Disambiguating Distributional Neighbors using a Lexical Substitution Dataset

François Morlane-Hondère Cécile Fabre Nabil Hathout Ludovic Tanguy

CLLE-ERSS: CNRS & University of Toulouse, France
{firstname.lastname}@univ-tlse2.fr

Abstract. This paper addresses the issue of polysemy in a distributional thesaurus. In such resources, distributional neighbors can relate indistinguishably to various senses. We propose a method to cluster the neighbors of a target word with respect to its senses, i.e. to attribute one sense to each neighbor. This is made possible by the use of a lexical substitution dataset, to which the distribution of the neighbors are compared.

Keywords: distributional semantics, word sense disambiguation, lexical substitution

1 Introduction

Many NLP applications need to know whether a word A is semantically more related to B than to C. Unsupervised corpus-based approaches to similarity have been widely used over the past 20 years. They are usually based on the *distributional hypothesis* (Harris, 1954): semantically related words tend to share many of their contexts (and vice versa). These approaches first collect information about the contexts in which words appear in a corpus, then measure the degree of relatedness between their distributions. The output is a *distributional thesaurus*, i.e. a list of word pairs – or *neighbors* – rated by a similarity score (Lin, 1998; Mohammad and Hirst, 2006). *Distributional thesauri* have become very popular in a wide range of NLP tasks (Weeds, 2003; Baroni and Lenci, 2010) because of the ever-increasing availability of textual data. However, their potential is still far from being fully exploited because of the fuzziness of the concept of *semantic relatedness* (Padó and Lapata, 2003).

In this paper, we address the issue of polysemy in distributional thesauri. For example, a polysemic word like *mouse* appears in different contexts according to its senses: *eat*_OBJ or *hunt*_OBJ in the ANIMAL sense and *plug*_OBJ or *cursor*_COMP in the DEVICE sense. These contexts are all mixed in the context vector of the word *mouse* when a large non-specific corpus is used. Therefore, the neighbors of *mouse* refer to both senses as in the following example taken from the Distributional Memory of Baroni and Lenci (2010), with the similarity score indicated for each word (in decreasing order):

> *rat* (0.542), *animal* (0.466), *rabbit* (0.449), **cursor** (0.440), *monkey* (0.424), *cat* (0.421), *pig* (0.413), **joystick** (0.384), *dog* (0.375), *human* (0.373)...

The example shows that the ANIMAL sense is dominant in ukWaC (McCarthy et al., 2004) (the large web-based corpus from which the distributional data was extracted): it is referred to by eight out of the ten nearest neighbors of *mouse*; only *cursor* and *joystick* refer to the DEVICE sense.

Semantic relatedness can vary in both nature and strength. Wordnet-like resources focus on synonymy and hypernymy, but sense disambiguation can be extended to other semantic relations (such as, for the word *mouse*, meronymy: *fur* vs *cord*; co-hyponymy: *rat* vs *keyboard*) and general semantic relatedness (*cheese* vs *scrollbar*). We specifically address this question and show that disambiguation can be extended from standard synonymy to semantically more distant words using a distributional approach.

For that purpose, we designed an experiment whose setting is illustrated in figure 1. The objective is to cluster the neighbors of a target word with respect to its senses, i.e. to attribute one sense to each neighbor. For each sense of the target word we have a list of manually-validated similar words, more precisely words that have been proposed as suitable substitutes for the target word in different contexts (from the Semdis task, see section 3). The relatedness of a neighbor with a sense is thus estimated by the similarity of the former with the substitutes for the latter. The selection of the sense a neighbor is related to is performed in two steps :

Ⓐ The neighbor is compared to the substitutes of each of the target word senses.
Ⓑ The sense with the most similar substitutes is selected.

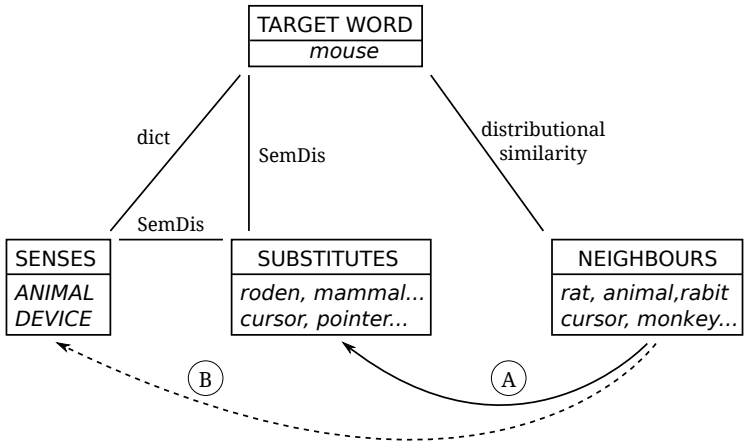

Fig. 1. Method overview

The experiment described here has been performed on French, but can of course be easily ported to any language, provided that similar data are available.

