Adapting the RASP System for the CoNLL07 Domain-Adaptation Task

Rebecca Watson and Ted Briscoe
Computer Laboratory
University of Cambridge
FirstName.LastName@cl.cam.ac.uk

Abstract

We describe our submission to the domain adaptation track of the CoNLL07 shared task in the open class for systems using external resources. Our main finding was that it was very difficult to map from the annotation scheme used to prepare training and development data to one that could be used to effectively train and adapt the RASP system unlexicalized parse ranking model. Nevertheless, we were able to demonstrate a significant improvement in performance utilizing bootstrapping over the PBIOTB data.

1 Introduction

The CoNLL07 domain adaptation task was created to explore how a parser trained in one domain might be adapted to a new one. The training data were drawn from the PTB (Marcus *et al.*, 1993) reannotated with dependency relations (Johansson and Nugues, 2007, hereafter DRs). The test data were taken from a corpus of biomedical articles (Kulick *et al.*, 2004) and the CHILDES database (Brown, 1973; MacWhinney, 2000) also reannotated with DRs (see Nivre *et al.*, 2007) for further details of the task, annotation format, and evaluation scheme. The development data consisted of a small amount of annotated and unannotated biomedical and conversational data.

The RASP system (Briscoe *et al.*, 2006) utilizes a manually-developed grammar and outputs grammatical bilexical dependency relations (see Briscoe, 2006 for a detailed description, hereafter GRs). Watson *et al.* (2007) describe a semi-supervised, bootstrapping approach to training the parser which utilizes unlabelled partially-bracketed input with respect to the system derivations. For the domain adaptation task we retrained RASP by mapping our GR scheme to the DR scheme and annotation format, and used this mapping to select a derivation to train the unlexicalized parse ranking model from the annotated PTB training data. We also performed similar partially-supervised bootstrapping over the 200 annotated biomedical sentences in the development data. We then tried unsupervised bootstrapping from the unannotated development data based on these initial models.

As the parser requires input to consist of a sequence of one of 150 CLAWS PoS tags, we also utilize a first-order HMM PoS tagger which has been trained on manually-annotated data from the LOB, BNC and Susanne Corpora (see Briscoe, 2006 for further details). Accordingly, we submitted our results in the open class.

2 Training and Adaptation

The RASP parser is a generalized LR parser which builds a non-deterministic generalized LALR(1) parse table from the grammar (Tomita, 1987). A context-free 'backbone' is automatically derived from a unification grammar. The residue of features not incorporated into the backbone are unified on each reduce action and if unification fails the associated derivation paths also fail. The parser creates a packed parse forest represented as a graph-structured stack.

Inui *et al.* (1997) describe the probability model

Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, pp. 1170–1174,
Prague, June 2007.

utilized in the system where a transition is represented by the probability of moving from one stack state, σ_{i-1}, (an instance of the graph structured stack) to another, σ_i. They estimate this probability using the stack-top state s_{i-1}, next input symbol l_i and next action a_i. This probability is conditioned on the type of state s_{i-1}. S_s and S_r are mutually exclusive sets of states which represent those states reached after shift or reduce actions, respectively. The probability of an action is estimated as:

$$P(l_i, a_i, \sigma_i | \sigma_{i-1}) \approx \left\{ \begin{array}{ll} P(l_i, a_i | s_{i-1}) & s_{i-1} \in S_s \\ P(a_i | s_{i-1}, l_i) & s_{i-1} \in S_r \end{array} \right\}$$

Therefore, normalization is performed over all lookaheads for a state or over each lookahead for the state depending on whether the state is a member of S_s or S_r, respectively. In addition, Laplace estimation can be used to ensure that all actions in the table are assigned a non-zero probability.

These probabilities are estimated from counts of actions which yield derivations compatible with training data. We use a confidence-based self-training approach to select derivations compatible with the annotation of the training and development data to train the model. In Watson *et al.* (2007), we utilized unlabelled partially-bracketed training data as the starting point for this semi-supervised training process. Here we start from the DR-annotated training data, map it to GRs, and then find the one or more derivations in our grammar which yield GR output consistent with the GRs recovered from the DR scheme. Following Watson *et al.* (2007), we utilize the subset of sentences in the training data for which there is a single derivation consistent with this mapping to build an initial trained parse ranking model. Then we use this model to rank the derivations consistent with the mapping in the portion of the training data which remains ambiguous given the mapping. We then train a new model based on counts from these consistent derivations which are weighted in some manner by our confidence in them, given both the degree of remaining ambiguity and also the ranking and/or derivation probabilities provided by the initial model.

