A Word Clustering Approach to Domain Adaptation: Robust Parsing of Source and Target Domains *Unedited final version. To appear shortly*

Djamé Seddah[(1,2)], Marie Candito[(1)] & Enrique Henestroza Anguiano[(1)]
[(1)] Alpage (Univ. Paris Diderot & INRIA), 175 rue du Chevaleret, 75013 Paris, France
[(2)] Univ. Paris Sorbonne, 28, rue Serpente, 75006 Paris, France
djame.seddah@paris-sorbonne.fr
marie.candito@linguist.jussieu.fr, henestro@inria.fr

December 11, 2012

Abstract

We present a technique to improve out-of-domain statistical parsing by 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 target-domain data. We also investigate the impact of guiding out-of-domain parsing with predicted part-of-speech tags. We provide an evaluation for French, and obtain improvements in performance for both non-technical and technical target domains. Though the improvements over a strong baseline are slight, an interesting result is that the proposed techniques also improve parsing performance on the source domain, contrary to techniques such as self-training, thus leading to a more robust parser overall. We also describe new target domain evaluation treebanks, freely available, that comprise a total of about 3,000 annotated sentences from the medical domain, regional newspaper articles, French Europarl and French Wikipedia.

1 Introduction

If Natural Language Processing were the Olympics, statistical parsing would be the combination of the long jump and the 100-meter dash: 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 subject 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., 2003a), 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 the use for domain adaptation 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). We also investigate the impact of guiding out-of-domain parsing with predicted part-of-speech tags.

We present in the next section the target domains used for evaluating our techniques, and the new gold standard data set that we had manually validated for that purpose. We then describe in section 3 the proposed lexical domain adaptation technique. Section 4 presents our experiments, results and discussion. We then compare with self-training approaches in section 5, and we provide concluding remarks in section 6.

2 Target Domains 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[1]. For the *target domains*, we used four corpus sources:

- the *L'Est Républicain* corpus, a corpus of regional news, freely available through the CNRTL[2];
- the French Wikipedia;
- the French part of the Europarl parallel corpus[3];
- biomedical texts from the European Medicines Agency, specifically the French part of the EMEA section[4] of the OPUS corpus (Tiedemann, 2009).

As regional newspaper, the first of these target domains can be considered closest to the source domain (articles from the national newspaper *Le Monde*). The French Wikipedia and the French Europarl constitute two other target domains that potentially potentially have more differences. Finally, the biomedical texts are intended to provide a technical domain that differs greatly in both syntax and vocabulary from journalistic texts.

For these four domains, we use both unlabeled data (to compute unsupervised word clusters) and manually validated parses (to evaluate domain adaptation parsing experiments). We describe both kinds of data in the next two sections.

2.1 Unlabeled target domain corpus

2.1.1 *L'Est Républicain* newspaper

The *L'Est Républicain* corpus contains three years of news from this regional newspaper. We performed tokenization, segmentation into sentences, and recognition of multiword expressions using the bonsai package[5] (hereafter *bonsai preprocessing*), in order to obtain tokenized text that resembles the tokenization of the FTB. This results in a 150 million token corpus.

[1] We use the treebank as released in 2007, and keep the 12,351 sentences that contain functional annotations — though we discard the functional annotations for our experiments — because the syntactic annotations of these sentences are known to be more reliable. The currently released version of the FTB contains around 4,000 additional sentences.

[2] *http://www.cnrtl.fr/corpus/estrepublicain*

[3] http://www.statmt.org/europarl/

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

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

2.1.2 Medical domain

The EMEA corpus, included in the OPUS corpus (Tiedemann, 2009), contains documents related to medicinal products: it mostly consists of summaries of European public assessment reports (EPAR), each on a specific medicine. The French part we used (hereafter EmeaFr) was taken from the English-French aligned bitext of the EMEA corpus, which consists of raw text converted from PDF files. We estimate that the French part contains around 1,000 documents. According to the Standard Operating Procedure of the EMEA for EPARs[6], 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 administer 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).

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 applied preprocessing using bonsai, and then 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 267,000 sentences.

Corpus Extension In order to obtain a bigger unlabeled corpus for this domain, we used pages from the *www.doctissimo.fr* web site, which is dedicated to health and well-being for the general public. More precisely we used the pages describing medicines[7] and the glossary pages[8], in order to stick to the medicine-oriented

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

[7] http://www.doctissimo.fr/html/medicaments/medicaments.htm

[8] http://www.doctissimo.fr/html/sante/encyclopedie

domain. After extracting the raw text from the XML source documents, we applied the same preprocessing steps as before and then concatenated the doctissimo corpus to the EmeaFrU corpus, finally removing duplicate sentences. In the end we obtained a 12.1 million token corpus for the medical domain, mainly focused on the descriptions of medicines (usage, operation, counter-indications).

2.1.3 French Europarl

The French part of the Europarl parallel corpus constitutes another target domain. Because Europarl mainly contains translations of read dialogues, its syntactic characteristics may differ from journalistic text. We preprocessed the corpus using bonsai, and then retained 12 million tokens so as to use the same amount of unlabeled data as that for the medical domain.