The remainder of the paper is structured as follows: the task is described in Section 2. In Section 3, we focus on the sets of data we have used in this experiment. We

then present the way we compute the similarity between the neighbors and the senses of the target words (Section 4). The details of the experiment are described in Section 5. We then give the results in Section 6.

2 The task

Sorting out polysemy among pairs of distributional neighbors can be considered in relation to the word sense disambiguation (WSD) task (Agirre and Edmonds, 2006). The purpose of WSD is to identify the sense of a polysemous word used in a given context. Our approach differs from traditional WSD in that word disambiguation relies on its surrounding context as in (1), while what we want to do instead is to relate the distributional neighbors of a word to its different senses, as in (2).

(1) A little [mouse] is running around the room.

(2) *rat* (ANIMAL), *mammal* (ANIMAL), *rabbit* (ANIMAL), *cursor* (DEVICE)

WSD systems usually require two components. First, they need an inventory of the potential senses of the words. Most approaches rely on repositories such as WordNet (Fellbaum, 1998). However, when available, wordnets have been found to be too fine-grained to be useful (Ide and Véronis, 1998; Navigli, 2009), although efforts were made to make coarser the level of granularity of their sense distinctions (McCarthy, 2006). Then, most WSD systems need corpora. In many approaches, a sense-tagged corpus – like SemCor (Miller et al., 1994) – is used to train a classifier. Because no such a resource exists for French, we used an untagged corpus composed of French Wikipedia articles.

Our approach consists in computing a *second-order affinity* measure (Grefenstette, 1994) between the substitutes and the words for which we want to characterize the meaning. In the above example, we expect a higher similarity score between the vectors of *rabbit* and the substitutes for the sense ANIMAL than between *rabbit* and DEVICE. Karov and Edelman (1998) took a comparable approach, except that the meaning vectors – *feedback sets* – of their target words are made up of the words extracted from dictionary definitions. They call this method *similarity-based disambiguation*. For other examples of the use of second-order context vectors for WSD, see Schütze (1998) and Pucci et al. (2009).

In Section 6.2 we compare this approach to a first-order affinity baseline. This approach is used by Yarowsky (1995) where the substitutes are seed words which the author considered representative of each target word sense. Both these approaches have been evaluated against a gold standard consisting in a manual annotation of each considered neighbor, for which the semantically closer meaning has been decided (see Section 5).

3 The data

In this section, we present the datasets we have used in the experiment, namely a set of lexical substitutes grouped by senses, and a set of distributional neighbors for 11 target words.

3.1 The *SemDis 2014* substitution dataset

The dataset used in this study was created as a gold standard for a French lexical substitution task presented at the SemDis 2014 workshop (Fabre et al., 2014). This task was inspired by the SemEval 2007 English lexical substitution task (McCarthy and Navigli, 2009). Therefore, their goals and settings are quite similar: participants were asked to develop systems that give the best substitutes for a set of 30 target words – 10 nouns, 10 verbs and 10 adjectives – each contextualized in 10 sentences. The target words are polysemous words, chosen in a dictionary, and the sentences were picked in the FrWaC corpus (Baroni et al., 2009). Each sentence is associated to one of the word senses.The evaluation is performed by comparing the substitutes proposed by the participants' systems to a gold standard provided by the manual annotation of the sentences. The annotators – 7 different annotators for each sentence – were asked to produce single word substitutes for the target words in the sentence. From 0 to 3 substitutes could be given; for the 300 sentences, a total of 3961 substitutions were gathered (of which 1098 unique substitutes). The overall inter-annotator agreement was similar to what had been observed for this kind of task (25 % average pairwise agreement).

Below is an example of the resulting list of substitutes for the word *affection* 'affection', which may refer to either ILLNESS or LOVE. 5 sentences for each sense were proposed, and the corresponding substitutes were then merged, along with calculating the number of cases in which an annotator proposed each substitute.

