Reducing VSM data sparseness by generalizing contexts: application to health text mining

Amandine Périnet
INSERM, U1142, LIMICS, Paris, France
Sorbonne Universités, UPMC Univ Paris 06, Paris, France
Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France
amandine.perinet@edu.univ-paris13.fr

Thierry Hamon
LIMSI-CNRS, Orsay, France
Université Paris 13, Sorbonne Paris Cité
Villetaneuse, France
hamon@limsi.fr

Abstract

Vector Space Models are limited with low frequency words due to few available contexts and data sparseness. To tackle this problem, we generalize contexts by integrating semantic relations acquired with linguistic approaches. We use three methods that acquire hypernymy relations on a EHR corpus. Context Generalization obtains the best results when performed with hypernyms, the quality of the relations being more important than the quantity.

1 Introduction

Distributional Analysis (DA) (Harris, 1954; Firth, 1957) computes a similarity between target words from the contexts shared by those two words. This hypothesis is applied with geometric methods, such as the Vector Space Model (VSM) (Turney and Pantel, 2010). The advantage of the VSM is that the similarity of word meaning can be easily quantified by measuring their distance in the vector space, or the cosine of the angle between them (Mitchell and Lapata, 2010). On the other hand, a major inconvenience is data sparseness within the matrix that represents the vector space (Turney and Pantel, 2010). The data sparseness problem is the consequence of the word distribution in a corpus (Baroni et al., 2009): in any corpus, most of the words have a very low frequency and appear only a few times. Thus, those words have a limited set of contexts and similarity is difficult to catch. Thus, methods based on DA perform better when more information is available (Weeds and Weir, 2005; van der Plas, 2008) and are efficient with large corpora of general language. But with specialized texts, as EHR texts that are usually of smaller size, reducing data sparseness is a major issue and methods need to be adapted.

Semantic grouping of contexts should decrease their diversity, and thus increase the frequency of the remaining generalized contexts. We assume that generalizing contexts may influence the distributional context frequencies. Information for generalization can be issued from existing resources or can be computed by linguistic approaches. In this paper, we propose to use semantic relations acquired by relation acquisition methods to group words in contexts. We define a method that switches words in DA contexts for their hierarchical parent or morphosyntactic variant that have been computed on the corpus with linguistic approaches before applying the VSM method.

In the following, we first present the related work, then our method and we finally describe the different experiments we led. The results obtained on the EHR corpus are then evaluated in terms of precision and MAP, and analyzed.

2 Related work

Our approach relates with works that influence distributional contexts to improve the performance of VSMs. Some of them intend to change the way to consider contexts; Broda et al. (2009) do not use the raw context frequency in DA, but they first rank contexts according to their frequency, and take the rank into account. Other models use statistical language models to determine the most likely substitutes to represent the contexts (Baskaya et al., 2013). They assign probabilities to arbitrary sequences of words that are then used to create word pairs to feed a co-occurrence model, before performing a clustered algorithm (Yuret, 2012). The limit of such methods is that their performance is proportional to vocabulary size and requires the availability of training data.

Influence on contexts may also be performed by embedding additional semantic information. The semantic relations may be issued from an existing resource or automatically computed. With a method based on bootstrapping, Zhitomirsky-Geffet and Dagan (2009) modify the weights of

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 90–95,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

the elements in contexts relying on the semantic neighbors found with a distributional similarity measure. Based on this work, Ferret (2013) uses a set of examples selected from an original distributional thesaurus to train a supervised classifier. This classifier is then applied for reranking the neighbors of the thesaurus selection. Within Vector Space Model, Tsatsaronis and Panagiotopoulou (2009) use a word thesaurus to interpret the orthogonality of terms and measure semantic relatedness.

With the same purpose of solving the problem of data sparseness, other methods are based on dimensionality reduction, such as Latent Semantic Analysis (LSA) in (Padó and Lapata, 2007) or Non-negative Matrix Factorization (NMF) (Zheng et al., 2011). Matrix decomposition techniques are usually applied to reduce the dimensionality of the original matrix, thereby rendering it more informative (Mitchell and Lapata, 2010).

Our approach differs from the aforementioned ones in that we add semantic information in contexts to reduce the number of contexts and to increase their frequency. Contrary to these latter approaches, we do not reduce the contexts by removing information but by generalyzing information and integrating extra semantic knowledge.

3 VSM and context generalization

The contexts in which occurs a target word have associated frequencies which may be used to form probability estimates. The goal of our method is to influence the distributional context frequencies by generalizing contexts.