2.1.4 French Wikipedia

The last target domain is from the French Wikipedia. We selected Wikipedia entries already gathered in the PASSAGE corpus (Villemonte De La Clergerie et al., 2008), so as to obtain 12 million tokens after preprocessing using bonsai.

2.2 Manually validated treebanks for target domains

To evaluate parsing performance, we manually annotated extracts of our four target domains. This constitutes the *Sequoia* corpus[9].

For the medical domain, we extracted two EPAR documents from the EmeaFrU corpus, to use for development and for final tests, respectively. From the French Wikipedia, the annotated corpus is made of 19 Wikipedia entries, each concerning famous political affairs. From the *L'Est Républicain* corpus, we chose to manually validate 39 articles that were selected from within the ANNODIS project[10], a project dedicated to discourse annotation. Finally, for Europarl we randomly took speaker turns among those included in the PASSAGE corpus (Villemonte De La Clergerie et al., 2008).

2.2.1 Annotation scheme

In order to obtain evaluation treebanks compatible with parsers trained on the FTB, we have used the annotation scheme of the FTB, and followed as much as possible the corresponding annotation guides for morphology, syntactic structure and

[9]Named after the project (SEQUOIA ANR-08-EMER-013) the funding was taken from. The manually-annotated corpus will be freely available.

[10]http://w3.erss.univ-tlse2.fr/annodis

functional annotations (Abeillé and Clément, 2006; Abeillé et al., 2004; Abeillé, 2004).

More precisely, we targeted the annotation scheme of the FTB-UC (Candito et al., 2010a), which was obtained by automatic modification of the FTB. The modifications concern compound words (or multi-word units), the tagset, and the standardization of preposition and complementizer projections:

- Compounds: These occur frequently in the FTB, with 17% of tokens belong to a compound. Compounds range from very frozen multi-word expressions like *y compris* (literally *there included*, meaning *including*) to named entities. They also include syntactically regular compounds with compositional semantics, such as *loi agraire* (*land law*), that are encoded as compounds because of a non-free lexical selection. These syntactically regular compounds tend to be inconsistently encoded in the FTB [11]. Further, the FTB includes "verbal compounds" that are potentially discontinuous, which leads to inconsistent annotations. In the FTB-UC, syntactically regular compounds are mapped to a regular syntactic representation. We followed this rule for the annotation of the Sequoia corpus. This has the virtue of uniformity, but clearly requires further treatment to spot clear cases of compounds (with non-compositional semantics).

- Tagset: the tagset includes 28 POS tags, originally tuned by (Crabbé and Candito, 2008) to optimize parsing, which are a combination of one of 13 coarse-grained categories and verbal mood information in addition to some other distinctions (proper versus common nouns, wh-feature, etc.) that are encoded in the FTB as features.

- Prepositions and complementizers: In our annotations, prepositions project a PP independently of the category of their object, contrary to the FTB, where prepositions with nominal objects project a PP but those with infinitival objects do not. Further, we systematically use a sentential phrase as sister node to complementizers, contrary to the flat structure approach favored by the FTB [12].

[11] For instance, *pays industrialisés* (*industrialized countries*) appears twice as a compound and 41 times as two words; *taux d'intérêt* (*interest rate*) appears 80 times as a compound and 25 times as two words.

[12] For the target domains manually validated treebanks, we provide the FTB-UC format and a version complying with the original FTB annotation scheme concerning prepositions and complementizers, obtained using the stanford tsurgeon tool (Levy and Andrew, 2006).

2.2.2 Annotation methodology

To obtain the manually validated Sequoia treebank, we have alternated steps of automatic preprocessing and manual validation. At each step, annotators were able to modify the choices made at previous stages. The steps were as follows:

- Automatic preprocessing using bonsai: segmentation into sentences, tokenization and non-contextual recognition of multi-word units;
- Automatic part-of-speech tagging using MElt (Denis and Sagot, 2009);
- Manual validation of the preceding steps by a single expert annotator. We removed section numbers starting or ending sentences, table cells, and also a few obviously incomplete sentences;
- Automatic parsing using two statistical parsers trained on the FTB-UC, provided with the manually validated tagging. The parsers were the Berkeley parser (Petrov and Klein, 2007) and the Charniak parser (Charniak, 2000), both adapted to French;
- Independent manual validation of the outputs from the two parsers, using the WordFreak graphical tool (Morton and LaCivita, 2003), followed by adjudication;
- Automatic annotation of grammatical functions for dependents of tensed verbs using the annotator integrated in bonsai;
- Validation via Wordfreak by two annotators, and adjudication;
- Systematic verification of distinctions known to be difficult, and of the coherence of multi-word unit annotations.

2.2.3 Corpus characteristics

We provide quantified characteristics for the Sequoia treebank in table 1, and compare them to the development and training sets[13] of the FTB-UC (hereafter FTB-dev and FTB-train).

We have approximately 500 annotated sentences from Europarl and *L'Est Républicain* (EstRép), and almost 1,000 for the medical domain (EMEA-dev and EMEA-test) and from the French Wikipedia (FrWiki). The average sentence length varies

[13] These are defined using the standard split into training (80%), dev (10%), and test (10%) split, containing respectively 9881, 1235 and 1235 sentences.

