Compounds and Distributional Thesauri

Olivier Ferret
CEA, LIST, Vision and and Content Engineering Laboratory
Gif-sur-Yvette, F-91191 France
olivier.ferret@cea.fr

Abstract

The building of distributional thesauri from corpora is a problem that was the focus of a significant number of articles, starting with (Grefenstette, 1994) and followed by (Lin, 1998), (Curran and Moens, 2002) or (Heylen and Peirsman, 2007). However, in all these cases, only single terms were considered. More recently, the topic of compositionality in the framework of distributional semantic representations has come to the surface and was investigated for building the semantic representation of phrases or even sentences from the representation of their words. However, this work was not done until now with the objective of building distributional thesauri. In this article, we investigate the impact of the introduction of compounds for achieving such building. More precisely, we consider compounds as undividable lexical units and evaluate their influence according to three different roles: as features in the distributional contexts of single terms, as possible neighbors of single term entries and finally, as entries of a thesaurus. This investigation was conducted through an intrinsic evaluation for a large set of nominal English single terms and compounds with various frequencies.

Keywords: Lexical semantics, distributional thesaurus, compounds

1. Introduction

Distributional semantic resources are used in a comprehensive set of tasks, ranging from relation extraction (Min et al., 2012) to machine translation (Marton, 2013) or question-answering (van der Plas, 2008). This article focuses more particularly on the automatic building of distributional thesauri from corpora. This problem, which was already tackled by much work for single terms, such as (Grefenstette, 1994), (Lin, 1998), (Curran and Moens, 2002) or (Heylen and Peirsman, 2007), has received less attention for compounds. From a more global perspective, the study of compounds from a semantic viewpoint is far from being an unexplored land. A significant part of the work in this area aims at interpreting the semantic relations that underlie the components of a compound (Girju et al., 2007). Another one, which is more recent, takes place in the distributional paradigm and, following the semantic compositionality hypothesis, aims at building the distributional representation of compounds, or larger units (Grefenstette et al., 2011), from the distributional representation of their components (Mitchell and Lapata, 2008; Guevara, 2011). The verification of the validity of the compositionality hypothesis for a particular compound or phrase was also addressed by some work (Biemann and Giesbrecht, 2011; Schulte im Walde et al., 2013) and strongly exploited in work about bilingual terminology extraction, especially from comparable corpora (Morin and Daille, 2012). While all this work focuses on the semantic analysis of compounds, the behavior of compounds taken as lexical units does not seem to have been much investigated in a distributional perspective. This is the problem we consider in this article.

More precisely, starting from the framework set up in (Ferret, 2010), we have studied the impact of the use of nominal compounds in the building of distributional thesauri of nouns from three different perspectives:

- compounds as features: following work such as (Curran and Moens, 2002), each entry of a distributional thesaurus is characterized by a vector of co-occurrents. As entries, these co-occurrents, which can be viewed as features, are generally single terms. We have extended (Ferret, 2010) by taking also nominal compounds as features;
- compounds as neighbors: in (Ferret, 2010) as in previous similar work, a distributional thesaurus associates each entry with a set of semantic neighbors, which are single terms. In this work, we have tested the use of nominal compounds as possible neighbors;
- compounds as entries: conversely, we have introduced nominal compounds as entries of a distributional thesaurus.

We first describe briefly our initial framework in the next section. Then, we say a few words about the way we identify compounds in texts before reporting the results of the three studies mentioned above.

