A Word Clustering Approach to Domain Adaptation: Effective Parsing of Biomedical Texts

Marie Candito†, Enrique Henestroza Anguiano† & Djamé Seddah†‡
† Alpage (Univ. Paris Diderot & INRIA), 175 rue du Chevaleret, 75013 Paris, France
‡ Univ. Paris Sorbonne, 28, rue Serpente, 75006 Paris, France
marie.candito@linguist.jussieu.fr, henestro@inria.fr, djame.seddah@paris-sorbonne.fr

Abstract

We present a simple and effective way to perform out-of-domain statistical parsing by drastically reducing lexical data sparseness in a PCFG-LA architecture. We replace terminal symbols with unsupervised word clusters acquired from a large newspaper corpus augmented with biomedical target-domain data. The resulting clusters are effective in bridging the lexical gap between source-domain and target-domain vocabularies. Our experiments combine known self-training techniques with unsupervised word clustering and produce promising results, achieving an error reduction of 21% on a new evaluation set for biomedical text with manual bracketing annotations.

1 Introduction

If Natural Language Processing were the Olympics, statistical parsing would be the combination of "long jump" and "100 meters": a discipline where performance is evaluated in light of raw metric data in a very specific arena. Leaving aside this far-fetched metaphor, it is a fact that statistical constituent-based parsing has long been subjected to an evaluation process that can almost be qualified as *addicted* to its own test set (Gildea, 2001; McClosky et al., 2006; Foster, 2010). However, the gap between this intrinsic evaluation methodology, which is only able to provide a ranking of some parser/treebank pairs using a given metric, and the growing need for accurate wide coverage parsers suitable for coping with an unlimited stream of new data, is currently being tackled more widely. Thus, the task of parsing out-of-domain text becomes crucial.

Various techniques have been proposed to adapt existing parsing models to new genres: domain adaptation via self training (Bacchiani et al., 2006; McClosky et al., 2006; Sagae, 2010), co-training (Steedman et al., 2003), treebank and target transformation (Foster, 2010), source-domain target data matching prior to self-training (Foster et al., 2007), and recently, uptraining techniques (Petrov et al., 2010). Although very diverse in practice, these techniques are all designed to overcome the syntactic and lexical gaps that exist between source domain and target domain data. Interestingly, the lexical gap found for English (Sekine, 1997) can only be wider for out-of-domain parsing of languages that are morphologically richer. Indeed, the relatively small size of their annotated treebanks and their levels of lexical variation are already a stress case for most statistical parsing models, without adding the extreme challenges caused by lexical out-of-domain variation.

In this paper, we take the PCFG-LA framework (Petrov and Klein, 2007), implemented by Attia et al. (2010), and explore a combination of known self-training techniques with a novel application of unsupervised word clustering (Koo et al., 2008) that was successfully used to reduce lexical data sparseness for French parsing (Candito and Crabbé, 2009; Candito and Seddah, 2010).

2 Target Domain Corpus

For our work on domain adaptation, we used the French Treebank (FTB) (Abeillé and Barrier, 2004) as the *source domain* corpus, which consists of 12,351 sentences from the *Le Monde* newspaper. For the *target domain*, we used biomedical texts from the European Medicines Agency, specifically the French part of the EMEA section[1] of the OPUS corpus (Tiedemann, 2009). Although we chose the biomedical domain for this paper, our approach can be used for any target domain.

2.1 Corpus Characteristics

The EMEA corpus includes documents related to medicinal products: it mostly consists of summaries of European public assessment reports (EPAR), each on a specific medicine. The French

[1] opus.lingfil.uu.se/EMEA.php

Proceedings of the 12th International Conference on Parsing Technologies, pages 37–42,
October 5-7, 2011, Dublin City University.

part we used (hereafter EmeaFr) was taken from the English-French aligned bi-text of the EMEA corpus, which consists of raw text converted from PDF files. We estimate that the French part contains around 1000 documents. According to the Standard Operating Procedure of the EMEA for EPARs[2], these documents are first written in English, in a "language understandable by someone not an expert in the field". The translation into all official European languages is managed by the Translation Centre for the Bodies of the European Union (CdT), with standardized terminology for biomedical lay language. As far as we can judge, the quality of the French translation is very good.

