Using subcategorization frames to improve French probabilistic parsing

Anthony Sigogne
Université Paris-Est, LIGM
sigogne@univ-mlv.fr

Matthieu Constant
Université Paris-Est, LIGM
mconstan@univ-mlv.fr

Abstract

This article introduces results about probabilistic parsing enhanced with a word clustering approach based on a French syntactic lexicon, the Lefff (Sagot, 2010). We show that by applying this clustering method on verbs and adjectives of the French Treebank (Abeillé et al., 2003), we obtain accurate performances on French with a parser based on a Probabilistic Context-Free Grammar (Petrov et al., 2006).

1 Introduction

Dealing with data sparseness is a real challenge for Probabilistic Context-Free Grammar parsers [PCFG], especially when the PCFG grammar is extracted from a small treebank[1]. This problem is also lexical because the richer the morphology of a language is, the sparser the lexicons built from a treebank will be for that language. Nevertheless, the effect of lexical data sparseness can be reduced by word clustering algorithms. Inspired by the clustering method of (Koo et al., 2008), (Candito and Seddah, 2010) have shown that by replacing each word of the corpus by automatically obtained clusters of words, they can significantly improve a PCFG parser on French. Recently, (Sigogne et al., 2011) proposed a clustering method based on a French syntactic lexicon, the Lexicon-Grammar [LG] (Gross, 1994). This method consists in replacing each word of the corpus by the combination of its part-of-speech tag and its cluster, pre-computed from the lexicon. A cluster corresponds to a class of the lexicon that gathers items sharing several syntactic properties. They applied this method on verbs only and reported significant gains.

In this article, we propose a clustering method of verbs and adjectives based on another French lexicon, the Lefff (Sagot, 2010). This lexicon does not offer a classification of items as in the LG but for each entry, information about subcategorization frame is available. Clusters of words are now computed by aggregating items that have a similar frame, a frame being reduced to a vector of syntactic functions linked to possible syntactic arguments.

In sections 2 and 3, we describe the probabilistic parser and the treebank used in our experiments. In section 4, we describe more precisely previous work on clustering methods. Section 5 introduces the syntactic lexicon, the Lefff, and then we present the clustering approach based on this lexicon. In section 6, we describe our experiments and discuss the obtained results.

2 Berkeley Parser

The probabilistic parser, used in our experiments, is the Berkeley Parser[2] [BKY] (Petrov et al., 2006). This parser is based on a PCFG model which is non-lexicalized. The main problem of non-lexicalized context-free grammars is that nonterminal symbols encode too general information which weakly discriminates syntactic ambiguities. The benefit of BKY is to try to solve the problem by generating a grammar containing

[1] Data sparseness implies the difficulty of estimating probabilities of rare rules extracted from the corpus.

[2] http://code.google.com/p/berkeleyparser/

complex symbols, following the principle of latent annotations introduced by (Matsuzaki et al., 2005). Parameters of the latent grammar are estimated with an algorithm based on Expectation-Maximisation [EM]. In the case of French, (Seddah et al., 2009) have shown that BKY produces *state-of-the-art* performances.

3 French Treebank

For our experiments, we used the French Treebank[3] (Abeillé et al., 2003) [FTB]. It is composed of articles from the newspaper *Le Monde* where each sentence is annotated with a constituent tree. Currently, most papers about parsing of French use a specific variant of the FTB, namely the FTB-UC described for the first time in (Candito and Crabbé, 2009). It is a partially corrected version of the FTB that contains 12.351 sentences and 350.931 tokens with a part-of-speech tagset of 28 tags and 12 nonterminal symbols[4].

