An empirical study of non-lexical extensions to delexicalized transfer

Anders Søgaard and Julie Wulff
Center for Language Technology
University of Copenhagen
DK-2300 Copenhagen S
soegaard@hum.ku.dk

ABSTRACT
We propose a simple cross-language parser adaptation strategy for discriminative parsers and apply it to easy-first transition-based dependency parsing (Goldberg and Elhadad, 2010). We evaluate our parsers on the Indo-European corpora in the CoNLL-X and CoNLL 2007 shared tasks. Using the remaining languages as source data we average under-fitted weights learned from each source language and apply the resulting linear classifier to the target language. Of course some source languages and some sentences in these languages are more relevant than others for the target language in question. We therefore explore improvements of our cross-language adaptation model involving source language and instance weighting, as well as unsupervised model selection. Overall our cross-language adaptation strategies provide better results than previous strategies for direct transfer, with near-linear time parsing and much faster training times than other approaches.

KEYWORDS: cross-language dependency parsing, regularization, importance weighting, typological information.

Proceedings of COLING 2012: Posters, pages 1181–1190,
COLING 2012, Mumbai, December 2012.

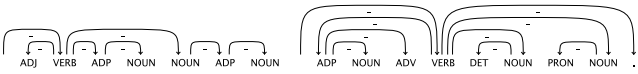

Figure 1: Bulgarian and German dependency structures, delexicalized

1 Introduction

High-quality syntactic parsing is important for advanced language technologies such as question-answering, machine translation between distant languages, and sentiment analysis. State-of-the-art parsers provide accurate syntactic analyses in languages for which annotated resources known as *treebanks* exist, although significant and sometimes prohibitive performance drops are observed when parsing text that differs in domain or genre from the available treebank(s). However, there are still many languages for which no treebanks exist, and for which we therefore do not have parsers available.

Unsupervised parsing has seen considerable progress over the last ten years (Gelling et al., 2012), but recently several authors have demonstrated that better results can be achieved transferring linguistic knowledge from treebanks from other languages rather than inducing this knowledge from unannotated text (Zeman and Resnik, 2008; Smith and Eisner, 2009; Spreyer and Kuhn, 2009; Søgaard, 2011; Cohen et al., 2011; McDonald et al., 2011; Täckström et al., 2012; Naseem et al., 2012). Some of these authors have projected syntactic structures across word aligned parallel text (Smith and Eisner, 2009; Spreyer and Kuhn, 2009; McDonald et al., 2011), while others have used a much simpler technique sometimes referred to as *delexicalized transfer* (Zeman and Resnik, 2008; Søgaard, 2011; Cohen et al., 2011; McDonald et al., 2011; Täckström et al., 2012; Naseem et al., 2012). This work presents a new approach to delexicalized transfer and explores possible improvements.

Sect. 2 covers related work on delexicalized transfer for cross-language parser adaptation. We explore delexicalized transfer in the context of easy-first transition-based dependency parsing (Goldberg and Elhadad, 2010), which is introduced in Sect. 3. Sect. 4 introduces our implementation of delexicalized transfer which differs from other approaches by doing model averaging of under-fitted models rather than concatenating data when learning from multiple source languages. Sect. 4 also introduces the possible improvements of this model we explore in our experiments. Sect. 5 and 6 present the experiments and results.

2 Cross-language adaptation with delexicalized transfer

Delexicalized transfer refers to a simple idea first introduced in Zeman and Resnik (2008). In unsupervised parsing, you hope to learn that nouns tend to attach to verbs, determiners tend to attach to nouns, adverbs to verbs, etc. Many of these tendencies are cross-linguistic tendencies, a fact also exploited in unsupervised parsing by Naseem et al. (2010) and Søgaard (2012). Since this knowledge is reflected in most treebanks, can't we extract this knowledge from a treebank in *one* language and apply the resulting model to another language for which we do not have a treebank? There are certainly differences between distant languages in, for example, how likely adjectives are to modify adverbs rather than nouns, but as mentioned in McDonald et al. (2011) using multiple source languages may reduce such biases on average.

Zeman and Resnik (2008), in their seminal paper, considered a pair of closely related languages, namely Danish and Swedish. They removed the words from the source treebank and

learned a parsing model from the distribution of parts of speech (POS) only. In order to parse the target language they devised a mapping of the different POS tag sets into a common feature representation.

Consider the two delexicalized dependency structures in Figure 1 to see how this makes sense. The left structure is from the Bulgarian treebank, and the right one from the German. However, the left structure contains many edges that also occur in the right structure, e.g. from root to verb, from verb to noun, and from adposition to noun. A dependency parser can learn such dependencies are likely in Bulgarian, but apply this knowledge when parsing German.

