Comparing the Use of Edited and Unedited Text in Parser Self-Training

Jennifer Foster, Özlem Çetinoğlu, Joachim Wagner and Josef van Genabith
National Centre for Language Technology/Centre for Next Generation Localisation
School of Computing
Dublin City University
Ireland
{jfoster,ocetinoglu,jwagner,josef}@computing.dcu.ie

Abstract

We compare the use of edited text in the form of newswire and unedited text in the form of discussion forum posts as sources for training material in a self-training experiment involving the Brown reranking parser and a test set of sentences from an online sports discussion forum. We find that grammars induced from the two automatically parsed corpora achieve similar Parseval f-scores, with the grammars induced from the discussion forum material being slightly superior. An error analysis reveals that the two types of grammars do behave differently.

1 Introduction

There have been several successful attempts in recent years to employ automatically parsed data in semi- and unsupervised approaches to parser domain adaptation (McClosky et al., 2006b; Reichart and Rappaport, 2007; Huang and Harper, 2009; Petrov et al., 2010). We turn our attention to adapting a Wall-Street-Journal-trained parser to user-generated content from an online sports discussion forum. The sentences on the discussion forum are produced by a group of speakers who are communicating with each other about a shared interest and are discussing the same events, but, who, given the open, unedited nature of the medium itself, do not follow an in-house writing style. Our particular aim in this paper is to compare the use of discussion forum comments as a source of unlabelled training material to the use of edited, professionally written sentences on the same theme. We hypothesise that the well-formed sentences will be more suitable as training material since they are likely to be closer syntactically to the source domain Wall Street Journal (WSJ) sentences than the noisier discussion forum sentences, while at the same time, remaining lexically close to the target domain, thus acting as a type of "self-training bridging corpus" (McClosky et al., 2006b).

2 Related Work

McClosky et al. (2006b) demonstrate that a WSJ-trained parser can be adapted to the fiction domains of the Brown corpus by performing a type of self-training that involves the use of the two-stage Brown reranking parser (Charniak and Johnson, 2005). Their training protocol is as follows: sentences from the LA Times are parsed using the first-stage parser (Charniak, 2000) and reranked in the second stage. These parse trees are added to the original WSJ training set and the *first-stage* parser is retrained. The sentences from the target domain, in this case, Brown corpus sentences are then parsed using the newly trained first-stage parser and reranked using the original reranker, resulting in a Parseval f-score increase from 85.2% to 87.8%.

McClosky and Charniak (2008) later show that the same procedure can be used to adapt a WSJ-trained parser to biomedical text. They also try an experiment which is very similar to the experiment described in this paper. Instead of using Medline abstracts as training material, they use sentences from a biology textbook under the assumption that the parses produced for these sentences will be more accurate (and thus more reliable as training data) than the sentences in the abstracts since they are closer to the source domain. They find, however, that the textbook sentences are less effective than the target domain material. We attempt to repeat the experiment with Web 2.0 data, believing that the two setups are sufficiently different for our experiment to be worthwhile — our bridging corpus is closely related in subject matter to our target corpus (both referring to the same events) but quite different in form (professionally edited versus an unedited mix of writing styles), whereas their bridging corpus is less closely related in con-

Proceedings of the 12th International Conference on Parsing Technologies, pages 215–219,
October 5-7, 2011, Dublin City University. © 2011 Association for Computational Linguistics

tent (biology textbooks versus Medline abstracts) and more closely related in form (both professionally edited and syntactically well-formed).

3 Data

Our dataset, summarised in Table 1, consists of a small treebank of hand-corrected phrase structure parse trees and two larger corpora of unannotated sentences.

Discussion Forum Treebank The treebank is an extension of that described in Foster (2010). It contains 481 sentences taken from two threads on the BBC Sport 606 discussion forum in November 2009.[1] The discussion forum posts were split into sentences by hand. The sentences were first parsed automatically using an implementation of the Collins Model 2 generative statistical parser (Bikel, 2004). They were then corrected by hand using as a reference the Penn Treebank (PTB) bracketing guidelines (Bies et al., 1995) and the PTB trees themselves (Marcus et al., 1994). For more detail on the annotation process, see Foster et al. (2011). The development set contains 258 sentences and the test set 223. The experiments in this paper are carried out on the development set (which we refer to as *FootballDev*).

Discussion Forum Corpus The same discussion forum used to create the treebank was scraped during the final quarter of 2010. The content was stripped of HTML markup and passed through an in-house sentence splitter and tokeniser, resulting in a corpus of 1,009,646 sentences. We call this the *FootballTrainDiscussion* corpus.