	Sequoia Corpus					FTB	
	Medical		Neutral				
	EMEA dev	EMEA test	Est Rép.	Euro Parl	Fr Wiki	dev	train
# Sentences	574	544	529	561	996	1235	9881
Avrg. length	16,3	22,0	21,0	26,3	22,2	29,6	28,1
Std. deviation	14,7	15,0	12,9	15,0	18,0	16,0	16,5
Counts using any token type (including punctuation)							
Vocabulary size	1916	1737	3337	3300	4687	7222	24110
% of unknown	41.4	35.8	29.2	20.6	34.2	22.5	-
# occurrences	9343	11964	11114	14745	22080	36508	278k
% of unknown	23.0	19.7	11.2	6.6	12.9	5.2	-
% of occ. of proper nouns	1.7	2.7	5.1	2.9	9.7	4.1	4.0
Counts using lower-cased alphanumerical tokens							
Vocabulary size	1695	1599	3173	3165	4410	6904	22526
% of unknown	36.6	34.0	28.0	20.1	32.6	21.6	-
# occurrences	8107	10451	9552	13073	18619	30940	235k
% of unknown	23.2	20.9	12.1	7.0	13.8	5.7	-

Table 1: Lexical characteristics of the manually annotated extracts. *Unknown* refers to tokens that are absent in the FTB-train.

for the different domains, but it can be noted that for each domain the standard deviation is high. Surprisingly, the FTB has the longest sentences on average (29.6 for FTB-dev and 28.1 for FTB-train), longer even than Europarl (26.3).

Table 1 also provides the vocabulary size (the number of token types) for each subcorpus, along with the proportion among these that are unknown in the FTB-train. We give the figures computed both using every token and using alphanumerical tokens mapped to lower-case, the latter providing a better evaluation of lexical diversity. We note that the medical corpus has a smaller vocabulary than the other subcorpus. For instance, EMEA-dev and EstRép have a comparable number of occurrences (around 11,000 occurrences), but EMEA-dev's vocabulary has half the size of the vocabulary of EstRép Yet, the medical domain has by far the vocabulary most different from the FTB (more than one third of the vocabulary is unknown in the FTB). This indicates that the unknown words in EMEA might be frequently used words in this domain. For EMEA-dev, the higher proportion of unknown token types (41.4%) is due to a high number of tokens appearing fully in uppercase (this proportion decreases to 36.6% when punctuation is ignored and tokens are

mapped to lower-case). For the FrWiki corpus, the high proportion of unknown token types (34.2%) could be due to the high frequency of proper nouns in this corpus (cf. the row *% of occ. of proper nouns*: approximately one occurrence out of ten is a proper noun in FrWiki).

Looking at the proportion of unknown token occurrences, as opposed to token types, we note that unknown tokens are quite frequent in EMEA: one occurrence out of five (and almost one out of four for EMEA-dev) is unseen in the FTB-train. One can also note the low percentages of proper noun occurrences for EMEA (1.7% in EMEA-dev and 2.7% in EMEA-test). On the other hand, for the FrWiki corpus, the proportion of unknown decreases drastically when computed within occurrences (from 34.2% of the vocabulary to 12.9% of the occurrences), which indicates that the majority of token types unknown in the FTB-train are not frequent in FrWiki. Finally, the corpus whose vocabulary is closest to the FTB seems to be Europarl. Indeed, the Europarl extract seems to be lexically no more different from the FTB-train than the FTB-dev is.

3 Lexical Domain Adaptation

In our approach to domain adaptation, we first adapt a method based on word clustering that has proven useful for source domain parsing through the reduction of lexical data sparseness. Building on the work of (Koo et al., 2008), in which unsupervised word clusters are used as features in a discriminative dependency parser, Candito and Crabbé (2009) proposed to use clusters as word substitutes in a generative constituency parser. The process is to: (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.

Two kinds of clusters are used:

- in the *disinflected* mode, the clustering is performed in two steps. Word forms are first grouped into morphological clusters using a morphological lexicon. Any word form known in the lexicon is mapped to a *disinflected form*, namely another existing word form with possibly less morphological marking, but with exactly the same part-of-speech ambiguity as the original form. The point is to reduce lexical sparseness, without committing as far as tagging is concerned. Plural and feminine suffixes are removed from word forms and past/future tenses are mapped to present tense, provided original ambiguity is retained.[14] Disinflected forms play the role of morphological

[14]For instance, the word form *entrées* is ambiguous between the plural form of the feminine noun

word clusters, which are in turn clusterized using an unsupervised clustering algorithm, run on a large disinflected corpus.

- in the *lemma-tag* mode (Candito and Seddah, 2010), text is first automatically tagged and lemmatized, and unsupervised clusters are computed over lemma+tag pairs. The lemmatization provides lexical reduction, and the tagging provides the linguistic distinctions crucial to parsing, such as whether a verb is finite or infinitive.

