Distributional context generalisation and normalisation as a mean to reduce data sparsity: evaluation of medical corpora

Amandine Périnet[1] and Thierry Hamon[2,3]

[1]LIMICS, INSERM, U1142, Université Paris 13, UPMC Univ Paris 06, Sorbonne Paris Cité, Villetaneuse, France
[2]LIMSI-CNRS, Orsay, France [3]Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France

[1]amandine.perinet@edu.univ-paris13.fr, [2]hamon@limsi.fr

Context

Vector Space Models (VSM)

- Geometric methods using the Distributional analysis
  Words in a similar context tend to be semantically close (Harris, 1954; Firth, 1957)
- Similarity between target words = a closeness in an *n*-dimension space
- Each target word is a point defined according to its distributional properties
- Similarity: distance between two target words

Data sparseness

- The consequence of the word distribution in a corpus (Baroni et al., 2009)
- Most of the words
  - very low frequency & appear only a few times
  - limited set of contexts ⇒ similarity is difficult to catch
- Methods need to be adapted for specialized texts, as EHR texts

Material

Corpus

Menelas corpus (Zweigenbaum1994):
French medical text collection on coronary diseases
84,839 words

- reference manual on coronography and coronary diseases
- anonymised patient records
- letters written by hospital doctors to doctors in charge of patients

Automatically acquired relations

Hypernyms for generalisation

- Lexico-syntactic Patterns (LSP): 98 hypernymy relations
  artère coronaire (coronary artery) → *veine (vein)*, *artère (artery)*, *vaisseau (vessel)*
- Lexical Inclusion (LI): 7,187 hypernymy relations
  restriction → *restriction du débit coronaire (restriction of coronary output)*
- Terminological Variation (TV): 171 hypernyms
  lésion significative (significant lesion) → *lésion coronaire significative (significant coronary lesion)*

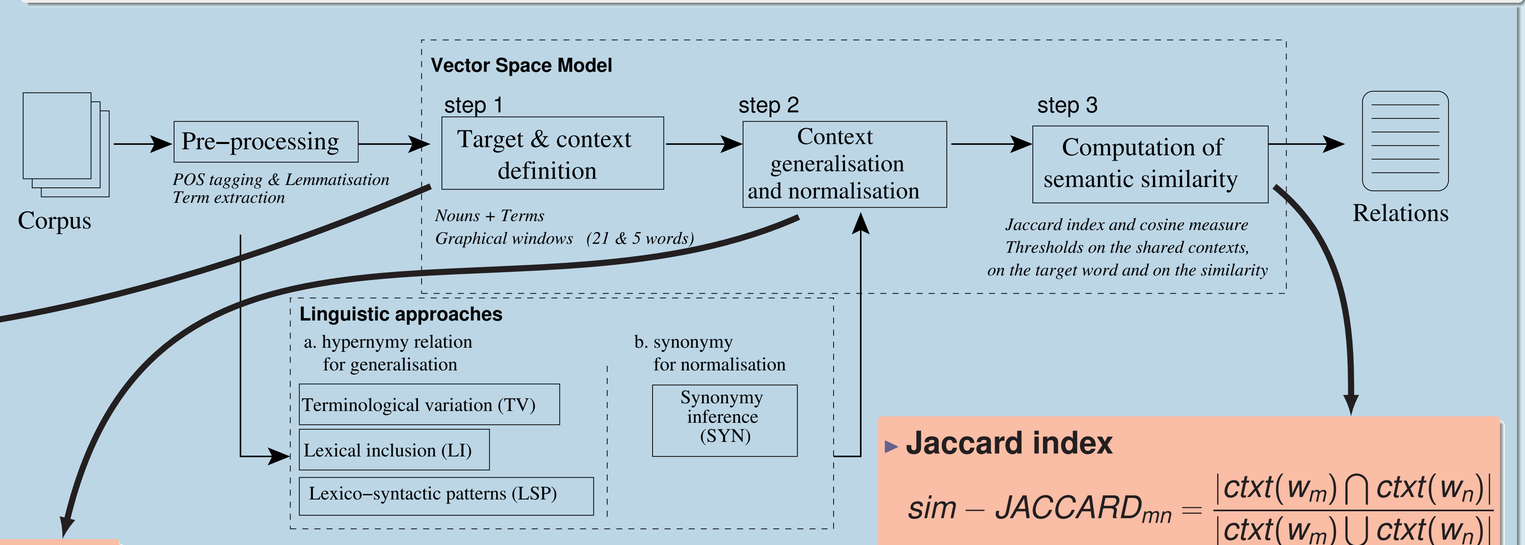
Synonyms for normalisation (SYN)

- Inferred relation throught semantic compositionality: 168 synonymy relations between complex terms
  infection de blessure (wound infection) ↔ *septicité de blessure (wound sepsis)*

Method: context generalisation within the VSM

Influence the distributional context frequencies by generalising and normalising contexts

- Decrease diversity in contexts ⇒ increase their frequency
- Switch words in contexts for their
  - hypernyms $\mathbb{H}_s(ctxt_i(w)) = \{H_1, \ldots, H_n\}$
  - synynoms $\mathbb{S}(R) = \{S_1, \ldots, S_n, R\}$