Independently of each other, three papers revisited the idea of delexicalized transfer in 2011. Søgaard (2011) used the tag set mappings in Zeman and Resnik (2008), but also used instance weighting to do a form of outlier detection. In a way similar to Jiang and Zhai (2007) he did not make use of the actual weights, but simply used all labeled instances with weights greater than some fixed threshold. McDonald et al. (2011) used the more recent tag set mappings by Petrov et al. (2011), explored combinations of delexicalized transfer and structure projection (Smith and Eisner, 2009; Spreyer and Kuhn, 2009) and were able to improve delexicalized transfer averaging across several languages. They also were the first to explicitly introduce the idea of using multiple source languages as a kind of regularization. Finally, Cohen et al. (2011) used the delexicalized transfer models to initialize unsupervised parameter estimation on unlabeled target data.

Naseem et al. (2012) subsequently explored a more complex transfer model where only hierarchical information is transferred directly to reflect that languages have very different word orders.

Täckström et al. (2012) augment delexicalized transfer with bilingual clusters, while Durrett et al. (2012) use a bilingual dictionary to project lexical features. Täckström (2012) used self-training to supply the bilingual word clusters with monolingual clusters, but evaluated the idea on named entity recognition rather than cross-language parser adaptation.

3 Easy-first transition-based parsing with averaged perceptron

The parser we apply in this study is a non-directional easy-first transition-based dependency parser (Goldberg and Elhadad, 2010). The parser consecutively applies one of two actions, AttachLeft(i) and AttachRight(i), to a list of partial structures initialized as the words in the sentence. Each action connects the heads of two neighboring structures, making one the head of the other. The dependent partial structure is removed from the list. The parsing algorithm is obviously projective.

The next action is chosen by a score function $score(\text{Action})(i)$ that assigns a weight to all pairs of actions and locations. The scoring function ideally ranks possible actions from easy to hard. The scoring function is learned from data using a variant of the averaged perceptron learning algorithm (Freund and Schapire, 1999; Collins, 2002) similar to the one used in Shen et al. (2007). While the ordering from easy to hard is not known in advance, the ordering is implicitly learned by decreasing weights associated with invalid actions and increasing weights associated with the currently highest scoring valid action.

The major advantage of using easy-first parsing is the efficient $\mathcal{O}(n \log n)$ parsing algorithm, but training is also a lot faster than with comparable dependency parsers (Goldberg and Elhadad, 2010); e.g. training an experiment for Spanish-Italian on a Macbook Pro takes less than a

minute. The easy-first learning algorithm is used throughout our experiments, except that we modify the update function when using easy-first with importance weighting.

4 Cross-language adaptation of easy-first parsing

The easy-first dependency parser (Goldberg and Elhadad, 2010) by default does 20 passes over the data and returns an averaged weight vector as our parsing model. Since our training data in cross-language adaptation is heavily biased, we do not want to over-fit our models to source data and only do a single round over data for each source language. This hyper-parameter is kept constant in all experiments. We use the feature model proposed for English in Goldberg and Elhadad (2010) for all languages (see Discussion). There are no other hyper-parameters to the parsing model. In our experiments we consider some possible extensions of this simple model. The extensions are discussed in the following subsections:

4.1 Language-level weighted learning

Intuitively some languages are more relevant as source languages for some language than others. While results in the literature show that good source languages may be geographically distant and unrelated to the target language (e.g. Arabic and Danish (Søgaard, 2011)), genealogically related languages should in general be better source languages for each other. This idea was first explored by Berg-Kirkpatrick and Klein (2010) who used genealogical relations to impose contraints on models in multi-lingual grammar inductions.

In the experiments below we use a language genealogy to take a weighted average of the models obtained from our set of source languages. Each model is weighted by $4-d$ where d is the distance between the source language and a node dominating the target language node in a genealogical tree (see Figure 2). If two languages belong to the same subfamily such as the Western Germanic languages Dutch and German, $d = 1$, but for Dutch to Greek, for example, $d = 3$.

We also try using a typological database to weight languages by their typological properties. The database lists basic typological properties of languages such as order of plural and noun, or whether the language has anti-passive constructions, and we simply let the weight of each target language be the inverse of the Hamming distance between the source language property vector and its own property vector.

$$f_w(L_S) = \frac{5}{H(L_S, L_T)}$$

The constant was chosen such that the number of weight vectors used in averaging was comparable to the experiments using linguistic genealogy to weight source languages.

