Generalising and normalising distributional contexts to reduce data sparsity: application to medical corpora

Amandine Périnet
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 implement the distributional hypothesis. They are based on the repetition of information occurring in the contexts of words to associate. However, these models suffer from a high number of dimensions and data sparsity in the matrix of contextual vectors. This is a major issue with specialised corpora that are of much smaller size and with much lower context frequencies. We tackle the problem of data sparsity on specialised texts and we propose a method that allows to make the matrix denser, by generalising and normalising distributional contexts. Generalisation gives better results with the Jaccard index, narrow sliding windows and relations of lexical inclusion. On the other hand, normalisation has no positive effect on the relation extraction, with any combination of distributional parameters.

1 Introduction

Distributional Analysis (DA) assumes that words occurring in a similar context tend to be semantically close (Harris, 1954; Firth, 1957). This hypothesis is usually applied through vector space models (VSM) where vectors represent the contextual information and distributional statistical data (Sahlgren, 2006). Each target word in a text is represented as a point defined according to its distributional properties in the text (Turney and Pantel, 2010; Lund and Burgess, 1996). Thus, the semantic similarity between two words is defined as a closeness in an n-dimension space, where each dimension corresponds to some potential shared contexts. The VSMs easily quantify the semantic similarity between two words by measuring the distance between the two corresponding vectors within this space, or the cosine of their angle. On the other hand, besides the high number of dimensions required (for example, Sahlgren (2006) uses VSMs with up to several millions of dimensions), VSMs also suffer from data sparseness within the matrix representing the vector space (Chatterjee and Mohan, 2008): many elements are equal to zero because only few contexts are associated to a target word. This disadvantage is partly due to word distribution in corpora: whatever the corpus size, most words have low frequencies and a very limited set of contexts compared to the number of words in the corpora. These last two elements make the similarity between two words hard to compute. Hence, methods based on the distributional hypothesis show better results when much information is available and especially with general corpora, usually of great size (Weeds and Weir, 2005; van der Plas, 2008). But the reduction of data sparseness is still an important aspect with general corpora. It is as well a major issue when working with specialised corpora. Indeed, these corpora are characterised by smaller sizes, and with frequencies and a number of different contexts especially lower. We focus here on this last point. We propose a rule-based method that aims at reducing context diversity by generalising contexts. The frequency of the obtained distributional contexts is then increased and, consequently, data sparseness and the dimensions of the vector space model are reduced. We present here a generalisation of the distributional contexts thanks to semantic relations acquired on corpora. The parameters of the distributional method are tuned to specialised corpora, especially in integrating those generalised contexts.

We first present a state of the art on data sparsity reduction within distributional methods. Then we describe the proposed context generalisation and normalisation method as well as the experiments

Proceedings of the 4th International Workshop on Computational Terminology, pages 1–10,
Dublin, Ireland, August 23 2014.

we performed to evaluate its impact on specialised corpora. Results are evaluated and analysed with precision, R-precision and MAP.

2 Related work

Reducing data sparsity is a main issue in distributional analysis. The existing methods aim at influencing the selection of useful contexts or at integrating semantic information to modify context distribution. Thus, contrary to the common usage, Broda et al. (2009) propose to weight contexts by first ranking contexts according to their frequency, and then take the rank into account to weight contexts. Other approaches rely on statistical language models to determine the most likely substitutes to represent contexts (Baskaya et al., 2013). These models assign probabilities to arbitrary sequences of words based on their co-occurrence frequencies in a training corpora (Yuret, 2012). These substitutes and their probabilities are then used to create word pairs to feed a co-occurrence model and to cluster the word list. The limit of such methods is their performance which depends on vocabulary size and requires an increasing amount of training data. Influence on contexts may also be done by incorporating additional semantic information: it has been shown that such information used to modify the standard distributional method can improve its performance (Tsatsaronis and Panagiotopoulou, 2009). This semantic information, in particular semantic relations, may be automatically computed or issued from an existing resource. Thus, with a bootstrap method, Zhitomirsky-Geffet and Dagan (2009) modify the context weights with the semantic neighbours proposed by a distributional similarity measure. Based on this latter work, Ferret (2013) addresses the problem of low frequency words. To better consider this information, a set of positive and negative examples are selected with an unsupervised classifier. A supervised classifier is then applied for re-ranking the semantic neighbours. The method allows to improve the quality of the similarity relation between nouns with low or mid frequency.

