COMBINING DISCRIMINATIVE RE-RANKING AND CO-TRAINING FOR PARSING MANDARIN SPEECH TRANSCRIPTS

Wen Wang

Speech Technology and Research Laboratory, SRI International, USA

wwang@speech.sri.com

ABSTRACT

Discriminative reranking has been able to significantly improve parsing performance, and co-training has proven to be an effective weakly supervised learning algorithm to bootstrap parsers from a small in-domain seed labeled corpus using a large amount of unlabeled in-domain data. In this paper, we present systematic investigations on combining discriminative reranking and co-training, including co-training reranked parsers and co-training rerankers. We show that combining discriminative reranking and co-training could improve the F-measure by 1.8%-2% absolute compared to co-training two state-of-the-art Chinese parsers without reranking, for parsing Mandarin broadcast news and conversation transcripts.

Index Terms— parsing, Mandarin speech, discriminative re-ranking, co-training, conversational speech

1. INTRODUCTION

Parsing aims at resolving structural ambiguity. State-of-the-art statistical parsers require treebanks to estimate their parameters, but their performance degrades when there is mismatch on genres/domains between the training treebank and the data to parse. Furthermore, creating high-quality in-genre/in-domain treebank for the data to parse is expensive and difficult. However, under the Defense Advanced Research Projects Agency (DARPA) Global Autonomous Language Exploitation (GALE) program[1], there are new genres besides newswire text, namely, broadcast news (BN), broadcast conversation (BC), newsgroup (NG), and web log (WB). Generating high-quality parse trees for Chinese data in these genres can be useful for various tasks within GALE, including syntax-guided translation and reordering models for Chinese-to-English machine translation (MT), name entity detection, and structured language modeling for automatic speech recognition (ASR) on Mandarin BN and BC audio. In our previous research [1], we employed the weakly supervised co-training technique on two state-of-the-art parsers, Charniak's parser and the Berkeley parser, to bootstrap them from a newswire Chinese treebank and a small amount of BN and BC seed annotated treebank with a large amount of unlabeled BN and BC transcripts, in order to achieve high parsing accuracy on Mandarin BN and BC transcripts. By employing co-training, we obtained 2.2% to 2.6% absolute improvement on F-measure [2] for parsing BN and BC transcripts. On the other hand, discriminative reranking for parsers [2, 3] has produced significant improvement on parsing accuracy. In this paper, we explore the effectiveness of combining discriminative reranking and co-training to further improve parsing performance on Mandarin BN and BC transcripts. Section 2 describes discriminative reranking. Section 3 describes the co-training algorithm. Section 4 proposes two approaches to combine discriminative reranking and co-training. Section 5 describes the available treebanks, the small seed annotated corpora, and the large unlabeled corpora for co-training. Experimental results, discussions, and conclusions appear in Section 6.

[1]The goal of the GALE program is to develop computer software techniques to analyze, interpret, and distill information from speech and text in multiple languages.

2. DISCRIMINATIVE RE-RANKING

We first describe our use of the RankBoost-based discriminative reranking approach that was originally developed by Collins and Koo [2] for parsing. This approach allows us to investigate the impact of various features on Mandarin parsing performance. The reranking algorithm takes as input a list of candidates produced by a Chinese parser and reranks these candidates based on a set of features. For training the reranker for the parsing task, there are n sentences $\{s_i : i = 1, \cdots, n\}$, each with n_i candidates $\{x_{i,j} : j = 1, \cdots, n_i\}$ along with the log-probability $L(x_{i,j})$ produced by the parser. Each parsing candidate $x_{i,j}$ in the training data has a score $Score(x_{i,j})$ that measures the similarity between the candidate and the gold reference. For parsing, we use parse accuracy as the similarity measure. Without loss of generality, we assume that $x_{i,1}$ has the highest score, i.e., $Score(x_{i,1}) \geq Score(x_{i,j})$ for $j = 2, \cdots, n_i$. A set of indicator functions $\{h_k : k = 1, \cdots, m\}$ is used to extract binary features $\{h_k(x_{i,j}) : k = 1, \cdots, m\}$ on each example $x_{i,j}$. Each indicator function h_k is associated with