2. Initial framework

Our work is based on the framework defined in (Ferret, 2010) for building and evaluating a state-of-the-art distributional thesaurus. This framework is set on the AQUAINT-2 corpus, a middle-size corpus made of around 380 million words coming from news articles in English. The preprocessing of this corpus was deliberately restricted to part-of-speech tagging and lemmatization, performed with *TreeTagger* (Schmid, 1994), to make possible the transposition of the framework to a large set of languages. Concerning the extraction of distributional data and the characteristics of the measure we have used for building our thesauri, we have adopted the parameters selected by (Ferret, 2010) from the results of an extended version of the TOEFL test proposed in (Freitag et al., 2005):

- distributional contexts made of graphical co-occurrents: co-occurrents collected in a fixed-size

freq.	ref.	#eval. words	#syn. / word	recall	R-prec.	MAP	P@1	P@5	P@10	P@100
all	W	10,473	2.9	24.6	8.2	9.8	11.7	5.1	3.4	0.7
# 14,670	M	9,216	50.0	9.5	6.7	3.2	24.1	16.4	13.0	4.8
high	W	3,690	3.7	28.3	11.1	12.5	17.2	7.7	5.1	1.0
# 4,378	M	3,732	69.4	11.4	10.2	4.9	41.3	28.0	21.9	7.9
middle	W	3,732	2.6	28.6	10.4	12.5	13.6	5.8	3.7	0.7
# 5,175	M	3,306	41.3	9.3	6.5	3.1	18.7	13.1	10.4	3.8
low	W	3,051	2.3	11.9	2.1	3.3	2.6	1.2	0.9	0.3
# 5,117	M	2,178	30.1	2.8	1.2	0.5	2.5	1.5	1.5	0.9

Table 1: Evaluation of thesaurus building for and with single terms only

window centered on each occurrence of the target word. These co-occurrents were restricted to content words, *i.e.* nouns, verbs and adjectives, whose frequency was higher than 10 in the AQUAINT-2 corpus;

- size of the window = 1 word before and 1 word after the target word, *i.e.* very short range co-occurrents;
- lenient filtering of contexts: removal of co-occurrents with only one occurrence;
- weighting function of co-occurrents in contexts = *Pointwise Mutual Information* between the target word and the co-occurrent;
- similarity measure between contexts, for evaluating the semantic similarity of 2 words = *Cosine* measure.

The building of a distributional thesaurus is performed straightforwardly: the selected similarity measure is computed between the target noun and each of its possible neighbors. These neighbors are then ranked in the decreasing order of the values of this measure and the first N ($N = 100$ here) neighbors are kept as the semantic neighbors of the target noun. Both entries and possible neighbors were made of the AQUAINT-2 nouns whose frequency was higher than 10.

Table 1 shows the results of the evaluation of this building process when single terms only are used as features, entries and neighbors. This evaluation was achieved by comparing the extracted semantic neighbors with two complementary reference resources: WordNet 3.0 synonyms [W] (Miller, 1990), which characterizes a semantic similarity based on paradigmatic relations, and the Moby thesaurus [M] (Ward, 1996), which gathers a larger set of types of relations and is more representative of *semantic relatedness*. The 4^{th} column of Table 1, which gives the average number of synonyms and similar words in our references for the AQUAINT-2 nouns, also illustrates the difference of these two resources in terms of richness. As our main objective is to evaluate the extracted semantic neighbors and not to rebuild the reference resources, these resources were filtered to discard entries and synonyms that are not part of the AQUAINT-2 vocabulary (see the difference between the number of words in the 1^{st} column and the number of evaluated words of the 3^{rd} column). In distributional approaches, the frequency of words related to the size of the corpus is an important factor. Hence, we give our results globally but also for three ranges of frequencies that split our vocabulary into roughly equal parts[1]: *high* frequency nouns (freq. > 1000), *middle* frequency nouns ($100 <$ freq. ≤ 1000) and *low* frequency nouns ($10 <$ freq. ≤ 100).

These results take the form of several measures and start at the 5^{th} column by the proportion of the synonyms and similar words of our references that are found among the first 100 extracted neighbors of each noun. As these neighbors are ranked according to their similarity value with their target word, the evaluation measures can be taken from Information Retrieval by replacing documents with synonyms and queries with target words (see the three last columns of Table 1). The R-precision (R-prec.) is the precision after the first R neighbors were retrieved, R being the number of reference synonyms; the Mean Average Precision (MAP) is the average of the precision value after a reference synonym is found; precision at different cut-offs is given for the 1, 5, 10 and 100 first neighbors. All these values are given as percentages.

