough noisy dimensions, while excessively low k values may not have enough dimensions to generate a proper representation. In this context, we hypothesize that when out-of-domain documents are discarded, the number of dimensions needed to represent the data should be lower, thus, k must be decreased.

Regarding word2vec, Cardellino and Alemany (2017) and Dusserre and Padró (2017) have shown that word2vec trained with a specific corpus can produce better performance in semantic tasks than when it is trained with a bigger and general corpus. Despite these works point out the relevance of domain-specific corpora, they do not study the specificity in isolation, as they compare corpus from different sources.

In this article, we set to investigate the effect of the specificity and size of training corpus in word-embeddings, and how this interacts with the number of dimensions. To measure the semantic representations quality we have used two different tasks: the TOEFL exam, and a categorization test. The corpus evaluation method consists in the comparison between two ways of progressive elimination of documents: the elimination of random documents vs the elimination of out-of-domain documents (unrelated to the specific task). In addition, we have varied k within a wide range of values.

As we show, LSA's dimensionality plays a key role in the LSA representation when the corpus analysis is made. In particular, we observe that both, discarding out-of-domain documents and decreasing the number of dimensions produces an increase in the algorithm performance. In one of the two tasks, discarding out-of-domain documents without the decrease of k results in the complete opposite behavior, showing a strong performance reduction. On the other hand, word2vec shows in all cases a performance reduction when discarding out-of-domain, which suggests an exploitation of higher-order word co-occurrences.

Our contribution in understanding the effect of out-of-domain documents in word-embeddings knowledge acquisitions is valuable from two different perspectives:

- From an operational point of view: we show that LSA's performance can be enhanced when: (1) its training corpus is *cleaned* from out-of-domain documents, and (2) a reduction of LSA's dimensions number is applied. Furthermore, the reduction of both the corpus size and the number of dimensions tend to speed up the processing time. On the other hand, word2vec can take advantage of all the documents, obtaining its best performance when it is trained with the whole corpus.

- From a cognitive modeling point of view: we point out that LSA's word-knowledge acquisition does not take advantage of indirect learning, while word2vec does. This throws light upon models capabilities and limitations in modeling human cognitive tasks, such as: human word-learning (Landauer and Dumais, 1997; Lemaire and Denhiere, 2006; Landauer, 2007), semantic memory (Denhière and Lemaire, 2004; Kintsch and Mangalath, 2011; Landauer, 2007) and words classification (Laham, 1997).

2 Methods

We used TASA corpus (Zeno et al., 1995) in all experiments. TASA is a commonly used linguistic corpus consisting of more than 37 thousand educational texts from USA K12 curriculum. We word-tokenized each document, discarding punctuation marks, numbers, and symbols. Then, we transformed each word to lowercase and eliminated stopwords, using the stoplist in NLTK Python package (Bird et al., 2009). TASA corpus contains more than 5 million words in its cleaned version.

In each experiment, the training corpus size was changed by discarding documents in two different ways:

- *Random documents discarding:* The desired number of documents (n) contained in the

subcorpus is preselected. Then, documents are randomly eliminated from the original corpus until there are exactly *n* documents. If any of the test words (i.e. words that appear in the specific task) does not appear at least once in the remaining corpus, one document is randomly replaced with one of the discarded documents that contains the missing word.

- *Out-of-domain documents discarding:* The desired number of documents (*n*) contained in the subcorpus is preselected. Then, only documents with no test words are eliminated from the original corpus until there are exactly *n* documents. Here, *n* must be greater than or equal to the number of documents that contain at least one of the test words.

Both, LSA and Skip-gram word-embeddings were generated with Gensim Python library (Řehůřek and Sojka, 2010). In LSA implementation, a Log-Entropy transformation was applied before the truncated Singular Value Decomposition. In Skip-gram implementation, we discarded tokens with frequency higher than 10^{-3}, and we set the window size and negative sampling parameters to 15 (which were found to be maximal in two semantic tasks over TASA corpus (Altszyler et al., 2017)). In all cases, word-embeddings dimensions values were varied to study its dependency.