The sparseness problem may also be tackled from the algorithmic point of view by limiting the dimensions of the context matrix, especially by smoothing it in order to reduce the number of vector compounds (Turney and Pantel, 2010). Thus, Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997; Padó and Lapata, 2007) implements a factorization method of matrices by Singular Value Decomposition (SVD). The original data in the context matrix is abstracted in independant linear components, that allow to reduce noise and to highlight the main elements. Besides reducing the computational cost, dimension reduction significantly improves precision in LSA applications. For instance, the use of the SVD to compute word similarity allows to obtain scores equivalent to human scores in a TOEFL test with multiple choice questions of synonymy (Landauer and Dumais, 1997). As for low frequency, the SVD is a way to counterbalance the lack of data (Vozalis and Margaritis, 2003). Also, some methods, as the non-negative matrix factorization (Lee and Seung, 1999), allow to better model word frequency. But, when it comes to the acquisition of semantic relations, performances do not seem better than the ones obtained with the LSA (Turney and Pantel, 2010; Utsumi, 2010). Furthermore, the dimension reduction makes easier to treat context vectors, but it does not solve the initial issue of building a huge co-occurence matrix. Random Indexing (RI) (Kanerva et al., 2000) may be considered as a solution to this problem, as it incrementally builds the context matrix according to an index vector of the target word randomly generated, as well as reducing the matrix dimension. RI and LSA have similar performance when identifying synonyms in a similar way than the TOEFL test (Karlgren and Sahlgren, 2001). Recently, the selection of the best contexts combined with a normalisation of their weights allows to improve the quality of a SVD reduced matrix (Polajnar and Clark, 2014). In the context of definition retrieval and phrase similarity computation, their impact depends on the compositional semantics operators used.

As above work, we aim at incorporating semantic information within distributional contexts, but by reducing the number of contexts and increasing their frequency. Contrary to SVD based methods that limit the contexts by removing information, here we both generalise and normalise contexts through the integration of additional semantic knowledge computed from our corpora.

3 Material

In this section, we present the corpus we use. We also describe the approaches used to acquire the semantic relations integrated in our method for context generalisation/normalisation.

Corpora To evaluate our approach, we use the Menelas corpus (Zweigenbaum, 1994). It consists in a medical text collection, in French, on the topic of coronary diseases. The corpus contains 84,839 words. It has been analysed through the Ogmios platform (Hamon et al., 2007). The linguistic analysis includes a morphosyntactic tagging and a lemmatisation of the corpus, with TreeTagger (Schmid, 1994), and a term extraction with YaTeA (Aubin and Hamon, 2006). This last step allows to identify terminological entities (both single word units, for eg. *artery*, and complex terms (i.e. multi-word expressions), for eg. *coronary disease*, that denote the domain concepts).

Semantic relations acquisition Our generalisation and normalisation method of distributional contexts is based on semantic relations acquired from the entire corpus. We use several classical approches that allow to acquire semantic relations between terms. For context generalisation, we use lexico-syntactic patterns, lexical inclusion and terminological variation rules. Context normalisation is based on a rule-based synonymy acquisition.

- **Lexico-syntactic patterns (LSP)** We use the patterns defined by (Morin and Jacquemin, 2004) to detect 98 hypernymy relations between simple or complex terms, for instance: {`some |several etc.`} `SN: LIST` or {`other`}`? SN such as LIST`. The relations acquired with such patterns are usually relevant but the pattern coverage remains low.

- **Lexical Inclusion (LI)** This approach is based on the hypothesis that the lexical inclusion of a term (ex: *infarctus* in another (*infarctus du myocarde (myocardial infarction)* convey a hypernymy relation between those terms (Grabar and Zweigenbaum, 2003). We constrain the method by exploiting the term syntactic analysis provided by YaTeA. We obtain 7,187 relations between the complex term and its head. This approach is known to acquire relations with high precision.