- *affection* as ILLNESS: *maladie* (22), *trouble* (9), *pathologie* (7), *mal* (5), *dysfonctionnement* (3), *atteinte* (2), *problème* (2), *anomalie* (2), *condition* (1)...
- *affection* as LOVE: *amour* (26), *tendresse* (16), *attachement* (6), *amitié* (4), *sentiment* (4), *attention* (2), *lien* (2), *intimité* (1), *sollicitation* (1), *proximité* (1)...

This dataset is very limited in terms of coverage, but it has a number of advantages over more traditional dictionary- or wordnet-based data:

- the substitutes are synonyms in a broader sense of the relation;
- the semantic relatedness is valued (number of annotators);
- all the substitutes are contemporary, commonly used words;
- the different senses are easily distinguished from one another.

The gold standard data is freely available from the SemDis workshop website[1].

In this study, we chose to first focus on a subset of 11 out of the 30 SemDis target words, those for which only two senses have been identified:

- 4 nouns: *affection* (LOVE vs ILLNESS), *débit* (DEBIT vs OUTPUT), *don* (TALENT vs PRESENT), *montée* (RISE vs ASCENT)
- 4 verbs: *entraîner* (TRAIN vs LEAD TO), *fonder* (CREATE vs BASE ON), *interpréter* (UNDERSTAND vs PERFORM), *maintenir* (HOLD vs ASSERT)
- 3 adjectives[2]: *aisé* (EASY vs RICH), *mince* (THIN vs POOR), *riche* (RICH vs WEALTHY)

[1] `http://www.irit.fr/semdis2014/fr/task1.html`

[2] There were only 3 two-senses adjectives in the 30 target words

3.2 The distributional thesaurus

We use a distributional thesaurus generated from a 262-million words corpus of French Wikipedia articles. The thesaurus contains the nouns, verbs and adjectives which occur in at least 5 different contexts in the corpus. Syntactic dependencies were used as contexts using the Talismane parser (Urieli and Tanguy, 2013). The weighting of the context relations was made using the pointwise mutual information and the cosine measure was used to compute the similarity between the context vectors.

Distributional information is used in two ways:

- it provides a list of its distributional neighbors for each of the 11 target words;
- it gives access to syntagmatic information in the Wikipedia corpus, corresponding to the syntactic contexts that characterize the substitutes and the neighbors (Section 2).

4 Method

In this section, we describe the way we compute the similarity between the neighbors and the senses of the target words. As mentioned in Section 2, for each target word, we compare the distribution of each neighbor with the overall distribution of the set of substitutes associated to each sense of this word.

So, what we first need to do is to build the context vectors of each neighbor as well as the context vectors of the 2 senses of each target word – or *sense vectors*. These are computed as follows:

1. we build a semantic space whose dimensions are the syntactic contexts of the target word in the Wikipedia corpus;
2. we compute the vectors of each substitute in this space, the values being the positive pointwise mutual information[3] (PPMI) between the substitute and a context;
3. we average the vectors of the substitutes of each sense to obtain a vector per sense. The weight for each substitute is the number of times it was proposed by the annotators, thus favoring the more common and natural substitutes.

In a second step, the cosine measure is used to compute the similarity between the vector of each neighbor and the two sense vectors. For example, we can see in table 1 that the correct sense systematically gets higher similarity scores (in bold).

5 Experiment

5.1 Gold standard for neighbor disambiguation

To evaluate the ability of our method to disambiguate the target words, we performed a manual annotation of the 10 nearest neighbors for each of the 11 SemDis target words we defined in Section 3. Neighbors that were also substitutes were excluded to avoid trivial configurations. For the purpose of this annotation, the two senses of each target

[3] A variant of the PMI where negative values are replaced by zero.

Table 1. Results of the similarity measure for the neighbors of the target word *affection*.

	senses	
	LOVE	ILLNESS
complication 'complication'	0.110	**0.590**
lésion 'lesion'	0.086	**0.744**
sympathie 'sympathy'	**0.680**	0.062
admiration 'admiration'	**0.776**	0.079
infection 'infection'	0.126	**0.691**
tumeur 'tumor'	0.052	**0.520**
symptôme 'symptom'	0.228	**0.542**
estime 'esteem'	**0.300**	0.024
manifestation 'manifestation'	0.134	**0.603**
épilepsie 'epilepsy'	0.083	**0.324**

word were arbitrarily assigned a number – *1* or *2* – and the most frequently proposed substitute was used as a gloss for each sense (like LOVE vs ILLNESS for the word *affection*).