Naseem et al. (2012) explore a similar idea (using the same typological database we do), but (a) they only use a subset of features (without specifying how this subset was selected), and (b) they use the typological features in a generative model rather than distances between property vectors.

4.2 Unsupervised model selection

Our models for the source languages are comparable weight vectors, i.e. the ith weight encodes the importance of the same feature across all models. So models with similar weights will lead

to similar decisions. Like with data points in graph-based semi-supervised learning, we can build a graph out of our models with edges representing similarity of models. We can then use random walk algorithms to select diverse or homogeneous subsets of models.

In our experiments we explore the following idea: Using a random walk we compute the probability of reaching a node in our graph of source language models. We then compute the average of the three lower quartiles in order to optimize diversity and thereby regularization in the final model. In principle we want to optimize diversity and individual accuracy (Brown et al., 2005), but this method does unsupervised model selection not taking accuracy into account. Combining this technique with the weighting techniques introduced here which are intended to optimize for individual accuracy is theoretically an appealing option.

4.3 Sentence-level weighted learning

Some sentences in a source language may be more like the target language than others. Søgaard (2011) introduces a very simple idea to reflect this intuition. He uses a language model over POS sequences to remove outliers, discarding the 10% source language labeled sentences with highest perplexity per word according to a target language model. We go beyond Søgaard (2011) by learning from *weighted* data, where each weight estimates the relevance of a labeled sentence by the perplexity by word of the corresponding POS sequence in a target language model over its perplexity by word in a source language model.

In weighted perceptron learning (Cavallanti et al., 2006), we make the learning rate dependent on the current instance, using the following update rule on $\mathbf{x}_n$:

$$\mathbf{w}^{i+1} \leftarrow \mathbf{w}^i + \beta_n \alpha(y_n - \text{sign}(\mathbf{w}^i \cdot \mathbf{x}_n))\mathbf{x}_n \quad (1)$$

where $\beta_1, \ldots, \beta_n$ are importance weights.

Søgaard and Haulrich (2011) present an application of an importance weighted version of the MIRA algorithm (Crammer and Singer, 2003) and apply it to dependency parsing. Huang et al. (2007) present an instance-weighted learning algorithm for support vector machines. Under the assumption that differences between source and target distributions are due to sample bias only (which is clearly not the case here) we should weight a data point by its probability in the target domain over its probability in the source domain (Shimodaira, 2000), but since it is not possible to estimate densities in our case, we resort to a heuristic combining the insights from Shimodaira (2000) and Søgaard (2011), weighting each sentence in every source language treebank by:

$$f_w(\mathbf{x}) = \frac{\sqrt{ppw_S(\mathbf{x})}}{\sqrt{ppw_T(\mathbf{x})}}$$

where $ppw_D(\cdot)$ is the perplexity per word given a language model trained on a corpus sampled from domain D.

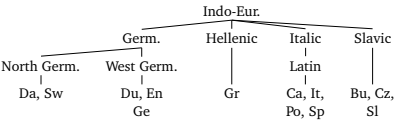

Figure 2: Language genealogy.

5 Experiments

Our parser is a modification of the publicly available implementation of the easy-first parser.[1] In our main experiments, we used the English feature model *as is* with no modifications. This is not entirely meaningful as the feature model refers to POS tags specific to the English Penn Treebank (PTB) (see Discussion). We used the datasets from the CoNLL-X (Buchholz and Marsi, 2006) and CoNLL 2007 (Nivre et al., 2007) shared tasks with standard train-test splits, but mapped all POS tags into Google's universal tag set (Petrov et al., 2011). See the shared task descriptions for dataset characteristics.

We used a publicly available language genealogy[2] and a publicly available database of typological properties[3] to obtain our weights. See Figure 2 for the linguistic genealogy. In the typological table, we disregarded phonological properties (properties 1–20) when computing Hamming distances between languages. We also report results obtained using voting. These results are obtained using the reparsing technique first described in Sagae and Lavie (2006) with our various weighted parsers as committee members.

6 Results

We note that our macro-average results are the best reported results for a fully delexicalized model, i.e. without lexical projections or bilingual word clusters. That being said recent results show that using cross-language projection or bilingual clustering to obtain lexical knowledge is beneficial, and we will explore different ways of augmenting the models presented here with such knowledge.

7 Discussion

We have tried to prevent over-fitting doing only a single pass over the data. While the averaged perceptron is less prone to over-fitting than the original perceptron learning algorithm (Rosenblatt, 1958), there is no guarantee that it does not over-fit the training data, and in our case where there is a considerable bias in the training data, over-fitting is more likely to happen. Averaging over several source languages provides implicit regularization.