- **Terminological variation (TV)** Terminological variant acquisition method proposed by (Jacquemin, 2001) exploits morphosyntactic transformation rules, as the insertion or the permutation, (*chirurgie coronarienne (coronary surgery) / chirurgie de revascularisation coronarienne (Coronary revascularisation surgery)*) to identify semantic relations between terms.
  The terminological variation rules, essentially the insertion on our French corpus, allow to acquire 171 hypernymy relations.

- **Semantic compositionality (Syn)** For context normalisation, we use 168 synonymy relations acquired with the method defined in (Hamon et al., 1998). Based on the semantic compositionality principle, a synonymy relation is inferred between complex terms, if at least one of their component are synonyms (*infection de blessure (wound infection)* and *septicit de blessure (wound sepsis)*).

4 Distributional context generalisation and normalisation

A solution to the problem of data sparsity on specialised corpora or smaller size corpora consists in increasing the density of the context matrix by disregarding superficial variations of contexts that are not strongly statistically significant or that result from the noise of the distributional method. Thus, we generalise (conceptual abstraction) and normalise (abstraction of minor lexical variations) contexts using the semantic information extracted from our corpus. In that respect, we use semantic relations automatically acquired with standard methods on specialised corpora. After a brief description of the distributional analysis we performed on specialised corpora, we present the distributional context generalisation and normalisation.

4.1 Distributional method

We focus on the extraction of relations between nouns, tokens tagged as nouns by TreeTagger, and terms, specific terminological entities extracted during the linguistic analysis of the corpus by YaTeA(see section 3). These semantic relations are crucial in specialised language. Nouns and terms are our targets. The distributional contexts of these targets correspond to adjectives, nouns, verbs and terms co-occurring with the target within a sliding window. A context is for us one element (a word or MWE), and it corresponds to one dimension in the vector space. For both targets and contexts, we consider the lemmas.

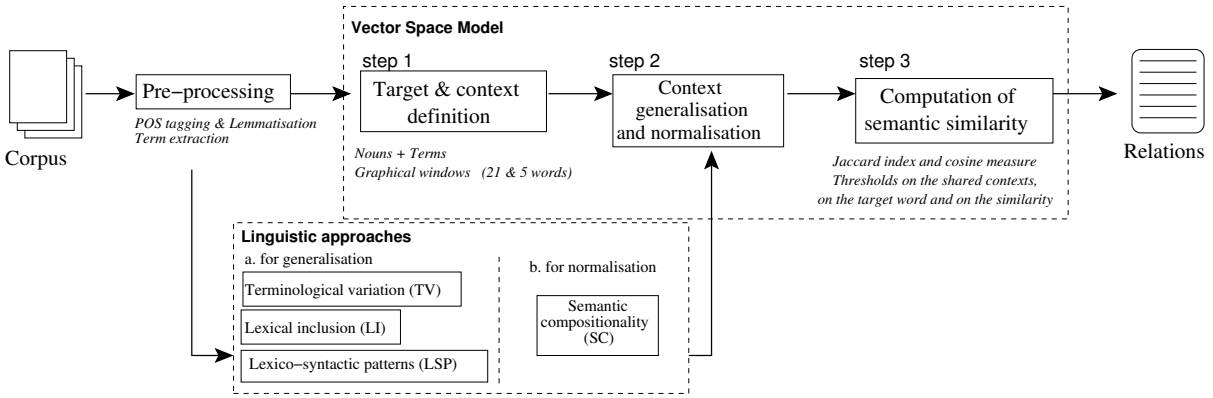

Figure 1: Process of distributional analysis

The overview of the distributional method is given in figure 1. The context generalisation and normalisation step takes place after the distributional context definition (see section 4.2). After extracting, generalising or normalising contexts, we compute a similarity score between each pair of target words, considering their shared contexts. We use the Jaccard index, recognised as suitable to specialised corpora (Grefenstette, 1994). The Jaccard index normalises the number of contexts shared by two words, w_m and w_n, by the total number of contexts of those two words, $ctxt(w)$. We also use the cosine of the context vector. In order to grant more or less importance to the information in contexts, we use these two measures combined with a weight function. With Jaccard, we use the relative frequency, that allows to consider the importance of a target, compared to the total number of contexts for the target. And with cosine we use Mutual Information (MI) (Fano, 1963). While these scores quantify the similarity between target words, it is necessary to apply thresholds to limit the proposed number of distributional relations and discard potentially noisy relations. We also intend to make the context matrix denser by applying thresholds on three distributional parameters: the number of shared contexts, the frequency of shared contexts and the frequency of target words. For each parameter, a threshold is automatically computed as the mean of the values taken by each parameter on the whole corpus. During the experiments (section 5), we test the impact of these thresholds on the results.

