Bigger does not mean better!

We prefer specificity

Emmanuelle Dusserre, Muntsa Padró

Eloquant, Grenoble, France
{emmanuelle.dusserre, muntsa.padro}@eloquant.com

Abstract. This paper studies the applicability of word2vec to the task of extracting similar words from small, domain-specific data. Results show that, even though the general tendency of the community is to focus on using more and more data, the specificity of the corpus has much more influence on word2vec results than its size. Actually, when the goal is to automatically detect similar words that are domain specific, it is necessary to have a corpus that correctly represents the use of those specific words more than to have huge amounts of data unrelated to the targeted language.

Keywords: word2vec; semantic extraction; taxonomy; domain-specific data; customer relation management

1. Introduction

Dealing with the automatic extraction of related terms is a trending topic on Natural Language Processing (NLP) area. From synonym extraction, ontology creation or automatic gazetteer building, this is a challenging task approached by many in many publications and shared tasks.

This paper presents a set of experiments on finding similar words in very specific domains. This work is framed on a bigger project on performing classification of customer reviews for different companies in the Customer Relationship Management (CRM) domain. To enrich the classification system, a taxonomy that assigns a semantic tag to the terms that are relevant to the domain was developed. Currently, this is a manual work that is very time consuming, especially given that CRM domain is in fact a combination of sub-domains, or business sectors. This means that every time the data from a company operating in a new sector is to be treated, the taxonomy needs to be enriched to cover the terms specific to the new sector. Doing that manually is demanding in time and resources, and it is difficult to assure a good coverage, so the present study explores how to automatize this step. Thus, this paper proposes to use a small existing taxonomy developed by hand, and to automatically enrich it with terms that are semantically similar to the words already present in the taxonomy, we call them the seed words.

This paper presents an approach to extract related terms in domain-specific corpora by using distributional hypothesis [6], which permits to extract words which share similar contexts and consequently same senses. Specifically, word2vec ([13]) caught our attention because of its impressive performances in semantic extraction tasks in many works of NLP.

The key point of this study is the very reduced size of the domain-specific corpus. Even though it is a limitation for this kind of tool, these experiments show that size is not the only parameter that matters. Indeed, in the present case, we obtained better results with a small, specific corpus than with a huge, general domain amount of words.

2. Related work

The automatic extraction of similar terms is a task widely covered in the literature. Word-context matrices based on vector space models proved their efficiency. For instance, [10] proposed Latent Semantic Analysis (LSA), and obtained high results on the Test of English as a Foreign Language (TOEFL). This approach was used by many afterwards, such as [11] or [7].

Vector space models also contributed a lot to automatic thesaurus generation. The pioneer was probably [3] who generated automatically global thesauri by using a discrimination value model of [14] and the complete-link clustering algorithm. Also, we can cite [5], who introduced an automatic method to create a thesaurus from a raw corpus.

However, word-context matrices techniques are well-known to need a lot of data to obtain good results. Most of experiments have been realized on huge corpora surpassing billions of words. Nevertheless, in [5] the authors built a method using specific statistical analysis techniques for small datasets combining it with a system of semantic class constitution and topic detection. The goal of their study was to achieve automatically lexical semantic information on small corpora to help languages with few resources.

In the recent last years, the apparition of word2vec [13] permitted to create vectors using artificial neural networks. This allowed to fasten the system and to use more data. Many works have been conducted with word2vec with the intention of finding related words ([2, 12]).

Some works tried to compare traditional methods, such as LSA, with word2vec. For instance, [1] observed that word2vec is better than LSA on bigger corpora from a dream database. But, since they started to reduce corpora, LSA outperformed word2vec.

The work presented in this paper follows the previously mentioned lines on distributional hypothesis to extract related words. However, we do not know antecedent works using distributional hypothesis based on artificial neural networks on such small, domain-specific corpora. Indeed, one important factor in this study is the super-specificity of the data we dealt with.

3. Methodology

3.1. Finding most related words

As explained before, in order to improve document classification for customer reviews in different sectors, the classifier is enriched with semantic information. In previous work, a taxonomy has been developed to tag terms that are relevant to the CRM domain. For example, “operator” or “advisor” are tagged as *Interlocutor* or “phone” is tagged as *Product* in a telecommunications sector while “tire” has the same tag in a car-related activity. This semantic information is used as a feature in order to build a generic classification system based on the semantic tags more than on the lemmas.