Two main observations arise from these results. First, their level heavily depends on the frequency of entries: the best results are obtained for high frequency entries while the measures significantly decrease for low frequency entries. Second, the differences of the reference resources both in terms of their number of neighbors by entry and the type of their relations have a significant impact on results: precisions at different cut-offs have a significantly higher value with Moby, which provides a large set of various kinds of neighbors, than with WordNet, which only gives a restricted set of synonyms, while MAP and R-precision show an opposite tendency.

3. Identification of compounds

The first step of our study was the modification of the preprocessing of the AQUAINT-2 corpus to identify nominal compounds in documents. This identification was done as a controlled indexing process: a reference set of nominal compounds was first built and then used to identify compounds in documents according to a longest-match strategy

[1] The exact number of entries for each range is given in the first column as #number.

freq.	ref.	recall	R-prec.	MAP	P@1	P@5	P@10	P@100
all	W	24.8	8.2 (0.0)	9.8 (0.0)	11.5 (-0.2)	5.2 (0.1)	3.4 (0.0)	0.7 (0.0)
	M	9.5	6.6 (-0.1)	3.2 (0.0)	24.0 (-0.1)	16.3 (-0.1)	12.8 (-0.2)	4.8 (0.0)
high	W	29.1	11.4 (0.3)	12.8 (0.3)	17.4 (0.2)	8.0 (0.3)	5.2 (0.1)	1.1 (0.1)
	M	11.4	10.2 (0.0)	4.9 (0.0)	41.7 (0.4)	28.0 (0.0)	21.8 (-0.1)	7.9 (0.0)
middle	W	28.3	10.0 (-0.4)	12.2 (-0.3)	13.2 (-0.4)	5.6 (-0.2)	3.6 (-0.1)	0.7 (0.0)
	M	9.0	6.2 (-0.3)	3.0 (-0.1)	18.1 (-0.6)	12.7 (-0.4)	9.9 (-0.5)	3.7 (-0.1)
low	W	11.3	1.9 (-0.2)	3.1 (-0.2)	2.2 (-0.4)	1.1 (-0.1)	0.8 (-0.1)	0.3 (0.0)
	M	2.8	1.2 (0.0)	0.5 (0.0)	2.5 (0.0)	1.5 (0.0)	1.4 (-0.1)	0.8 (-0.1)

Table 2: Evaluation of thesaurus building for single term entries and single/multi-term features

but without taking into account possible ambiguities. This reference set was defined as follows: first, two reference lexicons of nominal compounds were merged; then, the resulting set was filtered to keep only compounds whose frequency is higher than 10 in the AQUAINT-2 corpus. One of these two lexicons was WordNet 3.0 and the other one was the DELAC dictionary (Silberztein, 1999). The first one contains 60,292 nominal compounds while the second one gathers 64,479 nominal compounds. These two resources have only 14,618 compounds in common, which shows that each lexicon or dictionary generally covers only a small part of the large number of possible compounds. Finally, 33,003 compounds were kept after the frequency filtering step. The large majority of these compounds (around 74%) have one of the three following structures:

[1] <noun>$_{modifier}$ <noun>$_{head}$
[2] <adjective>$_{modifier}$ <noun>$_{head}$
[3] <noun>$_{head}$ <preposition><noun>$_{modifier}$

Typical examples for each these three structures are:

[1] cruise ship, movie industry, oil refinery, knowledge base, lecture room
[2] visual cortex, medical instrument, commercial law, educational program
[3] sense of duty, director of research, cash in hand, commission on human rights

4. Compounds as features for single term entries

As we consider compounds in this work as indivisible lexical units, similarly to single terms, the process for building distributional thesauri described in Section 2. was not modified by the introduction of nominal compounds. The only consequence is the presence of these compounds in the distributional contexts of words and their presence as neighbors or entries in the thesauri.