The semantic similarity (S) of two words was calculated using the cosine similarity measure between their respective vectorial representation ($\mathbf{v_1}$,$\mathbf{v_2}$),

$$S(\mathbf{v_1}, \mathbf{v_2}) = cos(\mathbf{v_1}, \mathbf{v_2}) = \frac{\mathbf{v_1}.\mathbf{v_2}}{\|\mathbf{v_1}\|.\|\mathbf{v_2}\|} \quad (1)$$

The semantic distances between two words $d(\mathbf{v_1}, \mathbf{v_2})$ is calculated as 1 minus the semantic similarity ($d(\mathbf{v_1}, \mathbf{v_2}) = 1 - S(\mathbf{v_1}, \mathbf{v_2})$).

Word-embeddings knowledge acquisition was tested in two different tasks: a semantic categorization test and the TOEFL test.

2.1 Semantic categorization test

In this test we measured the capabilities of the model to represent the semantic categories used by Patel et al. (1997) (such as drinks, countries, tools and clothes). The test is composed of 53 categories with 10 words each. In order to measure how well the word i is grouped vis-à-vis the other words in its semantic category we used the Silhouette Coefficients, $s(i)$ (Rousseeuw, 1987),

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (2)$$

where $a(i)$ is the mean distance of word i with all other words within the same category, and $b(i)$ is the minimum mean distance of word i to any words within another category (i.e. the mean distance to the neighboring category). In other words, Silhouette Coefficients measure how close is a word to its own category words compared to the closeness to neighboring words. The Silhouette Score is computed as the mean value of all Silhouette Coefficients. The score takes values between -1 and 1, higher values reporting localized categories with larger distances between categories, representing better clustering.

The high number of test words (530) and the high frequency of some of them leaves only a few documents with no test words. This makes varied corpus size range in the *out-of-domain documents discarding* very small. To avoid this, we tested only on the 10 least frequent categories. The frequency of a question is measured as the number of documents in which at least one word from this category appears.

2.2 TOEFL test

The TOEFL test was introduced by Landauer and Dumais (1997) to evaluate the quality of semantic representations. This test consists of 80 multiple-choice questions, in which it is requested to identify the synonym of a target word between 4 options. For example: *select the most semantically similar to "enormously" between this words: "tremendously", "appropriately", "uniquely" and "decidedly"*. The performance of this test was measured by the percentage of correct responses.

Again, The high number of test words (400) and the high frequency of some of them leaves few documents with no test words. So we performed the test only on the 20 least frequent questions in order to have out-of-domain documents to discard.

3 Results

3.1 Semantic categorization Test

In Figure 1 we show the LSA (top panel) and word2vec (bottom panel) categorization performance with both documents discarding methods.

For each corpus size and document discarding method, we took 10 subcorpus samples (in total we consider 90 subcorpus + the complete corpus). In each corpus/subcorpus, we trained LSA and word2vec with a wide range of dimension values, using in each case the dimension that produces the best mean performance.

In both cases, performance decreases when documents are randomly discarded (orange dashed lines). However, LSA and word2vec have different behavior in the out-of-domain document discarding method (blue solid lines). While LSA produces better scores with increasing specificity, the word2vec performance decreases in the same situation.

LSA's maximum performance is obtained using 20 dimensions and removing all out-of-domain documents in the training corpus. While, when all the corpus is used the best number of dimensions is 100. These results show that performance for a specific task may be increased by "cleaning" the training corpus of out-of-domain documents. But, in order to enhance the performance, the elimination of out-of-domain documents should be accompanied by a decrease of the number of LSA dimensions. For example, fixing the number of dimensions to 100 the performance result in a reduction of 55%. We also point out that this technical subtlety has not been taken into account in previous results that reported the presence of indirect learning in LSA (Landauer and Dumais, 1997; Lemaire and Denhiere, 2006).

Figure 2 shows the results disaggregated by number of dimensions. It can be seen that in all cases the performance decreases when documents are randomly discarded (bottom panels). However, in the case of LSA, the dependency with the number of out-of-domain documents varied with the number of dimensions (top left panel). In the cases of 300, 500 and 1000 dimensions, the performance decreases when out-of-domain documents are eliminated. In contrast, we obtain the opposite behavior in the cases of 5, 10, 20 dimensions, in which the elimination of out-of-domain documents increases LSA's categorization performance.

