Data point selection for cross-language adaptation of dependency parsers

Anders Søgaard
Center for Language Technology
University of Copenhagen
Njalsgade 142, DK-2300 Copenhagen S
soegaard@hum.ku.dk

Abstract

We consider a *very* simple, yet effective, approach to cross language adaptation of dependency parsers. We first remove lexical items from the treebanks and map part-of-speech tags into a common tagset. We then train a language model on tag sequences in otherwise unlabeled target data and rank labeled source data by perplexity per word of tag sequences from less similar to most similar to the target. We then train our target language parser on the most similar data points in the source labeled data. The strategy achieves much better results than a non-adapted baseline and state-of-the-art unsupervised dependency parsing, and results are comparable to more complex projection-based cross language adaptation algorithms.

1 Introduction

While unsupervised dependency parsing has seen rapid progress in recent years, results are still far from the results that can be achieved with supervised parsers and not yet good enough to solve real-world problems. In this paper, we will be interested in an alternative strategy, namely cross-language adaptation of dependency parsers. The idea is, briefly put, to learn how to parse Arabic, for example, from, say, a Danish treebank, comparing unlabeled data from both languages. This is similar to, but more difficult than most domain adaptation or transfer learning scenarios, where differences between source and target distributions are smaller.

Most previous work in cross-language adaptation has used parallel corpora to project dependency structures across translations using word alignments (Smith and Eisner, 2009; Spreyer and Kuhn, 2009; Ganchev et al., 2009), but in this paper we show that similar results can be achieved by much simpler means. Specifically, we build on the cross-language adaptation algorithm for closely related languages developed by Zeman and Resnik (2008) and extend it to much less related languages.

1.1 Related work

Zeman and Resnik (2008) simply mapped part-of-speech tags of source and target language treebanks into a common tagset, delexicalized them (removed all words), trained a parser on the source language treebank and applied it to the target language. The intuition is that, at least for relatively similar languages, features based on part-of-speech tags are enough to do reasonably well, and languages are relatively similar at this level of abstraction. Of course annotations differ, but nouns are likely to be dependents of verbs, prepositions are likely to be dependents of nouns, and so on.

Specifically, Zeman and Resnik (2008) trained a constituent-based parser on the training section of the Danish treebank and evaluated it on sentences of up to 40 words in the test section of the Swedish treebank and obtained an F_1-score of 66.40%. Danish and Swedish are of course *very* similar languages with almost identical syntax, so in a way this result is not very surprising. In this paper, we present similar results (50-75%) on full length sentences for very different languages from different language families. Since less related languages differ more in their syntax, we use data point selection to find syntactic

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 682–686,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

constructions in the source language that are likely to be similar to constructions in the target language.

Smith and Eisner (2009) think of cross-language adaptation as *unsupervised projection* using word aligned parallel text to construct training material for the target language. They show that hard projection of dependencies using word alignments performs better than the unsupervised dependency parsing approach described in Klein and Manning (2004), based on EM with clever initialization, and that a quasi-synchronous model using word alignments to reestimate parameters in EM performs even better. The authors report good results (65%-70%) for somewhat related languages, training on English and testing on German and Spanish, but they modified the annotation in the German data making the treatment of certain syntactic constructions more similar to the English annotations.

Spreyer and Kuhn (2009) use a similar approach to parse Dutch using labeled data from German and obtain good results, but again these are *very* similar languages. They later extended their results to English and Italian (Spreyer et al., 2010), but also modified annotation considerably in order to do so.

Finally, Ganchev et al. (2009) report results of a similar approach for Bulgarian and Spanish; they report results with and without hand-written language-specific rules that complete the projected partial dependency trees.

We will compare our results to the plain approach of Zeman and Resnik (2008), Ganchev et al. (2009) without hand-written rules and two recent contributions to unsupervised dependency parsing, Gillenwater et al. (2010) and Naseem et al. (2010). Gillenwater et al. (2010) is a fully unsupervised extension of the approach described in Klein and Manning (2004), whereas Naseem et al. (2010) rely on hand-written cross-lingual rules.

2 Data

We use four treebanks from the CoNLL 2006 Shared Task with standard splits. We use the tagset mappings also used by Zeman and Resnik (2008) to obtain a common tagset.[1][2] They define tagset mappings for Arabic, Bulgarian, Czech, Danish, Portuguese and Swedish. We only use four of these treebanks, since Bulgarian and Czech as well as Danish and Swedish are very similar languages.

The four treebanks used in our experiments are thus those for Arabic, Bulgarian, Danish and Portuguese. Arabic is a Semitic VSO language with relatively free word order and rich morphology. Bulgarian is a Slavic language with relatively free word order and rich morphology. Danish is a Germanic V2 language with relatively poor morphology. Finally, Portuguese is a Roman language with relatively free word order and rich morphology. In sum, we consider four languages that are less related than the language pairs studied in earlier papers on cross-language adaptation of dependency parsers.