We first study the impact of the presence of nominal compounds in the distributional contexts of words. This presence has two main consequences: the size of the contexts tends to grow, which is difficult to analyze in terms of effect, and their new elements are less ambiguous than single terms, which should be *a priori* a positive factor. The study is summed up by Table 2, in which evaluated entries and reference thesauri are identical to those of Table 1. Each evaluation measure is given with the difference with the corresponding value in Table 1 (in brackets). Globally, Table 2 shows that nominal compounds do not have a major impact, as features, on the building of distributional thesauri. Moreover, it shows that this impact is positive for high-frequency entries and for neighbors that are synonyms of these entries. It is negative for average-frequency entries, whatever the kind of relations with their neighbors, and tends to be neutral for low-frequency entries. Moreover, this global low impact is confirmed by the fact that the results of the similarity measure used for building the thesaurus of Table 2 to the extended TOEFL test of (Freitag et al., 2005) are nearly identical (accuracy equal to 71.7 vs. 71.6) to those of the measure of Table 1. One explanation of this low impact is the fact that our reference nominal compounds only represent 8.7% of occurrences of all content terms in the AQUAINT-2 corpus.

5. Compounds as neighbors for single term entries

The second step of our study is to introduce nominal compounds as possible neighbors of single term entries of a distributional thesaurus while they already appear in distributional contexts of words. More precisely, this is done in our case by adding, for each entry of the thesaurus, our set of reference compounds to the list of words whose similarity with this entry is computed.

The results of this introduction are given in Table 3. They clearly show that adding compounds as possible neighbors of entries leads to lower results in all cases. This could be considered as surprising: compounds tend to be semantically less ambiguous than single terms and we might expect that they would have been easier to find as neighbors, especially when they are synonyms of the target entry. However, the addition of a large set of candidate neighbors for each entry seems to add a significant number of semantically unrelated words among its highest-ranked neighbors. Moreover, this effect has clearly a larger negative impact for similar words (see Moby's results) than for synonyms, probably because of the limited semantic ambiguity of compounds.

freq.	ref.	recall	R-prec.	MAP	P@1	P@5	P@10	P@100
all	W	21.0	7.1 (-1.1)	8.4 (-1.4)	10.6 (-1.1)	4.7 (-0.4)	3.1 (-0.3)	0.6 (-0.1)
	M	7.8	5.5 (-1.2)	2.5 (-0.7)	21.3 (-2.8)	14.5 (-1.9)	11.3 (-1.7)	4.1 (-0.7)
high	W	25.2	10.1 (-1.0)	11.2 (-1.3)	16.4 (-0.8)	7.7 (0.0)	5.0 (-0.1)	1.0 (0.0)
	M	10.0	9.0 (-1.2)	4.1 (-0.8)	38.4 (-2.9)	26.1 (-1.9)	20.4 (-1.5)	7.3 (-0.6)
middle	W	23.9	8.7 (-1.7)	10.8 (-1.7)	12.1 (-1.5)	5.1 (-0.7)	3.2 (-0.5)	0.6 (-0.1)
	M	6.6	4.7 (-1.8)	2.1 (-1.0)	14.9 (-3.8)	10.2 (-2.9)	8.0 (-2.4)	2.9 (-0.9)
low	W	8.2	1.6 (-0.5)	2.3 (-1.0)	1.9 (-0.7)	0.9 (-0.3)	0.6 (-0.3)	0.2 (-0.1)
	M	1.6	0.7 (-0.5)	0.3 (-0.2)	1.5 (-1.0)	1.1 (-0.4)	0.9 (-0.6)	0.5 (-0.4)

