Using distributional thesauri for sorting synonyms

François MORLANE-HONDÈRE
CLLE-ERSS, University of Toulouse, France
francois.morlane@univ-tlse2.fr

1 - Introduction

A well-known way to assess the semantic similarity between distributional neighbours is to compare distributional data with gold standards. In this study, we compared two distributional thesauri and a French dictionary of synonyms (the *Electronic Dictionary of synonyms*, or *EDS* ; Manguin 2002).

The results show that the nature of the synonyms extracted for a given word varies between the two thesauri. Thus, we assume that automatically generated distributional thesauri can be used to improve the sorting of the synonyms contained in the EDS.

2 - The distributional model

Corpora

The two distributional thesauri we used were respectively constructed from two French corpora:

- a 262 million word corpus of Wikipedia articles ;
- a 200 million word corpus of newspaper (*Le Monde*).

Extraction of the distributional neighbours

1. both corpora are parsed using the syntactic analyzer Syntex (Bourigault 2007) ;
2. word contexts are weighted using mutual information ;
3. the program Upéry (Bourigault 2002) calculates the similarity between the context vectors using Lin's score (> 0,1).

Distributional thesauri

- the *Wikipedia's Distributional Neighbours* (WDN)
  - 3 922 657 neighbours
- the *Le Monde's Distributional Neighbours* (LMDN)
  - 5 525 480 neighbours

Examples from the WDN

neighbours	common contexts	score
voiture/automobile *car/automobile*	coffre_DE, accident_DE, voler_OBJ… *trunk_COMP, accident_COMP, steal_OBJ…*	0,161
choisir_OBJ/sélectionner_OBJ *choose_OBJ/select_OBJ*	option, candidat, cible… *option, candidate, target…*	0,196
calme/tranquille *calm/quiet*	endroit_MOD, eau_MOD, vie_MOD… *place_MOD, water_MOD, life_MOD…*	0,208
courage/bravoure *courage/bravery*	combattre_AVEC, preuve_DE, admirer_OBJ… *fight_WITH, proof_OF, admire_OBJ…*	0,427

3 - Comparison with the EDS

We measure the overlap between the two distributional thesauri and the dictionary of synonyms by considering only the words shared by the compared resources. The results show that **the overlap between the thesauri and the dictionary of synonyms is low :**

41,6 % and 33,7 % of the synonyms of the EDS are distributional neighbours in the WDN and the LMDN thesauri

This raises some questions :

- why do most of the synonyms of the EDS are not distributional neighbours ?
- what kind of synonyms are not substitutable in our corpora ?
- are those synonyms the same in both corpora ?

4 - Analysis of the results

In order to investigate the questions raised above, we consider a sample of 30 nouns, 30 verbs and 30 adjectives :

- with at least 10 synonyms in the EDS ;
- for which synonyms were not extracted as distributional neighbours.

We take a corpus linguistic approach. Based on our analysis of the contexts of the synonyms in the corpora, we were able to identify several reasons explaining why the synonyms of our sample do not share enough contexts to be extracted as distributional neighbours. One of them is the polysemy. We illustrate this case below with the word *fraternité* (*brotherhood*).

Polysemy : *fraternité* (*brotherhood*) means both (1) "a community of people" and (2) "the relationship between brothers".

- Contexts in the Wikipedia corpus : *rejoindre*_OBJ (*join*_OBJ), *membre*_DE (*member*_OF), *fondateur*_DE (*founder*_OF)…
- Distributional neighbours : *phalange* (*phalanx*), *parti ouvrier* (*workers party*), *congrégation* (*congregation*)…

The distribution of *fraternité* (*brotherhood*) shows that only the first sense of the word is expressed in the Wikipedia corpus : it cannot be substituted with synonyms like *solidarité* (*solidarity*) or *générosité* (*generosity*).

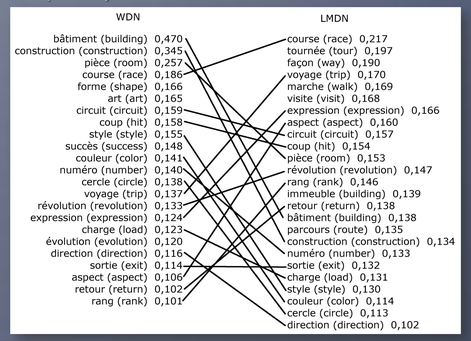
The results show that the corpus-based information provided by the distributional thesauri allows us to filter the synonyms of a given word according to the way the word is used in text.

5 - Sorting the synonyms

Based on the previous results, we wanted to see how much the *selection* of the synonyms varies depending on the distributional thesaurus. The results show that the synonyms selected by the WDN and the LMDN are mostly the same. However, **we assume that the similarity score of the synonyms in a distributional thesaurus is a good indicator of the relevance of the synonyms of a word for a given corpus.**

Indeed, we can see below that the ranking of the synonyms of a highly polysemic word like *tour* – which, in French, means *tower*, *circuit*, *rook*, *walk*, *trick*… – differs highly according to the distributional thesaurus :

- the most similar synonyms of *tour* in the WDN are *bâtiment* (*building*) and *construction* (*construction*), which refer to the sense *tower*. These synonyms are ranked way lower in the LMDN, which means that this sense of *tour* is less frequent in the newspaper corpus ;
- in the LMDN, the most similar synonyms of *tour* tend to refer to the action of moving : *course* (*race*), *tournée* (*tour*), *voyage* (*trip*), *marche* (*walk*)…



6 - Conclusion

Distributional neighbours reflect the way words behave in texts. In this study, we showed the potential benefits of adding corpus-based distributional information to a dictionary of synonyms. We assume a correlation between the distributional similarity score of two synonyms and their relevance regarding users' needs. However, this hypothesis has yet to be proven.

References

BOURIGAULT D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle*, p. 75–84, Nancy.
BOURIGAULT D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire d'habilitation à diriger des recherches. Université Toulouse II
MANGUIN J.-L. (2002). Le dictionnaire électronique des synonymes du CRISCO. In *Colloque Sciences humaines et nouvelles technologies*, Tunis.
MORLANE-HONDÈRE F. (2013), Utiliser une base distributionnelle pour filtrer un dictionnaire de synonymes, In *Actes de SemDis 2013*, Les Sables d'Olonne, 21 juin 2013.