This corpus is challenging for domain adaptation: though it contains well-formed sentences, it uses specialized terminology (protocols to test and administrate medicines, and descriptions of diseases, symptoms and counter-indications), and its writing style is very different from that used in the journalistic domain. There are many uses of imperative verbs (in the instructions for use), numerous dosage descriptions, and frequent information within brackets (containing abreviations, glosses of medical terms, and frequency information).

2.2 Corpus Preprocessing

The original EmeaFr corpus contains approximately 14 million words. We corrected some obvious errors from the PDF to text conversion, such as missing quotes after elided tokens (j' for elided "I", n' for elided "not", etc.). We then performed tokenization, segmentation into sentences, and recognition of multiword expressions using the BONSAI package[3], in order to obtain tokenized text that resembles the tokenization of the FTB. Finally we removed lines (sentences) not containing any alphabetical character, as well as duplicated sentences (we kept only one occurrence of each unique tokenized sentence). This resulted in a drastic reduction of the corpus, as many sentences provide general information or recommendations that are repeated in every EPAR document. In the end, the resulting preprocessed corpus (hereafter EmeaFrU) contains approximately 5.3 million tokens and 267 thousand sentences.

2.3 Manual Bracketing Annotation

To evaluate parsing performance, we manually annotated two extracts of the EmeaFrU corpus, corresponding to two EPAR documents: one for development and one for final tests. We removed them from EmeaFrU. In order to obtain gold parses for the development and test sets, we first parsed them using the BONSAI package, which contains the Berkeley parser (Petrov and Klein, 2007), and a French model as described in (Candito et al., 2010). We retained only the POS tags, and had them validated by an expert. Then we reparsed the sets in pre-tagged mode, and had them validated by the same expert, using the WORDFREAK tool (Morton and LaCivita, 2003) that we adapted to French. We removed section numbers starting or ending sentences, table cells, and also a few obviously incomplete sentences.[4]

	Test Set	**Dev Set**
# of sentences	544	574
avg sent. length	21.5	16.2
# of tokens	11,679	9,346
Stats for any type of token		
# of tokens (% unknown)		9,346 (23%)
# of types (% unknown)		1,917 (42%)
Stats for alpha-lc tokens		
# of tokens (% unknown)		8,109 (22%)
# of types (% unknown)		1,608 (36%)

Table 1: Statistics on the EMEA dev and test sets. *alpha-lc* stands for tokens converted to lowercase and containing at least one letter. *Unknown* tokens/types are those absent from the FTB training set.

Table 1 shows a few statistics for the evaluation sets, and a comparison of the dev set vocabulary with that of the FTB standard training set. Focusing on non-punctuation, non-numeric tokens, we see that more than $1/3$ of the vocabulary is unknown (36%), representing 22% of the token occurrences. This strongly motivates a domain adaptation technique focused on lexical variation between the source domain and the target domain.

3 Lexical Domain Adaptation

In our approach to domain adaption, we use unsupervised word clustering performed on a mixture of target-domain (biomedical) and source-domain (journalistic) text. The objective is to obtain clusters grouping together source-domain and target-domain words, thus bridging the two vocabularies.