Thereby, when the classifier needs to be adapted to a new sector, the relevant terms need to be added to the correct branch of the taxonomy, to be assigned to the correct semantic tag. To extract lists of similar words we use word2vec, and aim at placing similar words in the same branch of the taxonomy.

The basic idea is to compute vectors representing the context of the words in the corpus, and then compute the distance between each word and the seed words using the cosine of these vectors. Then, a threshold is applied and all words with a cosine above the threshold are considered as neighbors of the seed words, and thus, related words.

In this work, word2vec is ran to obtain the vectors that allow the computation of the closest neighbors of the seed words. To build the matrices and compare the results, the study is conducted on two corpora, presented in next section.

To evaluate performances of the method a gold standard was elaborated (section 3.3), and compared with the Random Indexing (RI) [8], a classical distributional algorithm.

3.2. Corpora

Two corpora are used in this work to build word embeddings using word2vec:

a. A CRM domain-specific corpus constituted with more than 35, 000 French customer reviews about a telecommunication company, amounting a total of 557, 676 words. Idiosyncrasies of this corpus are typical of CRM domain: texts are very short, one or two sentences per review, and the language used contains abbreviations and many spelling mistakes which add complexification to the treatment.

b. A word2vec model elaborated by J.P. Fauconnier[1] was used with the intention to compare domain specific and generic corpora on frWac corpus. It contains 1.6 billion of words crawled from the Web and POS-tagged and lemmatized with TreeTagger[2].

3.3. Gold standard

To evaluate the results of the experiments, a gold-standard was developed. It contains a set of seed words and their closest neighbors, and allows us to calculate the precision, the recall and the f-measure for each word2vec output.

To create the gold-standard, the first step was to pre-select set of nouns belonging to the taxonomy, this is the seed words. These nouns follow several criteria: (a) Belonging to the telecommunication sub-domain. (b) Having a frequency superior to five in the telecommunication corpus. (c) Select only one seed word for semantically close words. For instance, we choose *téléphone* and did not selection *mobile* because they can be synonyms.

The second step was to manually choose, with two linguists, the closest words to each seed word. These words are synonyms or also orthographical variations, and have a frequency superior to five in the telecommunication corpus too.

The totality of the gold-standard contains 97 words, with 18 seed words. A seed word can have one to nine neighbors. Even though it is difficult to say that the gold-standard is exhaustive and it is quite small, it contains the most relevant words for the sector so it allows us to study the behavior of the proposed method in the task of enriching our domain-dependent taxonomy.

Table 1 shows an extract of the gold-standard, in the first column there are the seed words, and in the second column there are their closest neighbors chosen by the two linguists.

Table 1. Extract of the gold-standard

Seed words	Neighbors
télé	*télévision tv tele television*
magasin	*boutique agence magazin*
message	*sms mail mms commentaire texto*
téléphone	*fixe portable mobile smartphone phone telephone tel*

3.4. Experiment

Word2Vec can be used with two architectures: Skip Gram Negative Sampling (SGNS) and Continuous Bag of Words (CBOW), both based on a prediction system that works with a neural network where words are represented by vectors. In the present case, experiments were conducted with SGNS architecture that is, according to [13], better for semantic relations.

Word2vec disposes of several parameters that can be adapted to improve the results such as the window size, the layer size or the number of iteration on the corpus. For the present experiments, it was decided to use a window size of 2 words, 400 dimensions for the vectors and 5 iterations on the corpus.

Several experiments were performed to study the performances and limitations of word2vec when applied to CRM domain. (a) Compare lemmatized corpus with PoS tagged corpus. (b) Test different sizes of the telecommunication corpus (using only part of the corpus to simulate lack of data). (c) Compare results on the telecommunication corpus with the frWac corpus to see the impact of the corpus size and corpus specificity. (d) Compare results obtained with Word2Vec and Random Indexing.

For each configuration, the list of closest neighbors for seed words was generated with word2vec. Different tests were conducted with different thresholds between 0.1 and 0.6, meaning that all words which are above the threshold are considered neighbors of the seed word (thus, semantically close). The different configurations performances were computed by using the gold-standard, over which was calculated micro-averaged Precision, Recall and F1.