We apply an unsupervised word clustering technique to lexical domain adaptation, similar to that used by the previous works mentioned above, with the difference being that clusters are learned over a mixture of source-domain and target-domain text (hereafter *mixed clusters*). The objective is to obtain clusters grouping together source-domain and target-domain words, thus bridging the two vocabularies. Another of our contributions is that we investigate the impact of part-of-speech tagging for parsing target domains; we thus test the use of predicted part-of-speech tags in both the training corpus and the test corpus to guide parsing, for both word clustering modes.

4 Parsing Experiments

In this section, we present the main results of a set of experiments we conducted to study the impact of various forms of word clustering on parsing of French out-of-domain data. Our experiments can be summarized as a work conducted on three axes: (i) the variation of type of word clustering (morphological, semi-supervised) ; (ii) the variation of the unlabeled data used to generate word[15] clusters; and (iii) the variation of parsing mode (no part-of-speech (POS) supplied, predicted POS), each corresponding to a different setting of the POS's training set.

4.1 Experimental Protocol

Parser For our parsing experiments, we use the Lorg parser (Attia et al., 2010), which is an implementation of the PCFG with Latent Annotations (PCFG-LA)

entrée (*entrance*) or the feminine plural past participle of the verb *entrer* (*to enter*). The disinflection removes the plural mark (*s*), leading to the form *entrée*, which has the same noun/verb ambiguity. The verbal interpretation could be further simplified by removing the final feminine morpheme *e*, leading to singular masculine past participle *entré*, but this is not performed because the nominal interpretation would be lost. Conversely, the lexicon is also used to stop the disinflection if it would result in the addition of an ambiguity.

[15]Here, word must be understood in the general "token" sense.

algorithm of Petrov and Klein (2007). Our experiments are run using five split-merge cycles and tuned suffixes for handling French unknown words (setup named FRENCHIG by Attia et al. (2010)) for all token configurations except for the <lemma,pos> ones, which do not use any sophisticated unknown words processing (i.e., GENERIC mode). The threshold under which tokens are considered unknown is 1 (only true unknowns and hapaxes). Experiments using higher thresholds (5 and 10) showed a lower baseline and led to slightly inferior, or at best comparable, results in most configurations.
In a previous version of this work, experiments were conducted within a PCFG-LA framework using one grammar extracted with a random seed of 29 and 5 split-merge cycles. However, Petrov (2010) showed that the random nature of the EM training in PCFG-LA could lead to substantial variation in performance. For that reason, experiments presented in this work were conducted using a product of four grammars with random seed ranging from 1 to 4.

Evaluation Metrics We provide POS tagging accuracy, as well as F-measure of labeled brackets precision and labeled brackets recall. F-measure was computed using EVALB (Black et al., 1991), ignoring punctuation tokens, and calculated over sentences of less than 40 tokens[16].

Data Set We use the FTB-train as training set (cf. the split mentioned in section 2.2.3). For testing parsing performance for the target domains, we consider the medical domain as an example of far-from-source target domain, and we gather FrWiki, EstRep and Europarl as an example of non-technical target domain, closer to our source domain (cf. the lexical characteristics of table 1). For development we use FTB DEV for the source domain, EMEA DEV for the medical domain, and SEQUOIA DEV for the non technical target domain, obtained by concatenating the first halves of the annotated sentences of FrWiki, EstRép and Europarl (1,043 sentences, among which 891 are of length < 40).

Token Types and word clusters To produce **disinflected** tokens (Candito et al.; Candito and Seddah, 2010), we use the disinflection tool embedded in the BONSAI parsing chain (Candito et al., 2010b), that uses the Le*fff* (Sagot et al., 2006) as morphological lexicon.

[16]When running EVALB, we encountered many length mismatch problems within certain settings, due to tokens tagged as punctuation in the test set and not in the gold reference or VICE-VERSA. To circumvent this problem, we ran EVALB with gold punctuation tags reinserted in lieu of the wrong tag, leaving the F-measure unchanged. For the opposite case, in which a token was wrongly tagged as a punctuation, a special tag was inserted, keeping the POS accuracy unchanged.

To produce **lemma+tag** tokens, the MORFETTE morphological analyser (Chrupała et al., 2008), adapted to French by Seddah et al. (2010) is used. It is based on joint modeling of lemmatisation and POS tagging that use the Le*fff* lexicon for additional features.
For the two cluster modes (**disinflected** and **lemma+tag**), unsupervised clusters are computed using the class-based algorithm of Brown et al. (1992), implemented by (Liang, 2005). For cluster generation, we always asked for 1,000 clusters and included only units (either disinflected forms or lemma+tag pairs) appearing more than 100 times in the unlabeled corpus. **Source clusters** are obtained using the 150 million token *L'Est Républicain* corpus (section 2.1.1), which is considered as being close to the source domain. **Bridge clusters** are obtained using a concatenation of this corpus and the three unlabeled corpora of 12 million tokens each, from the medical domain, French Europarl and French Wikipedia (Section 2.1), for a total of 186 million tokens[17]. Token types are presented Table 2.[18] In this table and in the following sections, we refer to the various experiments using the following conventions: token type in training and testing sets are either original tokens (WORD), disinflected forms (DFL), predicted lemma-tag pairs (MORF) or clusters (K-). Clusters are further distinguished as unsupervised clusters over disinflected forms (K-DFL) or over lemma-tag pairs (K-MORF), and as being computed over an unlabeled corpus from the source domain (K-DFL-SOURCE and K-MORF-SOURCE) or from a corpus of both source and target domains (K-DFL-BRIDGE and K-MORF-BRIDGE).