We build on the work of Candito and Crabbé (2009), who proposed a technique to improve in-domain parsing by reducing lexical data sparse-

[2]Document 3131, at: www.ema.europa.eu

[3]alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

[4]We plan to make the manually-annotated corpus freely available, following a final validation step.

ness: (i) replace tokens with unsupervised word clusters both in training and test data; (ii) learn a grammar from the word-clustered sentences in the training set; (iii) parse the word-clustered sentences in the test set; (iv) reintroduce the original tokens into the test sentences to obtain the final parsed output. The clustering is performed in two steps: (i) a morphological clustering is applied using the Le*fff* morphological lexicon (Sagot et al., 2006), where plural and feminine suffixes are removed from word forms and past/future tenses are mapped to present tense (provided this does not change the part-of-speech ambiguity of the form); (ii) an unsupervised clustering algorithm (Brown et al., 1992) is run on a large unlabeled corpus to learn clusters over the desinflected forms. Both clustering steps proved to be beneficial for parsing in-domain French text using the Berkeley parser.

We apply a similar unsupervised word clustering technique to lexical domain adaptation, with the difference being that clusters are learned over a mixture of source-domain and target-domain text (hereafter *mixed clusters*). We test this technique when training a parser on the FTB training set as well as in self-training mode (McClosky and Charniak, 2008), where the parser is trained on both the source-domain training set and automatically parsed sentences from the target domain.

4 Experiments

For our parsing experiments, we used the PCFG-LA algorithm of Petrov and Klein (2007), implemented by Attia et al. (2010).[5] The treebank used was the FTB (cf. Section 2). More precisely, we used the version of the treebank as defined by Candito and Crabbé (2009), which has a 28 POS tagset and some multiword expressions replaced by regular syntactic structures. We used the standard training (80%), dev (10%), and test (10%) split, containing respectively 9881, 1235 and 1235 sentences from the *Le Monde* newspaper.

For unsupervised clustering, we first systematically applied the desinflection process of Candito and Crabbé (2009), using the Bonsai tool. We obtained *source clusters* by applying Percy Liang's implementation[6] of the Brown clustering algorithm to the *L'Est Républicain* corpus (hereafter ER), a 125 million word journalistic corpus, freely available at CNRTL[7]. Though this newspaper is less formal than *Le Monde*, it is still journalistic, so we consider it as being in the source domain. The *mixed clusters* were obtained by concatenating the *L'Est Républicain* corpus and the EmeaFrU (cf. Section 2), herafter ER+EMEA. We did not investigate any weighting techniques for building the source corpus for mixed clusters. On both the ER and ER+EMEA corpora, we ran Brown clustering with 1000 clusters for the desinflected forms appearing at least 60 times.

	F-Measure on EMEA test set (≤ 40)	
Symbols	**No self-training**	**200k self-training**
raw	81.25	84.75
dfl	81.82	84.72
clt-er	82.65	85.09
clt-er-emea	83.53	85.19

Table 2: F-Measure for sentences $\leq$ 40 tokens on the EMEA test set, both with self-training (200k auto-parsed sentences from EmeaFrU) and without.

Having performed desinflection and different types of clustering, we trained PCFG-LA grammars on the FTB training set using four settings for terminal symbols: *raw* uses original word forms; *dfl* uses desinflected word forms; *clt-er* uses clusters of desinflected forms computed over the ER corpus, with a process described in detail by Candito and Crabbé (2009); *clt-er-emea* is the same as *clt-er*, but with mixed clusters over the ER+EMEA corpus. Having obtained these initial grammars, we used each to parse the EmeaFrU unlabeled corpus (with appropriate desinflection and clustering preprocessing for each of the four terminal symbol settings). We then performed self-training experiments adding up to 200k predicted parses from EmeaFrU to the FTB training set, and training new grammars for each such enlarged training set.

4.1 Results

Figure 1 shows the effect of self-training on parsing the EMEA and FTB dev sets. Unsurprisingly, the baseline parser (*raw* setting without self-training) has a 5 point drop in F-measure when parsing the EMEA compared to the FTB. Consistent with previous results on English biomedical texts (Lease and Charniak, 2005; McClosky and Charniak, 2008), self-training helps in parsing the EMEA, with more predicted parses generally leading to better performance on the EMEA (and worse performance on the FTB).

[5] Our experiments were run using five split-merge cycles and tuned suffixes for handling French unknown words.

