Sentence-Level Instance-Weighting for Graph-Based and Transition-Based Dependency Parsing

Anders Søgaard
Center for Language Technology
University of Copenhagen
soegaard@hum.ku.dk

Martin Haulrich
ISV Computational Linguistics Group
Copenhagen Business School
mwh.isv@cbs.dk

Abstract

Instance-weighting has been shown to be effective in statistical machine translation (Foster et al., 2010), as well as cross-language adaptation of dependency parsers (Søgaard, 2011). This paper presents new methods to do instance-weighting in state-of-the-art dependency parsers. The methods are evaluated on Danish and English data with consistent improvements over unadapted baselines.

1 Introduction

The default assumption in theoretical machine learning is that training and test data are independently and identically (iid) drawn from the same distribution. If the distributions differ, we face what is referred to as sample selection bias in the statistical literature. Sample selection bias is typically ignored in machine learning, but it occurs often in practice.

In natural language processing, the problem shows up in almost any real-world application. Machine translation systems are trained on large amounts of parallel text, but typically this text comes from a small set of sources or institutions, e.g. the Europarl corpora of transcribed debates from the European Parliament (Koehn, 2005). Machine translation systems are used to translate many different kinds of texts, however. In machine translation, which can be seen as a structured learning problem of predicting target sentence y given a source sentence x, we typically see a bias in $P(y)$ and $P(\mathbf{x})$, but not in $P(y|\mathbf{x})$. Statistical parsers for English are typically trained on annotated text from the Wall Street Journal corpus of newspaper articles (Marcus et al., 1993), but are used to process many different kinds of text. Since the problem of sample selection bias in natural language processing is typically related to differences in textual domains, computational linguists typically refer to the problem as domain adaptation.

Domain adaptation is one of the most fundamental yet-to-be-solved problems in natural language processing. While statistical parsers have accuracies of 90-92% parsing newspaper articles, accuracy on transcribed telephone conversations or child-directed speech often drop to 60-70% (Nivre et al., 2007a). Domain adaptation is therefore also receiving more and more attention, and it has recently been studied in the context of named entity recognition (Daume III, 2007), sentiment analysis (Blitzer et al., 2007), dependency parsing (Sagae and Tsujii, 2007; Kawahara and Uchimoto, 2009), text classification (Chen et al., 2009), context-free parsing (McClosky et al., 2010) and machine translation (Foster et al., 2010).

Domain adaptation is the problem of learning a target distribution from a labeled sample of source data with a similar, but different distribution. The problem comes in two variants; one where we also have a small amount of labeled target domain data, and one in which we only have labeled source domain data and must rely on unlabeled source and target domain data to do the actual adaptation of the model that can be learned from source domain data. Much work in natural language processing has assumed a small amount of labeled target domain data (Daume III, 2007; Foster et al., 2010), but we consider the more difficult case where none is available. This is sometimes referred to as unsupervised domain adaptation.

How domain adaptation is tackled depends much on the assumptions we may have about the similarities and differences between the two distributions. One line of approaches to domain adaptation is to change the feature representation of the source domain data, typically focusing on the features that are also predictive in the target domain (Ben-David et al., 2007). Such approaches assume a bias in $P(\mathbf{x})$, but may also try to deal with sce-

Proceedings of the 12th International Conference on Parsing Technologies, pages 43–47,
October 5-7, 2011, Dublin City University. © 2011 Association for Computational Linguistics

narios where there is a bias in $P(y|\mathbf{x})$. Others have proposed using priors to encode knowledge about one domain in a model induced from data in another domain, or they have promoted frequent target domain classes if they were less frequent in the source domain. Such approaches assume a bias in $P(y)$ and have become popular in word sense disambiguation (Zhu and Hovy, 2007), for example, where a particular reading of *bank* may be much more frequent in some domains rather than others. Classes can be promoted using instance weighting, but instance weighting can also be used to change the marginal distribution of data. The first case is typically referred to as solving class imbalance, while the second case is called covariate shift (Shimodaira, 2000). We will, assuming a bias in $P(\mathbf{x})$, consider the covariate shift scenario. A fourth line of research in domain adaptation applies semi-supervised or transductive learning algorithms to domain adaptation problems, using unlabeled data from the target domain.