[2]F-measure is based on labeled Precision (LP) and labeled Recall (LR). LP is the number of correct constituents divided by the number of constituents found by the parser, and LR is the number of correct constituents divided by the number of constituents in the gold parse. F-measure is defined as $F_1 = \frac{2PR}{P+R}$.

a weight parameter α_k that is real valued. In addition, a weight parameter α_0 is associated with the log-probability $L(x_{i,j})$. The ranking function of candidates $x_{i,j}$ is defined as $\alpha_0 L(x_{i,j}) + \sum_{k=1}^{m} \alpha_k h_k(x_{i,j})$. The objective of the training process is to set the parameters $\bar{\alpha} = \{\alpha_0, \alpha_1, \cdots, \alpha_m\}$ to minimize the loss function $Loss(\bar{\alpha})$ (which is an upper bound on the training error) as $Loss(\bar{\alpha}) = \sum_i \sum_{j=2}^{n_i} S_{i,j} e^{-M_{i,j}(\bar{\alpha})}$, where $S_{i,j}$ is the weight function that gives the importance of each example, and $M_{i,j}(\bar{\alpha})$ is the margin [2]. All the α_i's are initially set to zero. Then a greedy sequential optimization method is used in each boosting round to select the feature that has the most impact on reducing the loss function and then update its weight parameter accordingly.

Collins' method allows multiple updates to the weight of a feature. Huang et al. [4] found that for those strong features, Collins' weight update formula can increase their weight (in absolute value) in only one direction. Although these features are strong and useful, setting weights too large can be harmful in that it limits the use of other features for reducing the loss. Based on this analysis, Huang et al. [4] have developed an *update-once* method, in which the weight update is limited so that once a feature is selected in a certain iteration and its weight parameter is updated, no update will be conducted on it again. In this way, the weights of the strong features will not be allowed to prevent other features from being considered during the training procedure. Huang et al. observed that the *update-once* method could select significantly more features compared to Collins' original method and produce better reranking performance. In this paper, we employed this *update-once* strategy for updating feature weights.

For the work described in this paper, we employed the features described in [2]. Note that before generating these features, we applied headword percolation on the trees output by parsers [2]. Features include **rules** (all context-free rules in the tree), **bigrams** (adjacent pairs of non-terminals to the left and right of the head), **grandparent rules** (same as rules, but also including the non-terminal above the rule), **head-modifiers** (all head-modifier pairs, also including the grandparent non-terminal), and **PPs** (lexical trigrams involving the heads of arguments of prepositional phrases) etc. More details appear in [2].

3. CO-TRAINING

Co-training was first introduced by Blum and Mitchell [5] as a weakly supervised learning method and can be used for bootstrapping a model from a seed corpus of labeled examples, which is typically quite small, augmented with a much larger amount of unlabeled examples, by exploiting redundancy among multiple statistical models that generate different *views* of the data. Informally, co-training can be described as picking multiple classifiers ("views") of a classification problem, building models for each view and training these models on a small set of labeled data, then on a large set of unlabeled data, sampling a subset, labeling them using the models, selecting examples from the labeled results, adding them to the training pool, and iterating this procedure until the unlabeled set is all labeled.

In [1], we systematically investigated applying weakly supervised co-training approaches to improve parsing performance for parsing Mandarin BN and BC transcripts, by iteratively retraining two competitive Chinese parsers, Charniak's reranking parser [3] and the Berkeley parser [6], from a small set of treebanked data and a large set of unlabeled data. Compared to parsers trained only on the small in-domain seed labeled corpus, the parsers resulting from co-training could gain 6.8% absolute on BN and 7.3% absolute on BC based on the F-measure. Overall, compared to parsers trained on all available treebank data including in-domain and out-of-domain treebanks, co-training yields a 2.2% – 2.6% absolute gain on BN and 2.4% – 2.5% absolute gain on BC based on the F-measure (and 1.5% – 1.9% absolute gain on BN and 1.7% – 2.0% absolute gain on BC over self-training [1]). In this paper, we investigate the combination of discriminative reranking and co-training on Charniak's maximum-entropy inspired parser [7] (i.e., the parser without reranking compared to the parser in [3]) and the Berkeley parser (also originally without reranking). For co-training parsers, we employed the *max-t-min-s* example selection approach developed in [1], as it is computationally inexpensive and also produced the best performance.

