Unsupervised Linguistically-Driven Reliable Dependency Parses Detection and Self-Training for Adaptation to the Biomedical Domain

Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
Via G. Moruzzi, 1 – Pisa (Italy)
{felice.dellorletta,giulia.venturi,simonetta.montemagni}@ilc.cnr.it

Abstract

In this paper, a new self–training method for domain adaptation is illustrated, where the selection of reliable parses is carried out by an unsupervised linguistically–driven algorithm, ULISSE. The method has been tested on biomedical texts with results showing a significant improvement with respect to considered baselines, which demonstrates its ability to capture both reliability of parses and domain–specificity of linguistic constructions.

1 Introduction

As firstly demonstrated by (Gildea, 2001), parsing systems have a drop of accuracy when tested against domain corpora outside of the data from which they were trained. This is a real problem in the biomedical domain where, due to the rapidly expanding body of biomedical literature, the need for increasingly sophisticated and efficient biomedical text mining systems is becoming more and more pressing. In particular, the existence of natural language parsers reliably dealing with biomedical texts represents the prerequiste for identifying and extracting knowledge embedded in them. Over the last years, this problem has been tackled within the biomedical NLP community from different perspectives. The development of a domain–specific annotated corpus, i.e. the Genia Treebank (Tateisi, Yakushiji, Ohta, & Tsujii, 2005), played a key role by providing a sound basis for empirical performance evaluation as well as training of parsers. On the other hand, several attempts have been made to adapt general parsers to the biomedical domain. First experiments in this direction are reported in (Clegg & Shepherd, 2005) who first compared the performance of three different parsers against the Genia treebank and a sample of the Penn Treebank (PTB) (Mitchell P. Marcus & Santorini, 1993) in order to carry out an inter–domain analysis of the typology of errors made by each parser and demonstrated that by integrating the output of the three parsers they achieved statistically significant performance gains. Three different methods of parser adaptation for the biomedical domain have been proposed by (Lease & Charniak, 2005) who, starting from the results of unknown word rate experiments carried out on the Genia treebank, adapted a PTB–trained parser by improving the Part–Of–Speech tagging accuracy and by relying on an external domain–specific lexicon. More recently, (McClosky, Charniak, & Johnson, 2010) and (Plank & van Noord, 2011) devised adaptation methods based on domain similarity measures. In particular, both of them adopted lexical similarity measures to automatically select from an annotated collection of texts those training data which is more relevant, i.e. lexically closer, to adapt the parser to the target domain.

A variety of semi–supervised approaches, where unlabeled data is used in addition to labeled training data, have been recently proposed in the literature in order to adapt parsing systems to new domains. Among these approaches, the last few years have seen a growing interest in self–training for domain adaptation, i.e. a method for using automatically annotated data from a target domain when training supervised models. Self–training methods proposed so far mainly differ at the level of the selection of parse trees to be added to the in–domain gold trees as further training data. Depending on whether or not external supervised classifiers are used to select the parses to be added to the gold–training set, two types of methods are envisaged in the literature. The first is the case, among others, of: (Kawahara & Uchimoto, 2008), using a machine learning classifier to predict the reliability of parses on the basis of different feature types; or (Sagae & Tsujii, 2007), selecting

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 45–53,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

identical analyses for the same sentence within the output of different parsing models trained on the same dataset; or (McClosky, Charniak, & Johnson, 2006), using a discriminative reranker against the output of a n–best generative parser for selecting the best parse for each sentence to be used as further training data. Yet, due to the fact that several supervised classifiers are resorted to for improving the base supervised parser, this class of methods cannot be seen as a genuine istance of self–training. The second type of methods is exemplified, among others, by (Reichart & Rappoport, 2007) who use the whole set of automatically analyzed sentences, and by (McClosky & Charniak, 2008) and (Sagae, 2010) who add different amounts of automatically parsed data without any selection strategy. Note that (McClosky & Charniak, 2008) tested their self–training approach on the Genia Treebank: they self–trained a PTB–trained costituency parser using a random selection of Medline abstracts.

In this paper, we address the second scenario with a main novelty: we use an unsupervised approach to select reliable parses from automatically parsed target domain texts to be combined with the gold–training set. Two unsupervised algorithms have been proposed so far in the literature for selecting reliable parses, namely: PUPA (*POS–based Unsupervised Parse Assessment Algorithm*) (Reichart & Rappoport, 2009) and ULISSE (*Unsupervised LInguiSTically–driven Selection of dEpendency parses*) (Dell'Orletta, Venturi, & Montemagni, 2011). Both algorithms assign a quality score to each parse tree based on statistics collected from a large automatically parsed corpus, with a main difference: whereas PUPA operates on costituency trees and uses statistics about sequences of part–of–speech tags, ULISSE uses statistics about linguistic features checked against dependency–based representations. The self–training strategy presented in this paper is based on an augmented version of ULISSE. The reasons for this choice are twofold: if on the one hand ULISSE appears to outperform PUPA (namely, a dependency–based version of PUPA implemented in (Dell'Orletta et al., 2011)), on the other hand the linguistically–driven nature of ULISSE makes our self–training strategy for domain adaptation able to capture reliable parses which are also representative of the syntactic peculiarities of the target domain.