In dependency parsing, domain adaptation received attention in the CoNLL 2007 Shared Task. While semi-supervised learning and structural correspondence learning were used by participants in the CoNLL 2007 Shared Task, none of the participants used instance-weighting techniques. In this paper, we follow suggestions in the related literature on learning under sample selection bias to transform the density ratio estimation problem in co-variate shift into a problem of predicting whether an instance is from the source domain or from the target domain (Zadrozny, 2004; Bickel and Scheffer, 2007). We show how to do this in the context of graph-based and transition-based dependency parsing.

Related work includes Søgaard (2011) who uses perplexity per word to select the source data most similar to the target data, so a form of instance weighting with weights 0 and 1, but applies the technique to cross-language adaptation of dependency parsers; but also Plank and van Noord (2011) who in a similar fashion use topic similarity measures to select articles rather than sentences.

Our instance-weighted parsers are evaluated primarily on a new data set, namely a partitioning of the Danish treebank (Buch-Kromann, 2003) into four different textual domains. We do experiments with all pair-wise combinations of the four domain-specific treebanks. Our results are supplemented by a subset of the CoNLL 2007 Shared Task data. It has been noted in several places that there were annotation differences between the source and target data in the original data which makes domain adaptation almost impossible (Dredze et al., 2007). Consequently, we only use the three small target domain evaluation datasets, which were annotated more consistently, and do experiments with all pair-wise combinations of these datasets. Our experiments can also be seen as transductive learning experiments, since no target data other than the data used for evaluation is used.

2 Sentence-Based Instance-Weighting in Dependency Parsing

2.1 Using Text Classification for Instance-Weighting

The source and target plain text corpora are first extracted, and each sentence is assigned a label saying whether the sentence was sampled from source or target data. The idea is then to train a text classifier on the data and use the probability that a sentence comes from the target domain to weight the source instances. This is also the approach to learning under sample selection bias suggested by Zadrozny (2004).

Our text classifier was a logistic regression classifier implemented in Mallet. It represents each sentence by a vector representing occurrences of n-grams in the sentence ($n \leq 3$). No stop word lists were used. The text classifier was used to approximate the probability that each source sentence was sampled from the target domain. The weights are obtained using ten-fold cross-validation. We store one weight for each sentence in the labeled source data.

2.2 Graph-Based Dependency Parsing

Graph-based dependency parsing is a heterogeneous family of approaches to the dependency parsing algorithms, each of which couples a learning algorithm and a parsing algorithm. Some of these algorithms assume dependency trees are projective (Eisner, 1996), while others allow for non-projective dependency trees (McDonald et al., 2005).

One approach to graph-based parsing of non-projective dependency trees is applying minimum spanning tree algorithms to matrices of weighted head-dependent candidate pairs. The learning al-

	Malt-bl	**Malt-sys**	**MST-bl**	**MST-sys**
law-lit	63.55	**64.22**	62.57	**65.31**
law-magz	60.8	**61.34**	**58.65**	58.59
law-news	60.23	**60.58**	58.84	**62.07**
lit-laws	78.34	**79.31**	77.58	**78.06**
lit-magz	**80.22**	80.04	**80.61**	80.55
lit-news	77.31	**77.6**	79.79	**80.14**
magz-law	72.04	**73.98**	73.84	**74.74**
magz-lit	75.74	**76.63**	77.27	**77.78**
magz-news	**73.73**	73.42	**74.42**	73.91
news-law	77.85	**79.65**	80.69	**82.7**
news-lit	85.33	**85.49**	**88.25**	88.22
news-magz	84.93	**85.65**	87.81	87.81
AV	74.17	**74.86**	75.02	**75.82**

