Linguistic Preprocessing for Distributional Analysis Eff ciency: Evidence from French

Emmanuel Cartier (1) and Valeriya Vinogradova (2)

Université Paris 13 Sorbonne Paris Cité
(1) LIPN – RCLN, CNRS UMR 7030
(2) LDI, CNRS UMR 7187

Table of Contents

Motivations

Statistical paradigm:

- has prevailed in the NLP field for about twenty years.

- is grounded on the distributional hypothesis (Harris) and on the corpus linguistics works (Firth).

- results in convincing applications: multiword expression extraction, part-of-speech tagging, semantic relation identification and even probabilistic models of language.

- gives rise to interesting linguistics phenomena: collocations, “collostructions” (Stefanowitsch), “word sketches” (Kilgariff)

Vector Space Model:

- the prevailing computational model implementing the distributional analysis (Turney et Pantel, 2010 : 152)

- has been applied since to various areas in NLP (notably for MWE and semantic similarity).

- can be built from the raw text or from annotated text.

In this presentation we will try to show that the linguistic preprocessing can greatly impact on the accuracy of the results of VSM's exploitation.

Distributional Models for Linguistic Analysis

Distributional hypothesis (DH) for text analysis

- **Words that occur in similar contexts tend to have similar meanings** ((Turney & Pantel, 2010), (Harris, 1954; Firth, 1957; Deerwester et al., 1990).

As an example, here are the most similar words (i.e. words sharing the most contexts) with *outil* (tool), automatically processed with a computational implementation of the DH:

  - Outil(Tool)/NC: technique(technique)/NC, méthode(method)/NC, logiciel(software)/NC, outillage(tooling)/NC, support(support)/NC, procédé(process)/NC, moyen(medium)/NC, technologie(technology)/NC.

- It is intuitively obvious that the words share semantic relations with the source word

Distributional Models for Linguistic Analysis

Computational models for distributional analysis

- As far as computational linguistics is concerned, the Vector Space Models (VSM) has prevailed to implement the distributional hypothesis.
- Variations in terminology and techniques: Turney and Pantel, 2010, Lenci and al., 2010, Kiela and Clark, 2013-2014, Mikolov, 2014.
- Term-document matrix then term-term matrix application

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
against	0	0	0	1	0	0	3	2	3	0
age	0	0	0	1	0	3	1	0	4	0
agent	0	0	0	0	0	0	0	0	0	0
ages	0	0	0	0	0	2	0	0	0	0
ago	0	0	0	2	0	0	0	0	3	0
agree	0	1	0	0	0	0	0	0	0	0
ahead	0	0	0	1	0	0	0	0	0	0
ain't	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	0	0	0	0
aka	0	0	0	1	0	0	0	0	0	0

	against	age	agent	ages	ago	agree	ahead	ain.t	air	aka	al
against	2003	90	39	20	88	57	33	15	58	22	24
age	90	1492	14	39	71	38	12	4	18	4	39
agent	39	14	507	2	21	5	10	3	9	8	25
ages	20	39	2	290	32	5	4	3	6	1	6
ago	88	71	21	32	1164	37	25	11	34	11	38
agree	57	38	5	5	37	627	12	2	16	19	14
ahead	33	12	10	4	25	12	429	4	12	10	7
ain't	15	4	3	3	11	2	4	166	0	3	3
air	58	18	9	6	34	16	12	0	746	5	11
aka	22	4	8	1	11	19	10	3	5	261	9
al	24	39	25	6	38	14	7	3	11	9	861

Distributional Models for Linguistic Analysis

Typical steps for VSM building: from corpus to matrix

- Linguistic preprocessing: bag-of-words vs POS-tagging, stopwords removal, etc.
- Matrix building
- Frequency weighting (to give more weight to surprising events and less weight to expected events),
- Matrix smoothing (to reduce the vector dimensions),

Main areas of VSM exploitation

- Document Retrieval (Search Engine)
- MultiWord Expressions (MWE) extraction (Association measures)
- Semantic similarity assessment (Similarity measures)

Distributional Models for Linguistic Analysis

VSM building: parameters (Kiela and Clark, 2014) – as far as linguistic preprocessing is concerned:

- **stop words, high frequency cut-off:** removal of high-frequency words or "tool" words as they do not convey any useful information
- **window size:** context choice (1,2,3...5.. 7 words in the context)
- **type of linguistic analysis:** raw text vs POS-tagging vs dependency analysis
- **feature granularity:** taking into account only the word forms, or the lemmas, any other information (syntactic, semantic, etc.), or a combination of these

Distributional Models for Linguistic Analysis

VSM building: best parameters? (Kiela and Clark, 2014)

- **Stop-words, high-frequency words removal:** give better results, but only if no raw frequency weighting is applied to the results (in line with the conclusion of (Bulinaria and Levy, 2012).
- **Impact of dependency analysis** needs additional experiments, as several works (for example Pado and Lapata, 2007) made a contradictory conclusion.
- **As far as feature granularity** is concerned, the authors do not take into account a combination of features from different levels (i.e. form, lemma, POS-tag).