Consider the case when k is fixed in the value that maximizes the performance with the entire corpus (around $k = 100$). When the corpus is "cleaned" of out-of-domain documents, the remaining corpus will have not only fewer documents, but also less topic diversity among texts. Thus, the number of dimensions (k) needed to generate a proper semantic representation should be reduced. As k is fixed in high values, LSA may not eliminate enough noisy dimensions, leading to a decrease in the performance. This effect becomes

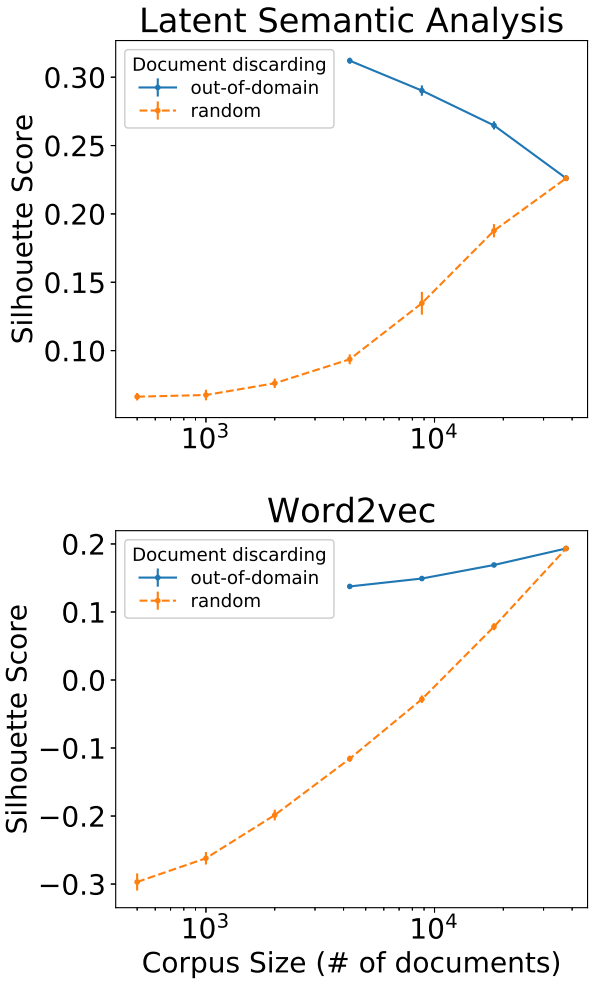

Figure 1: Semantic categorization test analysis. Silhouette Score vs corpus size for with both documents discarding methods: *random document discarding* (orange dashed lines) and *out-of-domain documents discarding* (blue solid lines). The shown Silhouette Score values and their error bars are, respectively, the mean values and the standard error of the mean of 10 samples. In most of the dots, the error bars are not visible, this is because their length is smaller than the dot size. The dimension was varied among {5, 10, 20, 50, 100, 300, 500, 1000} for LSA and among {5, 10, 20, 50, 100, 300, 500} for word2vec. Due to the high computational effort, in the case of word2vec we avoid using 1000 dimensions.