After introducing the in– and out–domain corpora used in this study (Section 2), we discuss the results of the multi–level linguistic analysis of these corpora carried out (Section 3) with a view to identifying the main features differentiating the biomedical language from ordinary language. In Section 4, the algorithm used to select reliable parses from automatically parsed domain–specific texts is described. In Section 5 the proposed self–training method is illustrated, followed by a discussion of achieved results (Section 6).

2 Corpora

Used domain corpora include *i)* the two out–domain datasets used for the "Domain Adaptation Track" of the CoNLL 2007 Shared Task (Nivre et al., 2007) and *ii)* the dependency–based version of the Genia Treebank (Tateisi et al., 2005). The CoNLL 2007 datasets are represented by chemical (CHEM) and biomedical abstracts (BIO), made of 5,001 tokens (195 sentences) and of 5,017 tokens (200 sentences) respectively. The dependency–based version of Genia includes ∼493k tokens and ∼18k sentences which was generated by converting the PTB version of Genia created by Illes Solt[1] using the (Johansson & Nugues, 2007) tool with the *-conll2007* option to produce annotations in line with the CoNLL 2007 data set[2]. As unlabelled data, we used the datasets distributed in the framework of the CoNLL 2007 Domain Adaptation Track. For CHEM the set of unlabelled data consists of 10,482,247 tokens (396,128 sentences) and for BIO of 9,776,890 tokens (375,421 sentences). For the experiments using Genia as test set, we used the BIO unlabelled data. This was possible due to the fact that both the Genia Treebank and the BIO dataset consist of biomedical abstracts extracted (though using different query terms) from PubMed.com.

As in–domain training data we have used the CoNLL 2007 dependency–based version of Sections 2–11 of the Wall Street Journal (WSJ) partition of the Penn Treebank (PTB), for a total of 447,000 tokens and about 18,600 sentences. For testing, we used the subset of Section 23 of WSJ consisting of 5,003 tokens (214 sentences).

All corpora have been morpho–syntactically tagged and lemmatized by a customized version

[1] http://categorizer.tmit.bme.hu/∼illes/genia_ptb/

[2] In order to be fully compliant with the PTB PoS tagset, we changed the PoS label of all punctuation marks.

of the pos–tagger described in (Dell'Orletta, n.d.) and dependency parsed by the DeSR parser using Multi–Layer Perceptron (MLP) as learning algorithm (Attardi, Dell'Orletta, Simi, & Turian, n.d.), a state–of–the–art linear–time Shift–Reduce dependency parser following a "stepwise" approach (Buchholz & Marsi, 2006).

3 Linguistic analysis of biomedical abstrats vs newspaper articles

For the specific concerns of this study, we carried out a comparative linguistic analysis of four different corpora, taken as representative of ordinary language and biomedical language. In each case, we took into account a *gold* (i.e. manually annotated) corpus, and an *unlabelled* corpus, which was automatically annotated. By comparing the results obtained with respect to gold and automatically annotated texts, we intend to demonstrate the reliability of features extracted from automatically annotated texts. As data representative of ordinary language we took *i*) the whole WSJ section of the Penn Treebank[3] including 39,285,425 tokens (1,625,606 sentences) and *ii*) the sections 2–11 of the WSJ. For what concerns the biomedical domain, we relied on the Genia Treebank in order to guarantee comparability of the results of our linguistic analysis with previous studies carried out on this reference corpus. As automatically annotated data we used the corpus of biomedical abstract (BIO) distributed as out–domain dataset used for the "Domain Adaptation Track" of the CoNLL 2007 Shared Task.

In order to get evidence of the differences holding between the WSJ newspaper articles and the selected biomedical abstracts, the four corpora have been compared with respect to a wide typology of features (i.e. raw text, lexical, morpho–syntactic and syntactic). Let us start from raw text features, in particular from average sentence length (calculated as the average number of words per sentence): as Figure 1 shows, both the corpus of automatically parsed newspaper articles (*WSJ_unlab*) and the manually annotated one (*WSJ_gold*) contain shorter sentences with respect to both the automatically parsed biomedical abstracts (*BIO_unlab*) and the manually annotated ones (*Genia_gold*), a result which is in line with (Clegg & Shepherd, 2005) findings. When we focus on the lexical level, *BIO_unlab* and *Genia_gold* appear to have quite a similar percentage of lexical items which is not contained in *WSJ_gold* (23.13% and 26.14% respectively) while the out–of–vocabulary rate of *WSJ_unlab* is much lower, i.e. 8.69%. Similar results were recorded by (Lease & Charniak, 2005) who report the unknown word rate for various genres of tecnical literature.