4. COMBINING RERANKING AND CO-TRAINING

Both Charniak and Berkeley parsers support generating N-best parses for reranking purposes. In fact, Charniak and Johnson have implemented a discriminative reranker using a MaxEnt estimator to find the feature weights and when using the reranker to rerank 50-best parses from Charniak's ME inspired parser, it improved F-measure by 1.3% absolute on sentences of length less than 100 words in Wall Street Journal Penn Treebank section 23 [3]. In this work, we adopted this reranker for Charniak's parser, implemented the RankBoost-based reranking algorithm described in Section 2 to rerank 50-best from the Berkeley parser, and then investigated two ways to combine reranking and co-training. The direct combination approach is for each iteration of co-training, instead of generating 1-best parse directly from the no-reranking, standard Charniak and Berkeley parsers, 50-best parses are generated from each parser and then reordered by their corresponding rerankers, respectively. Then the 1-best parses after reranking for the unlabeled data are selected and added to the training pool of the parsers. In this paper this approach is denoted **co-training reranked parsers**. Note that for this approach the features and feature weight parameters for rerankers remain the same during the co-training procedure.

Different from the original binary classification problems on which co-training was developed, parsing contains a number of smaller decisions about which constituents are probable, and inherently each parser includes good and bad decisions on how to create/attach different constituents. On

the other hand, reranking is closer to binary classification than parsing, as it tries to decide whether or not a parse hypothesis is the best parse for the sentence, so it is explicit to maximize agreement between rerankers, as the principled agreement-based example selection approach could be applied here, which could guarantee co-training to improve parsing accuracy. Hence, we hypothesize that co-training rerankers could better fit the co-training algorithm. For effectiveness, rerankers can consider features that span the entire tree of a parse (while parsers generally consider only local features). For efficiency, co-training rerankers requests unlabeled data to be parsed just once, compared to multiple parsing iterations for co-training reranked parsers. The output will be reranked many times but this is much more efficient than training and running parsers. Hence, in this work, we also investigated co-training our RankBoost-based reranker with Charniak's and Johnson's MaxEnt reranker and applied the co-trained rerankers to the two standard parsers. This second approach is denoted **co-training rerankers**.

5. DATA

For selecting parsers and also contributing to training parsers, we used Chinese Treebank 5.2 released by LDC (denoted as CTB). Chinese Treebank 5.2 contains 500K words, 800K characters, 18K sentences, and 900 data files. Under the GALE program, the BN genre follows its tradition and consists of "talking head" style broadcasts, i.e., generally one person reading a news script. The BC genre, by contrast, is more conversational and spontaneous, consisting of talk shows, interviews, call-in programs, and roundtables. The evaluation of co-training for parsing Mandarin BN and BC transcripts is conducted on the GALE OntoNotes released Mandarin BN and BC treebanks. The BN treebank is from the Mandarin TDT4 collection, and the BC treebank is from GALE Mandarin BC data and translations from English BC data. The Mandarin BN treebank includes 300K words and 814 data files, and the BC treebank 100K words and 16 data files. To create a seed corpus and a test set for evaluating parsing accuracy, for BN and BC respectively, we divided the whole BN/BC treebank into blocks of 10 files by sorted order. Within each block, the first file is used for co-training development, the second for testing parsing accuracy, and the remaining 8 files are used as part of the seed annotated corpus for co-training. The resulting BN test set is denoted **BN-test** and the seed annotated corpus **BN-seed**. The BC test set is denoted **BC-test** and the BC seed annotated corpus **BC-seed**. BN-test includes 31K words and 1,565 sentences. BC-test includes 11K words and 1,482 sentences. The large set of unlabeled data for BN parsing includes Hub4 1997 Mandarin BN acoustic transcripts, LDC Chinese TDT{2,3,4} corpora, Chinese Gigaword 3.0, and all GALE released BN audio transcripts, denoted **BN-unlabeled**. For BC parsing, we add all GALE released BC audio transcripts denoted **BC-unlabeled**. After word segmentation, **BN-unlabeled** comprises around 1.4 billion words while **BC-unlabeled** around 11 million words.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

