TALN 2011, Montpellier, 27 juin – 1er juillet 2011

FreDist : Automatic construction of distributional thesauri for French

Enrique Henestroza Anguiano & Pascal Denis
Alpage, INRIA Paris-Rocquencourt & Université Paris 7
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
{henestro, pascal.denis}@inria.fr

Résumé. Dans cet article, nous présentons FreDist, un logiciel libre pour la construction automatique de thésaurus distributionnels à partir de corpus de texte, ainsi qu'une évaluation des différents ressources ainsi produites. Suivant les travaux de (Lin, 1998) et (Curran, 2004), nous utilisons un corpus journalistique de grande taille et implémentons différentes options pour : le type de relation contexte lexical, la fonction de poids, et la fonction de mesure de similarité. Prenant l'EuroWordNet français et le WOLF comme références, notre évaluation révèle, de manière originale, que c'est l'approche qui combine contextes linéaires (ici, de type bigrammes) et contextes syntaxiques qui semble fournir le meilleur thésaurus. Enfin, nous espérons que notre logiciel, distribué avec nos meilleurs thésaurus pour le français, seront utiles à la communauté TAL.

Abstract. In this article we present FreDist, a freely available software package for the automatic construction of distributional thesauri from text corpora, as well as an evaluation of various distributional similarity metrics for French. Following from the work of (Lin, 1998) and (Curran, 2004), we use a large corpus of journalistic text and implement different choices for the type of lexical context relation, the weight function, and the measure function needed to build a distributional thesaurus. Using the EuroWordNet and WOLF wordnet resources for French as gold-standard references for our evaluation, we obtain the novel result that combining bigram and syntactic dependency context relations results in higher quality distributional thesauri. In addition, we hope that our software package and a joint release of our best thesauri for French will be useful to the NLP community.

Mots-clés : thésaurus distributionnel, similarité sémantique, méthodes non supervisées, lexique.

Keywords: distributional thesaurus, semantic similarity, unsupervised methods, lexicon.

1 Introduction

We present FreDist, software that implements methods for the automatic construction of distributional thesauri. Distributional lexical resources are appealing because they can be constructed automatically from raw text corpora, and are useful for alleviating data sparseness in many NLP applications (e.g. parsing and coreference resolution). Moreover, we believe that open software like FreDist can be useful to the NLP community by providing an easy way to generate distributional thesauri from any text corpus using adjustable settings.

We base our work on that of (Lin, 1998), which uses word context relations to calculate lexical distributional similarity, and the subsequent work of (Curran, 2004), which distinguishes between weight and measure functions and evaluates different functions on a semantic similarity task for English. We build on their work by considering the joint use of different types of context relations, and evaluating distributional similarity metrics for French.

Current lexical resources for French that have been semi-automatically created include the work of (Sagot, 2010) on the Le*fff*, a large-coverage morphosyntactic lexicon, and (Sagot & Fišer, 2008) on the WOLF, a semantic resource based on the Princeton WordNet. Our work differs by providing a fully automatic approach to the creation of a lexical resource. Previous work on distributional methods for French includes that of (Bourigault, 2002) on UPERY, a distributional analysis module that calculates proximities between words and their contexts, and the work of (Ferret, 2004), which uses distributional similarity to build word senses from a network of lexical co-occurrences. Our work differs by focusing on the construction and evaluation of distributional thesauri, combining different types of context relations, and making FreDist and our best distributional thesauri freely-available. [1]

1. `http://alpage.inria.fr/~henestro/fredist.html`

véhicule	voiture−0.546, camion−0.401, engin−0.301, camionnette−0.291, bus−0.276, fourgon−0.271, avion−0.269, appareil−0.254, tracteur−0.249, moto−0.248, fourgonnette−0.242, train−0.234, automobile−0.234, bateau−0.227, scooter−0.225, matériel−0.225, berline−0.224 ...
tragique	dramatique−0.331, cruel−0.269, douloureux−0.260, terrible−0.237, triste−0.230, sanglant−0.188, traumatisant−0.179, fatal−0.174, funeste−0.173, regrettable−0.173, effroyable−0.171, fâcheux−0.162, spectaculaire−0.160, violent−0.159, étrange−0.159, drôle−0.158, inquiétant−0.155 ...

FIGURE 1 – Distributional thesaurus entries for the noun *véhicule* and the adjective *tragique*.

Section 2 explains distributional lexical methods, Section 3 describes FreDist and the construction of our distributional thesauri, Section 4 evaluates different distributional similarity metrics, and we conclude in Section 5.