3 Experiments

3.1 Data point selection

The key idea in our experiments is that we can use a simple form of instance weighting, similar to what is often used for correcting sample selection bias or for domain adaptation, to improve the approach in Zeman and Resnik (2008) by selecting only sentences in the source data that are similar to our target domain or language, considering their perplexity per word in a language model trained on target data. The idea is that we order the labeled source data from most similar to least similar to our target data, using perplexity per word as metric, and use only a portion of the source data that is similar to our target data.

In cross-language adaptation, the sample selection bias is primarily a bias in marginal distribution $P(\mathbf{x})$. This is the covariate shift assumption (Shimodaira, 2000). Consequently, each sentence should be weighted by $\frac{P_t(\mathbf{x})}{P_s(\mathbf{x})}$ where P_t is the target distribution, and P_s the source distribution.

To see this let $\mathbf{x} \in \mathcal{X}$ in lowercase denote a specific value of the input variable, an unlabeled example. $y \in \mathcal{Y}$ in lowercase denotes a class value, and $\langle \mathbf{x}, y \rangle$ is a labeled example. $P(\langle \mathbf{x}, y \rangle)$ is the joint probability of the labeled example, and $\hat{P}(\langle \mathbf{x}, y \rangle)$ its empirical distribution.

In supervised learning with N labeled data points, we minimize the empirical risk to find a good model $\hat{\theta}$ for a loss function $l : \mathcal{X} \times \mathcal{Y} \times \Theta$:

[1] https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset

[2] We use the first letter in the common tag as coarse-grained part-of-speech, and the first three as fine-grained part-of-speech.

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{\langle \mathbf{x},y\rangle \in \mathcal{X}\times\mathcal{Y}} \hat{P}(\langle \mathbf{x},y\rangle) l(\mathbf{x},y,\theta)$$

$$= \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} l(\mathbf{x}_i, y_i, \theta)$$

In domain adaptation, we can rewrite this as:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{\langle \mathbf{x},y\rangle \in \mathcal{X}\times\mathcal{Y}} \frac{P_t(\langle \mathbf{x},y\rangle)}{P_s(\langle \mathbf{x},y\rangle)} \hat{P}_s(\langle \mathbf{x},y\rangle) l(\mathbf{x},y,\theta)$$

$$= \arg\min_{\theta\in\Theta} \sum_{i=1}^{N^s} \frac{P_t(\langle \mathbf{x}_i^s, y_i^s\rangle)}{P_s(\langle \mathbf{x}_i^s, y_i^s\rangle)} l(\mathbf{x}_i^s, y_i^s, \theta)$$

Under the covariate shift assumption $\frac{P_t(\langle \mathbf{x},y\rangle)}{P_s(\langle \mathbf{x},y\rangle)}$ for a pair $\langle \mathbf{x}, y\rangle$ can be replaced with $\frac{P_t(\mathbf{x})}{P_s(\mathbf{x})}$. We simplify this function further assuming that

$$\frac{P_t(\mathbf{x})}{P_s(\mathbf{x})} = \left\{ \begin{array}{l} 0 \text{ if } P_t(\mathbf{x}) \text{ is low} \\ 1 \text{ if } P_t(\mathbf{x}) \text{ is high} \end{array} \right\}$$

We use perplexity per word of the source language POS sequences relative to a model trained on target language POS sequences to guess whether $P_t(\mathbf{x})$ is high or low.

The treebanks are first delexicalized and all features except part-of-speech tags removed. The part-of-speech tags are mapped into a common tagset using the technique described in Zeman and Resnik (2008). For our main results, which are presented in Figure 1, we use the remaining three treebanks as training material for each language. The test section of the language in question is used for testing, while the POS sequences in the target training section is used for training the unsmoothed language model. We use an unsmoothed trigram language model rather than a smoothed language model since modified Knesser-Ney smoothing is not defined for sequences of part-of-speech tags.[3]

In our experiments we use a graph-based second-order non-projective dependency parser that induces models using MIRA (McDonald et al., 2005).[4] We do not optimize parameters on the different languages, but use default parameters across the board.

We present two results and a baseline for each language in Figure 1. Our baseline is the accuracy of our dependency parser trained on three languages and evaluated on the fourth language, where treebanks have been delexicalized, and part-of-speech tags mapped into a common format. This is the proposal by Zeman and Resnik (2008). We then present results using the 90% most similar data points and results where the amount of labeled data used is selected using 100 sentences sampled from the training data as held-out data. It can be seen that using 90% of the labeled data seems to be a good strategy if using held-out data is not an option. Since we consider the unsupervised scenario where no labeled data is available for the target language, we consider the results obtained using the 90% most similar sentences in the labeled data as our primary results.