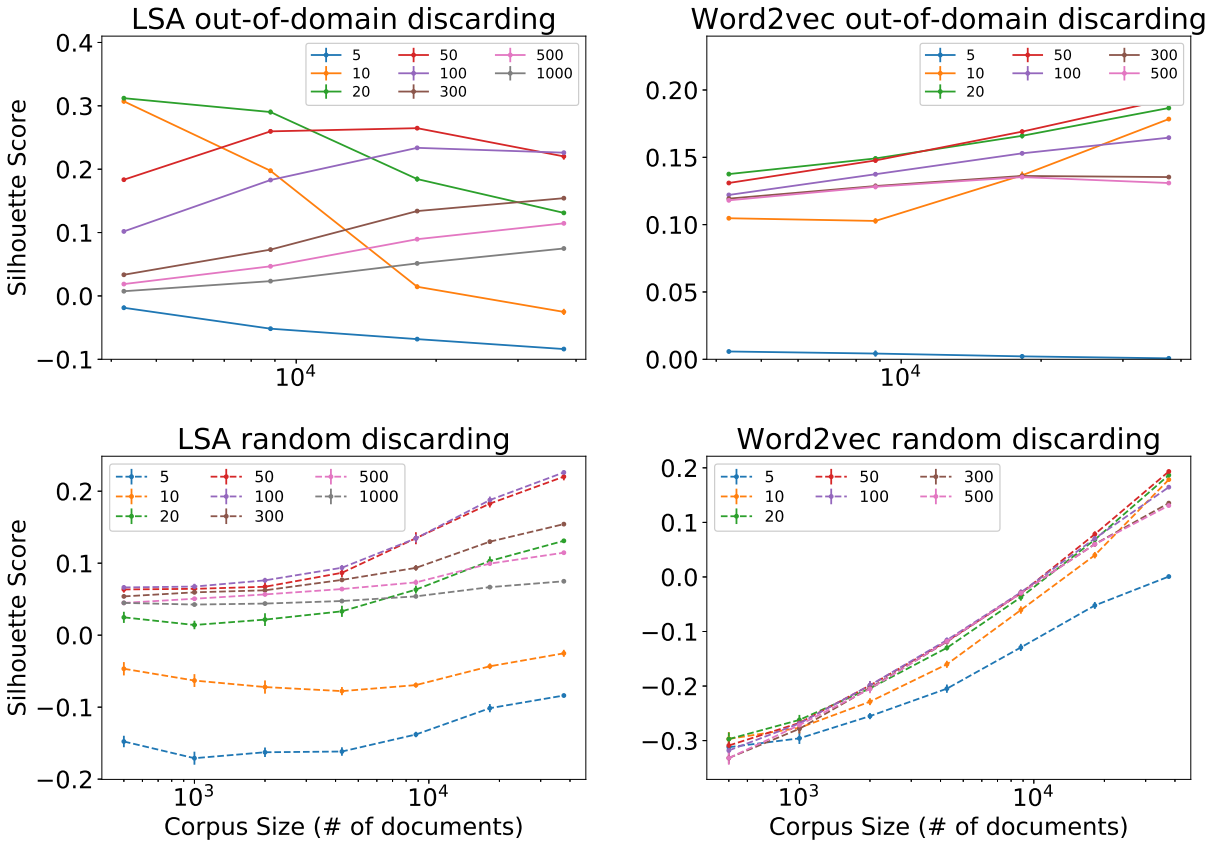

Figure 2: Semantic categorization test analysis disaggregated by number of dimensions. Categorization performance (Silhouette Score) vs corpus size, by number of dimensions. Both document variation methods are shown: *out-of-domain documents discarding* (top panels) and *random document discarding* (bottom panels). The shown scores values and their error bars are, respectively, the mean values and the standard error of the mean of 10 samples.

larger when the selected k is high, as it can be seen for $k = 300, 500, 1000$. On the other hand, consider the case when k is fixed in the value that maximizes the performance with the "cleaned" corpus (around $k = 20$). The presence of out-of-domain documents in the complete corpus increase the topic diversity. As k is fixed in low values, the LSA will not have enough dimensions to represent all the intrinsic complexity of the whole corpus. So, when the corpus is "cleaned" of out-of-domain documents, the performance should increase.

On the other hand, in the case of word2vec, the performance decrease, with almost all dimension values, when out-of-domain documents are eliminated. Moreover, the discarding of out-of-domain documents do not require a considerable decrease of the number of dimensions. These findings supports the idea that individual dimensions of word2vec do not encode *latent* semantic domains, however, more analysis must be done in these direction (see Baroni et al. (2014) discussion).

3.2 TOEFL Test

In Figure 3 we show the TOEFL correct answer fraction vs the corpus size. We varied the corpus size by both methods: the *out-of-domain documents discarding* and the *Random document discarding*. As in the categorization test procedure, a wide range of dimension values where tested, using in each case the dimension that produces the best mean performance.

In both models, performance decreases when documents are randomly discarded (orange dashed lines in figure 3). For LSA, the elimination of out-of-domain documents does not produce a significant performance variation, which shows that LSA can not take advantage of out-of-domain document. This results are in contradiction with Landauer and Dumais (1997) observation of indirect

