Evaluating an automatically extracted syntactic lexicon

Claire Gardent
CNRS/LORIA
Nancy
claire.gardent@loria.fr

October 12, 2009

Abstract

Previous work has shown that large scale subcategorisation lexicons could be extracted from parsed corpora with reasonably high precision. In this paper, we consider the lexicon resulting from applying a standard extraction procedure to a 100 millions words parsed corpus of french and focus on evaluation. We investigate several ways of evaluating this lexicon and argue that a good interpretation of the results is enhanced by a multifaceted evaluation scheme combining: evaluation against a gold standard, comparison with other existing lexicons and evaluation against a manually verified lexicon sample.

Keywords: Syntactic lexicon, Evaluation, Extraction

1 Introduction

A syntactic lexicon records for each verb, the nature and the type of its arguments. Additionally, it may contain information about the redistribution a verb might accept (e.g., passive, impersonal, reflexive usage) and about specific syntactico-semantic behaviour (e.g., control, raising). As has been repeatedly argued, a syntactic lexicon [7, 12] is an important resource for Natural Language Processing in that it provides valuable information e.g., for parsing, for machine translation or for surface realisation [8, 3]. As a result, corpus based, statistical methods have been proposed which permits the semi-automatic construction of subcategorisation lexicon and usually proceed in two steps [1, 10]. First, a large corpus is parsed and verb dependents are extracted from the available parse trees. Second a statistical filter is applied to determine which of the extracted hypotheses are plausible.

In this paper, we describe the acquisition of a subcategorisation lexicon for French (EASYLEX) and focus on the evaluation of the extracted lexicon. Specifically, we argue that a good interpretation of the results is enhanced by a multifaceted evaluation scheme combining an evaluation against against a reference, a comparison with other existing lexicons and an evaluation against a manually verified lexicon sample.

The structure of the paper is as follows. Section 2 describes the corpus, the parser and the extraction procedure used to extract a subcategorisation lexicon for French verbs from a parsed corpus. Section 3 concentrates on evaluation and presents the results of three distinct evaluation procedures. We show that these three procedures permits a clearer assesment of the usefulness and quality of the extracted lexicon than would a single one. Section 4 concludes.

2 Extraction procedure

As mentioned above, the extraction procedure proceeds in two steps. First, a large corpus is parsed and verb dependents are extracted from the available parse trees. Second several statistical and symbolic filters are applied to determine which of the extracted hypotheses are plausible.

2.1 Dependents extraction.

The experiment is based on the CPC corpus for French[1], a 100 million words corpus containing data extracted in equal proportion from wikisource, frwiki, EstRepublicain, JRCacquis and Europarl.

This corpus was parsed using Gil Francopoulo's TagParser [4], a parser which produces Passage conformant syntactic annotations [13]. Briefly, constituents are annotated using one of the following categories : NV (verbal nucleus), GN (Noun Phrase), GP (prepositional phrase), GA (adjectival phrase), GR (adverbial phrase) and PV (prepositional verb phrase). Further, verbs can be related to other words and/or constituents using one of the following relations : SUJ_V (subject), AUX_V (auxiliary), COD_V (object), CPL_V (prepositional object), MOD_V (verb modifier), ATB_SO (subject or object attribute).

The first step of the extraction procedure consists in extracting for each verb instance, all the dependents that can be arguments of this verb. We therefore store for each verb instance all information given by the parser for each constituent that is related to that verb by one of the following relations : SUJ_V, COD_V CPL_V, ATB_SO. For each such constituent, we store its category (NV, GN, GP, GA, GR or PV), the lemma of the head word and its part-of-speech.

2.2 Hypotheses filtering.

The second step of the extraction procedure groups together instances of the same verb/subcategorisation pair and applies several symbolic and statistical filters as described below.

Symbolic filter. This filter handles passives and repeated dependents. Verb/dependent patterns that indicate a passive are normalised to the corresponding active pattern. Additionally, multiple occurrences of dependents with identical function and category are reduced to one e.g., POBJ:PP[de]. The reason for this is (i) that verbs rarely subcategorise for two or more PPs introduced by the same preposition and (ii) that a given grammatical function need only be satisfied once for any given verb.