4.2 Rules of generalisation and normalisation of distributional contexts

Context generalisation and normalisation comes after the context definition step. It aims at reducing context sparsity and increasing the number of context occurrences. Generalisation and normalisation rules are separately applied and exploit additional semantic relations acquired on the corpus.

Rules of generalisation We generalise contexts with semantic relations automatically acquired from the corpus with hypernymy relations proposed by lexico-syntactic patterns and lexical inclusion (see section 3). While terminological variation approach does not propose semantically typed relations, the insertion rule is the only one used to acquire variants and it can be considered that the obtained relations are hypernymy relations. The hypernym and the hyponym are identified from the number of words present in each term: the shortest term then corresponds to the hypernym (*lésion significative (significant lesion)*), and the longest term is the hyponym (*lésion coronaire significative (significant coronary lesion)*).

We have then for each word $ctxt_i(w)$ in the context of the target word w three sets of hypernymy relations $\mathbb{H}_s(ctxt_i(w)) = \{H_1, \ldots, H_n\}$: $\mathbb{H}_{PLS}$, $\mathbb{H}_{IL}$ and $\mathbb{H}_{VT}$, with a hypernym set that may be empty. We define two substitution rules that allow to generalise contexts. Thus, for each word $ctxt_i(w)$ in the context of a target word w, we apply one of the following rules:

1. if $|\mathbb{H}_S(ctxt_i(w))| = 1$, then $ctxt_i(w) := H_1$, i.e. if the word in context corresponds to only one hypernym (H_1), acquired by one or several methods S, the word is replaced by this hypernym. For example, if lexical inclusion provides the relations *restriction / restriction du débit coronaire (restriction of coronary output)*, *restriction du débit coronaire* is replaced by *restriction*.

2. if $|\mathbb{H}_S(ctxt_i(w))| > 1$, $ctxt_i(w) = \text{argmax}_{|H_i|}(\mathbb{H}_S(ctxt_i(w)))$, i.e. if context corresponds to several hypernyms acquired by one or several methods S, we take into consideration the hypernym frequency $|H_1|, \ldots, |H_n|$ in the corpus, and we select the hypernym with the highest frequency. For example, for the word *artère coronaire (coronary artery)* in context, the lexico-syntactic patterns provide the following hypernyms: *veine (vein)*, *artère (artery)*, and *vaisseau (vessel)*, the one that is the most frequent is chosen and replaces *artère coronaire* in context.

4.2.1 Rules of normalisation

The normalisation rule aims at reducing semantic variation with automatically acquired synonymy relations. These relations are first organised in clusters of synonyms and a cluster representative is chosen: given the relations proposed by the acquisition method (section 3), the cluster representative corresponds to the most frequent word in the cluster. We have then for each target word $ctxt_i(w)$ in the context of the word w, a synonym cluster $\mathbb{S}_s(R) = \{S_1, \ldots, S_n, R\}$, with its representative R. We define one context normalisation rule applied for each word $ctxt_i(w)$ in the context of a word w to substitute the context word by the representative of the cluster it belongs: $if\ \exists R | ctxt_i(w) \in \mathbb{S}_S(R),\ then\ ctxt_i(w) := R$

5 Experiments and evaluation

5.1 Experiments

We performed several series of experiments on the Menelas corpus to evaluate the impact of both generalisation and normalisation rules on the quality of the acquired relations. Our baseline is the VSM without context substitution (VSMonly). First, we automatically compute the thresholds on the distributional parameters from the baseline (see section 4.1). The values of the thresholds are listed in table 1. We experiment two sliding window sizes ; a small window allows to detect classical types of relations (synonymy, meronymy, hypernymy, etc.) but increases the data sparseness problem. On the other hand, a large window provides more general relations, a contextual proximity.

