Linguistic Preprocessing for Distributional Analysis : Evidence from French

Emmanuel Cartier

Linguistic Preprocessing for Distributional Analysis Efficiency : Evidence from French

Emmanuel Cartier
Université Paris 13 Sorbonne Paris Cité
LIPN – RCLN – UMR 7030
emmanuel.cartier@lipn.univ-paris13.fr

1 Introduction

Distributional Hypothesis and Cognitive Foundations

For about fifteen years, the statistical paradigm, from the distributionalism hypothesis (Harris, 1954) and corpus linguistics (Firth, 1957), has prevailed in the NLP field, with a lot of convincing results : multiword expression, part-of-speech, semantic relation identification, and even probabilistic models of language. These studies have identified interesting linguistic phenomena, such as collocations, "collostructions" (Stefanowitsch, 2003), « word sketches » (Kilgariff et al., 2004).

Cognitive Semantics (Langacker, 1987, 1991; Geeraerts et al. 1994 ; Schmid, 2007, 2013), have also introduced novel concepts, most notably that of « entrenchment », which enables to ground the social lexicalization of linguistic signs and to correlate it with repetition in corpus.

Finally, Construction Grammars (Fillmore et al., 1988 ; Goldberg, 1995, 2003 ; Croft, 2001, 2004, 2007) have proposed linguistic models which reject the distinction lexicon (list of "words") - grammar (expliciting the combination of words) : all linguistic signs are constructions, from morphemes to syntaxical schemes, leading to the notion of "constructicon", as a goal for linguistic description.

Computational Models of the Distributional Hypothesis

As Computational Linguistics is concerned, the Vector Space Model (VSM) has prevailed to implement the distributional hypothesis, giving rise to continuous sophistication and several state-of-the-arts (Turney and Pantel, 2010; Lenci and al., 2010; Kiela and Clark, 2013; Clark, 2015). (Kiela and Clarke, 2014) state that the following parameters are implied in any VSM implementation: *vector size, window size, window-based or dependency-based context, feature granularity, similarity metric, weighting scheme, stopwords and high frequency cut-off*. Three of them are directly linked to linguistic preprocessing : *window-based or dependency-based context*, the second requiring a dependency analysis of the corpus; *feature granularity*, ie, the fact of taking into account either the raw corpus, or a lemmatized or pos-tagged one for n-gram calculus; *stopwords and high frequency cut-off* , ie removal of high-frequency words or "tool words". (Kiela and Clarke, 2014) conducted six experiments/tasks with varying values for each parameter, so as to assess the most efficient ones. They conclude that : dependency-based does not trigger any improvement over raw-text n-gram calculus; as for feature granularity, that stemming yields the better results; as for stopwords or high-frequency words removal, it does yield better results, but *only if* no raw frequency weighting is applied to the results; this is in line with the conclusion of (Bulinaria and Levy, 2012).

Nevertheless, these conclusions should be refined and completed :

1/ As feature granularity is concerned, the authors do not take into account a combination of features from different levels; (Béchet et al., 2012), for example, have shown that combining features from three levels (form, lemma, pos-tag) can result in better pattern recognition for specific linguistic tasks; such a combination is also in line with the Cognitive Semantics and the Construction Grammar hypothesis, that linguistic signs emerge as constructions combining schemes, lemmas and specific forms;

2/ The experiments on dependency need additional experiments, as several works (for example Pado and Lapata, 2007) made a contradictory conclusion.

3/ Stopwords or high-frequency words removal results in better results if no frequency weighting is applied; but the authors apply – as quasi all work in the field -, a brute-force removal either based on "gold standard" stopword lists, or on a arbitrary count to cut off results; this technique should be refined to remove only the noisy words or n-grams and should be linguistically motivated.

Linguistic Motivation for Linguistic Preprocessing

The hypothesis supported in this paper is that, if repetition of sequences is the best way to access usage and to induce linguistic properties, language users do not only rely on the sequentiality of language, but also on non-sequential knowledge thus untractable from the actual distribution of words. This knowledge is linked to the three classical linguistical units : lexical units, phrases and predicate structures, each being a combination of the preceding with language-specific rules for their construction. Probabilistic models of language have mainly focused until now on the lexical units level, but to leverage language, probabilistic research must also model and preprocess phrases and predicate structures.

The present paper will try to ground this hypothesis through an experiment, aimed at retrieving lexico-semantic relations in French, where we preprocess the corpus in three ways :

- morphosyntactic analysis
- peripheral lexical units removal
- phrases identification.