4.2 Baseline Experiments: Target domain parsing and Token Variation

This set of experiments is aimed at verifying the impact of each token variation, on each corpus (with the WORD setting constituting our primary baseline). The gold part-of-speech are used at training time, and part-of-speech tags are not supplied at testing time (the parser performs the tagging). Results are shown in Table 3 and exhibit a strong divergence with previously published results on French parsing with word clusters (Candito and Crabbé, 2009; Candito and Seddah, 2010).

In those previous works, a neat increase of parsing performance was noted between different mode of word clustering (WORD < DFL < K-DFL) when parsing

[17]When clustering over these corpora, we removed all sentences appearing in their corresponding annotated treebanks.

[18]Note that both types of generated clusters are appended with suffixes (morphological clues for disinflected clusters and 3 last letters for lemma+pos ones). We use a special unknown cluster, and such suffixes for replacing either disinflected forms or lemma+tag pairs that were not associated to any cluster (i.e. that did not appear more than 100 times in the unlabeled corpus used for clustering computation.

(Alias)	**Token Type**				
(-)	**Word**				
	la	CFDT	réclamera	des	augmentations
	the	*CFDT*	*will ask for*	*some*	*increases*
DFL	**Morpho clusters**				
	le	CFDT	réclamez	des	augmentation
K-DFL	**unsupervised clusters + _CAP + suffixes**				
	00	10101001_CAP	11100100_ez	1101110	10011001
MORF	**<lemma,pos> input**				
	le_DET	CFDT_NPP	réclamer_V	le_DET	augmentation_NC
K-MORF	**unsupervised clusters built on <lemma,pos>+ 3 token letters suffixes**				
	0110_DET	10110001_NPP	11100101_r_V	0110_DET	10011011_NC

Table 2: Tokens Type Illustrations

Cluster Type	N/A	N/A	N/A	BRIGDE	SOURCE	BRIGDE	SOURCE
Token Type	WORD	DFL	MORF	K-MORF	K-MORF	K-DFL	K-DFL
FTB DEV							
Bracketing FMeasure	87.68[a]	87.99[b]	86.09	87.02	87.12	**88.06**[c]	87.97[d]
Tagging accuracy	96.52	96.69	94.7	95.91	95.67	96.53	96.54
SEQUOIA DEV							
Bracketing FMeasure	86.79	87.83[e]	85.24	85.7	85.01	**87.84**[f]	87.62
Tagging accuracy	95.74	96.08	93.22	94	93.69	96.31	96.37
EMEA DEV							
Bracketing FMeasure	82.59	83.01	80.92'	81.90'	**83.76'**	82.19	82.74
Tagging accuracy	89.37	91	88.53	93.05	92.77	90.75	89.83

Table 3: Development Set: Baseline Results. *When a parsing run leads to more than one parse error, the corresponding result is quoted. Compared to (a), all* FTB DEV *results are statistically significant. Differences between (c) & (b), (c) & (d) and (e) & (f) are not.*

the FTB dev and test sections. Here, the differences between those three modes are barely noticeable for in-domain parsing. Different possible reasons may explain this discrepancy: (i) the implementations of the PCFG-LA framework are different, since we use here Lorg instead of the Berkeley parser; (ii) this work uses a product of four grammars; and (iii) the handling of unknown words is different because, for Lorg, tokens appearing a number of times less than or equal to a threshold (1 in our experiments) are mapped to classes of unknown words, according to their suffixes, the list of used suffixes being automatically learned. For the Berkeley parser, the same kind of handling is used for true unknown words, and a smoothing is performed for rare words (the threshold being 10). Indeed, when we use the Berkeley parser, and an average of four runs with four random seeds, we obtain for FTB-dev an average F-measure of 85.7, increased to 86.6 for the DFL mode, and to 87.3 for the K-DFL-BRIDGE mode. The baseline results are thus

much higher with the Lorg parser, and the DFL and cluster modes fail to improve greatly over a stronger baseline.
As expected from previous experiments on various language in out-of-domain parsing, parsing the target domain with word tokens exhibits a drop in performance when compared to the source domain baseline.

Focusing on source parsing results (i.e., FTB DEV), results show a slight and statistically significant[19] improvement over the baseline for the K-DFL-BRIDGE configuration (3% of error reduction). Note that other configurations built on DFL tokens (DFL, K-DFL-SOURCE) provide results with non statistically significant differences.
Regarding the out-of-domain corpora, results are encouraging as a clear improvement over the baseline is shown on the SEQUOIA DEV with all DFL configurations. Note than even though the performance of MORF clusters are disappointing at best in this baseline setting, the interest of using a bridge corpus to compute clusters is verified over the two clusters settings (K-DFL and K-MORF) on the SEQUOIA DEV.
The results on EMEA DEV show a different trend. First, using the bridge clusters seems to be detrimental compared to performance obtained with clusters computed over a source domain corpus. This is particularly true for the K-MORF-BRIDGE setting. Note though that the MORF and K-MORF-* experiments gave rise to 3 non parsed sentences, leading to results that are not directly comparable.