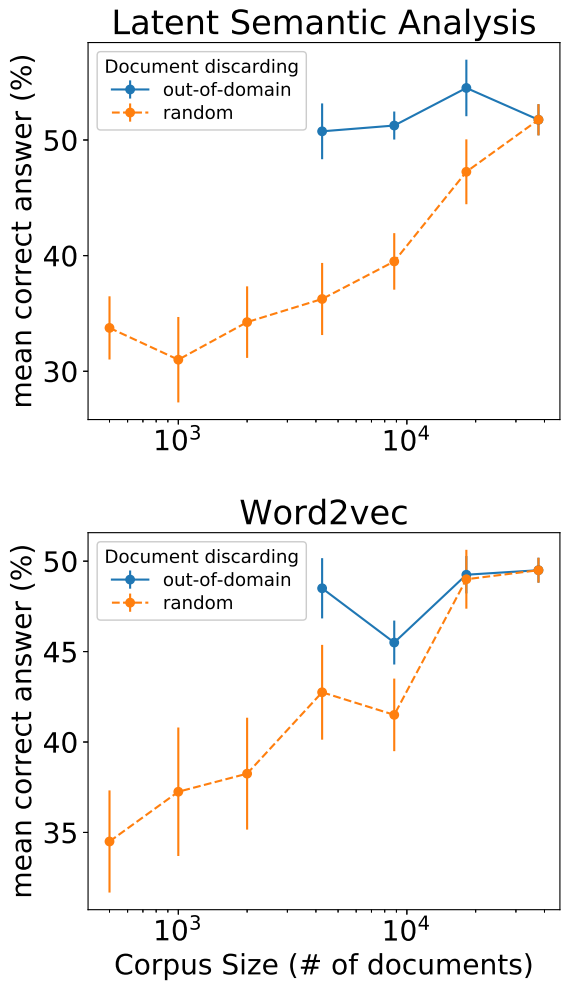

Figure 3: TOEFL test analysis. Correct answer percentage vs corpus size with both document variation methods: *Random document discarding* (orange dashed lines) and the *out-of-domain documents discarding* (blue solid lines). The shown Silhouette Score values and their error bars are,respectively, the mean values and the standard error of the mean of 10 samples. The dimension was varied among {5, 10, 20, 50, 100, 300, 500, 1000} for LSA and among {5, 10, 20, 50, 100, 300, 500} for word2vec. Due to the high computational effort, in the case of word2vec we avoid using 1000 dimensions.

learning. We believe that this difference is due to the lack of adjustment in the number of dimensions. On the other hand, word2vec has the same behaviour as in the categorization test. The performance when the out-of-domain documents are discarded show a small downward trend (not significant, with p-val=0.31 in a two-sided Kolmogorov-Smirnov test), but not as pronounced as in *random document discard* method. Unlike the categorization test, the performance measure in the TOEFL Test present a high variability (see Figure 4). This observation is consistent with the large fluctuations shown in Landauer and Dumais (1997). Despite this, we consider relevant to use this test in order to be able to compare with Landauer and Dumais (1997) results.

4 Conclusion and Discussion

Despite the popularity of word-embeddings in several semantic representation task, the way by which they acquire new semantic relations between words is unclear. In particular, for the case of LSA there are two opposite visions about the effect of incorporating out-of-domain documents. From one point of view, training LSA with a specific subcorpus, *cleaned* of documents unrelated to the specific task increases the performance (Stone et al., 2006). From the other point of view, the presence of unrelated documents improves the representations. The second view point is supported by the conception that the SVD in LSA can capture high-order co-occurrence words relations (Landauer and Dumais, 1997; Lemaire and Denhiere, 2006; Turney and Pantel, 2010). Based on this, LSA is used as a plausible model of human semantic memory given that it can capture indirect relations (high-order word co-occurrences).

In the present article we studied the effect of out-of-domain documents in LSA and word2vec semantic representations construction. We compared two ways of progressive elimination of documents: the elimination of random documents vs the elimination of out-of-domain documents. The semantic representations quality was measured in two different tasks: a semantic categorization test and a TOEFL exam. Additionally, we have varied a large range of word-embedding dimensions (k).

We have shown that word2vec can take advantage of all the documents, obtaining its best performance when it is trained with the whole corpus. On the contrary, LSA's word-representation quality increases with a specialization of the training corpus (removal of out-of-domain document) accompanied by a decrease of k. Furthermore, we have shown that the specialization without the decrease of k can produce a strong performance reduction. Thus, we point out the need to vary k when the corpus size dependency is studied. From a cognitive modeling point of view, we