Step 1: target and context definition During this step, we define targets and contexts, with different constraints for their extraction. To adapt our method to specialized texts, we identify terms (specific terminological entities that denote an event) with a term extractor (YATEA (Aubin and Hamon, 2006)). Target words are both nouns and terms (T). Their distributional contexts correspond to a graphical window of n number of words around the targets (Wilks et al., 1990; Schütze, 1998; Lund and Burgess, 1996). We consider two different window sizes defined in section 4.

Linguistic approaches During the generalization process, we use three existing linguistic approaches: two that acquire hypernymy relations and one to get morphosyntactic variants. Lexico-syntactic Patterns (LSP) acquire hypernymy relations. We use the patterns defined by (Hearst, 1992). Lexical Inclusion (LI) acquires hypernymy relations and uses the syntactic analysis of the terms. Based on the hypothesis that if a term is lexically included in another, generally there is a hypernymy relation between the two terms (*kidney transplant - cadaveric kidney transplant*) (Bodenreider et al., 2001). Terminological Variation (TV) acquires both hypernyms and synonyms. TV uses rules that define a morpho-syntactic transformation, mainly the insertion (*blood transfusion - blood cell transfusion* (Jacquemin, 1996).

Step 2: context generalization Once targets and contexts are defined, we generalize contexts with the relations acquired by the three linguistic approaches we mentioned. To integrate the relations in contexts, we replace words in context by their hypernym or morphosyntactic variant. We define two rules: (1) if the context matches with one hypernym, context is replaced by this hypernym. (2) if the context matches with several hypernyms or variants, we take the hypernym or variant frequency into account, and choose the most frequent hypernym/variant. The generalization step is individually or sequentially performed when several relation sets are available.

Step 3: computation of semantic similarity After the generalization step, similarity between target words is computed. As we previously decrease diversity in contexts, we choose a measure that favors words appearing in similar contexts. We use the Jaccard Index (Grefenstette, 1994) which normalizes the number of contexts shared by two words by the total number of contexts of those two words.

Parameter: thresholds The huge number of relations we obtain after computing similarity between targets leads us to remove the supposed wrong relations with three thresholds: (i) number of shared lemmatized contexts (2 for a large window, 1 for a small window) ; (ii) number of the lemmatized contexts (2 for a large window, 1 for a small window) ; (iii) number of the lemmatized targets (3 for both window sizes). For each parameter, the threshold is automatically computed, according to the corpus, as the mean of the values of parameters on the corpus. And we experiment two thresholds on similarity score we empirically defined : $sim > 0.001$ and $sim > 0.0005$.

4 Experiments

In this section, we present the material we use for the experiments and evaluation, and the distributional parameter values of the VSM automatically determined from the data. We then describe the generalization sets we experiment and the evaluation measures we used for evaluation.

4.1 Corpus

We use the collection of anonymous clinical English texts provided by the 2012 i2b2/VA challenge (Sun et al., 2013).

The corpus is pre-processed within the Ogmios platform (Hamon et al., 2007). We perform morphosyntactic tagging and lemmatization with Tree Tagger (Schmid, 1994), and term extraction with YATEA (Aubin and Hamon, 2006).

4.2 Distributional parameters

We consider two window sizes: a large window of 21 words (± 10 words, centered on the target, henceforth W21) and a narrow one of 5 words (± 2 words, centered on the target, W5).

The window size influences on the type, the volume and the quality of the acquired relations. Generally, the smaller windows allow to acquire more relevant contexts for a target, but increase the data sparseness problem (Rapp, 2003). They give better results for classical types of relations (eg. synonymy), whereas larger windows are more appropriate for domain relations (eg. collocations)(Sahlgren, 2006; Peirsman et al., 2008).

4.3 Generalizing distributional contexts

We define several sets of context generalization. We experiment in step 2 different ways of generalizing contexts. We use as a baseline the VSM without any generalization in the contexts (VSMonly), and compare the generalization sets to it.

Regarding context generalization, we first exploit the relations acquired from only one linguistic approach. We apply the method described at the section 3 (step 2) by separately using the three different sets of relations automatically acquired. Distributional contexts are replaced by their hypernym acquired with lexico-syntactic patterns (VSM/LSP) and lexical inclusion (VSM/LI), and by their morphosyntactic variants acquired with terminological variation (VSM/TV). Then, we replace contexts with relations acquired by two approaches (TV then LI, LSP then TV, etc.). This generalization is done sequentially: we generalize all the contexts with the relations acquired by one method (e.g. LI), and then with the relations acquired by another method (e.g. TV). And finally, similarly to what we perform with two methods, we experiment the generalization of contexts by relations acquired with the three different linguistic approaches (e.g. LSP then LI then TV). We experiment all the possible combinations. With both the single and multiple generalization, we aim at evaluating the contribution of each method but also the impact of the order of the methods.