Lexicon initialisation. This step includes the following operations. The information stored about multiple occurrences of the same (verb,dependents) pairs is gathered into a single entry. Special cases such as clitics, merged determiners[2] and relative pronouns are dealt with. Frequency and relative frequency are computed. The categories and relations are mapped to the Paris 7 treebank format with syntactic categories NP, PP[PREP], VPinf, VPpart, Ssub, AP, AdvP and grammatical relations SUJ, OBJ, ATS, ATO, AOBJ, DEOBJ, POBJ.

Frame filter. Based on a set of subcategorisation frames extracted from three existing subcategorisation lexicons for French (Dicovalence, TreeLex et Synlex), this filter removes from the extracted lexicon all lexical entries whose frame is unknown i.e., does not belong to this set.

Statistical filter. We observe that many of the frames produced by the extraction script are long and infrequent but include an acceptable frame. We therefore apply [15]'s iterative algorithm to filter out from each lexical entry, frames with low relative frequency. The algorithm works as follows. Frames whose relative frequency is below the set threshold are shortened by one argument and the resulting frame merged back into the current lexical entry. If the entry already contains this frame, the relative frequency of this frame and of the shortened frame are added. Else, the shortened frame inherits the relative frequency of the frames from which it is derived. The order of dependent is normalised and the argument removed is always the right most one that is, the most oblique one.

3 Evaluation

We applied the extraction procedure described in section 2 to the parsed corpus described in section 2.1

[1] `http://atoll.inria.fr/passage/ressources.en.html`

[2] Merged determiners result from the fusion of a preposition and a determiner. For instance, when preceding the determiner *le* (*the*), the preposition *de* (*of*) becomes *du* and *le* is droped.

and obtained a lexicon (called EASYLEX) for 4 847 verbs with an average of almost 6 frames per verb.

We now present three distinct evaluation schemes we used to assess the quality and coverage of this lexicon.

First, we perform a standard evaluation using an existing hand built lexicon called Dicovalence as a reference. This yields rather poor results.

Next we perform a comparative evaluation by evaluating two additional existing lexicons against Dicovalence and comparing their score with those obtained for EasyLex. This strongly suggests that Dicovalence, although hand written, is in fact incomplete thereby negatively impacting precision. This shows also that none of the existing lexicons have a very good recall thus justifying more work on automated lexical acquisition. Since automatically extracted lexicons have high recall, they could be used to extend existing lexicons by making available additional entries whose correctness can be checked using the corpus data which these entries were extracted from.

Finally, we evaluate EasyLex against a manually checked sample of 90 verbs equally distributed across two dimensions (frequency and number of frames) and three values (high, medium, low). This confirms the results of the comparative evaluation by showing that Dicovalence is indeed incomplete and that this incompleteness artificially decreases the precision of EasyLex.

Evaluation against a manually specified lexicon We begin by evaluating EasyLex against a hand build subcategorisation lexicon for French namely, Dicovalence [14]. Dicovalence records the subcategorisation frames of 3 700 verbs chosen amongst the most frequent and was built manually by a team of expert linguists. Dicovalence entries are factorised in that one entry contains information about several frames. In order to compare EasyLex to Dicovalence, we therefore unfolded the Dicovalence entries and converted them to the same format as used for EasyLex (same categories, same grammatical relations)[3]. We then computed the precision, recall and F1 measure of EasyLex with respect to Dicovalence taking into account only those verbs which were present in both lexicons. Precision is the proportion of the extracted entries (i.e., verb/frame pair) that are correct, recall is the proportion of corrected entries that have been extracted and the F1 score is the harmonic mean of precision and recall. The results are summarised in the following table.

V_{DV}	V_{EL}	$V_{DV \bigcap EL}$	P	R	F1
3 937	4 847	3 415	43.35	48.72	45.98

Compared to results reported for similar work on other languages, these results are rather low. For English for instance, [9] reports a precision, recall and F-mesure of 80.7%, 46.1% and 58.6% for a similar automated extraction approach[4].