Table 1 shows the parsing accuracy F-measure (%) on BN-test under various parser training conditions on Charniak's parser and the Berkeley parser without reranking. As can be seen from the table, training Charniak's parser and the Berkeley parser using only the small training set of BN treebank, i.e., BN-seed, resulted in relatively poor parsing performance, at 75.1% F1 for Charniak's parser and 75.2% for the Berkeley parser. Using the larger full CTB corpus for training improves parsing performance significantly and adding BN-seed to CTB brought additional gain. However, co-training using CTB plus BN-seed as the initial training pool significantly improved the performance of the two parsers over directly training on CTB plus BN-seed, with 1.9% absolute and 2.1% absolute improvement on F-measure for Charniak's parser and the Berkeley parser, respectively. For co-training carried out in these experiments, we used cache size as 10K sentences. Table 2 shows the F-measure from the two no-reranking parsers on BC-test under various training conditions. The condition **BN-co-trained** denotes the BN-seed treebank and the final annotated BN-unlabeled data after applying *m*ax-t-min-s co-training to the two parsers initialized on the BN-seed treebank. **BN-co-trained** significantly outperforms **CTB**, indicating greater similarity between the two speech genres compared to CTB vs. BC. Using CTB and BN-seed to initialize the two parsers and then co-training on the BN-unlabeled data achieved further gain on parsing performance, denoted by **(CTB+BN)-co-trained**. Consistent with Table 1, it is always helpful to add the small in-genre seed treebank into training, as **(CTB+BN)-co-trained+BC-seed** outperforms **(CTB+BN)-co-trained**. Co-training on BC-unlabeled also produced consistent improvement on F-measures. Overall, we gained 2.5% absolute on F-measure on BC-test over the two parsers from co-training. Using the same BC-unlabeled data for co-training, we also compared initializing the two parsers with the condition of CTB only and the condition of adding the small BN-seed and BC-seed corpora, and observed that adding this small in-genre seed corpus always outperforms initializing with CTB only, by 1% on BN and 1.4% on BC.

The results from the two approaches of combining discriminative reranking and co-training, as proposed in Section 4, are shown in Tables 3 and 4. The results of **co-training standard parsers** are the last rows in Tables 1 and 2. When co-training reranked parsers, the rerankers were trained on CTB+BN-seed for BN and CTB+BN-seed+BC-seed for BC and remained the same during co-training. When co-training rerankers, the rerankers were initialized on CTB+BN-seed for BN and CTB+BN-seed+BC-seed for BC and updated during co-training. For both combination approaches, co-training explored BN-unlabeled for BN and BC-unlabeled for BC as unlabeled data, respectively. As can be seen, **co-training reranked parsers** (using the *max-t-min-s* exam-

Table 1. Overall parsing accuracy F-measure (%) on the Mandarin BN treebank test set, BN-test, after applying co-training using Charniak's maximum-entropy inspired parser and the Berkeley parser, both without reranking.

Training Condition	F-measure (%)	
	Charniak	Berkeley
1. BN-seed	75.1	75.2
2. CTB	79.1	79.1
3. CTB+BN-seed	80.4	80.5
4. co-training initialized as Condition 3, max-t-min-s	**82.3**	**82.6**

Table 2. Overall parsing accuracy F-measure (%) on the Mandarin BC treebank test set, BC-test, after applying co-training using Charniak's maximum-entropy inspired parser and the Berkeley parser, both without reranking.