4.4 Evaluation

In order to evaluate the quality of the acquired relations, we compare our relations to the 53,203 UMLS relations between terms occurring in our EHR corpus. We perform the evaluation with the Mean Average Precision (MAP) (Buckley and Voorhees, 2005) and the macro-precision computed for each target word: semantic neighbors found in the resource by the total semantic neighbors acquired by our method. We consider three sets of neighbors: precision after examining 1 (P@1), 5 (P@5) and 10 (P@10) neighbors.

5 Results and discussion

Best results are obtained with a large window of 21 words, with a precision P@1 of 0.243 against 0.032 for a 5 word window, both for VSMonly, with a threshold of 0.001. Thus, a high threshold on the similarity score is not always relevant. We observe on this corpus that the generalization with the several linguistic approaches does not improve the results. For instance, VSM/LI obtains 0.250 of P@1 with a $>$ 0.001 threshold, and this precision is the same with VSM/LI+TV and with VSM/LI+LSP. This is an interesting behavior, different from what have been observed so far on more general French corpora that contains cooking recipes (Périnet and Hamon, 2013).

We discuss here the results we obtain for terms, for the two thresholds on the similarity score: a low and a higher thresholds, with relations with a similarity above 0.0005 and above 0.001. We observe that with a higher threshold, the precision is higher, with a P@1 of 0.243 against 0.187 for the lower threshold (when considering VSMonly). As for the number of relations acquired, with a lower threshold we obtain more relations (3,936 relations acquired for the baseline) than with a higher threshold (326 relations for the baseline).

We evaluate precison after examining three groups of neighbors. The best results are obtained with P@1, and in most cases, precision decreases when we consider more neighbors: the more neighbors we consider, the lower precision is. For a 0.001 threshold, the generalized experiment sets obtain a higher precision than VSMonly, in any case. While for a 0005 threshold, the use of LI to generalize contexts decreases the precision. We also observe that when considering generalisation with TV or LSP only, or their combination, the P@10 is slightly better than P@5.

The MAP values are higher when the thresold on the similarity measure is low, with 0.446 for VSM/LI against 0.089 with the > 0.001 threshold. It means that some correct relations are not well ranked with the similarity score, but are still present. We observe that the MAP values are always higher with the generalization sets than with the baseline with both thresholds: 0.089 for VSM/LI, 0.446 for VSM/LI+LSP, etc.

Comparison of the experimental sets When considering the relations found in the UMLS, we observe that the generalization with LSP brings the same relations that the baseline VSMonly plus 22 relations, the generalization with TV brings 16 more relations that VSMonly, and finally that the generalization with LI decreases the number of relations acquired. When the generalization of the contexts is performed with LI, only with LI or with LI combined to another method, it decreases the number of relations acquired as well as the number of relations found in the resource. On the contrary, generalizing contexts with LSP increases the number of relations acquired as well as the number of relations found in the UMLS resource. We obtain the highest number of relations when generalizing contexts with LSP, with 454 relations, and the highest precision with 0.273 for P@1.

Comparing those results with the relations acquired with the linguistic approaches on the EHR corpus shows a correlation between the quality of the relations acquired with the generalized sets and the relations used for generalization. Indeed, LI gives the highest number of relations with 14,437 relations, then TV gives 631 relations, and finally LSP acquires only three relations: *pancreatic complication - necrosis*, *pancreatic complication - abscess*, *gentle laxative - milk of magnesia*.

With these relations, if the second term (eg. *necrosis*) is found in the context, it is replaced by the first term (eg. *pancreatic complication*). These three relations used for generalization give better results in terms of precision that the many relations given by the two other approaches. We could deduce that the number of relations may not be as important as their quality when they are used for generalization. But when the LSP are used after TV or LI, they do not improve the results. From this observation, we make the hypothesis that these second terms may have already been replaced during the generalization with LI or TV. To confirm or reject this hypothesis, we look closer to the relations acquired with TV and LI. In TV, we find no relation including any of these second terms. On the contrary, with LI, we found the relation *milk - milk of magnesia* that inhibits one of the three relations acquired with the LSP.

We deduce that even if the quality of the relations used for generalization is more important than their number, the number of relations still matters. If generalization is first performed with a great number of relations, then a small number of relations used for generalization is not enough and does not improve the results.