Parameters	21 word window	5 word window
Similarity score	Jaccard: $sim > 0,000999$ Cosine: $sim > 0.9699$	Jaccard: $sim > 0,000999$ Cosine: $sim > 0.9699$
Number of contexts	2	1
Context frequency	3	2
Target frequency	3	3

Table 1: Definition of the threshold values on distributional parameters and on the similarity score according to the window width (21 and 5 words) and similarity measures (Jaccard index and Cosine)

We perform separately the experiments regarding generalisation and normalisation rules. With generalisation rules, in order to grasp the contribution of each linguistic method (see section 3), we define a set of experiments where context generalisation is performed using the hypernymy relations proposed individually by each method. The context generalisation rules $ctxt_i(w)$ are then applied separately using the sets $\mathbb{H}_{LSP}(ctxt_i(w))$ (VSM/LSP), $\mathbb{H}_{LI}(ctxt_i(w))$ (VSM/LI) and $\mathbb{H}_{TV}(ctxt_i(w))$ (VSM/TV). Then, sequentially, we apply generalisation rules by using the sets of hypernymy relations proposed by two linguistic approaches ($\mathbb{H}_{LSP}(ctxt_i(w))$ then $\mathbb{H}_{LI}(ctxt_i(w))$ – VSM/LSP+LI, $\mathbb{H}_{TV}(ctxt_i(w))$

then $\mathbb{H}_{LSP}(ctxt_i(w))$ – VSM/TV+LSP, etc.). All contexts are generalised following the relations proposed by one of the sets.Likewise, we combine the three sets of relations (for instance, $\mathbb{H}_{LSP}(ctxt_i(w))$ then $\mathbb{H}_{LI}(ctxt_i(w))$ then $\mathbb{H}_{TV}(ctxt_i(w))$ – VSM/LSP+LI+TV). By combining the hypernymy relation sources in several ways, we evaluate the complementarity of the approaches used for context generalisation. We also study the impact of the order of these methods in the generalisation sequence. All the hypernymy relations independently of the method used for their acquisition. We consider the set $H(ctxt_i(w)) = \mathbb{H}_{LSP}(ctxt_i(w)) \cup \mathbb{H}_{LI}(ctxt_i(w)) \cup \mathbb{H}_{TV}(ctxt_i(w))$ – VSM/ALL3. With context normalisation, we consider only one set of experiment, with normalisation of contexts (VSM/Syn).

All the experiments have been performed on both window sizes: 5 words ($\pm$ 2 words, centered on the target) and 21 words ($\pm$ 10 words, centered on the target). Indeed the window size influences the number and quality, but also the type of the relations acquired with distributional analysis. In general, a small window size (5 words) allows to have a highest number of relevant contexts for a given target word, but leads to more data sparsity than with a larger window (Rapp, 2003). Furthermore, the results obtained with small size windows are of greatest quality, especially for classical relations (synonymy, antonymy, hypernymy, meronymy, etc.), whereas larger windows are more adapted to the identification of domain specific relations (Sahlgren, 2006; Peirsman et al., 2008).

5.2 Evaluation

As usual to evaluate distributional methods, the obtained relations are considered as semantic neighbour sets associated to target words, and the quality of the neighbour sets is measured by comparing them to semantic relations issued from existing resources (Curran, 2004; Ferret, 2010). Thus, we compare the semantic relations acquired by our approach with the 1,735,419 relations in the French part of the UMLS metathesaurus[1]. The resource contains hypernyms, synonyms, co-hyponyms, meronyms and domain relations.

We used classical measures to evaluate the quality of our results: macro-precision (Sebastiani, 2002), the mean of the average precisions (MAP) (Buckley and Voorhees, 2005) and R-precision.

Macro-precision equally considers all target words whatever the number of semantic neighbours and provides a comprehensive quality of the results by computing the mean of the precision of each neighbour set. We consider one size of neighbour set for each target word: precision after examining 1 (P@1) semantic neighbour, the neighbour ranked first by its similarity score. Alternatively, we use R-precision that individually defines the size of the neighbour sets to examine as the number of correct neighbours expected for the corresponding target word (Buckley and Voorhees, 2005). To compute R-precision, we compare our results not to all the relations from the French part of UMLS, but to reference sets built from this resource, for each experiment. Thus, we have as many references as experiments. The mean of average precisions (MAP) is obtained taking in consideration the not interpolated precision of the semantic neighbours given their rank. It reflects the ranking quality and evaluates the relevance of the similarity measure used. Thus, the MAP favours the similarity measure that ranks all the correct semantic neighbours on top of the list. Reciprocally, adding noisy semantic neighbours at the end of the list does not discriminate against the method.