[1] Fauconnier, J.-P. web site : http://fauconnier.github.io/#data

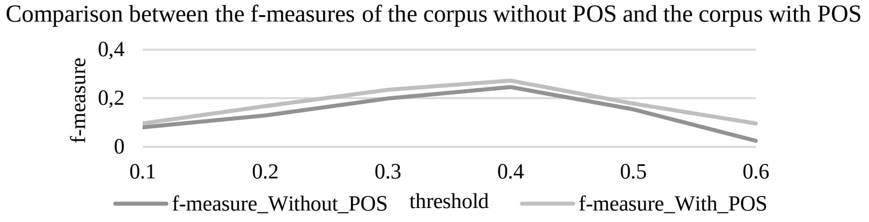
[2] TreeTagger: tool for annotating text with part-of-speech and lemma information. French version provided by Achim Stein. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

4. Results

4.1. Part-of-speech tagging

Adding PoS-tagging on the system allowed to conserve only nouns inside word2vec outputs, and to create finer grained contexts. Figure 1 shows the f-measure evolution in terms of the threshold, using the telecommunication corpus just lemmatized or PoS-tagged (via TreeTagger). As we can observe, adding PoS systematically lead to better performances, since it allows to keep in the output only the nouns, thus some noise was reduced.

Figure 1 Comparison of f-measures on entire corpus with and without PoS

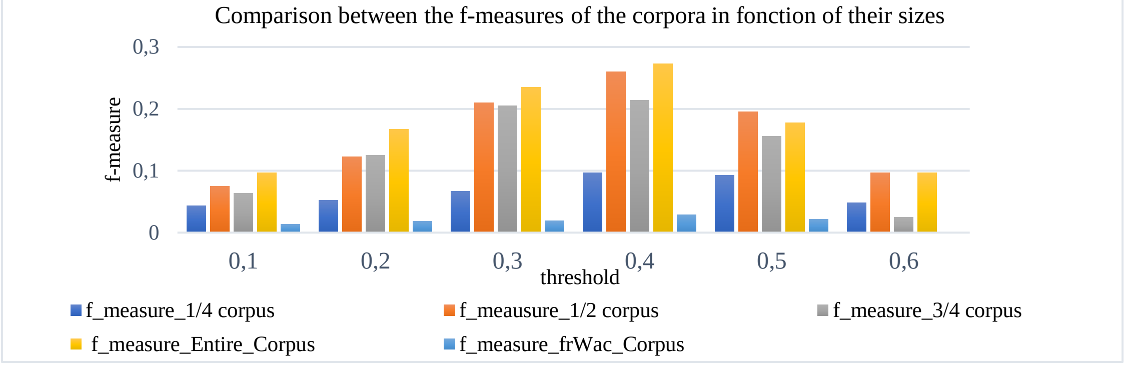


4.2. Corpus size and specificity

Two experiments were performed to see the influence of the corpus size on word2vec. (a) The telecommunication corpus was split in three sub-corpora, one quarter, half and three quarters. (b) The results were compared with the vectors learned with the frWac corpus, which contains much more data, but which is not specific to the studied domain.

Figure 2 shows, as expected, that reducing the telecommunication corpus produces a loss in the performance of the system. The smallest f-measure, by far, was surprisingly reached with the model built on frWac corpus, even if this corpus is almost 2,000 times bigger than the telecommunication corpus. Thus, in-domain corpus allows to better learn semantic similarity than a huge amount of words, so in this study, specificity is more important than size.

Figure 2 Comparison between the different sizes of corpus



5. Discussion

In Table 2, the best results obtained on the telecommunication and the frWac corpus, plus the results using Random Indexing algorithm on the telecommunication corpus are presented.

Table 2. Comparison between the best results of Word2vec and Random Indexing

	Threshold	Precision	Recall	F1
Word2Vec with Telecommunication corpus	0.40	0.34	0.23	0.27
Word2Vec with frWac corpus	0.30	0.02	0.03	0.02
RI with telecommunication corpus	0.80	0.06	0.32	0.10

Table 2 indicates that word2vec with telecommunication corpus generated the best results with a f-measure reaching 0.27 while other systems are much below. Note that best results for RI are obtained with a much higher threshold than word2vec. This is because RI has a different method to compute the vectors that generates high cosines, even if the number of returned neighbors is similar.

It seems important to explain why results obtained have a very low f-measure (never over 0.3). This is mainly due to the requirements of the gold standard. On the one hand, words were selected for the gold standard only if the linguists decided that they were very close to the seed words. Thus, it is possible that a seed word has only one neighbor in the gold standard. On the other hand, the gold standard contains low frequency words, only words with frequency below five were ignored, thus the method fails to extract words with low frequency what producing a low final performance.