As we will see, these steps enables to access more easily the predicate structures that the experiment aims at revealing, while using a VSM model on the resulting preprocessed corpus.

2 Evidence from French : Semantic Relations and Definitions

Definition model

Here we assume that a definitory statement is a statement asserting the essential properties of a lexical unit. It is composed of the definiendum (DFM), i.e. the lexical unit to be defined; the definiens (DFS), i.e. the phrasal expression denoting the essential properties of the DFM lexical unit; the definition relator (DEFREL), i.e. the linguistic items denoting the semantic relation between the two previous elements.

The traditional model of definition decomposes the DFS into two main parts : HYPERNYM + PROPERTIES.

Definitory statement can also comprise other information : enunciation components (*according to Sisley, a G-component is a ...*); domain restrictions (*in Astrophysics, a XXX is a YYY).*

Corpus

We use three corpora and retain only the nominal entries in each :

Trésor de la Langue Française (TLF) : 61 234 nominal lexical units, and 90 348 definitions;

French Wiktionary (FRWIK) : 140 784 nouns, for a total of 187 041 definitions.

Wikipedia (WIKP) : 610 013 glosses (ie first sentence of each article) from the French Wikipedia, using a methodology next to (Navigli and al, 2008)

The first two are dictionaries (TLF, FRWIK), the last one is an encyclopedia (WIKP). In the first case, definition obeys to lexicographic standards, whereas definitions are more "natural" in WIKP.

System Architecture

The system is composed of four steps :

- Morpho-syntactic analysis of the corpus
- Semantic Relation Trigger words Markup
- Sentence Simplification : this step aims at reducing, as much as possible, the sentences to the core semantic expressions of definition;
- Lexico-syntactic pattern-matching for semantic relations : relation(X,Y)

In the following, we will focus on the simplification step.

Sentence simplification

Sentence simplification has two main goals :

1. decompose any sentence into its main predicate-arguments structure, and remove and record peripherical elements if necessary;
2. Unify nominal phrases, as they are the target for hypernym relations and their sparsity complicate retrieval of patterns.

Take the following source definition :

en/P cuisine/NC ,/PONCT un/DET DEFINIENDUM être/V un/DET pièce/NC de/P pâte/NC aplatir/VPP ,/PONCT généralement/ADV au/P+D rouleau/NC à/P pâtisserie/NC ./PONCT ((cooking) an undercrust is a piece of dough that has been flattened, usually with a rolling pin.)

It will be reduced to :

un/DET DEFINIENDUM être/V un/DET pièce/NC de/P pâte/NC aplatir/VPP ,/PONCT au/P+D rouleau/NC à/P pâtisserie/NC ou/CC un/DET laminoir/NC ./PONCT

And we extract the domain restriction : *en/P cuisine/NC.*

Step 1 and 2 : Adverbials and subordinate clauses

The first linguistical sequences removed from the source sentence are adverbials and specific clauses. But we would like to remove only clauses dependent on the main predicate, not those dependent on one of its core components. For example, we remove the incidential clause in :

DEFINIENDUM (/PONCT parfois/ADV Apaiang/NPP ,/PONCT même/ADJ prononciation/NC)/PONCT être/V un/DET atoll/NC de/P le/DET république/NC du/P+D Kiribati/NPP ./PONCT

But relative clauses dependent on one of the definiens component should be first extracted:

DEFINIENDUM être/V du/P+ enzymes/NC qui/PROREL contrôler/V le/DET structure/NC topologique/ADJ de/P l'ADN/NC ...

To achieve this goal, we use distributional analysis on these clauses with (SDMC, Béchet et al., 2012) and human tuning to determine the most frequent patterns and trigger words of incidential clauses at specific locations in the sentence : beginning of the definition, between the definiendum and a definition relator.

Some incidential clauses convey a semantic relation, for example the synonymy relation :

DEFINIENDUM (/PONCT parfois/ADV Apaiang/NPP ,/PONCT même/ADJ prononciation/NC)/PONCT être/V/DEF_REL un/DET atoll/NC de/P le/DET république/NC du/P+D Kiribati/NPP ./PONCT (Wikipedia)

For these, we first extract the clause as a synonymy relation for the given definiendum.

Negative adverbials cannot not be removed, as they totally change the meaning of the sentence.

Other subordinate clauses denote a domain restriction : with SDMC, we identify the most frequent cases, which derive into the following two lexico-syntactic pattern, expressed in semi regular expression :

^((?:en|dans|à|sur|selon|pour|chez|par).{5,150}?)\t, \/PONCT\t/

DEFINIENDUM\t, \/PONCT\t((?:en|dans|à|sur|selon|pour|chez|par).{5,150}?)\t, \/PONCT

Adverbials and subordinate clauses removal obviously results in a simplification of sentences, easing the following extractions.

