Domain Adaptation for Dependency Parsing at Evalita 2011

Felice Dell'Orletta*, Simone Marchi*,
Simonetta Montemagni*, Giulia Venturi◇,
Tommaso Agnoloni†, and Enrico Francesconi†

*Istituto di Linguistica Computazionale "Antonio Zampolli", CNR - Pisa
◇Scuola Superiore Sant'Anna di Studi Universitari e di Perfezionamento - Pisa
†Istituto di Teoria e Tecniche dell'Informazione Giuridica, CNR - Firenze
{name.surname}@ilc.cnr.it
{name.surname}@ittig.cnr.it

Abstract. The domain adaptation task was aimed at investigating techniques for adapting state–of–the–art dependency parsing systems to new domains. Both the language dealt with, i.e. Italian, and the target domain, namely the legal domain, represent two main novelties of the task organised at Evalita 2011. In this paper, we define the task and describe how the datasets were created from different resources. In addition, we characterize the different approaches of the participating systems, report the test results, and provide a first analysis of these results.

Keywords: Dependency Parsing, Domain Adaptation, Self-training, Active Learning, Legal-NLP

1 Motivation

In spite of the fact that nowadays dependency parsing can be carried out with high levels of accuracy, the adaptation of parsers to new domains without target domain training data remains an open issue, as testified by several initiatives organised around this topic: e.g. the "Domain Adaptation Track" organized in the framework of the CoNLL 2007 Shared Task [11] and the Workshop on "Domain Adaptation for Natural Language Processing" (DANLP 2010) [5]. The domain adaptation (DA) task at Evalita 2011 aims to investigate techniques for adapting state–of–the–art dependency parsing systems to domains outside of the data from which they were trained or developed, with two main novelties: the language being dealt with, i.e. Italian, and the target domain, namely the legal domain. The motivations underlying the choice of the legal domain as a target are two–fold. From the linguistic point of view, the legal language is characterized by quite a peculiar distribution of morpho-syntactic as well as syntactic features with respect to the general language [12]. On the applicative front, it appears that a number of different legal text processing tasks could benefit significantly from the existence of dependency parsers adapted to the domain, e.g. legal argumentation, extraction of textual legal case elements and factors, legal text consolidation to mention only a few.

2 Definition of the Task

In the literature, work on domain adaptation falls roughly into two categories based on whether limited annotated resources for the target domain are available or not. If no annotated resources are available for the target domain, a large unlabeled corpus can be leveraged in adaptation: this was the scenario assumed in the Domain Adaptation Track at CoNLL 2007. For Evalita'11, we decided to organize the task into two different subtasks with the final aim of exploring a wider range of approaches to domain adaptation of dependency parsers. The two subtasks can be described as follows:
1) **minimally supervised domain adaptation** with limited annotated resources in the target domain and unlabeled corpora;
2) **unsupervised domain adaptation** with no annotated resources in the target domain, i.e. using only unlabeled target data.

Evaluation has been carried out in terms of standard accuracy dependency parsing measures, i.e. labeled attachment score (LAS) with respect to a test set of texts from the target domain.

3 Dataset

Different datasets have been distributed for the source and the target domains.

The source data is drawn from a corpus of news, the ISST–TANL corpus jointly developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) and the University of Pisa, exemplifying general language usage and consisting of articles from newspapers and periodicals, selected to cover a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.). This corpus represents a revised version of the ISST–TANL corpus used in the dependency parsing track of Evalita 2009 (pilot sub–task, [4]): the main revisions are concerned with the treatment of multi-word expressions, multi–rooted sentences as well as with revised annotation criteria for the treatment of sentential complements and the argument/adjunct distinction. For the source domain, two different datasets have been distributed to participants: a training corpus of ∼72,000 tokens and ∼3200 sentences and a development corpus (hereafter referred to as *SDevel*) of ∼5,000 tokens (∼250 sentences).

For the target domain we used three different data sets:

1. a target corpus (∼13 millions tokens and ∼ 620,000 sentences) drawn from an Italian legislative corpus, gathering laws enacted by different releasing agencies (European Commission, Italian State and Regions) and regulating a variety of domains, ranging from environment, equal opportunities for men and women, travel regulation, etc. The target corpus includes automatically generated sentence splitting, tokenization, morpho–syntactic tagging and lemmatization;
2. a manually annotated development set (hereafter referred to as *TDevel*), also including labeled dependency relations, consisting of 148 sentences for a total of 5,691 tokens;

3. a test set used for the evaluation (hereafter referred to as *Test*) constituted by 168 sentences for a total of 5,374 tokens and including labeled dependency relations.

