Wide-coverage parsing of speech transcripts

Jeroen Geertzen
Research Centre for English & Applied Linguistics
University of Cambridge, UK
jg532@cam.ac.uk

Abstract

This paper discusses the performance difference of wide-coverage parsers on small-domain speech transcripts. Two parsers (C&C CCG and RASP) are tested on the speech transcripts of two different domains (parent-child language, and picture descriptions).

The performance difference between the domain-independent parsers and two domain-trained parsers (MSTParser and MEGRASP) is substantial, with a difference of at least 30 percent point in accuracy. Despite this gap, some of the grammatical relations can still be recovered reliably.

1 Introduction

Even though wide-coverage, domain-independent[1] parser systems may perform sufficiently well for the task at hand, obtaining highly accurate parses of sentences in a particular domain usually requires the parser to be domain-trained. Training a parser requires a sufficient amount of labelled data (a gold standard), something that is only available for very few domains. When accurate parses of sentences in a new domain are desired, there are several ways to proceed. Hand labelling all data in the new domain is a consideration, but is usually unfeasible as manual annotation is a costly activity. Another possibility is to minimise the amount of annotation effort required to achieve good performance by resorting to semi-automatic annotation or domain adaptation methods. In any case, dedicated effort is still required to obtain highly accurate parses, even with recent automated domain adaptation methods (Dredze et al., 2007).

Work that requires parsing in a new domain as basis of further study or as part of a larger natural language processing system usually involves a domain-independent parser with the expectation that parses are sufficiently accurate for the specific purpose.[2] For instance, Bos and Markert (2005) use a wide-coverage CCG-parser (Clark and Curran, 2007) to generate semantic representations for recognising textual entailment. Geertzen (2009) uses a HPSG-based dependency parser (Bouma et al., 2001) to obtain the semantic content of utterances. And in the study of child language acquisition, Buttery and Korhonen (2007) use RASP, a wide-coverage dependency parser (Briscoe et al., 2006), to look at lexical acquisition.

The goal of this paper is to give an indication of wide-coverage, domain-independent parser performance on specific domains. Additionally, the study gives insight into RASP's performance on CHILDES, allowing to factor in parsing performance in the syntax-based study of Buttery and Korhonen (2007).

2 Parsing speech transcripts

Parsing performance of two domain-independent parsers, C&C CCG en RASP, is evaluated on two speech domains. The first domain, CHILDES, involves parent-child interactions; the second domain, CCC, involves a picture description task.

2.1 Parsing systems

Two wide-coverage parser systems are used. RASP (Briscoe et al., 2006) is a parsing system for

[1] In this paper, the terms 'wide-coverage' and 'domain-independent' are used synonymously.

[2] Without gold standard there is no way of knowing how well the parser component performs with respect to a desired outcome of syntactic structure. This may not necessarily be a problem, as parsing in such cases is paramount, and application-based evaluation is preferable. Moreover, it may be that using linguistically most desired parses does not result in best application performance.

Proceedings of the 11th International Conference on Parsing Technologies (IWPT), pages 218–221,
Paris, October 2009. ©2009 Association for Computational Linguistics

English that utilises a manually-developed grammar and outputs grammatical dependency relations. The C&C CCG parser (Clark and Curran, 2007) is a parsing system that is based on an automatically extracted grammar from CCG-Bank and uses discriminative training. Both systems are able to output the exact set of dependency relations, and in a comparison on a 560-sentence test set used by Briscoe and Carroll (2006), Clark and Curran (2007) report a micro-averaged F-score of 81.14 for the CCG parser, and 76.29 for RASP. [3] Both parsing systems utilise the Grammatical Relations (GR) annotation scheme proposed by Carroll et al. (1998). This scheme is intended to cater for parser evaluation, and extends the dependency structure based method of evaluation proposed by Lin (1998). For the parent-child interaction domain both parsing systems are compared with two syntactic dependency parsers that were specifically trained for CHILDES transcripts: MEGRASP (Sagae et al., 2007) and MST-parser (McDonald et al., 2005).

2.2 Speech phenomena

As CCC and CHILDES transcripts are describing spoken language, they contain various markers that encode speech phenomena, particularly disfluencies (e.g. filled pauses, partial words, false starts, repetitions) and speech repairs (e.g. retractions and corrections). Prior to parser evaluation, such disfluencies have been deleted from the transcripts, which slightly improves parser performance for all systems mentioned. Similar performance improvements are also reported in studies that address the effect of deletion of repairs and fillers on parsing (e.g. Charniak and Johnson (2001); Lease and Johnson (2006)).

2.3 CHILDES data

The major part of the evaluation is based on the parsing of parent-child interactions from the CHILDES database (MacWhinney, 2000). A large portion of CHILDES transcripts was recently parsed with a domain-specific parser (Sagae et al., 2007), allowing more reliable systematic studies of syntactic development in child language acquisition. Sagae et al. also released their gold standard data, allowing others to train and evaluate other parser systems.