Edited Text Corpus In order to compare the use of edited versus unedited text, we also collected a corpus of professionally written news articles on the same theme as the discussion forum sentences, namely, the English Premier League. Content was scraped from the online BBC sports site[2] and articles dating from April 2010 to February 2011 retrieved. Similar preprocessing was carried out on these as was carried out on the *FootballTrainDiscussion* content, i.e. HTML-stripping, sentence splitting and tokenisation. The resulting corpus, which we will refer to as *FootballTrainEdited*, contains 209,014 sentences.

4 Experiments

We retrain the Brown parser using the self-training protocol of McClosky et al. (2006b), that is, we retrain the first-stage parser using combinations of trees produced by the reranking parser for sentences from Sections 2 to 21 of the WSJ section of the Penn Treebank and from *FootballTrainEdited*|*Discussion*. We then parse the sentences in *FootballDev* using the retrained first-stage parser and the original reranker.

Because we have approximately five times the number of sentences in *FootballTrainDiscussion* than in *FootballTrainEdited*, we first train five different *FootballTrainDiscussion* grammars. The graph in Figure 1 shows the results on *FootballDev* when the training data contains disjoint subsections of *FootballTrainDiscussion*, each containing 200,000 sentences, along with varying amount of *WSJ2-21* trees. This gives us an idea of the amount of variation we might expect within one training set source — the f-score noise is roughly 1.5 points wide (= 3 boxes in the graph).

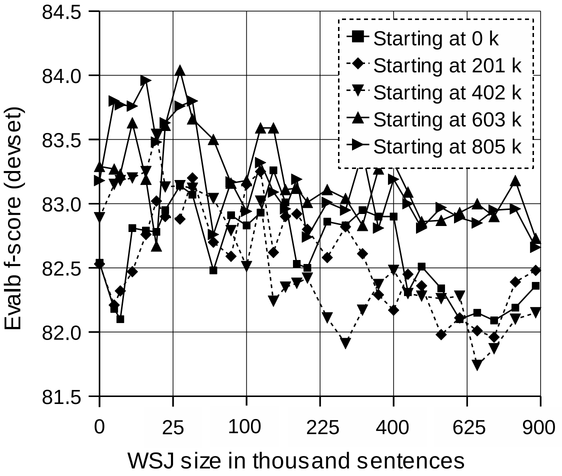

Figure 1: Comparing the performance of five grammars trained on disjoint 200k subsections of *FootballTrainDiscussion* in a Brown self-training experiment. Results are on *FootballDev*.

We now turn to the main experiment of the paper, i.e. the comparison of *FootballTrainDiscussion* and *FootballTrainEdited*. The graph in Figure 2 compares the performance of the *FootballTrainEdited* grammars with the performance average over the five types of *FootballTrainDiscussion*

[1] `http://www.bbc.co.uk/dna/606/F15264075?thread=7065503&show=50` and `http://www.bbc.co.uk/dna/606/F15265997?thread=7066196&show=50`

[2] `http://www.bbc.co.uk/search/sport/football` and `http://www.bbc.co.uk/search/news/football`

Corpus Name	*#Sen*	*SL Mean*	*SL Med.*	σ
FootballDev	258	17.7	14	13.9
FootballTest	223	16.1	14	9.7
FootballTrainDiscussion	1,009,646	15.4	12	13.3
FootballTrainEdited	209,014	17.7	17	11.4

Table 1: Basic Statistics on the Web 2.0 datasets: number of sentences, average sentence length, median sentence length and standard deviation

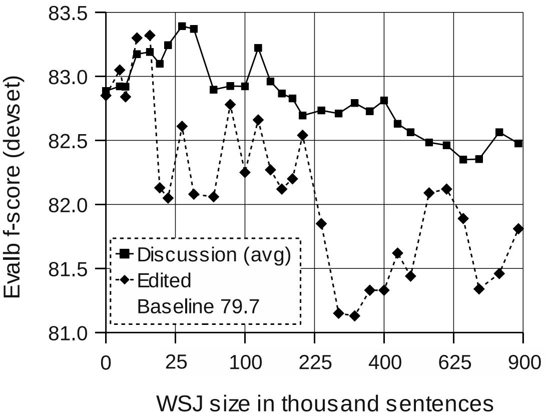

Figure 2: Comparing the use of discussion forum material (*FootballTrainDiscussion*) and newswire articles (*FootballTrainEdited*) in a Brown self-training experiment. Results are on *FootballDev*.

grammars on *FootballDev*. Note that a baseline grammar which is trained on one copy of *WSJ02-21* and no automatically parsed data achieves a Parseval f-score of 79.7. The results in Figure 2 appear to refute our original hypothesis, suggesting that there is very little difference between the two corpora, with the user-generated content of *FootballTrainDiscussion* emerging as slightly superior on our development set.[3] The only time that the *FootballTrainEdited* curve is above the *FootballTrainDiscussion* is when the size of the original WSJ training set is restricted. This is an intuitively appealing result — in this scenario, the sentences in the *FootballTrainEdited* corpus are making up for the lack of WSJ trained material, although it is not clear whether this is because the *FootballTrainEdited* sentences are slightly longer than the *FootballTrainDiscussion* sentences (see Table 1) or because they contain more WSJ-like constructions.