Distributed data adhere to the CoNLL 2007 tabular format used in the Shared Task on Dependency Parsing [9]. The morpho–syntactic and dependency tagsets were jointly developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) and the University of Pisa in the framework of the TANL (Text Analytics and Natural Language processing) project[1].

3.1 Source vs Target Domain Corpora Annotation Criteria

Note that in order to properly handle legal language peculiarities there are slight differences in the annotation criteria used for the corpora of the source and target domains. The main differences are concerned with both sentence splitting and dependency annotation. For sentence splitting, differently from the source domain in the target domain corpora sentence splitting is overtly meant to preserve the original structure of the law text. This entails that also punctuation marks such as ';' and ':', when followed by a carriage return, are treated as sentence boundary markers. For what concerns dependency annotation, it should be considered that legal texts are characterized by syntactic constructions hardly or even never occurring in the source domain corpora. In order to successfully cope with such peculiarities of legal texts, dependency annotation criteria have been extended to cover the annotation of legal language-specific *a)* elliptical constructions, *b)* participial phrases as well as *c)* long distance dependencies resulting in non–projective links to mention only a few. All these peculiar constructions are explicitly represented in the development target domain corpus.

4 Participation Results

The participants to the task were two, namely **Plank_Søgaard** and **Attardi_et-_al.**. They used two different dependency parsers and followed quite different approaches to domain adaptation. Both teams participated in the unsupervised domain adaptation subtask, while only the latter presented results for the minimally supervised domain adaptation subtask.

4.1 Base Parsing Models

Attardi_et_al. used *DeSR* [1], which is a Shift/Reduce deterministic transition–based parser that by using special rules is able to handle non–projective dependencies in linear time complexity. In Evalita 2011 the system is tested using a combination of three different configurations of two stage Reverse Revision parser [2], i.e. a stacked Righ-to-Left parser that uses hints produced by a first pass Left-to-Right parser.

[1] http://medialab.di.unipi.it/wiki/SemaWiki

Plank_Søgaard used the second-order projective model of MSTParser [8] with the on-line learning algorithm MIRA [7]. MSTParser is a graph-based parser which uses a maximum spanning tree algorithm for finding the highest scoring tree. In the second–order model, the parser factors the score of the tree into the sum of adjacent edge pair scores. They used the projective parsing algorithm of MSTParser that is unable to handle non-projective dependencies.

4.2 Domain Adaptation Strategies

Attardi_et_al. used a method based on active learning for both the minimally and unsupervised domain adaptation tasks. They followed a three–step incremental process where each step generates a new training corpus including manually revised dependency–annotated sentences from the target unlabeled corpus. Each step can be summarised as follows: a) DeSR with MLP (Multi Layer Perceptron Algorithm) [3] as learning algorithm is used to parse the unlabeled target corpus; b) perplexity measures based on the overall likelihood of the analysis of each sentence provided by DeSR are exploited to identify sentences with the highest perplexity and with a maximum length of n–tokens (where n differs for each step) and c) sentences selected during the previous step are manually revised and used to extend the training corpus in order to build a new parser model. At the end of the whole process, the base training set was augmented with 188 sentences. The augmented training set was used to parse the target test set. For the last run they used the parser system described in section 4.1 with SVM (Support Vector Machine) as learning algorithm.

Plank_Søgaard submitted two runs for the same unsupervised domain adaptation task, based on two different adaptation strategies both belonging to the class of self–training methods. The first adaptation strategy can be seen as a kind of "indirect" self–training approach. The unlabeled target corpus is parsed and statistics about non–lexicalised dependency triplets (i.e. <head_POS, dep_POS, dep_type>) are extracted. For each triplet, they calculate a normalized point-wise mutual information score ranging from 0 to 1. The triplet scores are put in bins to have binary-valued features that are used as new features by the parser. In the second strategy, the parser model results from the combination of dependency triplets features with a pure instance of self–training approach. They randomly selected from the parsed target corpus 12,800 parses with a maximum length of 100 tokens. These sentences were combined with the source training corpus in order to build a new parser model.

4.3 Results and Discussion

Table 4.3 reports the results for the minimally supervised domain adaptation task: Attardi_et_al.–Base(a) refers to the results of the base parser model using for training both the source training corpus and *TDevel* whereas Attardi_et_al.–DA(a) refers to the results achieved after domain adaptation.

Table 4.3 reports the results for the unsupervised domain adaptation task. Here, *system*–Base refers to the base parser model using only the source training

Table 1. LAS for the minimally supervised domain adaptation task.