Table 1: Unlabeled attachment scores for Danish.

gorithm used in McDonald et al. (2005) and the publicly available MSTParser[1] to learn candidate weights is MIRA (Crammer and Singer, 2003). The MIRA algorithm considers one sentence at each update of the weight vector, and the successive values of the vector are accumulated to later produce an averaged weight vector in a way similar to using averaged perceptron. Unlike using averaged perceptron, MIRA aggressively maximizes the margin between the correct dependency structure and the parser's prediction enforcing it to be larger than the loss of that prediction.

In our experiments we weight the margin such that a large margin between the correct and predicted structures is less aggressively enforced when learning from distant data points. This is achieved by weighting the loss of incorrect classifications by the probability that the sentence was sampled from the target domain.

2.3 Transition-Based Dependency Parsing

Transition-based parsing reduces the problem of finding the most likely dependency tree for a sentence to a series of classification problems by seeing parsing as transitions between configurations. Parsing is incremental and left-to-right. A configuration typically consists of the next couple of words to be read, the first couple words on a stack storing previously read words, and part of the dependency structure already build. Each configuration is a feature vector used to predict the parser's next transition. The guiding classifier is trained on canonical derivations of the dependency trees in the labeled training data.

The most widely used transition-based dependency parser is the MaltParser (Nivre et al., 2007b).[2] The parser comes with several parsing algorithms, but uses a projective and very efficient algorithm by default. MaltParser is bundled with LibSVM 2.91, implementing a wide range of support vector machine algorithms that are used to learn classifiers to guide parsing. LibSVM 2.91 does not allow for instance weighting. However, LibSVM 3.0 does. In our experiments with the MaltParser, we use LibSVM 3.0 in conjunction with the MaltParser providing it with sentence-level instance weights from our Mallet text classifier. This means that configuration-transition pairs in the canonical derivations of a sentence with weight w will have weight w when training the support vector machine used by our parser.

3 Data

We evaluate our instance-weighted parsers on two domain adaptation data sets from English and Danish annotated corpora, one of which (Danish) has not previously been used in the literature. The Danish corpus is a balanced corpus, annotated building the Danish Dependency Treebank (DD (Buch-Kromann, 2003) and used in the CoNLL-X Shared Task (Buchholz and Marsi, 2006). The DDT comes with metadata revealing the original source of each sentence. This metadata was used to split the DDT into four domains: law (77 sent.), literature (lit; 984 sent.), magazines (magz; 190 sent.) and newspapers (news; 5052 sent.).

The second dataset was also used for the CoNLL 2007 Shared Task on domain adaptation

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

[2] http://maltparser.org

	Malt-bl	**Malt-sys**	**MST-bl**	**MST-sys**
childes-pbiotb	43.11	**43.91**	46.03	**48.86**
childes-pchemtb	38.01	**39.69**	**44.89**	44.41
pbiotb-childes	**50.35**	49.91	59.07	**61.37**
pbiotb-pchemtb	**75.64**	75.26	**77.26**	77.16
pchemtb-childes	49.63	**50.69**	60.89	**60.91**
pchemtb-pbiotb	**75.28**	75.06	76.39	**76.73**
AV	55.34	**55.75**	60.76	**61.57**

Table 2: Unlabeled attachment scores for English.

for dependency parsers (Nivre et al., 2007a). In the shared task, the Penn-III treebank (Marcus et al., 1993) was used as source domain, and test domains were chemical and biomedical research articles and transcribed child-directed speech. The quality of the shared task data was questioned by participants (Dredze et al., 2007), and there is some consensus today that annotation styles were too different for evaluation results to be useful. We therefore use data from the target domains only: biomedical literature (160 sent.), chemical literature (195 sent.) and child-directed speech (666 sent.). We consider all pairwise combinations of datasets within the two languages.