Concerning tagging performance obtained through parsing, the general trend is that the DFL* treatments do not degrade tagging performance with respect to the baseline WORD setting, whereas the MORF* treatments do. This is probably due to a wrong combination of the suffixes used for the MORF* treatments with the OOV handling within the Lorg parser.[20] Therefore, one can assume that using an external tagging step will alleviate this effect (which is done in next section).[21]

4.3 Parsing SOURCE and TARGET Data with Predicted POS Tags

This experiment is meant to evaluate a pipeline architecture where the parser will have predicted pos tags as input. To make it more tolerant to tagging mistakes and

[19] Statistical significance calculated by Daniel Bikel's compare.pl script.

[20] See (Attia et al., 2010) for details on Lorg's unknown words treatment.

[21] One can note though that the K-MORF-BRIDGE results on EMEA DEV are counter-intuitive since the best tagging result (93.05) arises with the worst parsing Fmeasure. We have currently no explanation for this idiosyncratic result.

somehow more robust, we decide to perform a 10 fold jackknifing[22] to generate *noisy* automatically predicted POS tags to be reinserted in lieu of gold POS tags in the treebank training set. The tagger is trained on the regular training set to provide POS tags for our dev and test sets.

Cluster Type	N/A	N/A	N/A	BRIGDE	SOURCE	BRIGDE	SOURCE
Token Type	WORD	DFL	MORF	K-MORF	K-MORF	K-DFL	K-DFL
FTB DEV							
Bracketing FMeasure	88.1	88.33'	88.22	88.71[a]	88.45	88.56	**88.74**[b]
Tagging accuracy	97.05	97.06	97.06	97.06	97.06	97.06	97.06
SEQUOIA DEV							
Bracketing FMeasure	87.14'	87.65'	87.11	**87.94**[c]	87.14	87.57	87.64[d]
Tagging accuracy	96.56	96.56	96.56	96.56	96.56	96.57	96.57
EMEA DEV							
Bracketing FMeasure	83.1	83.12'	83.32	**84.25**[e]	84.21	83.43	84.09[f]
Tagging accuracy	92.74	92.75	92.79	92.75	92.75	92.75	92.75

Table 4: Development Set: Train set with predicted POS (JK) and devset with predicted POS supplied

Results (Table 4) indicate that using predicted POS with a jacknifed training set provides better performance in most configurations. As expected, part-of-speech tagging accuracy is increased in all configurations. Overall, parsing performance is improved compared to the baseline results presented in Table 3. It is better when using this mode on FTB DEV and EMEA DEV, while the situation is less bright on SEQUOIA DEV where all dfl configurations provide slightly inferior, or equivalent, performance. (eg. -0.3 with DFL, -0.3 K-DFL-BRIGDE, +0.02 K-DFL-SOURCE).
Regarding raw parsing performance of this mode in the baseline configuration (token=WORD), predicted POS clearly improve parsing performance (+0.5 in EMEA DEV, for instance, and +0.4 for the FTB DEV).
Looking at the impact of using clusters built from bridge or source corpora, it is difficult to draw conclusions as both cluster modes behave differently. In addition, most observed improvements are very small. However, two main tendencies emerge: (i) using a lexical bridge is only slightly efficient with K-MORF tokens, and it impacts negatively the K-DFL configuration ; (ii) almost all differences between the two cluster modes are alleviated. This also means that using predicted POS fills the gap between Lorg's different OOVs handling configurations.
The best configuration, K-MORF-BRIGDE, provides a clear improvement over both

[22]The objective of jackknifing is to obtain a treebank with tags predicted by a tagger that was not trained on the data it tags. In other words, we want the treebank to contain errors as close as possible to what will output a tagger trained on the unmodified treebank. To achieve this, the training data is split into 10 folds, and each fold is tagged with a tagger trained on the other 9 folds. The tagged corpus is then obtained by re-concatenating the 10 tagged folds.

WORD configuration baselines (Table 3 and Table 4) in all tested domains.

4.4 Additional self-training experiments

Because it increases both SOURCE and TARGET parsing performance, the architecture we presented is promising. It is therefore interesting to compare it to self-training methods which have been successfully proven effective on biomedical data (Lease and Charniak, 2005; McClosky and Charniak, 2008).
For this comparison, we will focus on biomedical data as those are the most complicated to parse for our architecture. Results of self-training parsing of the EMEA DEV are presented Table 5. 100,000 unlabeled sentences from the EMEA corpus were parsed in our baseline configuration (Train+gold POS, no predicted POS, product of four grammars, etc.). The resulting trees were then concatenated to the train set and the parser retrained on those data. For the second configuration, we tagged the same unlabeled set using MORFETTE trained on the FTB, and then parsed it, with predicted POS, using the WORD baseline model trained on the POS jackknifed FTB (Section 4.3). From the same configuration, we created a model with bridge clusters built on MORF tokens (K-MORF-BRIGDE) as input.

