FreDist: Automatic construction of distributional thesauri for French

Enrique Henestroza Anguiano & Pascal Denis
Alpage, INRIA Paris-Rocquencourt & Université Paris 7

Contributions

- A software package (FreDist) that implements methods for the automatic construction of distributional thesauri from raw or parsed text corpora
- An evaluation of different distributional thesauri settings for French
- The release of a distributional thesaurus for French constructed from a large corpus of 450 million words

Motivation

- Distributional lexical resources are appealing because they can be constructed automatically from raw or parsed text corpora
- Useful for alleviating data sparseness in many NLP applications (e.g. parsing and coreference resolution)
- There is a lack of freely-available software for distributional methods, or distributional thesauri for French

Context relations

w	r	w'	$f(w,r,w')$
imagination	← de ←	preuve	232
imagination	← à ←	laisser	205
imagination	←	être	180
imagination	← de ←	naître	128
imagination	← de ←	rivaliser	120
imagination	→	débordant	111
imagination	→	beaucoup	101
imagination	→	fértile	99

Extraction from a corpus

- Context relations were extracted from a 125 million word parsed corpus, the *L'Est Républicain*
- *bigram* contexts are terms w' that appear to the left or to the right of the primary term w
- *syntactic* contexts (pictured left) are terms w' that are governors or dependents of the primary term w

Similarity calculation

- A *weight* function (RelFreq, TTest, PMI) transforms the raw frequency $f(w,r,w')$ of each context relation by determining the informativeness of the context
- A *measure* function (Cosine, Jaccard, Lin) subsequently calculates the similarity between two weighted frequency vector representations of context relations

Weight and measure functions

RelFreq	$\frac{f(w,r,w')}{f(w,*,*)}$	Cosine	$\frac{\Sigma\, wgt(w_1,*_r,*_{w'}) \times wgt(w_2,*_r,*_{w'})}{\sqrt{\Sigma\, wgt(w_1,*,*)^2 \times \Sigma\, wgt(w_2,*,*)^2}}$
TTest	$\frac{p(w,r,w') - p(*,r,w')p(w,*,*)}{\sqrt{p(w,r,w')/f(*,*,*)}}$	Jaccard	$\frac{\Sigma\, min(wgt(w_1,*_r,*_{w'}), wgt(w_2,*_r,*_{w'}))}{\Sigma\, max(wgt(w_1,*_r,*_{w'}), wgt(w_2,*_r,*_{w'}))}$
PMI	$log\left(\frac{p(w,r,w')}{p(*,r,w')p(w,*,*)}\right)$	Lin	$\frac{\Sigma\, wgt(w_1,*_r,*_{w'}) + wgt(w_2,*_r,*_{w'})}{\Sigma\, wgt(w_1,*,*) + \Sigma\, wgt(w_2,*,*)}$

Software package

- Implemented in Python, with methods for: (1) extraction of context relations from a text corpus; (2) weighting of context relations; (3) generation of a term similarity matrix; (4) construction of a distributional thesaurus from a similarity matrix
- User options include: context relation type; weight function; measure function; frequency thresholding; etc.

Distributional thesaurus

tragique	dramatique−0.294, triste−0.253, cruel−0.243, terrible−0.226, douloureux−0.215, poétique−0.194, fantastique−0.187, cocasse−0.183, comique−0.183, étrange−0.180, drôle−0.179, mélancolique−0.178, romantique−0.177, émouvant−0.175 ...
absolument	tout_à_fait−0.408, totalement−0.333, si−0.324, tellement−0.302, vraiment−0.283, complètement−0.269, quasi−0.245, parfaitement−0.243, assez−0.242, quasiment−0.241, presque−0.232, pratiquement−0.212 ...
véhicule	voiture−0.490, camion−0.387, engin−0.307, avion−0.295, bus−0.256, navire−0.253, camionnette−0.253, train−0.250, automobile−0.248, appareil−0.244, tracteur−0.241, fourgon−0.241, bateau−0.238, moto−0.237 ...
calciner	carboniser−0.200, brûler−0.122, consumer−0.119, déchiqueter−0.110, éventrer−0.109, endommager−0.108, gésir−0.100, incendier−0.099, broyer−0.097, momifier−0.095, cribler−0.094, recouvrir−0.092, joncher−0.092 ...

Output of the system

- FreDist produced distributional thesauri corresponding to 27 setting combinations: type of context (bigram, syntactic, bigram+syntactic), weight function (RelFreq, TTest, PMI), and measure function (Cosine, Jaccard, Lin)
- Each distributional thesaurus in this paper's evaluation contained 4,126 adjectives, 802 adverbs, 10,997 nouns, and 3,562 verbs
- The system's runtime for each setting (from context relation extraction to distributional thesaurus output) was about 18 hours using a machine with a 2.4GHz processor

Thesaurus release

- A larger distributional thesaurus for French: 6,764 adjectives, 1,165 adverbs, 33,548 nouns, and 4,956 verbs

Experiments and results

- Distributional thesauri were compared to two wordnet references: FREWN for nouns and verbs; WOLF for adjectives and adverbs
- The inverse rank (InvR) evaluation metric is used, as in information retrieval: each w appearing in the thesaurus is a *query*, and its neighbors form a ranked list of *results*
- All terms appearing in a synset with w in the wordnet reference are considered *relevant*, while other terms are considered *irrelevant*
- InvR returns the sum, over relevant terms, of the inverse of that term's rank in the results for w; each distributional thesaurus is evaluated using the average InvR over all queries w
- The best setting is bigram+syntactic, PMI, and Cosine

Wordnet evaluation

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

Contact information

E-mail:
henestro@inria.fr
Webpage:
alpage.inria.fr/~henestro/fredist.html

Selected References

[1] Curran J. (2004). *From distributional to semantic similarity*. PhD thesis, University of Edinburgh.

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

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

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