4 Results

Our results for both Danish and English (see Table 1 and 2), reporting unlabeled attachment scores including punctuation, show rather consistent improvements across all pairwise combinations of datasets. Error reductions vary greatly from dataset to dataset, however. The average error reduction on the Danish data is $\geq 3\%$ for the instance-weighted MSTParser, and $\geq 2.5\%$ for the instance-weighted MaltParser.

It is interesting to note that there were no significant improvements when English input data was weighted by a text classifier trained on biomedical and chemical literature. These two text types are of course more similar to each other than to child-directed speech. This is reflected in the text classification accuracy, which is as high as 98–99% when comparing sentences sampled from technical literature and sentences sampled from child-directed speech, but considerably lower ($\sim$93.5%) when trying to differentiate biomedical sentences from chemical ones. Table 1 plots the correlation between system improvements and text classification accuracy for the Danish data. It is easy to see that high text classification accuracy is necessary for substantial improvements over the non-weighted baseline system.

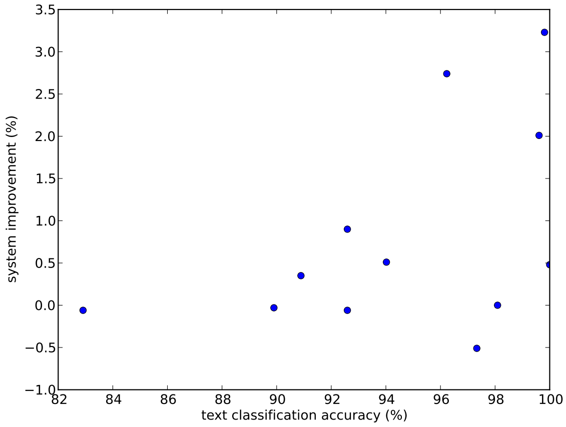

Figure 1: Correlation between text classification accuracy and system improvement (Danish).

Finally we note that we did similar experiments on the Penn-III treebank using the metadata also used by Webber (2009), with less robust results and smaller average improvements. The distribution of text types is very skewed in the Wall Street Journal, however, making text classification on this data alone a difficult job.

5 Conclusion

We have presented ways of implementing instance-weighting in transition-based and graph-based dependency parsing based on text classification and showed that this leads to consistent improvements over non-adapted baselines in domain adaptation scenarios, especially across very different domains.

References

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19:137–144.

Steffen Bickel and Tobias Scheffer. 2007. Dirichlet-enhanced spam filtering based on biased samples. *Advances in Neural Information Processing Systems*, 19:161–168.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*.

Matthias Buch-Kromann. 2003. The Danish Dependency Treebank and the DTAG Treebank Tool. In *TLT*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *CoNLL*.

Bo Chen, Wai Lam, Ivor Tsang, and Tak-Lam Wong. 2009. Extracting discriminative concepts for domain adaptation in text mining. In *KDD*.

Koby Crammer and Yoram Singer. 2003. Ultraconservative algorithms for multiclass problems. In *JMLR*.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *ACL*.

Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graca, and Fernando Pereira. 2007. Frustratingly hard domain adaptation for dependency parsing. In *EMNLP-CoNLL*.

Jason Eisner. 1996. Three new probabilistic models for dependency parsing. In *COLING*.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *EMNLP*.

Daisuke Kawahara and Kiyotaka Uchimoto. 2009. Learning reliability of parses for domain adaptation of dependency parsing. In *IJCNLP*.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *MT-Summit*.

Mitchell Marcus, Mary Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *NAACL-HLT*.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT-EMNLP*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007a. The CoNLL 2007 shared task on dependency parsing. In *EMNLP-CoNLL*.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007b. MaltParser: a language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *ACL*.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with lr models and parser ensembles. In *EMNLP-CoNLL*.

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

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *ACL*.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn Treebank. In *ACL-IJCNLP*.

Bianca Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. In *ICML*.

Jingbo Zhu and Eduard Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *EMNLP-CoNLL*.