6 Conclusion and perspectives

In this work, we face the problem of data spareseness of distributional methods. This problem especially arises from specialized corpora which have a smaller size and in which words and terms have lower frequencies.

To achieve this goal, we propose to generalize distributional contexts with hypernyms and variants acquired by three existing approaches. We focus on the acquistion of relations between terms. We experimented several generalization sets, using one, two or the three methods sequentially to replace words in context by their hypernym or variant. Evaluation of the method has been performed on an EHR English text collection. Generalization obtains the best results when realized with hypernyms. The quality of the relations matters much more than their number: few but good relations used to generalize contexts give better results than many relations of poorer quality. For future work, we plan to use for generalization relations issued from different distributional and terminological resources. Finally, we will intend to combine the methods before normalization.

References

Sophie Aubin and Thierry Hamon. 2006. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, number 4139 in LNAI, pages 380–387. Springer.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Osman Baskaya, Enis Sert, Volkan Cirik, and Deniz Yuret. 2013. Ai-ku: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Proceedings of SemEval - 2013*, pages 300–306, Atlanta, Georgia, USA. Association for Computational Linguistics.

Olivier Bodenreider, Anita Burgun, and Thomas Rindflesch. 2001. Lexically-suggested hyponymic relations among medical terms and their representation in the umls. In *TIA 2001*, pages 11–21.

Bartosz Broda, Maciej Piasecki, and Stan Szpakowicz. 2009. Rank-based transformation in measuring semantic relatedness. In Yong Gao and Nathalie Japkowicz, editors, *Canadian Conference on AI*, volume 5549, pages 187–190. Springer.

Chris Buckley and Ellen Voorhees. 2005. Retrieval system evaluation. In Ellen Voorhees and Donna Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3. MIT Press.

Olivier Ferret. 2013. Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. In *TALN 2013*, pages 48–61, Les Sables d'Olonne, France.

J.R. Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32.

Gregory Grefenstette. 1994. Corpus-derived first, second and third-order word affinities. In *Sixth Euralex International Congress*, pages 279–290.

T. Hamon, A. Nazarenko, T. Poibeau, S. Aubin, and J. Derivière. 2007. A robust linguistic platform for efficient and domain specific web content analysis. In *RIAO 2007*, Pittsburgh, USA.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *International Conference on Computational Linguistics*, pages 539–545, Nantes, France.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In *CoRR*, pages 425–438.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28:203–208.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Comput. Linguist.*, 33(2):161–199.

Yves Peirsman, Heylen Kris, and Geeraerts Dirk. 2008. Size matters. tight and loose context definitions in english word space models. In *ESSLLI Workshop on Distributional Lexical Semantics*, Hamburg, Germany.

Amandine Périnet and Thierry Hamon. 2013. Hybrid acquisition of semantic relations based on context normalization in distributional analysis. In *Proceedings of TIA 2013*, pages 113–120, Paris, France.

Reinhard Rapp. 2003. Word sense discovery based on sense descriptor dissimilarity. In *MT Summit'2003*, pages 315–322.

Magnus Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University, Stockholm, Sweden.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *New Methods in Language Processing*, pages 44–49, Manchester, UK.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *JAMIA*, 20(5):806–813.

George Tsatsaronis and Vicky Panagiotopoulou. 2009. A generalized vector space model for text retrieval based on semantic relatedness. In *EACL 2009*, pages 70–78, Stroudsburg, PA, USA.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *JAIR*, 37:141–188.

Lonneke van der Plas. 2008. *Automatic lexico-semantic acquisition for question answering*. Thèse de doctorat, University of Groningen, Groningen.

Julie Weeds and David Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Comput. Linguist.*, 31(4):439–475.

Yorick A. Wilks, Dan, James E. Mcdonald, Tony Plate, and Brian M. Slator. 1990. Providing machine tractable dictionary tools. *Journal of Machine Translation*, 2.

Deniz Yuret. 2012. Fastsubs: An efficient and exact procedure for finding the most likely lexical substitutes based on an n-gram language model. *IEEE Signal Process. Lett.*, 19(11):725–728.

Wenbin Zheng, Yuntao Qian, and Hong Tang. 2011. Dimensionality reduction with category information fusion and non-negative matrix factorization for text categorization. In Hepu Deng, Duoqian Miao, Jingsheng Lei, and Fu Lee Wang, editors, *AICI*, volume 7004 of *LNCS*, pages 505–512.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Comput. Linguist.*, 35(3):435–461.