Table 3 presents a sample of word2vec results. In the first column, there are the seed words, in the second and third columns there is the word2vec output: correctly extracted words (true positives) and words not expected by the gold standard (false positives). The last column shows the words expected in the gold standard not generated by word2vec (false negatives).

Table 3. Word2vec's outputs instances for a threshold equal to 0.4

Seeds	True Positives	False Positives	False Negatives
ligne			*adsl réseau*
problème	*panne probleme*		*pb question bug incident*
box	*décodeur boxe decodeur*	*adsl tv wifi pc*	
conseiller	*personne intervenant interlocuteur*		*conseillère équipe demoiselle personnel collaborateur correspondant*

As presented in the Table 3, some seed words, such as *ligne* [*line*], do not have any neighbors. This means that there is no word which have a cosine superior to 0.4 in the corpus according to word2vec, probably due to the lack of data. Otherwise, for words such as *box* all expected words are retrieved, with a perfect Recall, even though some false positives are also introduced. Note that nevertheless, those unexpected words are not semantically far from the seed word. This shows that word2vec is able to correctly capture in-domain, semantically related words.

Overall, the first steps of this study are satisfying. A manual analyze of the results points that even though the results are not perfect and the f-measure is low, the list of related words is quite adequate, and could be used as a good basis to enrich the existing taxonomy.

6. Conclusion and Further Work

In this work, word2vec was used to extract domain-specific related terms from very small corpora. The results obtained show that corpus specificity is an important parameter for extraction of neighbors when using word2vec, even more than the size of the corpus. Results suggest that it is possible to get satisfactory results with small data for domain-specific words. Even if the f-measure is low, results are promising and can be helpful to enrich the taxonomy. This brings a gain that is not negligible on this task and opens the door to quickly adapting the classification system to new sectors, even though some manual revision is advised in the current setup.

As further work, the same experiments will be conducted with other business sectors, as e-commerce or car insurance. Also, an extrinsic evaluation will be performed to enrich automatically the taxonomy with word2vec and then study the results before and after the enrichment. In this way, we will study the real impact of the automatic enrichment of the taxonomy. Also, the gold standard will be increased to have statistical results more reliable, and we plan to try the CBOW architecture which has, according to some studies[3], good results on small corpora.

[3] CBOW studies: https://www.tensorflow.org/tutorials/word2vec

7. References

1. Altszyler, E., Mariano, S., & Fernández Slezak, F.: Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. CoRR abs/1610.01520 (2016)
2. Baroni, M., Dinu, G. & Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics. (2014)
3. Crouch, C.-J.: A cluster-based approach to thesaurus construction. In Proceedings of the 11th Annual International ACM SIGIR Conference, pp. 309-320, Grenoble France. (1988)
4. Curran, J.-R., & Moens, M.: Improvements in automatic thesaurus extraction. In Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special In-terest Group on the Lexicon (SIGLEX), pp. 59-66, Philadelphia, PA. (2002)
5. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer. (1994)
6. Harris, Z.: Distributional structure. Word, 10(23):146–162. (1954)
7. Hofmann, T.: Probabilistic Latent Semantic Analysis, Uncertainty in Artificial Intelligence. (1999)
8. Karneva, P., Kristofersson, J. & Holst, A.: Random Indexing of text samples for Latent Semantic Analysis. Proceedings of the 22nd annual conference of the cognitive science society. New Jersey: Erlbaum. (2000)
9. Kato, R. & Goto, H.: Categorization of web news documents using word2vec and deep learning. Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia, March 8-10, (2016)
10. Landauer, T.-K. & Dumais, S.-T.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, Vol 104(2), Apr 1997, 211-240. (1997)
11. Landauer, T.-K., Foltz, P.-W. & et Laham, D.: « Introduction to Latent Semantic Analysis », Discourse Processes, vol. 25, p. 259-284, (1998)
12. Levy, O., Goldberg, Y. & Dagan, I..: Improving Distributional Similarity with Lessons Learned from Word Embeddings. Transactions of the Association for Computational Linguistics. (2015)
13. Mikolov, T., Corrado, G., Chen, K. & Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations, (ICLR 2013), p. 1–12. (2013a)

    Rossignol, M. & Sébillot, P.: Automatic acquisition of lexical semantic information using medium to small corpora. (2008)
14. Salton, G., Wong, A. & Yang, C.-S.: A vector space model for automatic indexing. Communications of the ACM, v.18 n.11, p.613-620, Nov. (1974)