In the annotation guidelines, we considered three options: a neighbor can be related to the first or the second sense of the target word (it will be annotated *1* or *2*), to both senses[4] (*0*) or to none (*-1*) when there is no identifiable semantic relation between the neighbor and the target word. Four judges have annotated all 110 neighbors. Overall inter-annotator agreement for this task, as measured by Fleiss' kappa, was 0.23, which traditionally indicates a "fair agreement". A majority vote was then applied to the four sets to produce a single annotation (only a few cases had to be decided through negotiation). Of the 110 neighbors, 77 were annotated *1* or *2*, 11 were annotated *0* and 22 were annotated *-1*.

5.2 Rule-based decision system

We designed a very simple rule-based decision system based on the similarity scores between the input neighbor and the two possible senses (see Section 4). Obviously, the higher similarity score should indicate the correct sense; however we also have the two other decisions to take into account. Our hypothesis is that if the two similarity values are too close, it can indicate either a relation with both senses (if both values are high), either no relation (if both values are low).

We therefore need to base our decision on two variables, by comparing them to threshold values:

- the ratio between the two similarity values (higher score divided by lower): (threshold σ_r);
- the lowest of the two similarity values (threshold σ_{low}).

[4] Mostly when the neighbor is morphologically related to the target word, like in *interpréter>interprétation*.

This rule is formalized in algorithm 1, where H is the sense for which the similarity value (sim_H) is higher and L (sim_L) the lower.

Algorithm 1 Decision rule:

```
if sim_H / sim_L > σ_r then
    output = sense H (1 or 2)
else
    if sim_L > σ_low then
        output = both senses (0)
    else
        output = no sense (-1)
    end if
end if
```

Both threshold values were estimated through a brute-force approach, and the highest accuracy was obtained with $\sigma_r = 1.08$ and $\sigma_{low} = 0.52$.

6 Results

6.1 Qualitative analysis of the results

The simple rule-based system presented in the previous section obtained an accuracy of 0.64 over the 110 neighbors. Details can be found in the confusion matrix (table 2). We can see that most of the errors are due to confusions with the categories *No sense* and *Both senses*.

Table 2. Confusion matrix for the rule-based system

Gold \ System	No sense	Both senses	Sense 1	Sense 2	Total
No sense	5	0	2	15	22
Both senses	0	1	8	2	11
Sense 1	1	0	36	0	37
Sense 2	4	3	4	29	40
Total	10	4	50	46	110

29When merging values for senses 1 and 2 (the distinction between the two categories being arbitrary) we reach a precision of 0.68, a recall of 0.84 and an f1-score of 0.75. As can be seen, there are only 4 cases out of 110 where the system chose one sense and the annotators chose the other.

We also found that the performance was rather unequal among the 11 target words. For example, the 10 neighbors of the word *affection* 'affection' were all related to their

correct sense whereas this was only the case for 4 of the 10 neighbors of *fonder* 'found'. This is due to the fact that the two senses of the latter are more related than the senses of the former. This can be demonstrated by looking at the sense vectors of these two words. We reported in table 3 the 10 contexts with the higher PPMI for the two meaning vectors of the words *affection* and *fonder*. We can see that the contexts in which the substitutes related to the meaning ILLNESS appear are clearly related to the medical terminology, which strongly contrasts with the contexts related to the meaning LOVE. This distinction is far from being as clearly marked for *fonder*. We can see that most of the contexts of the meanings CREATE and BASE ON refer to abstract concepts (the context *critère* 'criterion' is even present in the two sense vectors).