Table 3: Evaluation of thesaurus building for single term entries and single/multi-term features and neighbors

nuclear reactor	reactor [0.47], nuclear plant [0.35], nuclear power [0.29], research reactor [0.28], nuclear fuel [0.27], atomic reactor [0.24], weapons-grade [0.23], plutonium [0.22], enriched uranium [0.22] ...
stock exchange	stock market [0.32], index [0.30], share price [0.28], bourse [0.28], blue chip [0.27], new york stock exchange [0.27], share [0.25], trading [0.24], stock [0.24], stock index [0.23] ...
religious belief	religion [0.25], faith [0.22], belief [0.20], religious faith [0.18], freedom of religion [0.17], religious freedom [0.15], viewpoint [0.15], christianity [0.15], constitutional right [0.14] ...
academic program	low density [0.17], step aerobics [0.17], urban studies [0.17], miles-per-gallon [0.16], palisade [0.16], retirement-plan [0.16], alicia alonso [0.16], mutant gene [0.15], graduate program [0.15] ...

Table 4: First neighbors of some entries of the distributional thesaurus with compounds as entries

6. Compounds as thesaurus entries

The last part of our study does not take the results for single terms as reference but focuses specifically on nominal compounds as thesaurus entries. As a consequence, the sets of reference synonyms and similar words are different. Table 5 shows more specifically that WordNet is a significantly richer reference for nominal compounds than Moby (3rd and 4th columns). It also illustrates the fact that nominal compounds have less synonyms and similar words than single terms, probably because they are semantically less ambiguous.

Table 4 provides a qualitative view of this distributional thesaurus of nominal compounds by giving the first neighbors of some of its entries with their similarity value with their entry. The first three rows are illustrative of entries with rather good neighbors while the last row illustrates the fact that results are of course far from being perfect for a significant number of entries[2]. Table 4 also shows that the neighbors of the compound entries are rather balanced between single terms and compounds.

From a more quantitative viewpoint, although our three ranges of frequencies do not split our vocabulary of nominal compounds into well-balanced sets as for single terms, the results of Table 5 can be soundly compared to those of Table 1 and show two main trends. First, the synonyms of compounds are far easier to find than the synonyms of single terms. This is true in terms of both recall and measures such as R-precision and MAP, which means that these findings are not only explained by the small number of reference synonyms. They probably result from the nature of nominal compounds as lexical units with limited semantic ambiguity. Second, similar words are more difficult to find

[2] This is not the worst case as the first relevant neighbor, *graduate program*, appears at the 9th position.

freq.	ref.	#eval. words	#syn. / word	recall	R-prec.	MAP	P@1	P@5	P@10	P@100
all	W	6,350	1.6	29.6	8.8	11.6	10.7	4.3	2.7	0.5
# 16,154	M	906	19.8	6.4	2.8	1.6	6.3	4.8	3.6	1.3
high	W	584	1.8	48.9	21.9	26.0	27.2	10.3	6.3	0.9
# 1,173	M	104	19.4	18.5	9.7	5.9	22.1	17.3	12.5	3.6
middle	W	2,176	1.7	43.8	15.2	19.6	18.8	7.4	4.5	0.8
# 5,103	M	401	20.6	7.3	3.2	1.8	7.7	5.3	4.2	1.5
low	W	3,590	1.6	16.5	2.8	4.3	3.2	1.5	1.0	0.3
# 9,878	M	401	19.2	2.2	0.6	0.3	0.7	0.9	0.7	0.4

Table 5: Evaluation of thesaurus building for multi-term entries and single/multi-term features and neighbors

for compounds than for single terms. In this case, it is difficult to determine the origin of these findings as Moby is far richer for single term entries than for compounds and (Ferret, 2010) clearly shows that this kind of difference has a great influence on results. As a consequence, it is likely that a significant part of the low results of compounds for similar words is due to the sparsity of Moby for that kind of entries.