2 Distributional methods

The distributional hypothesis states that words occurring in the same contexts tend to have similar meanings, as posited by (Harris, 1954). We focus on methods that generate distributional thesauri from a large collection of lexical terms and contexts. An entry in a distributional thesaurus contains, for each lexical term, a list of neighboring terms ordered by similarity. Example entries are shown in Figure 1.

2.1 Context relations

Basing our terminology on the work of (Lin, 1998) and of (Curran, 2004), we define a *context relation* as the tuple (w, r, w'), where w is a primary lexical term (we use lemmas) that occurs in a particular context ; in our work, contexts consist of a relation r and a secondary lexical term w'. Commonly-used contexts for w include syntactic dependencies, fixed-size windows (such as bigrams), and bag-of-words representations of documents. The choice of context dictates the semantic relationship obtained between primary lexical terms : (Agirre *et al.*, 2009) find that syntactic dependencies and fixed-size windows best capture *semantic similarity* (synonymy and hypernymy/hyponymy), while bag-of-words approaches capture broader *semantic relatedness* (particularly shared topic). We use syntactic dependencies and linear bigrams, since we are interested in semantic similarity. If a relation r is bigram, it can take values of either -1 or $+1$, indicating that w' appears either before or after w. If r is syntactic, it can take values of either gov or dep, indicating that w either governs w' or is a dependent of w' : e.g. the context relation $(aboyer, gov, chien)$. After context relations are extracted from a corpus, each primary lexical term w can be represented as a frequency vector $v^w \in \mathbb{R}^d$, where d is the number of unique contexts appearing in the corpus, and $v^w_i = freq(w, r, w')$, where i corresponds to the context $c_i = (r, w')$.

2.2 Term similarity

Term similarity metrics are used to calculate similarities between pairs of primary lexical terms w_1 and w_2 using frequency vectors v^{w_1} and v^{w_2}. (Curran, 2004) breaks term similarity metrics down into two components : a *weight* function transforms the raw frequency of each context relation by determining the informativeness of the context, while a *measure* function subsequently calculates the similarity between two weighted frequency vectors.

We experiment with the following weight functions for a context relation (w, r, w') : relative frequency (RELFREQ), which normalizes the frequency of (w, r, w') with respect to the frequency of its primary lexical term w ; t-test (TTEST), where the null hypothesis states that the primary lexical term w and the context (r, w') are independent, and the test compares their product distribution to their observed joint distribution ; and the pointwise mutual information function (PMI), which calculates the mutual information between w and (r, w').

We experiment with the following measure functions for a pair of primary lexical terms w_1 and w_2 and their respective weighted vectors v^{w_1} and v^{w_2} : cosine similarity (COSINE), which measures the cosine of the angle between v^{w_1} and v^{w_2} ; the Jaccard measure (JACCARD), which compares the number of common contexts to the number of unique contexts between w_1 and w_2 ; and the similarity measure that (Lin, 1998) uses (LIN), which is an information theoretic measure to determine the similarity between w_1 and w_2.

RELFREQ	$\frac{f(w,r,w')}{f(w,*,*)}$
TTEST	$\frac{p(w,r,w')-p(*,r,w')p(w,*,*)}{\sqrt{p(w,r,w')/f(*,*,*)}}$
PMI	$log\left(\frac{f(w,r,w')}{f(*,r,w')f(w,*,*)}\right)$

TABLE 1 – Weight functions

COSINE	$\frac{\sum wgt(w_1,*_r,*_{w'})\times wgt(w_2,*_r,*_{w'})}{\sqrt{\sum wgt(w_1,*,*)^2\times\sum wgt(w_2,*,*)^2}}$
JACCARD	$\frac{\sum min(wgt(w_1,*_r,*_{w'}),wgt(w_2,*_r,*_{w'}))}{\sum max(wgt(w_1,*_r,*_{w'}),wgt(w_2,*_r,*_{w'}))}$
LIN	$\frac{\sum wgt(w_1,*_r,*_{w'})+wgt(w_2,*_r,*_{w'})}{\sum wgt(w_1,*,*)+\sum wgt(w_2,*,*)}$

TABLE 2 – Measure functions

The formulas for the weight and measure functions used in our experiments are listed in Tables 1 and 2. The symbol $*$ as an argument to a function is shorthand for taking the sum of the function over all possible values for that argument ; p denotes the probability of a context relation, where $p(w, r, w')$ is estimated as $f(w, r, w')/f(*, *, *)$; and wgt denotes the application of some weight function to a context relation count.