Results show that all self-training setups outperform our primary baseline on the EMEA DEV (resp. +0.5, +1.26 and +1.35) and slightly less when compared to the predicted pos input baseline (resp. +0.03, +0.76 and 0.84). Note that using self-training without predicted pos tags gives the same gain on the EMEA DEV than using predicted pos tags. Using K-MORF-BRIGDE tokens with self-training slightly alleviates the performance degradation in the SOURCE domain shown by the others self-training configurations. Nevertheless, the best self-training setup fails to outperform its non self-trained counterpart on the TARGET domain by a small margin (-0.3 points) while loosing 2 points on the SOURCE domain.
It is an open question to know whether adding more clustered unlabeled data will actually help TARGET domain parsing in our self-training setup. In fact, one of the goal of self-training is to reduce the lexical gap between domains, but, in our approach the lexicon is already drastically reduced by the clustering process. Furthermore, a significant part of the lexical handling is achieved by means of a purely supervised pos tagger. Building a specifically optimized pos tagger for biomedical data could probably help but this will contradict our goal of producing a model suitable for both SOURCE domain and TARGET domain parsing.

Settings	Baseline	Best-no-	Baseline+100K	Train with (JK)POS+100K	
Self-training	NONE	NONE	+100 K	+100 K	+100 K
POS Mode	NONE	PRED	NONE	PRED	PRED
Cluster Type	N/A	BRIDGE	N/A	N/A	BRIGDE
Token Type	WORD	K-MORF	WORD	WORD	K-MORF
FTB DEV					
Bracketing FMeas.	87.68	88.71	85.78	86.20	**86.64**
Tagging accuracy	96.52	97.06	96.37	97.08	97.06
EMEA DEV					
Bracketing FMeas.	82.59	84.25	83.13	83.86	**83.94**
Tagging accuracy	89.37	92.75	90.55	92.69	92.77

Table 5: Development Set: two first columns are a reminder from table 4 (baseline, and best result without self-training). Last three columns use self-training with +100k sentences from the EMEA corpus, with word form tokens, with or without predicted POS supplied, and with K-MORF-BRIDGE clusters as tokens.

5 Discussion

In the previous section, we demonstrated how the addition of a robust pos tagging step to a PCFG-LA based lexical bridge architecture could lead to a clear performance improvement in both in-domain and out-of-domain parsing. We compared our approach to three self-training architectures, including a self trained variant of our best setting, and noted that the resulting models still suffer from a clear degradation in SOURCE domain parsing while bringing less gain than our main approach on the TARGET domain.

The idea of pure self-training (retraining a parser on its own input, without a reranking step) is appealing, as one would only require a large amount of unlabeled data to provide efficient parsing model on different genre. The problem is that mitigated results have been reported over the years (Charniak, 1997; Steedman et al., 2003b; Plank, 2009), although Huang and Harper (2009); Sagae (2010) recently reported positive feedbacks on respectively SOURCE and TARGET domain parsing. Our own results confirm the positive effect of a self-training strategy noted by Sagae (2010) but we also show that a lexical bridge strategy within a PCFG-LA framework provides a parsing model better adapted for in and out-of-domain parsing.

Of course, this does not mean that mixing both strategies should not be pursued. We believe that the self-training approach is orthogonal to the lexical bridge approach we have presented. Promising solutions to prevent degradation of source parsing performance have been proposed, such as automatic domain adaptation (McClosky et al., 2010) or parsing with product grammars extracted from different target domains, as suggested by Huang and Harper (2009). The latter represents a

very interesting opportunity to mix grammars extracted from different types of text or from different types of tokens.

Our final Test set results are presented in Table 6, and are on par with the trend noticed in the preceding section. Those results are state-of-the-art regarding SOURCE and TARGET domain statistical parsing of French. Our best model on the FTB test set slightly outperforms the PCFG-LA and feature rich reranker approach of (Le Roux et al., 2011) with 89.18 vs 88.74 of F-measures (<40).[23] It is interesting to note that a technique initially developed to cope with the French language's rich morphology by reducing lexical data sparseness can be efficiently applied to robust out-of-domain parsing. This encourages us to continue pursuing this line of work, by tackling new text genres and, perhaps more urgently, by attacking noisy text such as user-generated content, micro-blogging, etc.

Settings	**Baseline**	**Train with JK Pos**		
Cluster Type	N/A	N/A	N/A	BRIGDE
Token Type	WORD	WORD	DFL	K-MORF
Pos supplied	NONE	PRED	PRED	PRED
FTB TEST				
Bracketing FMeasure	87.74	88.15	88.74	**89.18**
Tagging accuracy	96.92	97.26	97.26	97.26
SEQUOIA TEST				
Bracketing FMeasure	86.21	87.01	86.78	**87.13**
Tagging accuracy	95.28	96.13	96.13	96.13
EMEA TEST				
Bracketing FMeasure	83.27	84.16	84.45	**84.7**
Tagging accuracy	95.21	96.38	96.39	96.39

Table 6: Test Set: Final Results

6 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 and three different textual genres, regrouped into the SEQUOIA CORPUS, for which we constructed evaluation resources and provided state-of-the art results. The most important take away mes-

[23] Leroux et al.'s model is based on one grammar, it is very likely that a 4 grammars model will work even better with a reranker.

sage is the fact that the use of word clusters built on morphological token units leads to a robust and efficient parser suitable for target domain parsing while also improving source domain parsing, which stands in contrast to other techniques such as self-training.