7. Conclusion

In this work, we have studied the impact of nominal compounds on the building of distributional thesauri at three levels. As features for single terms, compounds do not globally have a strong impact but tends to favor synonyms for high-frequency entries. As candidate neighbors of single terms, they seem to bring more noise than meaningful neighbors. However, this phenomenon is probably not specific to compounds and happens when the set of candidate neighbors is large. Finally, the significance of nominal compounds in the building of distributional thesauri mainly appears through their role of entries as they tend to have more high-rank relevant neighbors than single terms, especially for synonyms.

One extension of this work might be to rely on this property to improve the quality of the neighbors of single terms. Conversely, methods for improving distributional thesauri made of single terms only, such as (Ferret, 2012) or (Ferret, 2013), could be tested when entries and neighbors include compounds. Finally, an analysis of our results according to the type of compounds we consider in terms of compositionality, through methods such as (Reddy et al., 2011), could be also interesting. In terms of resources, the most direct outcome of this study is a distributional thesaurus for English nominal compounds, called A2ST-COMP and available on demand.

8. References

Biemann, C. and Giesbrecht, E. (2011). Distributional semantics and compositionality 2011: Shared task description and results. In *Workshop on Distributional Semantics and Compositionality*, pages 21–28, Portland, Oregon, USA, June.

Curran, J. R. and Moens, M. (2002). Improvements in automatic thesaurus extraction. In *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 59–66, Philadelphia, USA.

Ferret, O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Ferret, O. (2012). Combining bootstrapping and feature selection for improving a distributional thesaurus. In *20th European Conference on Artificial Intelligence (ECAI 2012)*, pages 336–341, Montpellier, France.

Ferret, O. (2013). Identifying bad semantic neighbors for improving distributional thesauri. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 561–571, Sofia, Bulgaria.

Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., and Wang, Z. (2005). New experiments in distributional representations of synonymy. In *Ninth Conference on Computational Natural Language Learning (CoNLL)*, pages 25–32, Ann Arbor, Michigan, USA.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). Semeval-2007 task 04: Classification of semantic relations between nominals. In *Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic.

Grefenstette, E., Sadrzadeh, M., Clark, S., Coecke, B., and Pulman, S. (2011). Concrete sentence spaces for compositional distributional models of meaning. In *9th International Conference on Computational Semantics (IWCS 2011)*.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

Guevara, E. R. (2011). Computing semantic compositionality in distributional semantics. In *Ninth International Conference on Computational Semantics (IWCS 2011)*, pages 135–144, Oxford, UK.

Heylen, K. and Peirsman, Y. (2007). Automatic thesaurus extraction: Comparing context models. In *18th Computational Linguistics in the Netherlands (CLIN 2007)*, Nijmegen.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, pages 768–774, Montrï¿½al, Canada.

Marton, Y. (2013). Distributional phrasal paraphrase generation for statistical machine translation. *ACM Transactions on Intelligent Systems and Technology*, 4(3):1–32.

Miller, G. A. (1990). WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4).

Min, B., Shi, S., Grishman, R., and Lin, C.-Y. (2012). Ensemble semantics for large-scale unsupervised relation extraction. In *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 1027–1037, Jeju Island, Korea, July.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *ACL-08: HLT*, pages 236–244, Columbus, Ohio.

Morin, E. and Daille, B. (2012). Revising the compositional method for terminology acquisition from comparable corpora. In *24th International Conference on Computational Linguistics (COLING 2012)*, pages 1797–1810, Mumbai, India.

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 210–218, Chiang Mai, Thailand.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*.

Schulte im Walde, S., Müller, S., and Roller, S. (2013). Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *2nd*

Joint Conference on Lexical and Computational Semantics, pages 255–265, Atlanta, GA.

Silberztein, M. (1999). Les groupes nominaux productions et les noms composés lexicalisés. *Linguisticæ Investigationes*, 27(2):405–426.

van der Plas, L. (2008). *Automatic lexico-semantic acquisition for question answering*. Ph.D. thesis, University of Groningen.

Ward, G. (1996). Moby thesaurus. Moby Project.