There can be several reasons for this. The evaluation scheme and in particular, the reference lexicon might be incomplete or inappropriate, the parser might lack precision and/or the extraction procedure might be imperfect.

Indeed, it is clear that the latter two points deserve further work as (i) the parser we use is known to have a 60% precision on dependency labelling and (ii) the extraction procedure needs to be extended with some smoothing technique in order to handle the sparse data issue.

In this paper however, we focus on the evaluation scheme with the aim of better understanding both the source and the meaning of these first results. We aim in particular to answer the following questions :

- Does the reference lexicon lack correct entries? If this is the case, poor precision might be due to these missing entries.
- How does EASYLEXcompare with other existing subcategorisation lexicons for French in terms of precision, recall and coverage?

[3] As one of the referee remarked, this means that the comparison focuses on verbs rather than verb meanings and that therefore the information contained about verb meaning in Dicovalence but absent in the other lexicons is not taken into account.

[4] We consider here the figures the article gives for the approach most similar to ours namely, Lexicon 2 in Table 2, an approach based on a relative frequency threshold filtering with no smoothing.

- Is the information contained complementary or redundant with respect to these lexicons?

Comparison with other lexicons. A second way to evaluate Easylex consists in comparing it with other subcategorisation lexicons for French. We considered three such lexicons namely, TreeLex, SynLex and LexSchem. TreLex [11] was automatically extracted from a handbuilt treebank and manually validated. It therefore contains only correct entries. SynLex [5] was automatically extracted from the LADL table [6] and manually validated but is incomplete due to distribution restrictions on the LADL tables. It therefore contain correct entries but not necessary all the possible frames of a given verb. Finally, LexSchem [2] was automatically extracted from a parsed corpora using techniques similar to ours. It is therefore imperfect.

To better comprehend the limitations of our approach, we converted all three lexicons to the EasyLex format and evaluated them against Dicovalence. The results are listed in the table below.

Lexicon	Verbs	P	R	F1
REFERENCE	3 936			
TreeLex	1 598	**66.01**	*41.06*	50.63
Synlex	2 467	42.88	51.07	46.62
Lexschem	1 730	32.51	43.77	37.31
EasyLex	4 847	48.72	**43.35**	*45.98*

A first remark is that although, TreeLex was validated by hand and therefore only contains correct entries, its precision w.r.t. Dicovalence is of only 66.01%. Hence Dicovalence lacks correct entries. In that context, the 48.72% precision of EasyLex becomes more acceptable.

Second, we observe that the maximum recall for all 4 lexicons is 51.07% (SynLex). In other words, none of the existing lexicons for French are complete even with respect to a hand built reference. It is therefore useful to strive for ways of completing such lexicon in particular, by an automatic lexical acquisition process. As suggested above, such lexicons could be used to manually extend existing lexicons. Alternatively, more sophisticated lexical acquisition could be used such as e.g., the use of verb classes to support backoff smoothing explored in [10] thereby increasing precision and permitting a fully automated acquisition of large scale, high recall lexicons. To attain perfect precision, such lexicons would additionnally need to be hand validated.

Third, we note that EasyLex yields better results overall than Lexschem[5] with a higher F1, higher precision, higher recall and a higher coverage (4 847 verbs and a 92.7% coverage of the Dicovalence verbs for EasyLex against 1 730 verbs and a 43.95% coverage for LexSchem). The higher verb coverage can be explained by the fact that [2] filter out all verbs which occurs less than 100 times in the corpus and that they take into account only finite verbs[6]. The difference in precision and recall may be due to many factors (parser, corpus, extraction procedure) but one observable difference is that we use [15]'s iterative filtering algorithm whilst they do not.

Evaluation against a manually built sample lexicon. As mentioned in the previous sections, the comparative evaluation strongly suggests that Dicovalence is incomplete thereby negatively affecting precision. We wanted to get a clearer idea of what the precision of EasyLex would be on a reference that would be manually checked for missing entries. In practice, we build a reference lexicon for 90 verbs distributed equally along 2 dimensions (frequency and number of frames) and 3 values (high, medium, low). For these 90 verbs, we manually constructed a reference by adding to Dicovalence all frames that were manually found in corpus for the sample verbs and that were not present in Dicovalence. The results obtained by comparing EasyLex against this reference are given below (the DV column gives the results by comparison against Dicovalence, the DV+ column by comparison against Dicovalence extended with additional entries as described above).