[6] www.eecs.berkeley.edu/~pliang/software

[7] www.cnrtl.fr/corpus/estrepublicain

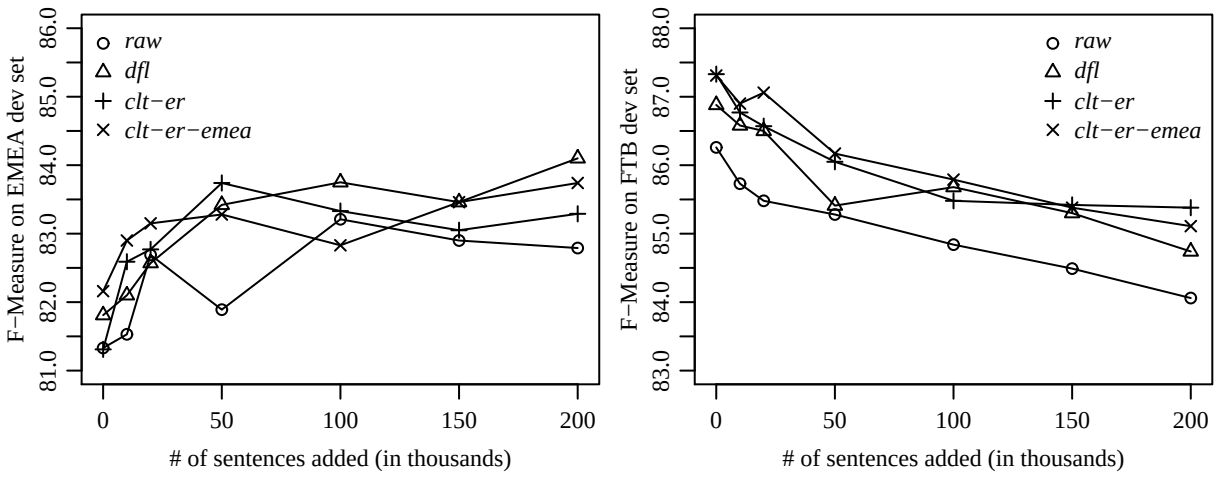

Figure 1: F-Measure for sentences of length ≤ 40 on the EMEA (left) and FTB (right) dev sets using self-training (with different amounts of auto-parsed sentences from EmeaFrU), by terminal symbol setting.

Concerning the different settings for terminal symbols, for source-domain data we reproduce (right side of Figure 1) the findings of Candito and Crabbé (2009): parsing desinflected forms (*dfl*) increases performance, and parsing unsupervised clusters of desinflected forms (*clt-er* and *clt-er-emea*) is even better. For target-domain data, we find that desinflection does help, and even achieves better performance than source clusters. This can be explained by the fact that the desinflection process provides forms that still follow French morphology, so the general handling of unknown words (with classes of suffixes) does apply. In contrast, terminological tokens are (hopefully) more frequently absent from the ER corpus than the ER+EMEA corpus, and they are replaced by the UNK token more often for source clusters (*clt-er*) than for mixed clusters (*clt-er-emea*). Indeed, for the EMEA dev set, 1,466 tokens are UNK in the *clt-er* setting, while only 729 tokens are UNK in the *clt-er-emea* setting.

Table 2 shows final parsing results on the EMEA test set for each of the four terminal symbol settings, with and without self-training (using 200k parses from EmeaFrU). We evaluated using F-Measure on labeled precision and recall, ignoring punctuation, and calculated the significance of differences between settings.[8] The *clt-er-emea* setting gives the best overall performance, with or without self-training. When comparing *clt-er-emea* with self-training (best overall) to *raw* without self-training (baseline), we obtain a 21% error reduction. This result is encouraging, given the small amount of raw target-domain data added to the ER corpus (5M added to 125M words). However, self-training produces the most pronounced increase in performance (statistically significant improvement over no self-training for each terminal symbol setting), and attenuates the improvement attained by clustering: while *clt-er-emea* is significantly better than *raw* or *dfl* without self-training, the differences are not significant with self-training. More raw target-domain data may be needed for mixed clusters to be fully effective.