4 Previous work on word clustering

Numerous works used a clustering approach in order to reduce the size of the corpus lexicon and therefore reduce the impact of lexical data sparseness on treebank grammars. Several methods have been described in (Candito and Seddah, 2010). The best one, called *Clust*, consists in replacing each word by a cluster id. Cluster ids are automatically obtained thanks to an unsupervised statistical algorithm (Brown et al., 1992) applied to a large raw corpus. They are computed on the basis of word co-occurrence statistics. Currently, this method permits to obtain the best results on the FTB-UC. Recently, (Sigogne et al., 2011) described a method, called *LexClust*, based on a French syntactic lexicon, the Lexicon-Grammar (Gross, 1994), that consists in replacing each verbal form of the corpus by the combination of its POS tag and its cluster. These clusters follow the particular classification of entries offered by this lexicon, that aggregates items sharing several syntactic properties (e.g. subcategorization information). For example, a class of this lexicon, called *31R*, indicates that all verbs belonging to this class are intransitive. By only modifying the verbs, this approach obtains significant results on the FTB-UC.

5 Word clustering based on a syntactic lexicon

5.1 A syntactic lexicon, the Lefff

The Lefff is a French syntactic and morphological wide-coverage lexicon (Sagot, 2010)[5] that contains 110.477 lemmatized forms (simple and compound) and 536.375 inflected forms. This lexicon describes for each lemmatized entry a canonical subcategorization frame, composed of all possible arguments of the entry, and a list of possible redistributions from this frame. Inflected entries are built from lemmatized form and for each possible redistribution. For each argument of a subcategorization frame, it is stated the mandatory nature, a syntactic function, syntagmatic productions (pronoun *cln*, noun phrase *np*, infinitive phrase *sinf*,...), and some semantic features (human, abstract,...). A syntactic function takes a value among a set of nine functions, *Suj* (subject), *Obj* (direct object), *Objà* (indirect object introduced by the preposition *à*), *Objde* (indirect object introduced by the preposition *de*), *Loc* (locative), *Dloc* (delocative), *Att* (attribute), *Obl* and *Obl2* (obliques). Figure 1 shows a simplified sample of the Lefff for an entry of the French verb *chérir* (to cherish). The frame of this entry is composed of two arguments, indicated by the two syntactic functions Suj and Obj. The coverage of the lexicon on the FTB-UC is high, with 99.0% and 96.4% respectively for verbs and adjectives, that are the only two grammatical categories that have available subcategorization frames in the Lefff.

$$\textit{chérir} \rightarrow Suj : (cln|sinf|sn), Obj : (cln|sn)$$

Figure 1: Sample of the Lefff for an entry of the verb chérir (to cherish).

[3] Available under licence at `http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php`

[4] There are also 7 possible syntactic functions attached to nonterminal nodes. Those annotations were removed for our experiments.

[5] `http://atoll.inria.fr/~sagot/lefff.html`

5.2 Word clustering based on the Lefff

The clustering method of verbs and adjectives that we propose in this paper follows the principle of the experiment *LexClust*. A word in the corpus is replaced by the combination of its part-of-speech tag and its cluster. These clusters are computed from the Lefff by exploiting subcategorization frames of entries. First, for each lemmatized form of the lexicon, we reduce its frame to the vector of syntactic functions linked to arguments. If a form appears in several entries (depending on meanings), we merge all vectors into a single one. Then, clusters are determined by grouping forms that have the same vector. Vectors are composed of syntactic functions taken from a subset of the seven most frequent ones, *Suj*, *Obj*, *Objà*, *Objde*, *Loc*, *Att* et *Obl*. This subset allows for creating less clusters and improving results. Table 1 shows an example of the clustering process on several verbs of the Lefff. Each verb is associated with its vector of syntactic functions and its cluster. In this example, vectors of verbs *abolir* and *cibler* are identical and are composed of a subject and a direct object. Therefore, they belong to the same verb cluster, while other verbs are associated with a distinct cluster. Table 2 shows a similar example for adjective clusters.

Verb	Vector	Cluster
abolir (to abolish)	Suj, Obj	1
cibler (to target)	Suj, Obj	1
prouver (to prove)	Suj, Obj, Objà, Obl	2
gratifier (to gratify)	Suj, Obj, Objde	3

Table 1: Verb clusters obtained from the *Lefff*.