System	SDevel	Test
Attardi_et_al.–Base(a)	82.09	80.29
Attardi_et_al.–DA(a)	82.34	81.39

corpus in the training phase. Attardi_et_al.–DA(b) refers to the results obtained after the active learning process, Plank_Søgaard-DA1 refers to the parser model using dependency triplets features and Plank_Søgaard-DA2 refers to the parser model combining dependency triplets features with the self–training process.

Table 2. LAS for the unsupervised domain adaptation task.

System	SDevel	Test
Attardi_et_al.–Base(b)	82.09	75.85
Attardi_et_al.–DA(b)	81.09	80.83
Plank_Søgaard-Base	80.19	74.62
Plank_Søgaard-DA1	80.87	74.02
Plank_Søgaard-DA2	80.31	74.30

Despite the fact that the results obtained by the two teams are not comparable due to the deep difference holding between the adopted DA strategies, we can observe that the active learning method by Attardi_et_al. shows a significant parsing improvement (i.e. 1.1% in the minimally supervised DA task and 4.98% in the unsupervised one), whereas no improvements could be detected with the self–training approaches experimented by Plank_Søgaard[2]. The reasons underlying the low performance of self–training methods need to be further investigated. Among the possible causes we should mention the syntactic peculiarities of legal texts. The good performance of the active learning method suggests that a small amount of new target data (188 sentences only) are enough to enable the parser to reliably handle new syntactic structures specific to the target domain. On the other hand, the self–training methods which have been experimented with do not appear to be able to detect such key sample data, despite the fact that 12,800 new sentences were combined with the source training corpus in order to build a new parser model. Incidentally, this could also explain the great improvement achieved by Attardi_et_al. in the minimally supervised task, where the base parser model shows already a significant improvement by enriching the source training set with TDevel in the training phase. As a last remark, it should be noted that Attardi_et_al.–DA tested on SDevel decreases of 1% with respect to the base model. The new target data (188 manually corrected sentences from the

[2] It is interesting to note that, contrary to the results achieved on *Test*, Plank_Søgaard registered some improvement on *TDevel* (see their Evalita 2011 report).

unlabeled corpus) are constituted by sentences showing the highest perplexity score, i.e. sentences characterised by peculiar linguistic features with respect to the source training data. This result is in line with what observed by Plank and van Noord ([10]) who proved that parsers trained on the *union* of more than one different *gold* corpora taken from different domains achieve lower accuracy with respect to the same parsers trained on data belonging to a single target domain.

5 Conclusions

The participant results demonstrated that the active learning strategy achieves a good performance in adapting dependency parsers to the legal domain, whereas no improvements could be achieved with self–training approaches. Since the results of this shared task are in contrast with other DA experiments (e.g. [6] and [11]) carried out on different target domains (e.g. biomedical, chemical, etc.), we believe that the low performance of self–training methods needs to be further investigated: this is possibly due to the linguistic peculiarities characterising the legal language which has been dealt with for the first time in a DA shared task.

References

1. Attardi, G.: Experiments with a multilanguage non-projective dependency parser. In: Proceedings of the Shared task CoNLL-X, New York City, pp. 166–170 (2006)
2. Attardi, G., Dell'Orletta, F.: Reverse revision and linear tree combination for dependency parsing. In: Proceedings of NAACL-HLT (2009)
3. Attardi, G., Dell'Orletta, F., Simi, M., Turian, J.: Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In: Proceedings of Evalita'09 (2009)
4. Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F., Lenci, A.: Evalita'09 Parsing Task: comparing dependency parsers and treebanks. In: Proceedings of Evalita'09, Reggio Emilia (2009)
5. Proceedings of the Workshop on Domain Adaptation for Natural Language Processing (2010), available at http://aclweb.org/anthology/W/W10/W10-2600.pdf
6. McClosky, D.: Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing. Ph.D. thesis, Brown University (2010)
7. McDonald, R., Crammer, K., Pereira, F.: Online large-margin training of dependency parsers. In: Proceedings of ACL (2005)
8. McDonald, R., Lerman, K., Pereira, F.: Multilingual Dependency Analysis with a Two-Stage Discriminative Parser. In: Proceedings of CoNLL, New York City (2006)
9. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 Shared Task on Dependency Parsing. In: Proceedings of the EMNLP-CoNLL, pp. 915–932 (2007)
10. Plank, B. and van Noord, G.: Effective Measures of Domain Similarity for Parsing. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, June 19-24, pp. 1566–1576 (2011)
11. Sagae K., Tsujii, J.: Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, pp. 1044–1050, Prague, June (2007)

12. Venturi, G.: Lingua e diritto: una prospettiva linguistico–computazionale. Ph.D. Thesis, Università di Torino (2011)