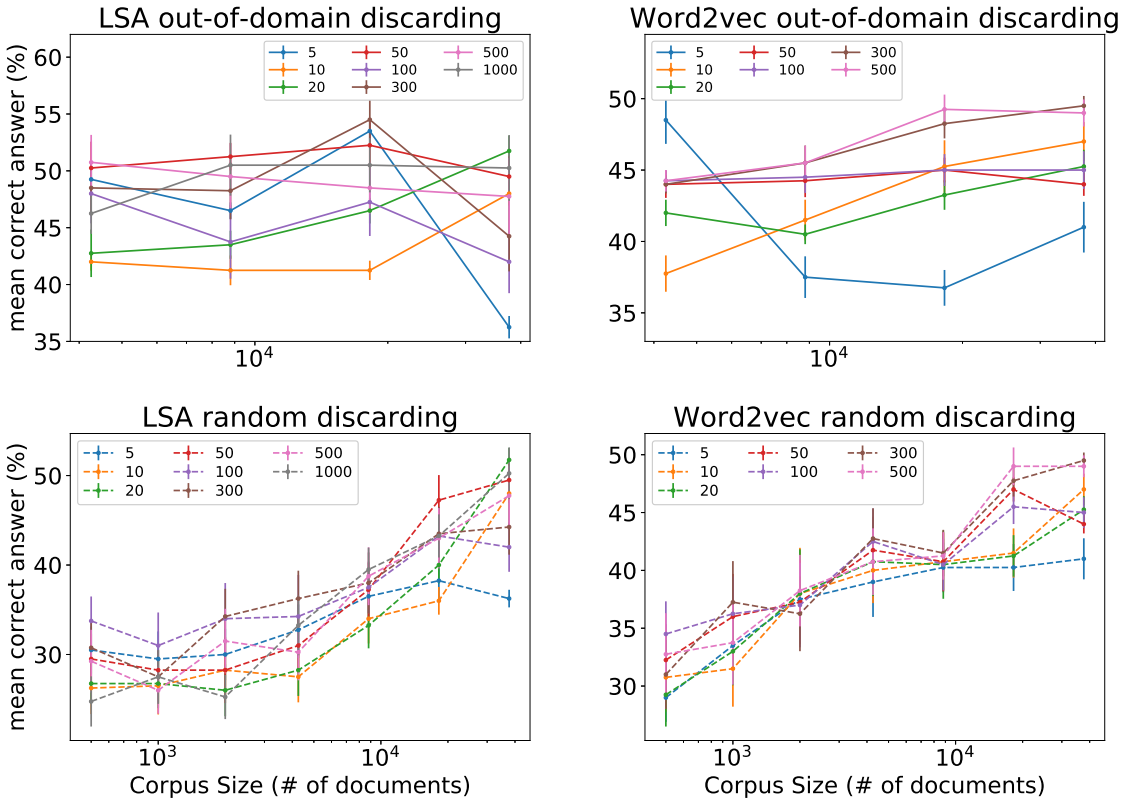

Figure 4: TOEFL test analysis disaggregated by number of dimensions. Correct answer percentage vs corpus size, by number of dimensions. Both document variation methods are shown: *out-of-domain documents discarding* (top panels) and *random document discarding* (bottom panels). The shown scores values and their error bars are, respectively, the mean values and the standard error of the mean of 10 samples.

point out that LSA's word-knowledge acquisitions does not take advantage of indirect learning (high-order word co-occurrences), while word2vec does. This throws light upon word-embeddings capabilities and limitations in modeling human cognitive tasks, such as: human word-learning (Landauer and Dumais, 1997; Lemaire and Denhiere, 2006; Landauer, 2007), semantic memory (Denhière and Lemaire, 2004; Kintsch and Mangalath, 2011; Landauer, 2007) and words classification (Laham, 1997).

Acknowledgments

This research was supported by Consejo Nacional de Investigaciones Cientficas y Tcni-cas (CONICET), Universidad de Buenos Aires, and Agencia Nacional de Promocin Cientfica y Tecnolgica. We also want to thank LSA and NLP Research Labs, University of Colorado at Boulder for shearing access to the TOEFL word set.

References

Edgar Altszyler, Ariel Berenstein, David Milne, Rafael A. Calvo, and Diego Fernandez Slezak. 2018. Using contextual information for automatic triage of posts in a peer-support forum. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Edgar Altszyler, Sidarta Ribeiro, Mariano Sigman, and Diego Fernández Slezak. 2017. The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and Cognition* https://doi.org/10.1016/j.concog.2017.09.004.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*. pages 238–247.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. " O'Reilly Media, Inc.".

Roger B. Bradford. 2008. An empirical study of required dimensionality for large-scale latent semantic indexing applications. *Proceedings of the 17th ACM, CIKM* pages 153–162. https://doi.org/10.1145/1458082.1458105.

Cristian Cardellino and Laura Alonso Alemany. 2017. Disjoint semi-supervised spanish verb sense disambiguation using word embeddings. In *ASAI, Simposio Argentino de Inteligencia Artificial*.