Adjective	Vector	Cluster
celtique (celtic)	Suj, Objde, Objà	1
censuré (censored)	Suj, Obl2	2
chanceux (lucky)	Suj, Objde, Objà	1
lavé (washed)	Suj, Obj, Obl2	2

Table 2: Adjective clusters obtained from the *Lefff*.

However, this approach requires a POS tagger and a lemmatizer in order to analyze a raw text (clusters being determined from lemmatized forms). Therefore, we chose one of the best tagger for French called *LGTagger* (Constant and Sigogne, 2011) which is based on a Conditional Random Field probabilistic model. Lemmatization is made with the *Bonsaï* tool[6] which is based on the Lefff and some heuristics in case of ambiguities.

6 Experiments and results

6.1 Evaluation metrics

As the FTB-UC is a small corpus, we used a *cross-validation* procedure for evaluation. This method consists in splitting the corpus into *p* equal parts, then we compute training on *p-1* parts and evaluations on the remaining part. We can iterate this process *p* times. This allows us to calculate an average score for a sample as large as the initial corpus. In our case, we set the parameter *p* to 10. Results on evaluation parts for all sentences are reported using several standard measures, the F_1score and *unlabeled attachment* scores. The labeled F_1score [F1][7] , defined by the standard protocol called PARSEVAL (Black et al., 1991), takes into account the bracketing and labeling of nodes. In order to establish the significance of results between two experiments, we used an unidirectional t-test for two independent samples[8]. The *unlabeled attachment score* [UAS] evaluates the quality of unlabeled dependencies between words of the sentence[9]. Punctuation tokens are ignored in all metrics.

6.2 Berkeley parser settings

We used a modified version of BKY enhanced for tagging unknown and rare French words (Crabbé and Candito, 2008)[10]. We can notice that BKY uses two sets of sentences at training, a learning set and a validation set for optimizing the grammar parameters. As in (Candito et al., 2010), we used 2% of each training part as a validation set and the remaining 98% as a learning set. The number of split and merge cycles was set to 5. The random seed was set to 8.

[6] `http://alpage.inria.fr/statgram/frdep/`

[7] *Evalb* tool available at `http://nlp.cs.nyu.edu/evalb/`

[8] Dan Bikel's tool available at `http://www.cis.upenn.edu/~dbikel/software.html`

[9] This score is computed by automatically converting constituent trees into dependency trees. The conversion procedure is made with the *Bonsaï* tool.

[10] Available in the *Bonsaï* package.

6.3 Clustering methods

We evaluated the impact of our clustering method on verbs and adjectives of the FTB-UC (respectively noted *Verb* and *Adj*). Those of each training part are replaced by the corresponding cluster and, in order to do it on the evaluation part, we used LGTagger and a lemmatizer. Tagging `TAG` and lemmatization `LEM` accuracies of these tools are reported in the Table 3 according to cross-validation on the FTB-UC. In addition to the overall score for all words in the corpus, F1 score is also reported for verbs and adjectives[11]. First, we can see that verbs are efficiently tagged and lemmatized. About adjectives, there is a greater number of errors (about 5%), and this is mainly because of the ambiguity involved with the past (31% of all errors).

	All	Verbs	Adjectives
TAG	97.75	97.83	94.80
LEM	96.77	97.15	95.84

Table 3: Tagging and lemmatization accuracies of LGTagger and Bonsaï lemmatizer according to cross-validation on the FTB-UC.

6.4 Results

The experimental results are shown in the Table 4[12]. The columns *#cls* and *#lex* respectively indicate the number of created clusters and the size of the FTB-UC lexicon according to clustering methods. Note that all results are significant compared to the baseline[13] (t-test$< 10^{-4}$). Absolute gains of experiment *Verb* are about +0.4 for both F1 and UAS. By just modifying verbs, we can drastically reduce the size of the corpus lexicon. About experiment *Adj*, despite lower tagging and lemmatization accuracies, clusters allow to obtain gains of about +0.3 for both F1 and UAS. However, combining *Adj* to *Verb* has no positive effect compared to *Verb* and *Adj*.