3 Resource construction

In order to generate distributional thesauri for French, we introduce the FreDist package for the Python programming language. FreDist provides functionality for each step of distributional thesauri construction : (1) Extraction of context relations from a text corpus in CONLL [2] format ; (2) Weighting of context relations according to a specified weight function ; (3) Generation of a similarity matrix for primary lexical terms according to a specified measure function ; (4) Construction of a distributional thesaurus from a similarity matrix. FreDist is highly flexible, with parameters including : context relation type(s), weight function, measure function, term frequency thresholding, part-of-speech (POS) restrictions, filtering of numerical terms, etc.

We now discuss the steps we took to build our distributional thesauri for French. We first describe the preprocessing used to tag and parse our chosen raw corpus. We then describe the settings we used for FreDist to extract context relations and perform similarity calculation.

3.1 Corpus preprocessing

We chose to use the freely-available *L'Est Républicain* corpus, which contains 125 million words of French journalistic text. The corpus was first preprocessed using simple tokenization and sentence segmentation tools. POS tagging was then conducted using MElt [3], a freely-available POS tagger for French. Subsequently, we performed lemmatization and morphological analysis using the previously mentioned Le*fff* [4] : the Le*fff* was queried with a word+POS pair to obtain a corresponding lemma (in case of ambiguity, the first lemma was chosen) as well as a set of morphological features. The corpus was then parsed with the MaltParser [5], a fast and highly accurate system for data-driven dependency parsing.

Both MElt and MaltParser were trained on the standard training section of the French Dependency Treebank (FTB) as described in (Candito *et al.*, 2010), which is based on the original French Treebank with constituent structure (Abeillé & Barrier, 2004). The FTB contains a total of 12,531 sentences from the *Le Monde* newspaper, and the annotation scheme contains 28 unique tags in its POS tagset. On the FTB development set, MElt obtains 97.7% tagging accuracy and MaltParser obtains 89.3% unlabeled attachment accuracy.

2. http://nextens.uvt.nl/depparse-wiki/DataFormat
3. http://gforge.inria.fr/frs/download.php/26999/melt-0.5.tar.gz
4. http://alpage.inria.fr/~sagot/lefff.html
5. http://www.maltparser.org. We used version 1.3.1, with features including lemma and morphological information.

3.2 Context relation extraction

Once the corpus had been automatically annotated with lemmas, POS tags, and syntactic dependency relations between words, we extracted context relations. Although the base lexical term used in distributional lexical methods is often the inflected form or the lemma, we chose to use a base lexical term consisting of lemma+POS to distinguish between homonyms (as in *dîner+noun* vs. *dîner+verb*) within contexts.

We extracted context relations exclusively for primary lexical terms w that had a POS tag of adjective, adverb, common noun, or verb, and that appeared at least 100 times in the corpus. A frequency threshold of 100 is often used in the literature, and is applied because distributional similarity methods are known to suffer degraded performance for terms that appear infrequently in a corpus (Gorman & Curran, 2006). Secondary lexical terms w' were also subject to the POS tag restriction, and contexts (r, w') were subject to the frequency thresholding. Additionally, lexical terms containing numbers were replaced with a *num* token. Primary lexical terms above the frequency threshold totaled 4,126 adjectives, 802 adverbs, 10,997 common nouns, and 3,562 verbs.

Bigram context relations were generated in a straightforward manner. For each token of a primary lexical term w in the corpus, we placed it in the relation -1 with the preceding token's lexical term (or a generic term *beg* if w was the first token in a sentence) and in the relation $+1$ with the subsequent token's lexical term (or a generic term *end* if w was the last token in a sentence). Unique bigram contexts above the frequency threshold totaled 6,680 for adjectives, 3,935 for adverbs, 6,134 for common nouns, and 6,436 for verbs.

Syntactic context relations were extracted in accordance with the FTB annotation style. For each occurrence of a primary lexical term w in the corpus, we placed it in a relation r with its governing lexical term (or a generic term *root* if w rooted a dependency tree), where r is the dependent relation *dep*. Then for each dependent of w, we placed w in a relation r with that dependent, where r is the governor relation *gov*.[6] Some dependencies were collapsed : for prepositional/coordinating phrases, we chose to fold the preposition/conjunction into r in order to include both the head of the prepositional/coordinated phrase and the head of the modified/preceding phrase. Some dependencies were ignored : none involving a punctuation mark was included, due to the underspecification of punctuation attachment in the FTB. Unique syntactic contexts above the frequency threshold totaled 6,389 for adjectives, 3,141 for adverbs, 33,881 for common nouns, and 25,624 for verbs.