Our perspectives for future work are to investigate: (i) using specialized lexicon-informed part-of-speech taggers; and (ii) supplementing our approach with other techniques such as reranking, which is known to improve self-training for domain adaptation (McClosky and Charniak, 2008), or uptraining (Petrov et al., 2010).

Acknowledgements

We are grateful to the anonymous reviewers for their valuable comments. 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

References

Abeillé, Anne. 2004. Annotation fonctionnelle, version du 1er mars 2004. `http://www.llf.cnrs.fr/Gens/Abeille`.

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

Abeillé, Anne and Lionel Clément. 2006. Annotation morpho-syntaxique, version du 10 nov. 2006. `http://www.llf.cnrs.fr/Gens/Abeille`.

Abeillé, Anne, François Toussenel, and M. Chéradame. 2004. Corpus le monde, annotation en constituants, guide pour les correcteurs, version du 31 mars 2004. `http://www.llf.cnrs.fr/Gens/Abeille`.

Attia, Mohammed, 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.

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

Black, Ezra, Steve Abney, Dan Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, Fred Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the 1991 DARPA Speech and Natural Language Workshop*. pages 306–311.

Brown, Peter F., 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.

Candito, M., B. Crabbé, and D. Seddah. ???? On statistical parsing of french with supervised and semi-supervised strategies. In *EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference*. page 49.

Candito, Marie 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)*. Association for Computational Linguistics, Paris, France, pages 138–141.

Candito, Marie, Benoit Crabbé, and Pascal Denis. 2010a. Statistical french dependency parsing : Treebank conversion and first results. In *Proceedings of LREC'2010*. Valletta, Malta.

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

Candito, Marie 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.

Charniak, E. 1997. Statistical techniques for natural language parsing. *AI magazine* 18(4):33.

Charniak, Eugene. 2000. A maximum entropy inspired parser. In *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*. Seattle, WA, pages 132–139.

Chrupała, Grzegorz, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with morfette. In *In Proceedings of LREC 2008*. ELDA/ELRA, Marrakech, Morocco.

Crabbé, Benoit and Marie Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *Proc. of TALN'08*. Senlis, France, pages 45–54.

Denis, Pascal and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proc. of PACLIC*. Hong Kong, China.

Foster, J., 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.

Foster, Jennifer. 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*. Association for Computational Linguistics, Los Angeles, California, pages 381–384.

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

Huang, Z. and M. Harper. 2009. Self-training pcfg grammars with latent annotations across languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, pages 832–841.

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

Le Roux, J., B. Favre, S. A. Mirroshandel, and A Nasr. 2011. Modèles génératif et discriminant en analyse syntaxique : expériences sur le corpus arboré de paris 7. In *In Proceedings of TALN 2011*. Montréal, Canada.

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

Levy, Roger and Galen Andrew. 2006. Tregex and tsurgeon: Tools for querying and manipulating tree data structures. In *Proc. of LREC'06*. Geneva, Italy.

Liang, Percy. 2005. Semi-supervised learning for natural language. In *MIT Master's thesis*. Cambridge, USA.

McClosky, D., 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*. Association for Computational Linguistics, pages 337–344.

McClosky, D., E. Charniak, and M. Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 28–36.

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

Morton, T. 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*. Association for Computational Linguistics, pages 17–18.

Petrov, S., 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*. Association for Computational Linguistics, pages 705–713.

Petrov, Slav. 2010. Products of random latent variable grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, pages 19–27.

Petrov, Slav 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*. Association for Computational Linguistics, Rochester, New York, pages 404–411.

Plank, B. 2009. A comparison of structural correspondence learning and self-training for discriminative parse selection. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing*. pages 37–42.

Sagae, Kenji. 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*. Association for Computational Linguistics, Uppsala, Sweden, pages 37–44.

Sagot, Benoît, 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* .

Seddah, Djamé, Grzegorz Chrupała, Ozlem Cetinoglu, Josef van Genabith, and Marie Candito. 2010. Lemmatization and statistical lexicalized parsing of morphologically-rich languages. In *Proceedings of the NAACL/HLT Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*. Los Angeles, CA.

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

Steedman, M., R. Hwa, S. Clark, M. Osborne, A. Sarkar, J. Hockenmaier, P. Ruhlen, S. Baker, and J. Crim. 2003a. 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*. Association for Computational Linguistics, pages 157–164.

Steedman, M., M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaier, P. Ruhlen, S. Baker, and J. Crim. 2003b. Bootstrapping statistical parsers from small datasets. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 331–338.

Tiedemann, J. 2009. News from opus - a collection of multilingual parallel corpora with tools and interfaces. *Recent advances in natural language processing V: selected papers from RANLP 2007* 309:237.

Villemonte De La Clergerie, Eric, Olivier Hamon, Djamel Mostefa, Christelle Ayache, Patrick Paroubek, and Anne Vilnat. 2008. Passage : from french parser evaluation to large sized treebank. In *Proceedings of the 6th International Conference on Languages Resources and Evaluation (LREC'2008)*.