Step 1: definition of targets and contexts

- Term identification with a term extractor (YaTeA (Aubin&Hamon-2006))
- Target words: nouns + terms
- Distributional contexts: *n* of words around the targets
  - large window (W21): 21 words (± 10 words, centered on the target word)
  - narrow window (W5): 5 words (± 2 words, centered on the target word)

Step 2:

- **Rules of context generalisation**
  1. Word in context ↔ several hypernyms/variants: the hypernym/variant frequency taken into account, and the most frequent hypernym/variant is chosen
     if $|\mathbb{H}_S(ctxt_i(w))| > 1$, $ctxt_i(w) := \text{argmax}_{|H_i|}(\mathbb{H}_S(ctxt_i(w)))$
  2. Word in context ↔ only one hypernym/variant: context replaced by this hypernym/variant
     if $|\mathbb{H}_S(ctxt_i(w))| = 1$, then $ctxt_i(w) := H_1$
- **Rule of context normalisation**
  - Word in context ↔ synynoms: context replaced by the representative of the synonym cluster
    if $\exists R | ctxt_i(w) \in \mathbb{S}(R)$, *then* $ctxt_i(w) := R$

- **Jaccard index**

$$sim - JACCARD_{mn} = \frac{|ctxt(w_m) \bigcap ctxt(w_n)|}{|ctxt(w_m) \bigcup ctxt(w_n)|}$$

- **Cosine**

Evaluation:

- R-Precision (UMLS/Fr – 1,735,419 relations)
- MAP (mean average precision)
  $MAP = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i} P(l_i^j)$,
  $P(l_i^j)$ is the not interpolated precision of the semantic neighbor l_i^j at the rank j
  N is the number of targets
  n_i is the number of semantic neighbors l_i^j of the target word R_i

Results

Results obtained with a large window (21 words)

	Acquired Rel.		Rel. in UMLS		MAP		R-precision	
	JACC	COS	JACC	COS	JACC	COS	JACC	COS
BASELINE	**406**	**428**	**4**	**44**	**0.406**	**0.188**	**0.250**	**0.118**
Generalisation								
Variants	472	424	8	40	0.280	0.188	0.143	0.118
Lexical inclusion	328	223	4	26	**0.454**	**0.110**	0.250	0.133
Patterns	398	381	6	36	0.219	**0.206**	0.000	0.000
Variants + lex. inclusion	328	223	4	26	**0.454**	0.110	0.250	0.000
Patterns + lex. inclusion	338	220	4	26	**0.454**	0.101	0.250	0.000
3 generalisation methods	336	243	4	26	0.414	0.108	0.250	0.000
Normalisation								
Synonyms (Syn)	474	424	8	40	0.280	0.188	0.143	0.118
Normalisation + generalisation								
Syn/Variants	474	419	8	22	0.280	0.189	0.143	0.118
Syn/Lexical inclusion	366	279	6	14	**0.440**	0.105	0.333	0.000
Syn/Patterns	394	377	6	**20**	**0.219**	**0.206**	**0.000**	**0.133**
Syn/Variants + lex. inclusion	324	223	4	14	**0.454**	0.110	0.250	0.000
Syn/Patterns + lex. inclusion	394	373	6	20	**0.219**	**0.206**	**0.000**	**0.133**
Syn/3 generalisation methods	370	280	**6**	14	**0.454**	0.105	**0.333**	0.000

- Best results for a large window (21 words)
- Jaccard index:
  - Best MAP and R-precision
  - but few acquired relations found in UMLS
- Cosine: lower results
  - but more acquired relations found in UMLS

Experiments

- **Baseline**: VSM without any context generalisation (VSMonly)
- **Normalising distributional contexts** (VSM/SYN)
- **Generalising distributional contexts**
  - Use of automatically acquired relations
    1 linguistic approach: VSM/LSP, VSM/LI, VSM/TV
    2 approaches, sequentially. eg: TV then LI (VSM/TV+LI)
    3 approaches sequentially. eg: LSP then LI then TV (VSM/LSP+LI+TV) and all together (VSM/ALL3)
- **Normalising then generalising distributional contexts** (VSM/SYN/LSP, VSM/SYN/TV+LI, VSM/SYN/ALL3, ...)

Impact of the generalisation and the normalisation

- Positive impact:
  - Jaccard: when generalising the contexts with hypernyms issued from lexical inclusion
  - Cosine: when normalising and generalising with hypernyms provided by lexical inclusion
- No influence of context normalisation in most of the cases
- Reduction of the number of relations acquired

General observations

- **Context generalisation**
  - Quality improvement
  - Reduction of the number of relations acquired
  - Best MAP with
    - Jaccard and generalising with lexical inclusion
    - Cosine with pattern generalisation
  - Sequential generalisation: no influence of the order of the methods
  - Generalisation with several linguistic approaches does not improve the results
- **Context normalisation**
  - Decrease of the results (Jaccard) or no effect (Cosine)
- **Context normalisation and generalisation**
  - Jaccard:
    Increase of the MAP and the R-precision when using all the hypernyms or variants and lexical inclusion
  - Cosine:
    Improvement of the results when using lexical inclusion for generalising contexts

Perspectives

1. Manual analysis of the relations acquired and of the impact of the generalisation and normalisation process
2. Comparison of our method with other dimension reduction methods as Random Indexing and LSA

PolTAL 2014 - 9th International Conference on Natural Language Processing
Warsaw, Poland, September 17-19, 2014