=> no evidence has emerged; linguistic preprocessing impact is still to be assessed.

Experiment in French: impact of linguistic preprocessing in a similarity task

Experiment summary

- In this experiment we have studied the *impact of linguistic preprocessing in a similarity task.*
- We have set up **five corpora from the French Wikipedia** applying different linguistic preprocessings and use the raw text as the baseline
- We have **compared the results to a manually set up gold standard** composed of 8 randomly selected words [rasoir (razor), femme (woman), biche (doe), arbre (tree), école (school), robe (dress), seau (bucket), outil (tool)] and their best similar words
- We have used the **word2vec tool (Mikolov, 2013)** that demonstrates the best results on various semantic tasks. We have used the following parameters: bag-of-words method, vector size: 200, window-size: max. 5 words.

Experiment in French: impact of linguistic preprocessing in a similarity task

Gold standard

- no existing gold standard for French
- eight words chosen:
- procedure:

1/ 40 best similar words with *word2vec* from each corpus

2/ Three linguistic experts assessed the similar nature of each pair

3/ Inter-agreement resolution: removal of pairs with no agreement resulting in 724 pairs for the gold standard.

Word/POS	Nb of distinct similar words
arbre/NC	96
biche/NC	103
femme/NC	81
outil/NC	83
rasoir/NC	108
robe/NC	85
seau/NC	86
école/NC	82

Experiment in French: impact of linguistic preprocessing in a similarity task

Corpora

- **French Wikipedia (5 August 2014):** 363 243 different words (types), 178 943 471 occurrences (tokens), a sufficiently representative corpus (Kiela and Clark, 2014).

- **Generic linguistic preprocessing:** one sentence per line; dates and figures reduced to a generic token NPDATE and NB; reduction of proper names to a generic annotation PN. (results in less word types: 186 195)

- **Specific linguistic preprocessing:**

corpus	description	Number of types	Number of tokens
Corpus 0	Raw text	363243	178943471
Corpus A	Raw text with generic linguistic preprocessing.	186195	178943471
Corpus B	Corpus annotated morpho-syntactically (Treetagger)	155316	157423881
Corpus C	Corpus annotated morpho-syntactically for plain words (nouns, personal pronouns, adjectives, verbs, prepositions), the other words are reduced to their part of the speech.	105050	85782609
Corpus D	Corpus annotated morpho-syntactically for plain words (nouns, personal pronouns, adjectives, verbs, prepositions), the other words are reduced to a single tag W.	105027	84732369
Corpus E	From corpus B; Removal of peripherical elements from the corpus (adverbials, determiners); Splitting of complex sentences into simple ones (relatives and subordinate clauses)	105027	84732369

Experiment in French: impact of linguistic preprocessing in a similarity task

Results

	Corpus 0		Corpus A		Corpus B		Corpus C		Corpus D		Corpus E	
Word/POS	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL
arbre/NC	0,53	0,13	0,77	0,35	1	0,45	1	0,45	1	0,45	1	0,65
biche/NC	0,72	0,21	0,9	0,37	1	0,41	0,95	0,39	0,95	0,39	1	0,58
femme/NC	0,76	0,32	1	0,49	1	0,49	1	0,49	1	0,49	1	0,59
outil/NC	0,64	0,32	1	0,48	0,97	0,47	1	0,48	0,97	0,47	0,97	0,67
rasoir/NC	0,21	0,12	0,27	0,26	0,57	0,54	0,45	0,42	0,55	0,52	0,63	0,74
robe/NC	0,65	0,22	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,61
seau/NC	0,54	0,21	1	0,46	1	0,46	1	0,46	1	0,46	1	0,64
école/NC	0,61	0,31	1	0,48	1	0,48	1	0,48	1	0,48	1	0,6
TOTAL	**0,5825**	**0,23**	0,86375	0,41875	**0,93875**	0,47	0,92125	0,45375	0,93	0,465	**0,94625**	**0,635**

Precision = number of good extracted pairs / number of good gold standard pairs
Recall = number of good extracted pairs / total number of extracted pairs

For each corpus (Corpus 0 and A to E, in column), the table shows the precision and recall for each word (in rows) and the average f ġure (last row).