5 Analysis

We next attempt to determine the strengths and weaknesses of the two types of training material by classifying our development set items into those that have improved as a result of self-training, those that have remained unchanged and those that have deteriorated. We examine all edited grammars shown in Figure 2, i.e. the thirty grammars obtained using 200,000 sentences from *FootballTrainEdited* and varying sized copies of *WSJ02-21*. For the discussion grammars, we examine the grammars trained using one of the five disjoint 200,000-sentence subsets of *FootballTrainDiscussion* and varying sized copies of *WSJ02-21* — we randomly choose the grammars marked with squares in Figure 1. Following McClosky et al. (2006a), we present a breakdown of our results according to sentence length, number of co-ordinating conjunctions (`CC`) in the sentence, and, number of unknown words[4] in the sentence. The results are shown in Tables 2, 3 and 4. Sentence counts are provided along with average f-score differences between a self-trained grammar and the baseline grammar.

It is not possible to discern strong patterns in the breakdown of results but we do observe the following subtle differences between the two types of grammars:

- The *FootballTrainEdited* grammars are more conservative than the *FootballTrainDiscussion* grammars, with a larger number of sentences unchanged by self-training.

- The *FootballTrainDiscussion* grammars outperform the *FootballTrainEdited* grammars for short sentences.

[3] Keeping the *WSJ02-21* dataset size constant, we test whether the difference between a *FootballTrainEdited* grammar and its five corresponding *FootballTrainDiscussion* grammars is statistically significant. Of the 150 pairs, 42 differences are statistically significant ($p < 0.05$).

[4] A word is considered to be unknown if it does not appear at all in *WSJ02-21*.

Table 2: Effect of Self-Training Broken Down by Sentence Length

Discussion	**1-9**		**10-19**		**20-29**		**30-39**		**40-59**		**>= 60**		**TOTAL**	
Better	339	(+21.4)	761	(+14.7)	643	(+11.0)	229	(+9.6)	271	(+12.5)	65	(+5.7)	2308	(+13.6)
No Change	1853		1760		444		119		27		7		4210	
Worse	118	(-26.6)	329	(-8.3)	443	(-7.7)	192	(-7.0)	92	(-7.1)	48	(-5.3)	1222	(-9.4)
TOTAL	2310	(+1.8)	2850	(+3.0)	1530	(+2.4)	540	(+1.6)	390	(+7.0)	120	(+1.0)	7740	(+2.6)
Edited	**1-9**		**10-19**		**20-29**		**30-39**		**40-59**		**>=60**		**TOTAL**	
Better	237	(+20.6)	667	(+12.3)	543	(+10.8)	218	(+10.4)	213	(+9.2)	97	(+7.5)	1975	(+12.1)
No Change	1985		1934		656		202		29		1		4807	
Worse	88	(-18.5)	249	(-6.8)	331	(-9.4)	120	(-5.8)	148	(-6.3)	22	(-2.5)	958	(-8.5)
TOTAL	2310	(+1.4)	2850	(+2.3)	1530	(+1.8)	540	(+2.9)	390	(+2.6)	120	(+5.6)	7740	(+2.0)

Table 3: Effect of Self-Training Broken Down by Number of Coordinating Conjunctions in a Sentence

Discussion	**0**		**1**		**2**		**3**		**4**		**TOTAL**	
Better	1286	(+16.5)	631	(+11.1)	312	(+9.1)	0	(—)	79	(+5.5)	2308	(+13.6)
No Change	3061		981		131		30		7		4210	
Worse	573	(-10.8)	518	(-8.3)	97	(-8.8)	0	(—)	34	(-4.6)	1222	(-9.4)
TOTAL	4920	(+3.0)	2130	(+1.3)	540	(+3.7)	30		120	(+2.3)	7740	(+2.6)
Edited	**0**		**1**		**2**		**3**		**4**		**TOTAL**	
Better	1052	(+14.5)	561	(+10.0)	234	(+9.8)	18	(+5.2)	110	(+6.3)	1975	(+12.1)
No Change	3414		1168		212		12		1		4807	
Worse	454	(-9.6)	401	(-8.0)	94	(-5.4)	0	(—)	9	(-3.6)	958	(-8.5)
TOTAL	4920	(+2.2)	2130	(+1.1)	540	(+3.3)	30	(+3.1)	120	(+5.5)	7740	(+2.0)