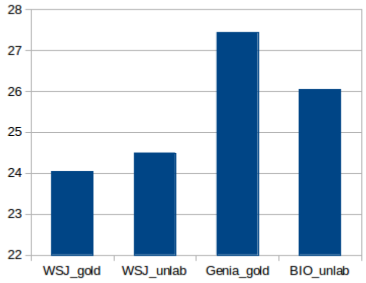
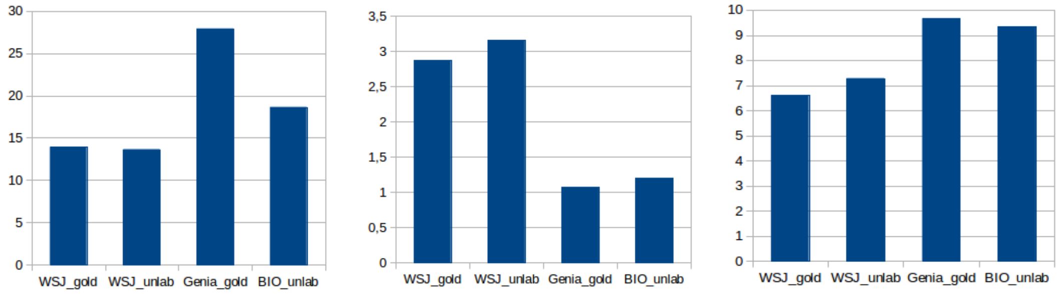


Figure 1: Average sentence length in biomedical and newspaper corpora.

Let us focus now on the morpho–syntactic level. If we consider the distribution of nouns, verbs and adjectives, three features typically representing stylistic markers associated with different linguistic varieties (Biber & Conrad, 2009), it can be noticed (see Figures 2(a) and 2(c)) that the biomedical abstracts contain a higher percentage of nouns and adjectives while showing a significantly lower percentage of verbs (Figure 2(b)). The syntactic counterpart of the different distribution of morpho–syntactic categories can be observed in Table 1, reporting the percentage distribution of the first ten Parts–of–Speech dependency triplets occurring in the biomedical and newspaper corpora: each triplet is described as the sequence of the PoS of a dependent and a head linked by a depedency arc, by also considering the PoS of the head father. It turned out that biomedical abstracts are characterized by *nominal* dependency triplets, e.g. two nouns and a preposition (NN–NN–IN) or noun, preposition, noun (NN–IN–NN) or adjective, noun and preposition (JJ–NN–IN), which occur more frequently than *verbal* triplets, such as determiner, noun and verb (DT–NN–VBZ) or a verbal root (.–VBD–ROOT)[4]. Interestingly, in *Genia_gold* no verbal triplet occurs within the top ten triplets, which cover the 21% of the total amount

[3] This corpus represents to the unlabelled data set distributed for the CoNLL 2007 Shared Task on Dependency Parsing, domain adaptation track.

[4] We named 'ROOT' the artificial root node.



(a) Distribution of Nouns

(b) Distribution of Verbs

(c) Distribution of Adjectives

Figure 2: Distribution of some Parts–of–Speech in biomedical and newspaper corpora.

of triplets occurring in the corpus. By contrast, the same top ten triplets represent only ∼11% in *WSJ_gold*, testifying the wider variety of syntactic constructions occurring in newspaper articles with respect to texts of the biomedical domain. This is also proved by the total amount of different PoS dependency triplets occurring in the two gold datasets, i.e. 7,827 in *WSJ_gold* and 5,064 in *Genia_gold* even though the Genia Treebank is ∼50,000–tokens bigger.

Further differences can be observed at a deeper syntactic level of analysis. This is the case of the average depth of embedded complement 'chains' governed by a nominal head. Figure 3(a) shows that biomedical abstracts are characterized by an average depth which is higher than the one observed in newspaper articles. A similar trend can be observed for what concerns the distribution of 'chains' by depth. In Figure 3(b) shows that *WSJ_unlab* and *WSJ_gold* 'chains', on the one hand, and *BIO_unlab* and *Genia_gold* 'chains', on the other hand, overlap. The corpora have also been compared with respect to *i)* the average length of dependency links, measured in terms of the words occurring between the syntactic head and the dependent (excluding punctuation marks), and *ii)* the average depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to a leaf. In both cases it can be noted that *i)* the biomedical abstracts contain much longer dependency links than newswire texts (Figure 3(c)) and *ii)* the average depth of *BIO_unlab* and *Genia_gold* parse trees is higher than in the case of the *WSJ_unlab* and *WSJ_gold* (Figure 3(d)). A further distinguishing feature of the biomedical abstracts concerns the average depth of 'chains' of embedded subordinate clauses, calculated here by taking into account both clausal arguments and complements linked to a verbal sentence root. As Figure 3(e) shows, both *BIO_unlab* and *Genia_gold* have shorter 'chains' with respect to the ones contained in the newspaper articles. Interestingly, a careful analysis of the distributions by depth of 'chains' of embedded subordinate clauses shows that the biomedical abstracts appear to have *i)* a higher occurrence of 'chains' including just one subordinate clause and *ii)* a lower percentage of deep 'chains' with respect to newswire texts. Finally, we compared the two types of corpora with respect to the distribution of verbal roots. The biomedical abstracts resulted to be characterised by a lower percentage of verbal roots with respect to newspaper articles (see Figure 3(f)). This is in line with the distribution of verbs as well as of the *nominal* dependency triplets observed in the biomedical abstracts at the morpho–syntactic level of analysis.