Note the particularly large number of syntactic contexts extracted for common nouns and verbs. This is perhaps indicative of participation in rich and varied long-range dependencies for common nouns and verbs, as opposed to more local dependencies for adjectives and adverbs that can be largely accounted for with bigram contexts.

3.3 Similarity calculation

For each of the four relevant POS categories, we applied weight and measure functions to its collection of context relations and pairs of primary lexical terms. The resulting thesauri provided the basis for our evaluation. Since we wanted to test each combination of weight and measure functions, as well as three settings for context relations (bigram, syntactic, bigram+syntactic), we generated in total 27 test thesauri for each relevant POS category.

4 Evaluation

In order to evaluate our distributional thesauri, we used two wordnets for French as references : the French Euro WordNet (FREWN)[7], which is manually validated, and the WOLF[8], which is not manually validated. While the FREWN covers only verbs and common nouns, the WOLF covers all four relevant POS categories. We thus evaluated verbs and common nouns using FREWN, and adjectives and adverbs using WOLF.

During evaluation, we considered only those primary lexical terms appearing in both the distributional thesauri and the wordnet reference. This reduction was carried out in order to prevent unnecessary penalization of the thesauri due to potential incompleteness of the wordnet references. This gave us 3,018 common nouns and 1,426 verbs for the FREWN evaluation, and 374 adjectives and 195 adverbs for the WOLF evaluation.

6. Dependency labels were not used in our evaluation for computational efficiency reasons, but the option is available in FreDist.
7. `http://www.illc.uva.nl/EuroWordNet`
8. `http://alpage.inria.fr/~sagot/wolf.html`

Nouns (FREWN)	
Setting	INVR
bigram+syntactic, PMI, COSINE	0.282
syntactic, PMI, COSINE	0.281
bigram, PMI, COSINE	0.273
syntactic, TTEST, COSINE	0.266
syntactic, TTEST, JACCARD	0.260
bigram+syntactic, TTEST, JACCARD	0.259
syntactic, PMI, JACCARD	0.259
bigram+syntactic, TTEST, COSINE	0.259
bigram+syntactic, PMI, JACCARD	0.259
linear, PMI, JACCARD	0.250

Verbs (FREWN)	
Setting	INVR
bigram+syntactic, PMI, COSINE	0.345
syntactic, PMI, COSINE	0.334
syntactic, TTEST, COSINE	0.332
syntactic, TTEST, JACCARD	0.330
bigram+syntactic, TTEST, JACCARD	0.322
bigram+syntactic, PMI, JACCARD	0.317
syntactic, PMI, JACCARD	0.312
bigram+syntactic, TTEST, COSINE	0.308
bigram, PMI, COSINE	0.297
syntactic, TTEST, LIN	0.281

Adjectives (WOLF)	
Setting	INVR
bigram+syntactic, PMI, COSINE	0.403
syntactic, PMI, COSINE	0.397
bigram+syntactic, PMI, JACCARD	0.373
syntactic, PMI, JACCARD	0.372
bigram, PMI, COSINE	0.348
bigram+syntactic, TTEST, JACCARD	0.342
bigram, PMI, JACCARD	0.338
syntactic, TTEST, JACCARD	0.330
bigram+syntactic, PMI, LIN	0.319
bigram+syntactic, TTEST, LIN	0.317

Adverbs (WOLF)	
Setting	INVR
bigram+syntactic, PMI, COSINE	0.548
bigram+syntactic, PMI, JACCARD	0.522
bigram, PMI, COSINE	0.520
bigram+syntactic, TTEST, COSINE	0.519
syntactic, PMI, COSINE	0.517
syntactic, TTEST, COSINE	0.502
bigram, PMI, JACCARD	0.494
bigram+syntactic, TTEST, JACCARD	0.491
syntactic, PMI, JACCARD	0.490
bigram, TTEST, COSINE	0.485

TABLE 3 – Average INVR evaluation scores for the top 10 distributional thesauri (out of 27) by POS category. Each setting name is a combination of the context relation type (bigram, syntactic, or bigram+syntactic), weight function (RELFREQ, TTEST, or PMI), and measure function (COSINE, JACCARD, or LIN).

For our evaluation metric, we chose to use average inverse rank (INVR), a standard information retrieval metric. For each term w, we considered all terms appearing in a synset with w in the wordnet reference to be $relevant$, while other terms were considered $irrelevant$. In the distributional thesaurus to be evaluated, the entry for w was treated as a ranked list of query results (neighbor terms ranked by descending similarity). The INVR metric returns the sum, over relevant neighboring terms, of the inverse of that term's rank in the list. The average INVR is taken over all terms to be evaluated, providing an evaluation metric for the quality of a distributional thesaurus. One downside to this evaluation approach is that we ended up evaluating our distributional thesaurus using strict synonymy, which ignores pairs of words that may be otherwise semantically similar. Due to the scarcity of appropriate and/or manually validated resources for evaluating distributional thesauri, we believe that a synset-based evaluation is nonetheless useful and allows for a comparison of the relative quality of different thesauri settings.