That we obtain good results training on all the three remaining treebanks for each language illustrates the robustness of our approach. However, it may in some cases be better to train on data from a single resource only. The results presented in Figure 2 are the best results obtained with varying amounts of source language data (10%, 20%, . . . , or 100%). The results are only explorative. In all cases, we obtain slightly results with training material from only one language that are better than or as good as our main results, but differences are marginal. We obtain the best results for Arabic training using labeled data from the Bulgarian treebank, and the best results for Bulgarian training on Portuguese only. The best results for Danish were, somewhat surprisingly, obtained using the Arabic treebank,[5] and the best results for Portuguese were obtained training only on Bulgarian data.

4 Error analysis

Consider our analysis of the Arabic sentence in Figure 3, using the three remaining treebanks as source data. First note that our dependency labels are all wrong; we did not map the dependency labels of the source and target treebanks into a common set of labels. Otherwise we only make mistakes about punctuation. Our labels seem meaningful, but come

[3] http://www-speech.sri.com/projects/srilm/

[4] http://sourceforge.net/projects/mstparser/

[5] Arabic and Danish have in common that definiteness is expressed by inflectional morphology, though, and both languages frequently use VSO constructions.

	Arabic		Bulgarian		Danish		Portuguese	
	≤ 10	∞	≤ 10	∞	≤ 10	∞	≤ 10	∞
Ganchev et al. (2009)	-	-	67.8	-	-	-	-	-
Gillenwater et al. (2010)	-	-	54.3	-	47.2	-	59.8	-
Naseem et al. (2010)	-	-	-	-	51.9	-	71.5	-
100% (baseline)	-	45.5	-	44.5	-	51.7	-	37.1
90%	48.3	48.4	**77.1**	**70.2**	**59.4**	51.9	**83.1**	**75.1**
Held-out %	-	**49.2**	-	**70.3**	-	**52.8**	-	**75.1**

Figure 1: Main results.

source/target	Arabic	Bulgarian	Danish	Portuguese
Arabic	-	45.8	**56.5**	37.8
Bulgarian	**50.2**	-	50.8	**76.9**
Danish	46.9	60.4	-	63.5
Portuguese	50.1	**70.3**	52.2	-

Figure 2: Best results obtained with different combinations of source and target languages.

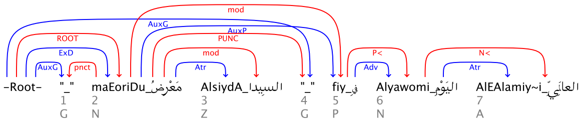

Figure 3: A predicted analysis for an Arabic sentence and its correct analysis.

from different treebanks, e.g. 'pnct' from the Danish treebank and 'PUNC' from the Portuguese one.

If we consider the case where we train on all remaining treebanks and use the 90% data points most similar to the target language, and compare it to our 100% baseline, our error reductions are distributed as follows, relative to dependency length: For Arabic, the error reduction in F_1 scores decreases with dependency length, and more errors are made attaching to the root, but for Portuguese, where the improvements are more dramatic, we see the biggest improvements with attachments to the roots and long dependencies:

Portuguese	bl (F_1)	90% (F_1)	err.red
root	0.627	0.913	76.7%
1	0.720	0.894	62.1%
2	0.292	0.768	67.2%
3–6	0.328	0.570	36.0%
7–	0.240	0.561	42.3%

For Danish, we see a similar pattern, but for Bulgarian, error reductions are equally distributed.

Generally, it is interesting that cross-language adaptation and data point selection were less effective for Danish. One explantation may be differences in annotation, however. The Danish dependency treebank is annotated very differently from most other dependency treebanks; for example, the treebank adopts a DP-analysis of noun phrases.

Finally, we note that all languages benefit from removing the least similar 10% of the labeled source data, but results are less influenced by how much of the remaining data we use. For example, for Bulgarian our baseline result using 100% of the source data is 44.5%, and the result obtained using 90% of the source data is 70.2%. Using held-out data, we only use 80% of the source data, which is slightly better (70.3%), but even if we only use 10% of the source data, our accuracy is still significantly better than the baseline (66.9%).

5 Conclusions

This paper presented a simple data point selection strategy for semi-supervised cross language adaptation where no labeled target data is available. This problem is difficult, but we have presented very positive results. Since our strategy is a parameter-free wrapper method it can easily be applied to other dependency parsers and other problems in natural language processing, incl. part-of-speech tagging, named entity recognition, and machine translation.

References

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *ACL*.

Jennifer Gillenwater, Kuzman Ganchev, Joao Graca, Fernando Pereira, and Ben Taskar. 2010. Sparsity in dependency grammar induction. In *ACL*.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *ACL*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *ACL*.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *EMNLP*.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244.

David Smith and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *EMNLP*.

Kathrin Spreyer and Jonas Kuhn. 2009. Data-driven dependency parsing of new languages using incomplete and noisy training data. In *CoNLL*.

Kathrin Spreyer, Lilja Øvrelid, and Jonas Kuhn. 2010. Training parsers on partial trees: a cross-language comparison. In *LREC*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJC-NLP*.