Averaging has several advantages over just concatenating data. First of all we assign equal weights to all source languages (unless taking a weighted average), which means our model is not biased by the size of the linguistic resources used. More importantly, if we want to deliberately under-fit our models, learning with concatenated data is risky, especially if we do not shuffle the data.

Arguably our method to prevent over-fitting is crude, and we would like to explore more

[1] http://www.cs.bgu.ac.il/~yoavg/software/easyfirst/
[2] http://andromeda.rutgers.edu
[3] http://wals.info

source/target	Bu	Ca	Cz	Da	Du	Ge	Gr	It	Po	Sl	Sp	Sw	AV	AV_M	p-value
Bulgarian	68.5	55.7	45.0	44.1	50.4	49.2	47.8	55.7	67.0	35.8	51.7	61.1			
Catalan	59.7	77.8	43.2	45.9	49.4	57.0	59.1	**74.4**	71.3	51.6	**70.2**	52.5			
Czech	29.6	24.8	33.8	28.1	23.6	24.0	26.2	26.0	23.0	23.3	22.7	21.9			
Danish	36.9	30.0	29.5	45.4	25.5	22.9	18.7	37.7	30.5	26.0	29.5	26.6			
Dutch	59.6	59.4	43.4	46.6	71.2	57.6	63.1	59.5	67.4	44.1	53.1	57.4			
German	54.2	51.7	41.7	40.8	43.7	81.2	53.4	55.5	55.0	41.3	49.3	57.9			
Greek	56.5	67.6	45.8	42.3	60.5	52.3	70.6	66.9	67.9	**55.2**	58.0	57.5			
Italian	61.6	**79.5**	41.1	47.0	49.6	55.1	61.9	77.0	71.6	52.2	67.5	51.7			
Portuguese	49.5	59.7	34.5	25.8	49.0	37.9	50.6	59.3	58.1	34.1	50.6	39.7			
Slovene	20.8	16.6	19.5	33.1	19.2	20.4	23.9	18.7	19.3	24.0	17.2	22.8			
Spanish	46.0	74.8	30.5	42.1	41.3	46.6	42.6	64.1	67.2	42.1	72.9	46.9			
Swedish	58.6	58.3	36.5	45.0	50.8	54.2	56.7	57.6	68.3	37.5	55.1	80.6			
flat	62.1	68.9	45.5	46.0	54.0	**58.0**	63.4	67.4	76.5	44.7	64.4	60.7	59.2	61.2	
gtree	62.1	67.0	**46.9**	45.8	54.3	57.7	63.4	69.4	75.5	47.4	66.0	61.6	59.8	61.7	
typology	62.1	68.9	45.5	46.0	54.3	56.3	62.8	67.7	76.5	44.7	64.7	58.4	59.0	60.8	
pr	62.6	69.3	46.1	45.8	53.2	57.4	63.2	68.0	76.0	46.4	65.3	60.1	59.4	61.1	
vote (a)	62.7	69.3	45.8	46.3	54.0	57.7	63.4	68.8	76.4	45.8	65.6	61.2	60.5	62.6	
weighted	64.6	68.5	**46.9**	48.6	57.2	56.3	65.0	67.2	**77.6**	44.6	62.9	62.4	60.1	62.1	~ 0.005
w-typ	64.6	68.5	**46.9**	48.6	58.6	55.7	**66.2**	68.3	**77.6**	44.6	62.8	61.4	60.3	62.4	< 0.001
w-gtree	64.2	69.1	46.7	48.3	57.2	56.3	65.0	68.3	76.9	46.1	63.8	63.1	60.4	62.4	< 0.001
vote (b)	64.2	68.9	46.9	48.6	57.3	57.4	65.3	69.0	77.3	45.7	63.6	62.6	60.6	62.6	< 0.001
MPH11(dir)	-	-	-	48.9	55.8	56.7	60.1	64.1	74.0	-	64.2	65.3		61.2	
MPH11(proj)	-	-	-	**49.5**	**65.7**	56.6	65.1	65.0	75.6	-	64.5	**68.0**		**63.8**	
TMU12(dir)	-	-	-	36.7	52.8	48.9	-	64.6	66.8	-	60.2	55.4		55.1*	
TMU12(clust)	-	-	-	38.7	54.3	50.7	-	68.8	71.0	-	62.9	56.9		57.6*	
NBO12(best)	**66.8**	71.8	44.6	-	55.9	53.7	67.4	65.6	73.5	-	62.1	61.5			

Figure 3: Results, incl. language-level weighting (gtree and typology), unsupervised model selection (pr), instance-weighted extensions (w–) and a comparison with recent work. AV is macro-average, and AV_M is macro-average on the 8 languages used in McDonald et al. (2011). The voted systems (a) and (b) take non-weighted votes over the systems in the above rows.

advanced methods in the future.