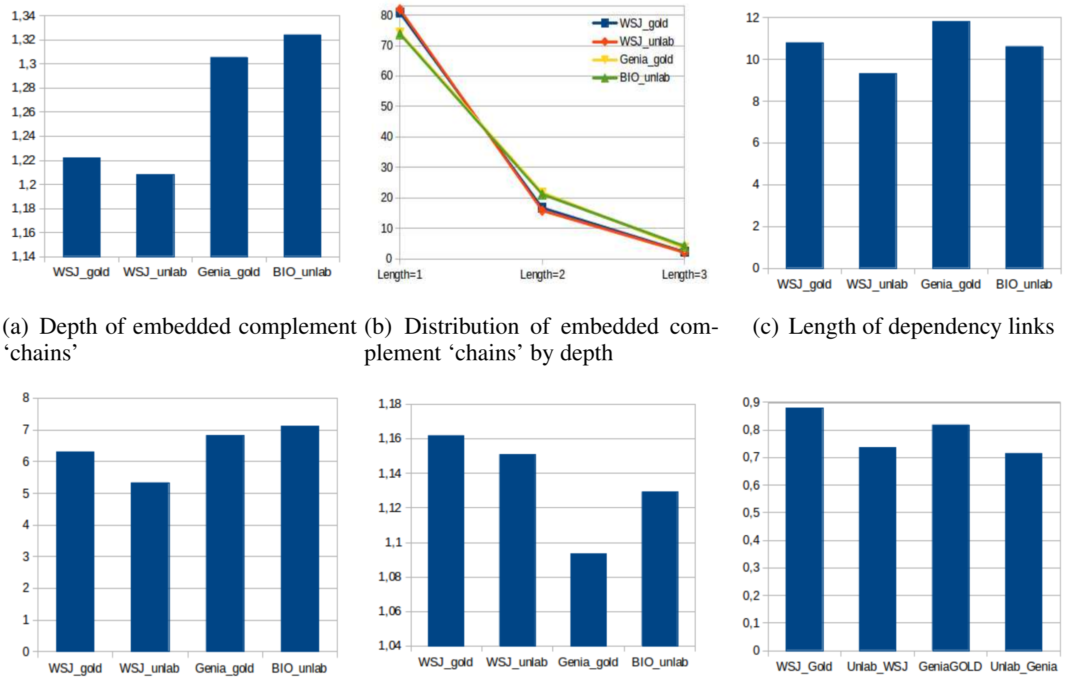
Interestingly, the results obtained with respect to automatically parsed and manually annotated data show similar trends for both considered in– and out–domain corpora, thus demonstrating the reliability of features monitored against automatically annotated data. In what follows, we will show how detected linguistic peculiarities can be exploited in a domain adaptation scenario.

4 Linguistically–driven Unsupervised Ranking of Parses for Self–training

In the self–training approach illustrated in this paper, the selection of parses from the automatically annotated target domain dataset is guided by an augmented version of ULISSE, an unsupervised linguistically–driven algorithm to select reliable parses from the output of dependency annotated texts (Dell'Orletta et al., 2011) which has shown a good performance for two different languages

WSJ_gold		WSJ_unlab		Genia_gold		BIO_unlab	
Triplet	% Freq	Triplet	% Freq	Triplet	% Freq	Triplet	% Freq
DT-NN-IN	2.03	DT-NN-IN	1.72	NN-NN-IN	3.66	DT-NN-IN	2.87
.-VBD-ROOT	1.61	.-VBD-ROOT	1.30	NN-IN-NN	2.93	NN-IN-NN	2.39
NN-IN-NN	1.11	JJ-NN-IN	0.99	DT-NN-IN	2.48	JJ-NN-IN	2.08
JJ-NN-IN	1.10	NN-IN-NN	0.97	JJ-NN-IN	1.96	NN-NN-IN	1.73
.-VBZ-ROOT	1.09	NNP-NNP-IN	0.87	NN-NNS-IN	1.88	IN-NN-IN	1.72
NNP-NNP-IN	0.95	DT-NN-VBD	0.85	JJ-NNS-IN	1.77	JJ-NNS-IN	1.36
DT-NN-VBZ	0.89	NN-VBD-ROOT	0.80	IN-NN-IN	1.65	.-VBD-ROOT	1.33
DT-NN-VBD	0.87	JJ-NNS-IN	0.79	NN-CC-IN	1.64	NNS-IN-NN	1.13
JJ-NNS-IN	0.87	NNP-NNP-VBD	0.78	NNS-IN-NN	1.56	NNP-NN-IN	1.03
IN-NN-IN	0.87	.-VBZ-ROOT	0.75	NN-NN-CC	1.47	NN-IN-VBN	0.93