6 Results and discussion

In this section, we present and discuss the results we obtain, first with the 21 word window and then with the 5 word window. For both sizes, we present the number of relations acquired (*Acq. Rel*), the number of relations found in the UMLS resource (*Rel. UMLS*), and the results in terms of MAP, R-precision and precision to 1 (P@1). Before discussing our results, we briefly present some results of similar work.

Results in existing similar work In order to understand better our results, we first provide some results obtained in similar work. Keep in mind that these results are not obtained with the same corpus. Indeed, a major problem is that the comparisons with reference resources are often given for large copora and very frequent words (Curran and Moens, 2002), or for different tasks than our task. An effective comparison

[1] `http://www.nlm.nih.gov/research/umls/`

is then difficult. We first present results obtained on a similar task, but with general copora, and then results obtained on similar documents (i.e. specialised medical texts) but in a different tasks. Despite this limit, we can still quote for comparison the values obtained by Ferret (2011) for the evaluation of semantic neighbours extraction on English general copora (of 380 million words). The parameters of his VSM are a small sliding window of 3 words ($\pm$ 1 word centered on the target), the cosine measure and mutual information. In his work, Ferret (2011) considers three sets of target words according to their frequency. As in the Menelas corpus, the highest frequency of a target word is 270 and that only a few frequencies are above 100, we may consider the set of low frequency words, that occur less than 100 times. The highest values he obtains for those target words are a MAP of 0.03, a P@1 of 0.026 and an R-precision of 0.02. For more frequent words, occurring between 100 and 1,000 times, the values are higher: a MAP of 0.125, a P@1 of 0.209 and an R-precision of 0.104.

For a comparison with specialised texts, we can look at Moen et al. (2014)'s work on document similarity between care episodes in a retrieval system. The matrix is then a term-document matrix, and not a term-context one, and the task is different. We do not know exactly how the comparison is effective, but they obtain for their best system a MAP of 0.326 and a precision at 10 neighbours of 0.515.

	Acquired Rel.		Rel. in UMLS		MAP		R-precision		P@1	
	JACC	COS	JACC	COS	JACC	COS	JACC	COS	JACC	COS
VSMonly (baseline)	406	9,154	4	46	0.406	0.105	0.250	0.000	0.250	0.000
VSM/TV	472	5,322	8	24	0.280	0.149	0.143	0.053	0.143	0.053
VSM/LI	324	2,844	4	18	0.454	0.232	0.250	0.167	0.250	0.200
VSM/LSP	398	4,684	6	18	0.219	0.154	0.000	0.071	0.000	0.071
VSM/TV+LI	324	2,844	4	18	0.454	0.232	0.250	0.167	0.250	0.200
VSM/TV+LSP	398	4,678	6	18	0.219	0.149	0.000	0.071	0.000	0.071
VSM/LSP+LI	336	2,748	4	14	0.454	0.263	0.250	0.208	0.250	0.250
VSM/ALL3	336	2,982	4	16	0.414	0.259	0.250	0.192	0.250	0.231
VSM/Syn	474	5,282	8	24	0.280	0.157	0.143	0.053	0.143	0.053

Table 2: Results obtained with the Jaccard index and Cosine measure for a 21 word window

Large window For the large window, we present the results obtained with Jaccard and Cosine. We do not present all the generalisation sets because adding more relations (more methods) in the generalisation process does not change the results: once we generalise with two methods, the results get stable. We first observe a different behaviour according to the similarity measure in terms of relations acquired: Cosine allow to acquire many more relations than Jaccard, and as well more relations acquired with Cosine are found in the UMLS. However results are in general much better with Jaccard. Quite similar results are observed between both similarity measures when generalisation is perfomed with all three linguistic methods at a time (VSM/ALL3). Using the three methods at the same time for generalisation does not provide better results. Indeed, it even decreases the number of relations found in the UMLS. As for generalisation/normalisation, it allows to decrease the number of relations acquired by two when Cosine is used, that is good because the number of relations acquired with Cosine is extremely high, but it also divides the number of relations found in the UMLS by two.