Since we average over languages - which in a domain adaptation setting corresponds to distinct source domains - our model is very similar to multi-source domain adaptation models that use mixtures of experts (McClosky et al., 2010; Spinello and Arras, 2012). In future work we would also like to explore the idea of using smoothness assumptions in the target domain to select models based on source languages (Gao et al., 2008). Another option would be to view multi-source language learning as multi-task learning and apply a multi-task perceptron learning algorithm (Cavallanti et al., 2008) rather than just averaged perceptron learning.

On a more technical note, as already mentioned, readers familiar with the easy-first parser may wonder what we did with the POS-tag specific features in the English feature model distributed with the parser. We did not change anything in the feature model specification, keeping features that refer to PTB-specific tags. Changing the PTB-specific tags to their Google tag set translations yielded worse results on average with only small improvements for four languages (Catalan, Italian, Slovene and Spanish). Instead of optimizing the feature model we therefore chose to keep the original feature specification file *as is* for reproducibility.

References

Berg-Kirkpatrick, T. and Klein, D. (2010). Phylogenetic grammar induction. In *ACL*.

Brown, G., Wyatt, J., and Tino, P. (2005). Managing diversity in regression ensembles. *Journal of Machine Learning Research*, 6:1621–1650.

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

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. (2006). Tracking the best hyperplane with a simple budget perceptron. In *COLT*.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. (2008). Linear algorithms for online multitask classification. In *COLT*.

Cohen, S., Das, D., and Smith, N. (2011). Unsupervised structure prediction with non-parallel multilingual guidance. In *EMNLP*.

Collins, M. (2002). Discriminative training methods for hidden markov models. In *EMNLP*.

Crammer, K. and Singer, Y. (2003). Ultraconservative algorithms for multiclass problems. In *JMLR*.

Durrett, G., Pauls, A., and Klein, D. (2012). Syntactic transfer using a bilingual lexicon. In *EMNLP*.

Freund, Y. and Schapire, R. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, 37:277–296.

Gao, J., Fan, W., Jiang, J., and Han, J. (2008). Knowledge transfer via multiple model local structure mapping. In *KDD*.

Gelling, D., Cohn, T., Blunsom, P., and Graca, J. (2012). The pascal challenge on grammar induction. In *WILS-NAACL*.

Goldberg, Y. and Elhadad, M. (2010). An efficient algorithm for easy-first non-directional dependency parsing. In *NAACL*.

Huang, J., Smola, A., Gretton, A., Borgwardt, K., and Schölkopf, B. (2007). Correcting sample bias by unlabeled data. In *NIPS*.

Jiang, J. and Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *ACL*.

McClosky, D., Charniak, E., and Johnson, M. (2010). Automatic domain adaptation for parsing. In *NAACL-HLT*.

McDonald, R., Petrov, S., and Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *EMNLP*.

Naseem, T., Barzilay, R., and Globerson, A. (2012). Selective shargin for multilingual dependency parsing. In *ACL*.

Naseem, T., Chen, H., Barzilay, R., and Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction. In *EMNLP*.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *EMNLP-CoNLL*.

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. CoRR abs/1104.2086.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Sagae, K. and Lavie, A. (2006). Parser combination by reparsing. In *HLT-NAACL*.

Shen, L., Satta, G., and Joshi, A. (2007). Guided learning for bidirectional sequence classification. In *ACL*.

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

Smith, D. and Eisner, J. (2009). Parser adaptation and projection with quasi-synchronous grammar features. In *EMNLP*.

Søgaard, A. (2011). Data point selection for cross-language adaptation of dependency parsers. In *ACL*.

Søgaard, A. (2012). Unsupervised dependency parsing without training. *Natural Language Engineering*, 18(1):187–203.

Søgaard, A. and Haulrich, M. (2011). Sentence-level instance-weighting for graph-based and transition-based dependency parsing. In *IWPT*.

Spinello, L. and Arras, K. (2012). Leveraging RGB-D data: adaptive fusion and domain adaptation for object detection. In *CRA*.

Spreyer, K. and Kuhn, J. (2009). Data-driven dependency parsing of new languages using incomplete and noisy training data. In *CoNLL*.

Täckström, O. (2012). Nudging the envelope of direct transfer methods for multilingual named entity recognition. In *WILS-NAACL*.

Täckström, O., McDonald, R., and Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL*.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *IJCNLP*.