Table 1: Frequency distribution of the first ten Parts–of–Speech dependency triplets in biomedical and newspaper corpora.



(a) Depth of embedded complement 'chains'

(b) Distribution of embedded complement 'chains' by depth

(c) Length of dependency links

(d) Parse tree depth

(e) Depth of embedded subordinate clauses 'chains'

(f) Distribution of verbal roots

Figure 3: Syntactic features in biomedical and newspaper corpora.

(English and Italian) against the output of two supervised parsers (MST, (McDonald, Lerman, & Pereira, 2006) and DeSR, (Attardi, 2006)) selected for their behavioral differences (McDonald & Nivre, 2007). ULISSE assigns to each dependency tree a score quantifying its reliability based on a wide range of linguistic features. After collecting statistics about selected features from a corpus of automatically parsed sentences, for each newly parsed sentence ULISSE computes a reliability score using the previously extracted feature statistics. In its reliability assessment, ULISSE exploits both global and local features, where global features (listed in Table 2 and discussed in Section 3) are computed with respect to each sentence and averaged over all sentences in the corpus, and the local features with respect to individual dependency links and averaged over all of them. Local features include the plausibility of a dependency link calculated by considering selected features of the dependent and its governing head as well as of the head father: whereas in ULISSE the selected features were circumscribed to part–of–speech information (so–called "ArcPOSFeat" feature), in this version of the algorithm a new local feature has been introduced, named "ArcLemmaFeat", which exploits lemma information. "ArcPOSFeat" is able to capture the different distribution of PoS dependency triplets (see Table 1), along with the type of dependency link, while the newly introduced "ArcLemmaFeat" is meant to capture the lexical peculiarities of the target domain (see Section 3). As demonstrated in (Dell'Orletta et al., 2011), both global and local linguistic features contribute to the selection of reliable parses. Due to the typology of linguistic features underlying ULISSE, selected reliable parses typically include domain–specific constructions. This peculiarity of the ULISSE algorithm turned out to be particularly useful to maximize the self–training effect in improving the parsing performance in a domain adaptation scenario.

The reliability score assigned by this augmented version of ULISSE to newly parsed sentences results from a combination of the weights associated with individual features, both global and local ones. In this study, the reliability score was computed as a simple product of the individual feature weights: in this way, one low weight feature is sufficient to qualify a parse as low quality and thus to exclude it from the self–training dataset[5].

Feature
Parse tree depth
Embedded complement 'chains' headed by a noun
- Average depth
- Distribution by depth
Verbal roots
Arity of verbal predicates
- Distribution by arity
Subordinate vs main clauses
- Relative ordering of subordinate clauses with respect to the main clause
- Average depth of 'chains' of embedded subordinate clauses
- Distribution of embedded subordinate clauses 'chains' by depth
Length of dependency links

Table 2: Global features underlying ULISSE.

5 Experimental set–up

In the reported experiments, we used the DeSR parser. Its performance using the proposed domain adaptation strategy was tested against *i)* the two out–domain datasets distributed for the "Domain Adaptation Track" of the CoNLL 2007 Shared Task and *ii)* the dependency–based version of the Genia Treebank, described in Section 2. For testing purposes, we selected from the dependency–based version of the Genia Treebank sentences with a maximum length of 39 tokens (for a total of 375,912 tokens and 15,623 sentences).

Results achieved with respect to the CHEM and BIO test sets were evaluated in terms of "Labelled Attachment Score" (LAS), whereas for Genia the only possible evaluation was in terms of "Unlabelled Attachment Score" (UAS). This follows from the fact that, as reported by Illes, this version of Genia is annotated with a Penn Treebank–style phrase–structure, where a number of functional tags are missing: this influences the type of evaluation which can be carried out against the Genia test set.

Test corpus	LAS	UAS
PTB	86.09%	87.29%
CHEM	78.50%	81.10%
BIO	78.65%	79.97%
GENIA	n/a	80.25%

Table 3: The *BASE* model tested on PTB, CHEM, BIO and GENIA.