When terminological variation is individually used, it always improves the quality of the results in terms of MAP, precision and R-precision for Cosine, whereas it always decreases the quality with Jaccard. The normalisation with synonyms with both similarity measures behaves similarly to generalisation with terminological variation: they both get the best recall.

With Jaccard, generalisation with lexical inclusion improves the MAP results, that means that the relations are better ranked. Lexical inclusion improves the results when used individually or within a combination. Lexico syntactic patterns with a large window have little (with Cosine) or negative impact on the results. Lexical inclusion allows to increase the MAP, R-precision and precision values when used after the lexico syntactic patterns with Cosine.

Finally, we can conclude that with a large window, the use of Jaccard and generalisation with lexical inclusion improves the quality of the relations acquired. But the recall is also really low. With Cosine, the recall is higher and generalisation with LI is also the best choice.

	Acquired Rel.		Rel. in UMLS		MAP		R-precision		P@1	
	JACC	COS	JACC	COS	JACC	COS	JACC	COS	JACC	COS
VSMonly (baseline)	1,882	16,178	6	60	0.502	0.118	0.333	0.054	0.333	0.048
VSM/TV	2,258	13,804	16	56	0.276	0.110	0.143	0.051	0.143	0.051
VSM/LI	976	6,172	2	38	0.536	0.132	0.500	0.067	0.500	0.067
VSM/LSP	2,112	12,656	16	50	0.187	0.106	0.071	0.057	0.071	0.057
VSM/TV+LI	976	6,172	2	38	0.536	0.132	0.500	0.067	0.500	0.067
VSM/TV+LSP	2,066	12,338	16	50	0.191	0.106	0.071	0.057	0.071	0.057
VSM/LSP+LI	934	5,996	4	38	0.378	0.135	0.250	0.067	0.250	0.067
VSM/ALL3	1,002	6,540	4	38	0.379	0.131	0.250	0.067	0.250	0.067
VSM/Syn	2,292	14,022	16	56	0.273	0.110	0.143	0.051	0.143	0.051

Table 3: Results obtained for a 5 word window– with thresholds on the distributional parameters

Narrow window With the 5 word window, the observations and results also differ according to the similarity measure used. The best results are also obtained with the Jaccard Index and the behaviour of both similarity measures is similar to the one observed with a large window: generalisation with lexical inclusion reduces by two the number of relations acquired but also the number of relations found in the UMLS. In order to better understand the behaviour of generalisation with lexical inclusion, and to improve the results in terms of recall without decreasing precision, manual evaluation is required. The results obtained with Cosine are lower than with a large window.

The choice of the similarity measure is a difficult choice and is linked to the other distributional parameters of the VSM. We can deduce that Jaccard with small corpora allows to get a better precision than Cosine, and obtains better results with a narrow window. Generalisation with lexical inclusion appears to be the best generalisation for both measures, and normalisation with synonymy relations does not improve the results.

But when LI is combined with relations acquired with lexico-syntactic patterns, its contribution decreases the results with the Jaccard index, and on the contrary improves the results with the Cosine. The order of the methods also matters and differs according to the similarity measure: with Jaccard, the generalisation with relations acquired with lexical inclusion before lexico-syntactic patterns has a lower precision than the inverse combination (i.e. VSM/LSP+LI).

7 Conclusion

In this paper, we address the reduction of data sparsity in matrices of context vectors used to implement the distributional analysis. We proposed to generalise and normalise the distributional contexts with synonymy and hypernymy relations acquired from our corpus. Words in contexts are considered as hyponyms and are replaced by hypernyms identified on the corpus, or are considered as members of a synonym set, and normalised with the cluster representative of this set. We performed some experiments on a French medical corpus combining several parameters. Even if the evaluation of distributional methods is difficult, we compare the results to the semantic relations proposed by the French UMLS. Several evaluation measures have been used to evaluate the impact of context generalisation and normalisation on distributional analysis. The analysis of the results show that when the size of the window that allow to produce distributional contexts is small and when the Jaccard index is used, it is better to generalise contexts with relations acquired with lexical inclusion. However, when the window is large, generalisation with lexical inclusion with the use of Jaccard index also improves the results. Normalisation seems to have no positive effect on relation extraction, with any combination of distributional parameters.