Thus, the first and hardest step was to map the DR scheme to our GR scheme. Issues concerning this mapping are discussed in section 4. Given this mapping, we determined the subset of sentences in the (PTB) training data for which there was a single derivation in the grammar compatible with the set of mapped GRs. These derivations were used to create the initial trained model (B) from the uniform model (A). To evaluate the performance of these and subsequent models, we tested them using our own GR-based evaluation scheme over 560 sentences from our reannotated version of DepBank, a subset of section 23 of the WSJ PTB (see Briscoe & Carroll, 2006). Table 1 gives the unlabelled precision, recall and microaveraged F_1 score of these models over this data. Model B was used to rerank derivations compatible with the mapped GRs recovered for the PTB training data. Model C was built from the weighted counts of actions in the initial set of unambiguous data and from the highest-ranked derivations over the training data (i.e. we do not include duplicate counts from the unambiguous data). Counts were weighted with scores ranging $[0 - 1]$ corresponding to the overall probability of the relevant derivation. The evaluation shows a steady increase in performance for these successive models. We also explored other variants of this bootstrapping approach involving use of weighted counts from the top *n* ranked parses derived from the initial model (see Watson *et al.*, 2007, for details), but none performed better than simply selecting the highest-ranked derivation.

To adapt the trained parser, we used the same technique for the 200 in-domain biomedical sentences (PBIOTB), using Model C to find the highest-ranked parse compatible with the annotation, and derived Model D from the combined counts from this and the previous training data. We then used Model D to rank the parses for the unannotated in-domain data (PBIOTB unsupervised), and derived Model E from the combined counts from the highest-ranked parses for all of the training and development data. We then iterated this process two more times over the unannotated datasets (each with an increasing number of examples though increasingly less relevant to the test data). The performance over our out-of-domain PTB-derived test data remains approximately the same for all these models. Therefore, we chose to use Model G for the blind test as it incorporates most information from the in-

Mdl.	Data	Init.	Prec.	Rec.	F_1
A	Uniform	-	71.06	69.00	70.01
	PTB				
B	Unambig.	A	75.94	73.16	74.53
C	Ambig.	B	77.88	75.11	76.47
	PBIOTB				
D	Supervised	C	77.86	75.09	76.45
E	Unsup. 1	D	77.98	75.25	76.59
F	Unsup. 2	E	77.85	75.19	76.50
G	Unsup. 3	F	77.76	75.09	76.41
	CHILDES				
H	Unsup. 1	C	78.34	75.59	76.94

Table 1: Performance of Successive Bootstrapping Models

	Score	Avg.	Std
PCHEMTB - labelled	55.47	65.11	09.64
PCHEMTB - unlab.ed	62.79	70.24	08.14
CHILDES - unlab.ed	45.61	56.12	09.17

Table 2: Official Scores

domain data. For the CHILDES data we performed one iteration of unsupervised adaptation in the same manner starting from Model C.

3 Evaluation

For the blind test submission we used Models G and H to parse the PCHEMTB and CHILDES data, respectively. We then mapped our GR output from the highest-ranked parses to the DR scheme and annotation format required by the CoNLL evaluation script. Our reported results are given in Table 2.

We used the annotated versions of the blind test data supplied after the official evaluation to assess the degree of adaptation of the parser to the in-domain data. We mapped from the DR scheme and annotation format to our GR format and used our evaluation script to calculate the precision, recall and microaveraged F_1 score for the unadapted models and their adapted counterparts on the blind test data, given in Table 3. The results for CHILDES show no evidence of adaptation to the domain. However, those for PCHEMTB show a statistically significant (Wilcoxin Signed Ranks) improvement over the initial model. The generally higher scores in

Model	Test Data	Prec.	Rec.	F_1
C	PCHEMTB	71.58	73.69	72.62
G	PCHEMTB	72.32	74.56	73.42
C	CHILDES	82.64	65.18	72.88
H	CHILDES	81.71	64.58	72.14

Table 3: Performance of (Un)Adapted Models

Table 3, as compared to Table 2, reflect the differences between the task annotation scheme and our GR representation as well as those of the evaluation schemes, which we discuss in the next section.

4 Discussion

The biggest issue for us participating in the shared task was the difficulty of reconciling the DR annotation scheme with our GR scheme, given the often implicit and sometimes radical underlying differences in linguistic assumptions between the schemes.

Firstly, the PoS tagsets are different and ours contains three times the number of tags. Given that the grammar uses these tags as preterminal categories, this puts us at a disadvantage in mapping the annotated training and development data to optimal input to train the (semi-)supervised models.

Secondly, there are 17 main types of GR relation and a total of 46 distinctions when GR subtypes are taken into account – for instance the GR **ncsubj** has two subtypes depending on whether the surface subject is the underlying object of a passive clause. The DR scheme has far fewer distinctions creating similar difficulties when creating optimal (semi-)supervised training data.

Thirdly, the topology of the dependency graphs is often significantly different because of reversed head-dependent bilexical relations and their knock-on effects – for instance, the DR **AUX** relation treats the (leftmost) auxiliary as head and modifiers of the verb group attach to the leftmost auxiliary, while the GR scheme treats the main verb as (semantic) head and modifiers of the verb group attach to it.

Fourthly, the treatment of punctuation is very different. The DR scheme includes punctuation markers in DRs which attach to the root of the subgraph over which they have scope. By contrast, the GR scheme does not output punctuation marks directly but follows Nunberg's (1990) linguistic analysis of punctuation as delimiting and typing text units or adjuncts (at constituent boundaries). Thus the GR scheme includes text (adjunct) relations and treats punctuation marks as indicators of such relations – for instance, for the example *The subject GRs – ncsubj, xsubj and csubj – all have subtypes.*, RASP outputs the GR **(ta dash GRs and)** indicating that the dash-delimited parenthetical is a text adjunct of *GRs* with head *and*, while the DR scheme gives **(DEP GRs and)**, and two **(P and –)** relations corresponding to each dash mark.