Achieved results were compared with two baselines, represented by: i) the *Baseline model* (*BASE*), i.e. the parsing model trained on the PTB training set only; ii) the *Random Selection* (*RS*) of parses from automatically parsed out–domain corpora, calculated as the mean of a *10*–fold cross–validation process. As proved by (Sagae, 2010) and by (McClosky & Charniak, 2008) for the biomedical domain, the latter represents a strong unsupervised baseline showing a significant accuracy improvement which was obtained by adding incremental amounts of automatically parsed out–domain data to the training dataset without any selection strategy.

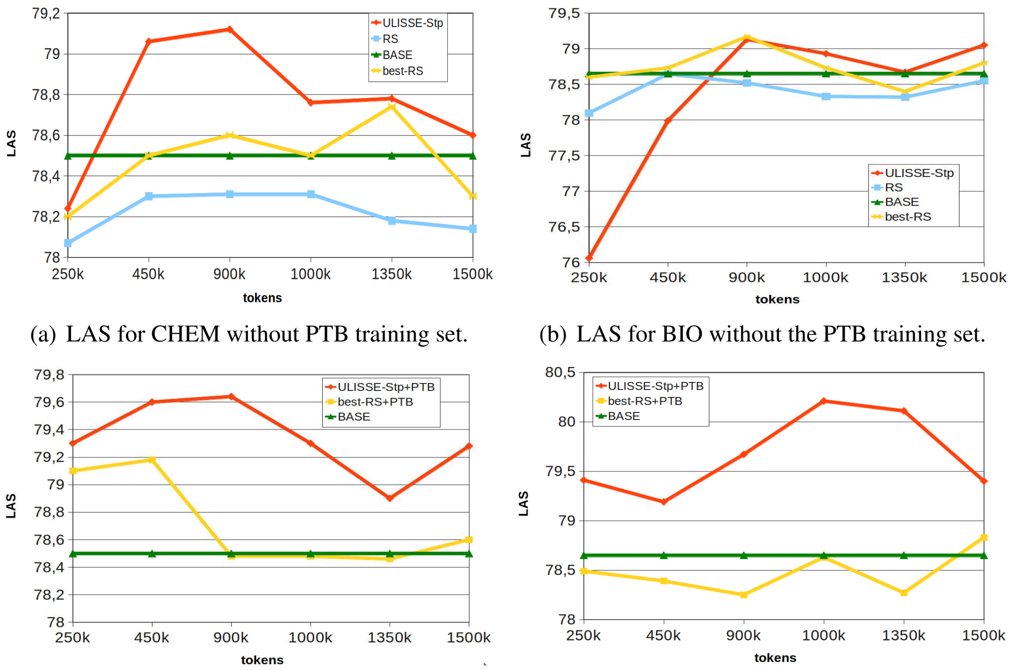
The experiments we carried out to test the effectiveness of our self–training strategy were organised as follows. ULISSE and the baseline algorithms were used to produce different rankings of parses of the unlabelled target domain corpora. From the top of these rankings different pools of parses were selected to be used for training. In particular, two different sets of experiments were carried out, namely: i) using only automatically parsed data as training corpus and ii) combining automatically parsed data with the PTB training set. For each set of experiments, different amounts of unlabelled data were used to create the self–training models.

6 Results

Table 3 reports the results of the *BASE* model tested on PTB, CHEM, BIO and GENIA. When applied without adaptation to the out–domain CHEM, BIO and GENIA test sets, the *BASE* parsing model has a drop of about 7.5% of LAS in both CHEM and BIO cases. For what concerns UAS, the drop is about 6% for CHEM and about 7% for BIO and GENIA.

The results of the performed experiments are shown in Figures 4 and 5, where each plot reports the accuracy scores (LAS and UAS respectively) of the self–trained parser using the ULISSE

[5] See (Dell'Orletta et al., 2011) for a detailed description of the quality score computation.



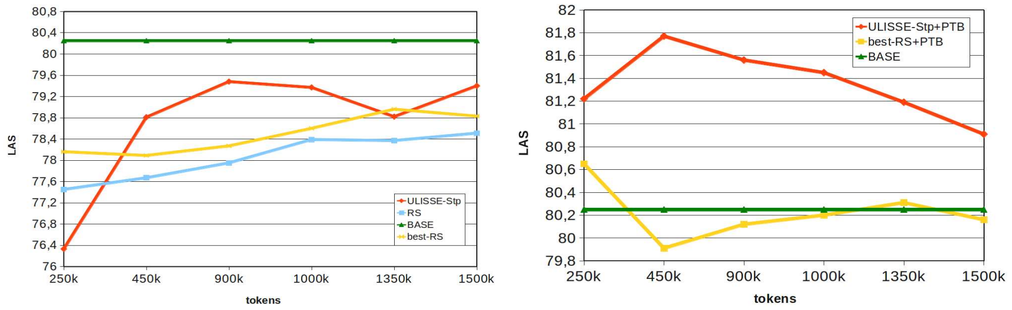
(a) LAS for CHEM without PTB training set.