Beside a manual analysis of the relations and of the impact of the process of generalisation and normalisation on manipulated data, these results open several perspectives. The hypernymy relations we used have been separately exploited. But these relations could be considered as a sketch towards a taxonomy and we plan to adapt the context generalisation method in order to consider this network of relations acquired from corpora. Furthemore, all the relations acquired from corpora may be noisy. We plan to use other sources of relations as the ones contained in terminologies. It could then be possible to evaluate the impact of generalisation and of the relations when their terminological type is known. Finally, we plan to

compare our method with two other dimension reduction methods, such as Random Indexing and LSA.

References

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

O. Baskaya, E. Sert, V. Cirik, and D. Yuret. 2013. Ai-ku: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Proc. of SemEval - 2013*, pages 300–306, Atlanta, USA. ACL.

B. Broda, M. Piasecki, and S. 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.

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

N. Chatterjee and S. Mohan. 2008. Discovering word senses from text using random indexing. In *Proc. of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'08, pages 299–310, Berlin, Heidelberg. Springer-Verlag.

J.R. Curran and M. Moens. 2002. Improvements in automatic thesaurus extraction. In *Workshop on Unsupervised lexical acquisition*, volume 9, pages 59–66, Morristown, NJ, USA. Association for Computational Linguistics.

J. R. Curran. 2004. *From distributional to semantic similarity*. Ph.D. thesis, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

R. Fano. 1963. *Transmission of Information: A Statistical Theory of Communications*. The MIT Press, Cambridge, MA.

O. Ferret. 2010. Similarit smantique et extraction de synonymes partir de corpus. In *TALN 2010*, Montral.

O. Ferret. 2011. Utiliser l'amorage pour améliorer une mesure de similarité sémantique. In Mathieu Lafourcade and Violaine Prince, editors, *TALN 2011*, volume 1, pages 155–160, Montpellier, France, juillet.

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.

N. Grabar and P. Zweigenbaum. 2003. Lexically-based terminology structuring. In *Terminology*, volume 10, pages 23–54.

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

T. Hamon, A. Nazarenko, and C. Gros. 1998. A step towards the detection of semantic variants of terms in technical documents. In *International Conference on Computational Linguistics (COLING-ACL'98)*, pages 498–504, Université de Montréal, Quebec, Canada.

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.

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

C. Jacquemin. 2001. *Spotting and discovering terms through natural language processing*. The MIT Press.

P. Kanerva, J. Kristofersson, and A. Holst. 2000. Random indexing of text samples for latent semantic analysis. In L.R. Gleitman and A.K. Josh, editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, volume 1036, Erlbaum, New Jersey.

J. Karlgren and M. Sahlgren. 2001. From words to understanding. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 294–308. Foundations of Real-World Intelligence.

T.K. Landauer and S.T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, 104(2):211.

D.D. Lee and H.S. Seung. 1999. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791.

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

H. Moen, E. Marsi, F. Ginter, L.-M. Murtola, T. Salakoski, and S. Salanterä. 2014. Care episode retrieval. In *Proc. of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 116–124, Gothenburg, Sweden. ACL.

E. Morin and C. Jacquemin. 2004. Automatic Acquisition and Expansion of Hypernym Links. *Computers and the Humanities*, 38(4):363–396.

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

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

T. Polajnar and S. Clark. 2014. Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of EACL 2014*. To appear.

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

M. 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 Univ., Sweden.

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

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

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

P.D. Turney and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

A. Utsumi. 2010. Evaluating the performance of nonnegative matrix factorization for constructing semantic spaces: Comparison to latent semantic analysis. In *Proceedings of SMC*, pages 2893–2900. IEEE.

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

E. Vozalis and K. G. Margaritis. 2003. Analysis of recommender systems' algorithms. In *The 6th Hellenic European Conference on Computer Mathematics & its Applications (HERCMA), Athens, Greece*.

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

D. 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.

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

P. Zweigenbaum. 1994. Menelas: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, 45.