[8]Significance at $p = 0.05$, using Bikel's Statistical Significance Tester: www.cis.upenn.edu/~dbikel/software.html

5 Conclusion

We have proposed a technique of parsing word clusters for domain adaptation, clustering together source and target-domain words. We have shown this to be beneficial for parsing biomedical French texts, though it did not provide significant additional improvement over self-training.

Our perspectives for future work are to investigate: (i) producing mixed clusters with a larger unlabeled target-domain corpus; (ii) using lexicon-informed part-of-speech taggers; (iii) supplementing our approach with other techniques like reranking, known to improve self-training for domain adaptation (McClosky and Charniak, 2008), or uptraining (Petrov et al., 2010).

Acknowledgements

Thanks to J. Foster, D. Hogan and J. Le Roux for making the LORG parser available to us and to the French National Research Agency (SEQUOIA project ANR-08-EMER-013).

References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a french treebank. In *Proc. of LREC'04*, Lisbon, Portugal.

Mohammed Attia, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi, and Josef van Genabith. 2010. Handling unknown words in statistical latent-variable parsing models for arabic, english and french. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

M. Bacchiani, M. Riley, B. Roark, and R. Sproat. 2006. Map adaptation of stochastic grammars. *Computer speech & language*, 20(1):41–68.

Peter F. Brown, Vincent J. Della, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 138–141, Paris, France, October. Association for Computational Linguistics.

Marie Candito and Djamé Seddah. 2010. Parsing word clusters. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, CA.

Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. 2010. Benchmarking of statistical dependency parsers for french. In *Proceedings of COLING 2010*, Beijing, China.

J. Foster, J. Wagner, D. Seddah, and J. Van Genabith. 2007. Adapting wsj-trained parsers to the british national corpus using in-domain self-training. In *Proceedings of the Tenth IWPT*, pages 33–35.

Jennifer Foster. 2010. "cba to check the spelling": Investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California, June. Association for Computational Linguistics.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 167–202.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08*, pages 595–603, Columbus, USA.

M. Lease and E. Charniak. 2005. Parsing biomedical literature. *Natural Language Processing–IJCNLP 2005*, pages 58–69.

David McClosky and Eugene Charniak. 2008. Self-training for biomedical parsing. In *Proceedings of ACL-08: HLT, Short Papers*, pages 101–104, Columbus, Ohio, June. Association for Computational Linguistics.

D. McClosky, E. Charniak, and M. Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 337–344. Association for Computational Linguistics.

T. Morton and J. LaCivita. 2003. Wordfreak: an open tool for linguistic annotation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4*, pages 17–18. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

S. Petrov, P.C. Chang, M. Ringgaard, and H. Alshawi. 2010. Uptraining for accurate deterministic question parsing. In *Proceedings of the*

2010 Conference on Empirical Methods in Natural Language Processing, pages 705–713. Association for Computational Linguistics.

Kenji Sagae. 2010. Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 37–44, Uppsala, Sweden, July. Association for Computational Linguistics.

Benoît Sagot, Lionel Clément, Eric V. de La Clergerie, and Pierre Boullier. 2006. The lefff 2 syntactic lexicon for french: Architecture, acquisition, use. *Proc. of LREC 06, Genoa, Italy*.

S. Sekine. 1997. The domain dependence of parsing. In *Proceedings of the fifth conference on Applied natural language processing*, pages 96–102. Association for Computational Linguistics.

M. Steedman, R. Hwa, S. Clark, M. Osborne, A. Sarkar, J. Hockenmaier, P. Ruhlen, S. Baker, and J. Crim. 2003. Example selection for bootstrapping statistical parsers. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 157–164. Association for Computational Linguistics.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.