(b) LAS for BIO without the PTB training set.

(c) LAS for CHEM with the PTB training set.

(d) LAS for BIO with the PTB training set.

Figure 4: LAS of the different self–training models in the two sets of performed experiments.



(a) UAS for GENIA without the PTB training set.

(b) UAS for GENIA with the PTB training set.

Figure 5: UAS of the different self–training models for GENIA.

algorithm (henceforth, ULISSE–Stp) and of the baseline models (*RS* and *BASE*). The parser accuracy was computed with respect to different amounts of automatically parsed data which were used to create the self–trained parsing model. For this purpose, we considered six different pools of 250k, 450k, 900k, 1000k, 1350k and 1500k tokens. Plots are organized by experiment type: i.e. the results in subfigures 4(a), 4(b) and 5(a) are achieved by using only automatically parsed data as training corpus, whereas those reported in the other subfigures refer to models trained on automatically parsed data combined with PTB. Note that in all figures the line named *best–RS* represents the best *RS* score for each pool of *k* tokens in the *10*–fold cross–validation process.

For what concerns BIO and CHEM, in the first set of experiments ULISSE–Stp turned out to be the best self–training algorithm: this is always the case for CHEM (see subfigure 4(a)), whereas for BIO (see subfigure 4(b)) it outperforms all baselines only when pools of tokens $>= 900k$ are added. When 900k tokens are used, ULISSE–Stp allows a LAS improvement of 0.81% on CHEM and of 0.61% on BIO with respect to *RS*, and of 0.62% on CHEM and of 0.48% on BIO with respect to *BASE*. It is interesting to note that ULISSE–Stp using only automatically parsed data for training achieves better results than *BASE* (using the PTB gold training set): to our knowledge, a similar result has never been reported in the literature. The behaviour is similar also when the experiments are evaluated in terms of UAS[6].

The results achieved in the first set of experiments carried out on the GENIA test set (see 5(a)) differ significantly from what we observed for CHEM and BIO: in this case, the *BASE* model appears to outperform all the other algorithms, with the ULISSE–Stp algorithm being however more effective than the *RS* baselines.

Figures 4(c), 4(d) and 5(b) report the results of the second set of experiments, i.e. those carried out by also including PTB in the training set. Note that in these plots the *RS+PTB* lines represent the score of the parser models trained on the pools of tokens used to obtain the *best–RS* line combined with the PTB gold training set. It can be observed that the ULISSE–Stp+PTB self–training model outperforms all baselines for CHEM, BIO and GENIA for all the different sizes of pools of selected tokens. For each pool of parsed data, Table 4 records the improvement and the error reduction observed with respect to the *BASE* model.

Pool of tokens	CHEM	Err. red.	BIO	Err. red.	GENIA	Err. red.
250k	0.8	3.72	0.76	3.55	0.97	4.91
450k	1.1	5.12	0.54	2.53	1.52	7.7
900k	1.14	5.3	1.02	4.77	1.31	6.63
1000k	0.8	3.72	1.56	7.29	1.2	6.08
1350k	0.4	1.49	1.46	6.82	0.94	4.76
1500k	0.78	3.62	0.75	3.37	0.66	3.34

Table 4: *%* improvement of ULISSE–Stp+PTB vs *BASE* reported in terms of LAS for CHEM and BIO and of UAS for GENIA.

Differently from (Sagae, 2010) (with a constituency–based parser), in this set of experiments the self–training approach based on random selection of sentences (i.e. the *best–RS+PTB* baseline) doesn't achieve any improvement with respect to the *BASE* model with only minor exceptions (observed e.g. with 250k and 450k pools of added tokens for CHEM and with 250k for GENIA). Moreover, even when the *best–RS* LAS is higher than the ULISSE–Stp score (e.g. in the first pools of *k* of Figure 4(b)), ULISSE–Stp+PTB turns out to be more effective than the *best–RS+PTB* baseline (Figure 4(d)). These results may follow from the fact that ULISSE–Stp is able to capture not only reliable parses but also, and more significantly here, parses which reflect the syntactic peculiarities of the target domain.

Table 5 shows the results of the different *ULISSE–Stp+PTB* models tested on the PTB test set: no LAS improvement is observed with respect to the results obtained with the *BASE* model, i.e. 86.09% (see Table 3). This result is in line with (McClosky et al., 2010) and (Plank & van Noord, 2011) who proved that parsers trained on the union of gold corpora belonging to different domains achieve a lower accuracy with respect to the same parsers trained on data belonging to a single target domain. Last but not least, it should be noted that the performance of ULISSE–Stp across the experiments carried out with pools of automatically parsed tokens of different sizes is in line with the behaviour of the ULISSE ranking algorithm (Dell'Orletta et al., 2011), where increasingly wider top lists of parsed tokens show decreasing LAS scores. This helps explaining why the performance of ULISSE–Stp starts decreasing after a certain threshold when wider top–lists of tokens are added to the parser training data.