Table 4: Effect of Self-Training Broken Down by Number of Unknown Words in a Sentence

Discussion	**0**		**1**		**2**		**3**		**4**		**5**		**>=6**		**TOTAL**	
Better	488	(+16.2)	865	(+10.8)	360	(+11.2)	319	(+19.1)	91	(+15.8)	107	(+11.9)	78	(+17.2)	2308	(+13.6)
No Change	2028		1277		584		223		58		30		10		4210	
Worse	274	(-14.0)	408	(-7.5)	406	(-9.5)	58	(-6.3)	31	(-5.3)	13	(-4.2)	32	(-5.7)	1222	(-9.4)
TOTAL	2790	(+1.5)	2550	(+2.5)	1350	(+0.1)	600	(+9.6)	180	(+7.1)	150	(+8.1)	120	(+9.7)	7740	(+2.6)
Edited	**0**		**1**		**2**		**3**		**4**		**5**		**>=6**		**TOTAL**	
Better	330	(+14.9)	758	(+11.1)	329	(+8.8)	328	(+15.7)	55	(+14.0)	94	(+9.5)	81	(+10.4)	1975	(+12.1)
No Change	2291		1352		800		241		91		32		0	(—)	4807	
Worse	169	(-13.5)	440	(-7.0)	221	(-8.6)	31	(-3.8)	34	(-7.4)	24	(-1.6)	39	(-11.2)	958	(-8.5)
TOTAL	2790	(+1.0)	2550	(+2.1)	1350	(+0.8)	600	(+8.4)	180	(+2.9)	150	(+5.7)	120	(+3.4)	7740	(+2.0)

- The *FootballTrainEdited* grammars appear to perform better than the *FootballTrainDiscussion* grammars when there are a relatively high number of coordinating conjunctions in a sentence (greater than two).

- Self-training with both *FootballTrainDiscussion* and *FootballTrainEdited* data tends to benefit sentences containing several unknown words, with the discussion grammars being superior.

6 Conclusion

We compare the use of edited versus unedited text in the task of adapting a WSJ-trained parser to the noisy language of an online discussion forum. Given the small size of our development set, we have to be careful how we interpret the results. However, they do seem to suggest that the two corpora are performing at similar levels of effectiveness but exhibit differences. For example, if we take the best performing *FootballTrainEdited* and *FootballTrainDiscussion* grammars from those used in our error analysis of Section 5, we get two grammars with a Parseval f-score of 83.2 on *FootballDev*. Assuming the existence of a perfect classifier, which, given an input sentence, can predict which of the two grammars will produce the higher-scoring tree, the f-score for *FootballDev* increases from 83.2 to 85.6. When we include the baseline grammar (f-score: 79.7), this increases to 86.4. This suggests that the next step in our research is to build such a classifier including as features the sentential properties we examined in Section 5, as well as the features described in McClosky et al. (2010) and Ravi et al. (2008).

Acknowledgements

This research has been supported by Enterprise Ireland (CFTD/2007/229) and by Science Foundation Ireland (Grant 07/CE/ I1142) as part of the CNGL (www.cngl.ie) at the School of Computing, DCU.

References

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for Treebank II style. Technical report, University of Pennsylvania.

Daniel Bikel. 2004. Intricacies of Collins parsing model. *Computational Linguistics*, 30(4).

Eugene Charniak and Mark Johnson. 2005. Course-to-fine n-best-parsing and maxent discriminative reranking. In *Proceedings of the 43rd ACL*.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL*.

Jennifer Foster, Özlem Çetinoğlu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. From News to Comment: Resources and benchmarks for parsing the language of Web 2.0. In *Proceedings of IJCNLP*.

Jennifer Foster. 2010. “cba to check the spelling” Investigating parser performance on discussion forum posts. In *Proceedings of HLT NAACL*.

Zhongqiang Huang and Mary Harper. 2009. Self-training PCFG grammars with latent annotations across languages. In *Proceedings of EMNLP*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Speech and Natural Language Workshop*.

David McClosky and Eugene Charniak. 2008. Self-training for biomedical parsing. In *Proceedings of ACL:HLT*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006a. Effective self-training for parsing. In *Proceedings of NAACL*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006b. Reranking and self-training for parser adaptation. In *Proceedings of ACL*.

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

Slav Petrov, Pi-Chuan Chang, Michael Ringgaard, and Hiyan Alshawi. 2010. Uptraining for accurate deterministic question parsing. In *Proceedings of EMNLP*.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic prediction of parser accuracy. In *Proceedings of EMNLP*.

Roi Reichart and Ari Rappaport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of ACL*.