Unification of nominal phrases:

Most of the time, the definiens is composed of a nominal phrase followed by complements (adjectival clauses or relative clauses). The first nominal element is therefore the hypernym of the definiendum. A series of phenomena complexify the identification of this nominal. Mainly : multiword determiners, (*a great variety of,...*) quantifiers (*three thousand ...*) and trigger words (*a kind of...*).

To overcome these cases, we rely on the tokenization process, which has recognized most of the multiword determiners, as well as trigger words, and unify only the remaining elements, based on an SDMC processing working on sequences beginning with a determiner and ending with a relative clause. We end up with three main lexico-syntactic patterns for identifying most of the nominal phrases :

N (ADJ) ? de/P N (ADJ) ?
N (ADJ){0,3}
PN+

Results

The linguistic preprocessing improves greatly the extraction process, as will be seen in table 1.

3 Conclusion and Future Work

In this contribution, we have shown through an experiment that the distributionalist hypothesis and the accompanying computational models, can benefit from a linguistic preprocessing of corpora, especially in tasks connected to predicate-arguments structures. That derives from the fact that language has not only a sequential structure but also a hierarchical one linking lexical units to phrases, phrases to predicate-argument structures and also essential versus peripherical elements at each level. Depending on the task, any probabilistic model should preprocess the peripherical elements to eliminate noisy analysis.

References

Baroni___M. and Alessandro Lenci, 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics* 36(4):673-721

Béchet N., Cellier P., Charnois T., and Crémilleux B., 2012. Discovering linguistic patterns using sequence mining. In Alexander F. Gelbukh, editor, *13th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2012*, volume 7181 of Lecture Notes in Computer Science, pages 154–165. Springer, 2012.

Blacoe W. and Mirella Lapata. 2012. A Comparison of Vector-based Representations for Semantic Composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Lan- guage Learning*, pages 546–556, Jeju Island, Korea, July. Association for Computational Linguistics.

Bullinaria John A. and Joseph P. Levy. 2012. Extracting Semantic Representations from Word Co- occurrence Statistics: Stop-lists, Stemming and SVD. *Behavior Research Methods*, 44:890–907.

Clark S. 2015. Vector Space Models of Lexical Meaning (to appear). In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics*. Wiley-Blackwell, Oxford.

Croft W. & Cruse D.A. 2004. *Cognitive Linguistics*. Cambridge UK : Cambridge University Press.

Croft, William A. 2007. Construction Grammar. In H. Cuyckens and D. Geeraerts (eds.), *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press, 463 508.

Croft,W. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Fillmore, Charles J., Paul Kay and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of Let alone. *Language* 64/3, 501-? 538.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pp. 1–32. Blackwell, Oxford.

Geeraerts, D., Grondelaers, S., & Bakema, P. 1994. *The structure of lexical variation. A descriptive framework for cognitive lexicology.* Berlin etc.: Mouton de Gruyter.

Goldberg, Adele E. 1995. *Constructions. A Construction Grammar Approach to Argument Structure.* Chicago: University of Chicago Press.

Goldberg, Adele. E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7/5, 219–224

Harris, Z. 1954. Distributional structure. *Word*, 10(2-3):1456–1162.

Kiela, D. and Stephen Clark, "A Systematic Study of Semantic Vector Space Model Parameters," in *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, 2014, pp. 21–30

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004) The Sketch Engine. In: Williams G. and S. Vessier (eds.), *Proceedings of the XI Euralex International Congress*, July 6-10, 2004, Lorient, France, pp. 105-111.

Langacker, R. W. 1987. *Foundations of cognitive grammar. Vol. 1, Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Langacker, R. W. 1991. *Foundations of cognitive grammar. Vol. 2, Descriptive application*. Stanford, CA: Stanford University Press.

Pado, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Schmid H.-J. 2007. Entrenchment, salience and basic levels. In: Dirk Geeraerts and Hubert Cuyckens, eds., *The Oxford Handbook of Cognitive Linguistics*, Oxford: Oxford University Press, 117-138.

Schmid H.-J. and Küchenhoff H. 2013. Collostructional

analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3), 531-577.

Stefanowitsch, Anatol, and Stefan Th. Gries. 2003. Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8/2, 209-? 243.

Turney, Peter D. & Patrick Pantel (2010), From frequency to meaning: Vector space models of semantics, *Journal of Artificial Intelligence Research* 37:141–188.