Facundo Carrillo, Mariano Sigman, Diego Fernández Slezak, Philip Ashton, Lily Fitzgerald, Jack Stroud, David J Nutt, and Robin L Carhart-Harris. 2018. Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistant depression. *Journal of Affective Disorders* 230:84–86.

Cheryl M Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C Javitt, Carrie E Bearden, and Guillermo A Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* 17(1):67–75.

Scott Deerwester, Susan T Dumais, Thomas Landauer, George Furnas, and Richard. Harshman. 1990. Indexing by latent semantic analysis. *JAsIs* 41(6).

G Denhière and B Lemaire. 2004. A Computational Model of Children's Semantic Memory. *Proc 26 thAnnual Meeting of the Cognitive Science Society* pages 297–302.

Carlos G. Diuk, D. Fernandez Slezak, I. Raskovsky, M. Sigman, and G. a. Cecchi. 2012. A quantitative philology of introspection. *Frontiers in Integrative Neuroscience* 6(September):1–12. https://doi.org/10.3389/fnint.2012.00080.

Susan Dumais. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers* 23(2):229–236. https://doi.org/10.3758/BF03203370.

Susan Dumais. 2003. Data-driven approaches to information access. *Cognitive Science* 27(3):491 – 524. https://doi.org/http://dx.doi.org/10.1016/S0364-0213(03)00013-2.

Emmanuelle Dusserre and Muntsa Padró. 2017. Bigger does not mean better! we prefer specificity. In *IWCS 201712th International Conference on Computational SemanticsShort papers*.

Z. Harris. 1954. Word Distributional structure 23(10):146162.

X Hu, Z Cai, P Wiemer-Hastings, a Graesser, and D McNamara. 2007. Strengths, limitations, and extensions of LSA. *Handbook of Latent Semantic Analysis* pages 401–426. https://doi.org/10.1164/rccm.201012-2079ED.

Walter Kintsch and Praful Mangalath. 2011. The construction of meaning. *Topics in Cognitive Science* 3(2):346–370. https://doi.org/10.1111/j.1756-8765.2010.01107.x.

D Laham. 1997. Latent Semantic Analysis approaches to categorization. *Proceedings of the 19th annual conference of the Cognitive Science Society* (1984):979.

Thomas K Landauer. 2007. Lsa as a theory of meaning. *Handbook of latent semantic analysis* pages 3–34.

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):211–240. https://doi.org/10.1037/0033-295X.104.2.211.

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics* 2:531–545.

Benoît Lemaire and Guy Denhi. 2004. Incremental Construction of an Associative Network from a Corpus. *26th Annual Meeting of the Cognitive Science Society* pages 825–830.

Benoit Lemaire and Guy Denhiere. 2006. Effects of High-Order Co-occurrences on Word Semantic Similarity. *Current psychology letters* 1(18):1–12.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed Representations of Words and Phrases and their Compositionality. *Nips* pages 1–9. https://doi.org/10.1162/jmlr.2003.3.4-5.951.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013b. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)* pages 1–12. https://doi.org/10.1162/153244303322533223.

Preslav Nakov, Elena Valchanova, and Galia Angelova. 2003. Towards a Deeper Understanding of the LSA Performance. *Proceedings of Recent Advances in Natural Language Processing* 2(2):311–318.

Brent A Olde, Donald R Franceschetti, Ashish Karnavat, Arthur C Graesser, and Tutoring Research Group. 2002. The right stuff: do you need to sanitize your corpus when using latent semantic analysis? *Proceedings of the 24th annual meeting of the cognitive science society* .

Malti Patel, John A. Bullinaria, and Joseph P Levy. 1997. Extracting semantic representations from large text corpora. *Proceedings of the 4th Neural Computation and Psychology Workshop* pages 199–212.

J. Quesada. 2011. *Creating your own LSA space*, Erlbaum, chapter 1.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50. http://is.muni.cz/publication/884893/en.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Benjamin P Stone, Simon J Dennis, Peter J Kwantes, Drdc Toronto, and Sheppard Ave W. 2006. A Systematic Comparison of Semantic Models on Human Similarity Rating Data : The Effectiveness of Subspacing pages 1813–1818.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37:141–188. https://doi.org/10.1613/jair.2934.

S. Zeno, S. Ivens, and R.and Duvvuri R. Millard. 1995. *The educator's word frequency guide*. Brewster.