Training Condition	F-measure (%)	
	Charniak	Berkeley
1. BC-seed	72.0	72.8
2. CTB	73.4	73.7
3. BN-co-trained	74.7	74.8
4. (CTB+BN)-co-trained	75.6	75.7
5. (CTB+BN)-co-trained+BC-seed	76.8	77.0
6. co-training initialized as Condition 5, max-t-min-s	**79.3**	**79.5**

ple selection approach) significantly outperforms co-training without reranking, by 1.5% absolute and 1.4% absolute gain on F-measure on the two parsers on BN-test, and 1.7% absolute and 1.6% absolute gain on F-measure on the two parsers on BC-test. For **co-training rerankers**, as discussed in Section 4, it is feasible now for us to employ the more principled agreement-based example selection approach during co-training since we can simply train each reranker multiple times on different subsets of the automatically labeled data and examine which partition of the data produced the maximum agreement among the rerankers. As a reminder, for **co-training reranked parsers**, we still used the *max-t-min-s* approach as it is computationally efficient and also proved to be very effective for co-training parsers [1]. As can be seen from the tables, **co-training rerankers** produced a small yet consistent gain over **co-training reranked parsers**, by 0.2% – 0.4% absolute improvement on BN-test and 0.4% – 0.5% absolute improvement on BC-test, raising the absolute improvement on F-measure up to 1.8% on BN-test and 2% on BC-test, from combining discriminative reranking and co-training compared to co-training only.

In conclusion, we have demonstrated that discriminative rerankers and co-trained models can work well across genres/domains. We investigated co-training reranked parsers and co-training rerankers and observed that co-training rerankers outperforms co-training reranked parsers and the former is also computationally more efficient. These results are quite encouraging. In future work, we will investigate other approaches for combining discriminative reranking and co-training and algorithms for parser adaptation across genres/domains.

Table 3. Overall parsing accuracy F-measure (%) on BN-test, after applying co-training using Charniak's maximum-entropy inspired parser and the Berkeley parser, both without reranking and with reranking.

Training Condition	BN-test F-measure (%)	
	Charniak	Berkeley
co-training standard parsers	82.3	82.6
co-training reranked parsers	83.8	84.0
co-training rerankers	**84.0**	**84.4**

Table 4. Overall parsing accuracy F-measure (%) on BC-test, after applying co-training using Charniak's maximum-entropy inspired parser and the Berkeley parser, both without reranking and with reranking.

Training Condition	BC-test F-measure (%)	
	Charniak	Berkeley
co-training standard parsers	79.3	79.5
co-training reranked parsers	81.0	81.1
co-training rerankers	**81.5**	**81.5**

7. ACKNOWLEDGMENTS

The authors thank Mary Harper and Zhongqiang Huang for discussions on Chinese parsing and discriminative reranking. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 (approved for public release, distribution unlimited). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

8. REFERENCES

[1] W. Wang, "Weakly supervised training for parsing Mandarin broadcast transcripts", *in Proceedings of Interspeech*, Brisbane, Australia, September 2008.

[2] M. Collins and T. Koo, "Discriminative reranking for natural language parsing", *Computational Linguistics*, vol. 31, pp. 25–70, 2005.

[3] E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking", *in Proceedings of the 43rd ACL*, 2005.

[4] Z. Huang, M. Harper, and W. Wang, "Mandarin part-of-speech tagging and discriminative reranking", *in Proceedings of EMNLP*, 2007.

[5] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training", *in Proceedings of COLT*, 1998.

[6] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, "Learning accurate, compact, and interpretable tree annotation", *in Proceedings of the 44th ACL*, pp. 433–440, Sydney, Australia, July 2006.

[7] E. Charniak, "A Maximum-Entropy-Inspired Parser", *in Proceedings of NAACL*, 2000.