So as to compare our results with previous work on word clustering, we report, in Table 5, results of the method *Clust* described in section 4. Moreover, we tried some combination of methods *Verb*, *Adj* and *Clust*. In this case, *Clust* only replaces words of other grammatical categories.

	#cls	#lex	F1	UAS
Baseline	-	27.143	84.03	89.58
Verb	96	20.567	**84.44**	**89.96**
Adj	16	23.982	84.30	89.79
Verb+Adj	112	17.108	84.42	89.92

Table 4: Results from cross-validation evaluation according to our clustering methods.

	#cls	#lex	F1	UAS
Verb	96	20.567	84.44	89.96
Clust	1000	1.987	**85.25**	**90.42**
Verb+Clust	1096	2.186	85.13	90.25
Verb+Adj+Clust	1112	730	84.93	89.98

Table 5: Results from cross-validation evaluation according to our clustering methods.

We can see that *Clust* obtains the best scores, with an absolute gain of +0.9 for F1 and +0.4 for UAS compared to *Verb*. Nevertheless, we obtain similar results to *Clust* when our verb clusters are combined with method *Clust*, applied on all other words of the corpus (t-test$>$0.2). Therefore, it would mean that verb clusters computed from a lexicon are as powerfull as clusters from a statistical model.

7 Conclusion

In this article, we have shown that by using information about verbs (and to a lesser extent, adjectives) from a syntactic lexicon, the Lefff, we are able to improve performances of a statistical parser based on a PCFG grammar. In the near future, we plan to reproduce experiments with other grammatical categories like nouns available in other French lexicons.

References

A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks: building and using parsed corpora*, Kluwer, Dordrecht.

E. Black, S.Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic cov-

[11] We can compute this score because words can be, for example, labeled incorrectly as a verb, or verbs may be labeled incorrectly.

[12] All experiments have a tagging accuracy of about 97%.

[13] Baseline experiment consists in training and evaluating BKY on FTB-UC with original words.

erage of english grammars. In *Proceedings of the DARPA Speech and Naturale Language Workshop*.

P. F. Brown, V. J. Della, P. V. Desouza, J. C. Lai, and R. L. Mercer. 1992. Class-based n-gram models of natural language. In *Computational linguistics, 18(4)*, pages 467–479.

M. Candito and B. Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of IWPT'09*, pages 138–141.

M. Candito and D. Seddah. 2010. Parsing word clusters. In *Proceedings of SPMRL'10*, pages 76–84.

M. Candito, B. Crabbé, and P. Denis. 2010. Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of LREC10*.

M. Constant and A. Sigogne. 2011. MWU-aware Part-of-Speech Tagging with a CRF model and lexical resources. In *ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*, France.

B. Crabbé and M. Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *Proceedings of TALN'08*.

M. Gross. 1994. Constructing Lexicon-grammars. In Atkins and Zampolli, editors, *Computational Approaches to the Lexicon*, pages 213–263.

T. Koo, X. Carreras, and M. Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08*.

T. Matsuzaki, Y. Miyao, and J. Tsujii. 2005. Probabilistic CFG with latent annotations. In *Proceedings of ACL-05*, pages 75–82, Ann Arbor, USA.

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of ACL'06*.

B. Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC'10*.

G. Sampson and A. Babarczy. 2003. A test of the leaf-ancestor metric for parsing accuracy. In *Natural Language Engineering, 9 (4)*, pages 365–380.

D. Seddah, M. Candito, and B. Crabbé. 2009. Adaptation de parsers statistiques lexicalisés pour le français : Une évaluation complète sur corpus arborés. In *Proceedings of TALN'09*, Senlis, France.

A. Sigogne, M. Constant, and E. Laporte. 2011. French parsing enhanced with a word clustering method based on a syntactic lexicon. In *Proceedings of SPMRL'11*, pages 22–27, Dublin, Ireland.