Although we attempted to derive an optimal and error-free mapping between the schemes, this was hampered by the lack of information concerning the DR scheme, lack of time, and the very different approaches to punctuation. This undoubtedly limited our ability to train effectively from the PTB data and to adapt the trained parser using the in-domain data. For instance, the mean average unlabelled F_1 score between the GRs mapped from the annotated PTB training data and closest matching set of GRs output by RASP for this data is 84.56 with a standard deviation of 12.41. This means that the closest matching derivation which is used for training the initial model is on average only around 85% similar even by the unlabelled measure. Thus, the mapping procedure required to relate the annotated data to RASP derivations is introducing considerable noise into the training process.

Mapping difficulties also depressed our official scores very significantly. In training and adapting we found that bootstrapping based on unlabelled dependencies worked better in all cases than utilizing the labelled mapping we derived. For the official submission, we removed all **ta**, **quote** and **passive** GRs and mapped all punctuation marks to the **P** relation with head **0**. Furthermore, we do not generate a root relation, though we assumed any word that was not a dependent in other GRs to have the dependent **ROOT**. In our own evaluations based on mapping the annotated training and development data to our GR scheme, we remove all **P** relations and map **ROOT** relations to the type **root** which we added to our GR hierarchy. We determined the semantic head of each parse during training so as to compare against the **root** GR and better utilize this additional information. In the results given in Table 1 over our DepBank test set, the effect of removing the **P** dependencies is to depress the F_1 scores by over 20%. For the CHILDES and PCHEMTB blind test data, our F_1 scores improve by over 7% and just under 9% respectively when we factor out the effect of **P** relations. These figures give an indication of the scale of the problem caused by these representional differences.

5 Conclusions

The main conclusion that we draw from this experience is that it is very difficult to effectively relate linguistic annotations even when these are inspired by a similar (dependency-based) theoretical tradition. The scores we achieved were undoubtedly further depressed by the need to use a partially-supervised boostrapping approach to training because the DR scheme is less informative than the GR one, and by our decision to use an entirely unlexicalized parse ranking model for these experiments. Despite these difficulties, performance on the PCHEMTB dataset using the adapted model improved significantly over that of the unadapted model, suggesting that bootstrapping using confidence-based self-training is a viable technique.

Acknowledgements

This research has been partially supported by the EPSRC via the RASP project (grants GR/N36462 and GR/N36493) and the ACLEX project (grant GR/T19919). The first author is funded by the Overseas Research Students Awards Scheme and the Poynton Scholarship appointed by the Cambridge Australia Trust in collaboration with the Cambridge Commonwealth Trust.

References

E. Briscoe (2006) *An introduction to tag sequence grammars and the RASP system parser,* University of Cambridge, Computer Laboratory Technical Report, UCAM-CL-TR-662.

E. Briscoe and J. Carroll (2006) 'Evaluating the Accuracy of an Unlexicalized Statistical Parser on the PARC DepBank', *Proceedings of the ACL-Coling'06,* Sydney, Australia.

Briscoe, E.J., J. Carroll and R. Watson (2006) 'The Second Release of the RASP System', *Proceedings of the ACL-Coling'06,* Sydney, Australia.

R. Brown (1973) *A First Language: The Early Stages,* Harvard University Press.

Inui, K., V. Sornlertlamvanich, H. Tanaka and T. Tokunaga (1997) 'A new formalization of probabilistic GLR parsing', *Proceedings of the 5th International Workshop on Parsing Technologies,* MIT, Cambridge, Massachusetts, pp. 123–134.

R. Johansson and P. Nugues (2007) *Extended Constituent-to-Dependency Conversion for English,* NODALIDA16.

S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein and L. Ungar (2004) 'Integrated Annotation for Biomedical Information Extraction', *Proceedings of the HLT-NAACL2004,* Boston, MA..

B. MacWhinney (2000) *The CHILDES Project: Tools for Analyzing Talk,* Lawrence Erlbaum.

M. Marcus, B. Santorini and M. Marcinkiewicz (1993) 'Building a Large Annotated Corpus of English: the Penn Treebank', *Computational Linguistics, vol.19.2,* 313–330.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel and D. Yuret (2007) 'The CoNLL 2007 Shared Task on Dependency Parsing', *Proceedings of the EMNLP-CoNLL2007,* Prague.

G. Nunberg (1990) *The Linguistics of Punctuation,* CSLI Publications.

Tomita, M. (1987) 'An Efficient Augmented Context-Free Parsing Algorithm', *Computational Linguistics, vol.13(1–2),* 31–46.

R. Watson, E. Briscoe and J. Carroll (2007) 'Semi-supervised Training of a Statistical Parser from Unlabeled Partially-bracketed Data', *Proceedings of the IWPT07,* Prague.