Table 3. Extracts of the meaning vectors of the target words *affection* and *fonder*.

affection 'affection'		*fonder* 'found'	
ILLNESS	LOVE	CREATE	BASE ON
*neurologique*_A	*éprouver*_V	*fait*_N	*occasion*_N
'neurologic'_A	'feel'_V	'fact'_N	'occasion'_N
*système nerveux*_NP	*particulier*_A	*principe*_N	*critère*_N
'nervous system'_NP	'particular'_A	'principle'_N	'criterion'_N
*grave*_A	*profond*_A	*hypothèse*_N	*guerre mondial*_NP
'severe'_A	'deep'_A	'hypothesis'_N	'world war'_NP
*diagnostic*_N	*grand*_A	*affirmation*_N	*XIXe siècle*_NP
'diagnosis'_N	'large'_A	'statement'_N	'XIXth century'_NP
*souffrir*_V	*vouer*_V	*critère*_N	*remplacement*_N
'suffer'_V	'vow'_V	'criterion'_N	'replacement'_N
*rénal*_A	*témoigner*_V	*théorie*_N	*emplacement*_N
'renal'_A	'show'_V	'theory'_N	'place'_N
*mental*_A	*sentiment*_N	*idée*_N	*modèle*_N
'mental'_A	'feeling'_N	'idea'_N	'model'_N
*soigner*_V	*manifester*_V	*décision*_N	*nom de institut*_NP
'cure'_V	'show'_V	'decision'_N	'institute name'_NP
*chronique*_A	*lien*_N	*témoignage*_N	*rive*_N
'chronic'_A	'link'_N	'gesture'_N	'bank'_N
*oculaire*_A	*paternel*_A	*raison*_N	*même époque*_NP
'ocular'_A	'paternal'_A	'reason'_N	'same era'_NP

6.2 Comparison with a baseline

In order to get a better evaluation of our method, we designed a baseline system in which second-order (distributional) similarity is replaced with first-order (cooccurrence). In other words, we used a similarity value between each neighbor and each substitute by calculating the PPMI based on their cooccurrence in a Wikipedia article (regardless of their frequency in the articles themselves). This PPMI score was then averaged as before to get a similarity score for both senses of the target words. This similarity values were

then processed by the same rule-based system described above, with its own pair of optimal threshold values.

This baseline reaches an overall accuracy of 50.91 %, and a f1-score of 0.60 for the "1 sense" target category. Both these scores are significantly lower than those obtained by our previous system.

7 Conclusion

The question of polysemy among the distributional neighbors addressed in this paper is part of a larger research effort concerning the description of the results provided by distributional semantics methods. The use of a lexical substitution dataset is an interesting alternative to the traditional wordnets and dictionaries, although more difficult to obtain in most cases. The idea of using the substitutes of a given sense to generate context vectors, and then to rely on these vectors to disambiguate the neighbors gives encouraging results considering the two modalities *0* and *-1*, usually absent from WSD tasks. Our method also sheds light on the fact that some meaning vectors have more discrimination power than others. A cosine measure between the meaning vectors of a given word could be helpful to measure its degree of polysemy (to merge the meanings of a fine-grained resource like WordNet, for example).

Bibliography

Agirre, E. and Edmonds, P. (2006). *Word Sense Disambiguation*. Springer.

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

Baroni, M. and Lenci, A. (2010). Distributional memory: a general framework for corpus-based semantics. *Comput. Linguist.*, 36(4):673–721.

Fabre, C., Hathout, N., Ho-Dac, L.-M., Morlane-Hondère, F., Muller, P., Sajous, F., Tanguy, L., and Van de Cruys, T. (2014). Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In *Actes de l'atelier SemDis*.

Fellbaum, C., editor (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Grefenstette, G. (1994). *Corpus-derived First, Second, and Third-order Word Affinities*. Rank Xerox Research Centre.

Harris, Z. S. (1954). Distributional structure. *Word*.

Ide, N. and Véronis, J. (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1).

Karov, Y. and Edelman, S. (1998). Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–59.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

McCarthy, D. (2006). Relating wordnet senses for word sense disambiguation. In *Proceedings of the ACL Workshop on Making Sense of Sense*, pages 17–24.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. A. (2004). Finding predominant word senses in untagged text. In *ACL*, pages 279–286.

McCarthy, D. and Navigli, R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. (1994). Using a semantic concordance for sense identification. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 240–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mohammad, S. and Hirst, G. (2006). Distributional measures as proxies for semantic distance: A survey. *Computational Linguistics*, 1(1).

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69.

Padó, S. and Lapata, M. (2003). Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Sapporo, Japan. Association for Computational Linguistics.

Pucci, D., Baroni, M., Cutugno, F., and Lenci, A. (2009). Unsupervised lexical substitution with a word space model. In *Proceedings of EVALITA 2009*.

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

Urieli, A. and Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 188–201, Les Sables d'Olonne, France.

Weeds, J. (2003). *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, Department of Informatics, University of Sussex.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.