Experiment in French: impact of linguistic preprocessing in a similarity task

Analysis

	Corpus 0		Corpus A		Corpus B		Corpus C		Corpus D		Corpus E	
Word/POS	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL
arbre/NC	0,53	0,13	0,77	0,35	1	0,45	1	0,45	1	0,45	1	0,65
biche/NC	0,72	0,21	0,9	0,37	1	0,41	0,95	0,39	0,95	0,39	1	0,58
femme/NC	0,76	0,32	1	0,49	1	0,49	1	0,49	1	0,49	1	0,59
outil/NC	0,64	0,32	1	0,48	0,97	0,47	1	0,48	0,97	0,47	0,97	0,67
rasoir/NC	0,21	0,12	0,27	0,26	0,57	0,54	0,45	0,42	0,55	0,52	0,63	0,74
robe/NC	0,65	0,22	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,61
seau/NC	0,54	0,21	1	0,46	1	0,46	1	0,46	1	0,46	1	0,64
école/NC	0,61	0,31	1	0,48	1	0,48	1	0,48	1	0,48	1	0,6
TOTAL	**0,5825**	**0,23**	0,86375	0,41875	**0,93875**	0,47	0,92125	0,45375	0,93	0,465	**0,94625**	**0,635**
			+0,28	**+0,19**	**+0,36**	**+0,24**	**+0,34**	**+0,22**	**+0,35**	**+0,24**	**+0,36**	**+0,41**

Precision = number of good extracted pairs / number of good gold standard pairs
Recall = number of good extracted pairs / total number of extracted pairs

Raw text generates the worst results among all the conf ġurations.
=> **impact of general linguistic preprocessing (reduction of proper names, dates and numbers) and POS-tagging in the similarity tasks**

Experiment in French: impact of linguistic preprocessing in a similarity task

Analysis

	Corpus 0		Corpus A		Corpus B		Corpus C		Corpus D		Corpus E	
Word/POS	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL
arbre/NC	0,53	0,13	0,77	0,35	1	0,45	1	0,45	1	0,45	1	0,65
biche/NC	0,72	0,21	0,9	0,37	1	0,41	0,95	0,39	0,95	0,39	1	0,58
femme/NC	0,76	0,32	1	0,49	1	0,49	1	0,49	1	0,49	1	0,59
outil/NC	0,64	0,32	1	0,48	0,97	0,47	1	0,48	0,97	0,47	0,97	0,67
rasoir/NC	0,21	0,12	0,27	0,26	0,57	0,54	0,45	0,42	0,55	0,52	0,63	0,74
robe/NC	0,65	0,22	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,61
seau/NC	0,54	0,21	1	0,46	1	0,46	1	0,46	1	0,46	1	0,64
école/NC	0,61	0,31	1	0,48	1	0,48	1	0,48	1	0,48	1	0,6
TOTAL	**0,5825**	**0,23**	0,86375	0,41875	**0,93875**	0,47	0,92125	0,45375	0,93	0,465	**0,94625**	**0,635**

Precision = number of good extracted pairs / number of good gold standard pairs
Recall = number of good extracted pairs / total number of extracted pairs

Corpus A (result of generic linguistic preprocessing, without POS-tagging) has slightly worse results (about -6%) in comparison to the annotated corpora (B, C, D, E) but is far better than raw text.

=> **impact of general linguistic preprocessing**

=> **impact of POS-annotation is not that impressive, at least in the similarity task**

Experiment in French: impact of linguistic preprocessing in a similarity task

Analysis

Word/POS	Corpus 0 PRECISION	RECALL	Corpus A PRECISION	RECALL	Corpus B PRECISION	RECALL	Corpus C PRECISION	RECALL	Corpus D PRECISION	RECALL	Corpus E PRECISION	RECALL
arbre/NC	0,53	0,13	0,77	0,35	1	0,45	1	0,45	1	0,45	1	0,65
biche/NC	0,72	0,21	0,9	0,37	1	0,41	0,95	0,39	0,95	0,39	1	0,58
femme/NC	0,76	0,32	1	0,49	1	0,49	1	0,49	1	0,49	1	0,59
outil/NC	0,64	0,32	1	0,48	0,97	0,47	1	0,48	0,97	0,47	0,97	0,67
rasoir/NC	0,21	0,12	0,27	0,26	0,57	0,54	0,45	0,42	0,55	0,52	0,63	0,74
robe/NC	0,65	0,22	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,61
seau/NC	0,54	0,21	1	0,46	1	0,46	1	0,46	1	0,46	1	0,64
école/NC	0,61	0,31	1	0,48	1	0,48	1	0,48	1	0,48	1	0,6
TOTAL	**0,5825**	**0,23**	0,86375	0,41875	**0,93875**	0,47	0,92125	0,45375	0,93	0,465	**0,94625**	**0,635**