[5] For the comparison, we used Version 1 of LexSchem downloaded in January 2009.

[6] Thierry Poibeau and Cedric Messiant personal communication.

	DV	DV+
P	0.42	0.59
R	0.65	0.75
F1	0.51	0.66

The results confirm our expectations. By manually checking the entries suggested by EasyLex for the 90 verbs, we found that Dicovalence was indeed incomplete[7]. By adding these entries to Dicovalence, we gain an increase in of 1.7 point reference. Recall also improves as some Dicovalence entries were removed for which no example could be found nor imagined. More generally, the results obtained on this sample are much closer to the P=80.7%, R=46.1% and F=58.6% figures cited by [9][8].

4 Conclusion

In this paper, we have presented an extraction procedure for the automatic acquisition of a subcategorisation lexicon for French verbs. The results obtained being rather low, we took a careful look at one of the factors that might impact the results namely, the particular evaluation scheme chosen. We showed that by combining an evaluation against an existing hand written lexicon with both a comparative evaluation and an evaluation against a manually checked representative sample, a clearer and more accurate interpretation of the results could be obtained.

In current work, we investigate other potentially impacting factors namely, the quality of the data produced by the parser and that of the extraction procedure. In particular, we are currently working on using the combined results of several parsers on the same corpus . Further, we plan to extend the extraction procedure with smoothing techniques as proposed in [10] and to better tailor it to the specificities of French such as in particular, the presence of clitics and relative pronouns which are strong (although ambiguous) indicators of a given grammatical function.

[7] We only considered verbs present in both Dicovalence and Easylex. Hence incomplete here means that for a given verb present in Dicovalence, frames are missing.

[8] The comparison is not entirely fair however since in [9] the evaluation is made with respect to a full size lexicon based on Comlex.

References

[1] E. Briscoe and J. Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC, 1997.

[2] T. Poibeau C. Messiant and A. Korhonen. Lexschem: a large subcategorization lexicon for french verbs. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

[3] J. Carroll and A. Fang. The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, pages 107–114, Sanya City, China, 2004.

[4] G. Francopoulo. Tagparser et technolangue-easy. In *Actes de l'atelier Easy, TALN*, 2005.

[5] C. Gardent, B. Guillaume, G. Perrier, and I. Falk. Extracting subcategorisation information from Maurice Gross' Grammar Lexicon. *Archives of Control Sciences*, 15(LI):253–264, 2005.

[6] M. Gross. *Méthodes en syntaxe*. Hermann, 1975.

[7] N. Ide and J. Veronis. Multext: Multilingual text tools and corpora. In *Proceedings of COLING 94*, Kyoto, 1994.

[8] V. Jijkoun, J. Mur, and M. de Rijke. Information extraction for question answering: Improving recall through syntactic patterns. In *COLING-2004*, 2004.

[9] A. Korhonen, Y. Krymolowski, and T. Briscoe. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th international conference on Language Resources and Evaluation*, Genova, Italy, 2006.

[10] Anna Korhonen. *Subcategorization Acquisition*. PhD thesis, University of Cambridge, 2002.

[11] A. Kupsc and A. Abeillé. Growing treelex. In Alexander F. Gelbukh, editor, *CICLing*, volume 4919 of *Lecture Notes in Computer Science*, pages 28–39. Springer, 2008.

[12] C. Macleod, R. Grishman, and A. Meyers. COMLEX syntax: Building a computational lexicon. In *Proceedings of COLING '94*, pages 268–272, 1994.

[13] P. Paroubek and G. Francopoulo. *Définition du formalisme d'annotation*. 2007. ANP Passage. Livrable 1.

[14] K. van den Eynde and P. Mertens. La valence: l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies 13, 63-104*, 2003.

[15] D. Zeman and A. Sarkar. Learning verb subcategorization from corpora: Counting frame subsets. In *In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000*, pages 227–233, 2000.