The top five scoring distributional thesauri for each wordnet+POS pairing are shown in Table 3. Our primary finding is that the combination of bigram+syntactic context relations, PMI weight function, and COSINE measure function consistently produces the best thesauri across all four POS categories (though the results were not tested for significance). The finding that PMI and COSINE outperform other combinations is consistent with a recent comparison of distributional similarity metrics for English (Ferret, 2010), although in that work a TOEFL test was used for evaluation. A possible explanation for PMI weight function performing best is that RELFREQ does not weight contexts on the basis of informativeness, and TTEST has the downside of erroneously assuming normal distributions for the probabilities of particular context relations appearing in a corpus (Church & Mercer, 1993). COSINE simply appears to work better empirically than the other measure functions. The finding that bigram and syntactic contexts are most effective when used together is remarkable : to our knowledge, the combination of different types of contexts has not been explored before. A possible explanation is that bigram and syntactic contexts provide different views of the distribution of lexical terms, resulting in better informed similarity estimates.

Finally, we also looked at the running time of FreDist. We used the best setting (bigram+syntactic contexts, PMI weight function, and COSINE measure function), starting from a parsed corpus and ending with distributional thesauri for the four relevant POS categories. When the input contained $1/2$ of the *L'Est Républicain* corpus (62.5 million words) the cpu time was just under 8 hours. For the full corpus (125 million words) the cpu time was just over 18 hours. Trials were run on a Linux machine with a 2.4GHz processor and 8GB of memory.

5 Conclusion

We have presented FreDist, a freely-available software package that implements lexical distributional methods for the automatic construction of distributional thesauri from text corpora. FreDist is highly customizable and can be used with any type of text corpus in a number of different languages. Its running time is currently acceptable for medium-sized corpora of up to a few hundred million words (with the particular settings we used in our evaluation), but we hope to optimize the code in future versions in order to handle larger corpora of 1 billion words or more. In addition to the software, we are also releasing distributional thesauri for French that were created using the best settings from our evaluation.

A second goal of this work was to evaluate distributional methods on French data, and we have obtained results similar to those of past work on English : our finding that the PMI weight function and COSINE measure function work best for French mirrors the results of (Ferret, 2010) for English. We also experimented with a novel approach that involved joint consideration of bigram and syntactic context relations, and found that it works better than either type of context relation on its own. Given this interesting result, an avenue for future work might be to include additional context relation types, such as document or paragraph co-occurrence, and determine if they can help further improve the quality of our distributional thesauri.

Acknowledgements

We would like to thank Marie Candito for helpful discussions. This work was partially funded by the ANR project Sequoia ANR-08-EMER-013.

References

ABEILLÉ A. & BARRIER N. (2004). Enriching a french treebank. In *Proceedings of LREC '04*, Lisbon, Portugal.

AGIRRE E., ALFONSECA E., HALL K., KRAVALOVA J., PAŞCA M. & SOROA A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL '09*, p. 19–27, Boulder, Colorado.

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de TALN '02*, p. 75–84, Nancy, France.

CANDITO M., CRABBÉ B. & DENIS P. (2010). Statistical french dependency parsing : Treebank conversion and first results. In *Proceedings of LREC '10*, Valetta, Malta.

CHURCH K. & MERCER R. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, **19**(1), 1–24.

CURRAN J. (2004). *From distributional to semantic similarity*. PhD thesis, University of Edinburgh.

FERRET O. (2004). Discovering word senses from a network of lexical cooccurrences. In *Proceedings of COLING '04*, p. 1326–1332, Geneva, Switzerland.

FERRET O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. In *Actes de TALN '10*, Montreal, Quebec.

GORMAN J. & CURRAN J. (2006). Scaling distributional similarity to large corpora. In *Proceedings of COLING-ACL '06*, p. 361–368, Sydney, Australia.

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

LIN D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL '98*, p. 768–774, Montreal, Quebec.

SAGOT B. (2010). The lefff, a freely available, accurate and large-coverage lexicon for french. In *Proceedings of LREC '10*, Valetta, Malta.

SAGOT B. & FIŠER D. (2008). Building a free french wordnet from multilingual resources. In *Proceedings of Workshop on OntoLex*, p. 14–19, Marrakech, Morocco.