Precision = number of good extracted pairs / number of good gold standard pairs
Recall = number of good extracted pairs / total number of extracted pairs

Among the various annotated versions, the best ones are: **Corpus B**: the annotated version preserving lemma and POS-tag for every word and **Corpus E**: the one implying additional linguistic preprocessing **=> tool-words removal (Corpus C and D) has NO positive effect on the results**

Experiment in French: impact of linguistic preprocessing in a similarity task

Analysis

	Corpus 0		Corpus A		Corpus B		Corpus C		Corpus D		Corpus E	
Word/POS	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL
arbre/NC	0,53	0,13	0,77	0,35	1	0,45	1	0,45	1	0,45	1	0,65
biche/NC	0,72	0,21	0,9	0,37	1	0,41	0,95	0,39	0,95	0,39	1	0,58
femme/NC	0,76	0,32	1	0,49	1	0,49	1	0,49	1	0,49	1	0,59
outil/NC	0,64	0,32	1	0,48	0,97	0,47	1	0,48	0,97	0,47	0,97	0,67
rasoir/NC	0,21	0,12	0,27	0,26	0,57	0,54	0,45	0,42	0,55	0,52	0,63	0,74
robe/NC	0,65	0,22	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,46	0,97	0,61
seau/NC	0,54	0,21	1	0,46	1	0,46	1	0,46	1	0,46	1	0,64
école/NC	0,61	0,31	1	0,48	1	0,48	1	0,48	1	0,48	1	0,6
TOTAL	**0,5825**	**0,23**	0,86375	0,41875	**0,93875**	0,47	0,92125	0,45375	0,93	0,465	**0,94625**	**0,635**

Precision = number of good extracted pairs / number of good gold standard pairs
Recall = number of good extracted pairs / total number of extracted pairs

The **recall rate** is stable in corpora A to D (41% to 46%) and achieves its highest point (63%) in corpus E.
The difference between Corpus B and E resides essentially in a gain in recall.

=> probably due to the specif ic linguistic preprocessing

Conclusions and perspectives

Conclusions

- Linguistic preprocessing has an evident impact on the accuracy of VSM in a similarity task
- This impact is mainly due to general linguistic preprocessing (reduction of proper names, dates and figures), the impact of POS-tagging is not that evident
- Specific linguistic preprocessing has an impact on recall which is probably due to the splitting of complex sentences into simple ones
- But: this splitting is not so easy to implement and should be tuned for every language

Perspectives

- Confirmation in other languages to be done
- Impact of linguistic preprocessing in other tasks (MWE extraction) as adequate linguistic preprocessing certainly depends on the task

THE END

Thanks a lot!

Any questions?

Appendix: Questions

Justify word2vec paramaters in the experiment

Best parameters according to Kiela and Clark 2014

Gold Standard limitation to 8 words

The experiment presented here is a f ṙst step: we chose common words linked to human activities

We intend to set up a more extensive gold standard in the near future

Appendix: Specif ic linguistic preprocessing

- **Adverbials and determiners removal (green) and decomposition of complex sentences (subordinate and relative clauses) (underlined red)**

Essentiellement territorial, le chat est un prédateur de petites proies comme les rongeurs ou les oiseaux. Les chats ont diverses vocalisations dont les ronronnements, les miaulements, les feulements ou les grognements, bien qu'ils communiquent principalement par des positions faciales et corporelles et des phéromones. Selon les résultats de travaux menés en 2006 et 2007, le chat domestique est une sous-espèce du chat sauvage (Felis silvestris) dont son ancêtre, le chat sauvage d'Afrique (Felis silvestris lybica) a vraisemblablement divergé il y a 130 000 ans.

=>

territorial, chat est prédateur de petites proies comme rongeurs ou oiseaux.

chats ont vocalisations dont ronronnements, miaulements, feulements ou grognements.

ils communiquent par positions faciales et corporelles et phéromones.

Selon résultats de travaux menés en DATE et DATE

chat domestique est sous-espèce du chat sauvage (Felis silvestris)

ancêtre, chat sauvage d'Afrique (Felis silvestris lybica) a divergé il y aNB ans.

Removal and decomposition are achieved with POS-tagging (adverbials) and through automatic pattern-matching (presence of a conjunction or relative pronoun at the beginning, rare sequence of pos-tagged words at the end of clause, or end of sentence)