Pool of tokens	CHEM	BIO
250k	83.53	85.55
450k	85.53	86.01
900k	85.95	84.79
1000k	86.03	85.45
1350k	85.49	85.71
1500k	85.67	86.39

Table 5: ULISSE–Stp+PTB on PTB test set with automatically parsed data.

Conclusion

In this paper we explored a new self–training method for domain adaptation where the selection of reliable parses within automatically annotated texts is carried out by an unsupervised linguistically–driven algorithm, ULISSE. Results achieved for the CoNLL 2007 datasets as well as for the larger test set represented by GENIA show a significant improvement with respect to considered baselines. This demonstrates a two–fold property of ULISSE, namely its reliablity and effectiveness both in capturing peculiarities of biomedical texts, and in selecting high quality parses. Thanks to these properties the proposed self–training method is able to improve the parser performances when tested in an out–domain scenario. The same approach could in principle be applied to deal with biomedical sub–domain variation: as reported by (Lippincott, Séaghdha, & Korhonen, 2011), biomedical texts belonging to different sub–domains do vary along many linguistic dimensions, with a potential negative impact on biomedical NLP tools.

[6] In this paper, for CHEM and BIO experiments we report only the LAS scores since this is the standard evaluation metric for dependency parsing.

References

Attardi, G. (2006). Experiments with a multi-language non-projective dependency parser. In *Proceedings of CoNLL-X '06* (pp. 166–170). New York City, New York.

Attardi, G., Dell'Orletta, F., Simi, M., & Turian, J. (n.d.). Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009.*

Biber, D., & Conrad, S. (2009). *Genre, register, style*. Cambridge: Cambridge University Press.

Buchholz, S., & Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL 2006.*

Clegg, A. B., & Shepherd, A. J. (2005). Evaluating and integrating treebank parsers on a biomedical corpus. In *In proceedings of the ACL 2005 workshop on software.*

Dell'Orletta, F. (n.d.). Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009.*

Dell'Orletta, F., Venturi, G., & Montemagni, S. (2011). Ulisse: an unsupervised algorithm for detecting reliable dependency parses. In *Proceedings of CoNLL 2011* (pp. 115–124). Portland, Oregon.

Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of EMNLP 2001* (p. 167-202). Pittsburgh, PA.

Johansson, R., & Nugues, P. (2007). Extended constituent-to-dependency conversion for english. In *Proceedings of NODALIDA 2007.* Tartu, Estonia.

Kawahara, D., & Uchimoto, K. (2008). Learning reliability of parses for domain adaptation of dependency parsing. In *Proceedings of IJCNLP 2008* (pp. 709–714).

Lease, M., & Charniak, E. (2005). Parsing biomedical literature. In *Proceedings of the second international joint conference on natural language processing (IJCNLP-05), Jeju Island, Korea* (pp. 58–69).

Lippincott, T., Séaghdha, D. Ó., & Korhonen, A. (2011). Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, *12*, 212–233.

McClosky, D., & Charniak, E. (2008). Self-training for biomedical parsing. In *Proceedings of ACL–HLT 2008* (pp. 101–104).

McClosky, D., Charniak, E., & Johnson, M. (2006). Reranking and self-training for parser adaptation. In *Proceedings of ACL 2006* (pp. 337–344). Sydney, Australia.

McClosky, D., Charniak, E., & Johnson, M. (2010). Automatic domain adaptation for parsing. In *Proceedings of NAACL–HLT 2010* (pp. 28–36). Los Angeles, California.

McDonald, R., Lerman, K., & Pereira, F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of CoNLL 2006.*

McDonald, R., & Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of EMNLP-CoNLL 2007* (p. 122-131).

Mitchell P. Marcus, M. A. M., & Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, *19*(2), 313–330.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007). The conll 2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL 2007* (pp. 915–932).

Plank, B., & van Noord, G. (2011). Effective measures of domain similarity for parsing. In *Proceedings of ACL 2011* (pp. 1566–1576). Portland, Oregon.

Reichart, R., & Rappoport, A. (2007). Self–training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of ACL 2007* (pp. 616–623).

Reichart, R., & Rappoport, A. (2009). Automatic selection of high quality parses created by a fully unsupervised parser. In *Proceedings of CoNLL 2009* (pp. 156–164).

Sagae, K. (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 (DANLP 2010)* (pp. 37–44). Uppsala, Sweden.

Sagae, K., & Tsujii, J. (2007). Dependency parsing and domain adaptation with lr models and parser ensemble. In *Proceedings of EMNLP–CoNLL 2007* (pp. 1044–1050).

Tateisi, Y., Yakushiji, A., Ohta, T., & Tsujii, J. (2005). Syntax annotation for the GENIA corpus. In *Proceedings of IJCNLP'05* (pp. 222–227). Jeju Island, Korea.