The gold standard data uses a GR scheme that is based on that of Carroll et al. (1998) but that differs in two respects: the scheme is extended to suit the specific need of the child language research community (cf. (Sagae et al., 2004)), and the scheme does not extensively and explicitly use the GR hierarchy.

To compare parsing performance, a mapping from RASP GRs to CHILDES GRs was manually constructed, containing 75 rules that involve the label and optional restrictions on the word or POS-tag of the head or dependent.

3 Parser evaluation

3.1 Measures

System performance is reported with accuracy measures for labelled and unlabelled dependencies resulting from 15-fold cross-validation.[4] The performance on each grammatical relation is expressed by precision, recall, and F_1-score. Punctuation has been excluded.

3.2 CHILDES

The gold-standard used for evaluation is based on 15 (out of 20) files in the Eve section of the Brown corpus. The annotations that are available were made with the CHILDES GR scheme, for which an inter-annotator percentage agreement of 96.5% ($N = 2$) has been reported by Sagae et al. (2004). From all manually annotated utterances initially available, duplicates, those with less than three tokens (about 30% of all), and those with missing or incomplete parses (1% of all) were removed, resulting in a set of 14.137 sentences, comprising 93,594 tokens with 4.5 tokens per utterance on average.

The performance scores that are obtained when the parsing systems are compared against the gold-standard are listed in the upper part of Table 1.

As can be seen from the accuracy scores, MEGRASP and the MSTParser perform with more than 30 percent point accuracy considerably better than the domain-independent parsers. However, the list of performance scores for each of the grammatical relations in Table 2 shows that some relations can be recovered with acceptable

[3] It should be remarked that such cross-formalism comparisons are difficult in nature. In this case, training data were different (RASP is not tuned to newspaper text), and CCG utilises a lexicalised parsing model where RASP does not.

[4] The exception being the MEGRASP, for which because of computation problems the full gold standard was used (7% larger than the other training sets), resulting in somewhat higher scores than expected with cross-validation.

Table 1: Parsing accuracy scores.

CHILDES	*labelled*	*unlabelled*
RASP	60.1	69.2
CCG parser	39.1	66.5
MSTParser	93.8	95.4
MEGRASP	90.7	93.5
CCC	*labelled*	*unlabelled*
RASP	66.7	72.3
CCG parser	60.2	68.5

F_1-scores, such as auxiliaries, determiners, subjects, and objects of prepositions.[5]

3.3 CCC

The Cambridge Cookie-theft Corpus (CCC, TO APPEAR, 2010) contains audio-recorded monologues of 196 subjects that were asked to fully describe a scene in a picture. As a result, the domain is small, but at the same time, sentence boundaries are difficult to indicate. From this corpus of 5,628 intonational phrases, a small evaluation set of 80 phrases has been manually annotated[6] with GRs. The performance scores for each of the parsers is listed in the lower part of Table 1. Accuracy scores are higher than those for CHILDES, and the difference in labelled accuracy between the domain-independent parsers is less than with CHILDES. Due to space restrictions it is not possible to present performance on individual GRs, but the GRs that are most reliably recovered are similar to those mentioned in Section 3.2.

4 Considerations

In the work reported here, performance of domain-independent parsers on narrow domains was calculated for two domains. The availability of more domain-specific datasets with manually supervised GR annotations would allow a better generalisation of parser performance. Unfortunately, datasets with manually verified annotations that use the same set of syntactic dependencies are rare.

The CHILDES figures show that the performance difference between domain-independent and domain-trained parsers is big. It should be noted that these results are obtained from speech, which is usually less syntactically well-formed than written language. For the speech data analysed, RASP performs better than the CCG parser, whereas Clark and Curran (2007) have shown that the CCG parser outperforms RASP on written text. To better explain this difference, it would be insightful to compare the confusion matrices of GR assignments. This would allow assessment on how the domain-independent parser errors compare to the domain-trained parser errors.

The mapping from RASP GRs to CHILDES GRs that was constructed is exhaustive, but there is still room for fine-tuning and more refined mappings, gaining up to about 2% accuracy by estimate.

5 Conclusions and future work

This paper has provided performance scores of wide-coverage parsers applied to narrow domain spoken language transcripts to assess the performance gap with domain-trained parsers. This gap appears to be considerable (more than 30 percent point for CHILDES), but a subset of GRs can still be recovered with fair accuracy.

We have not yet dealt with comparing domain-independent and domain-trained parser errors, which may provide additional insight into the strengths and weaknesses of wide-coverage parsers for narrow use.

Acknowledgements

This work is supported by UK EPSRC Grant EP/F030061/1.

References

Bos, J. and Markert, K. (2005). Recognising textual entailment with logical inference. In *Proceedings of the HLT and EMNLP conference*, pages 628–635.

Bouma, G., van Noord, G., and Malouf, R. (2001). Alpino: Wide-coverage computational analysis of dutch. In *Proceedings of the CLIN 2000*, pages 45–59.

Briscoe, T. and Carroll, J. (2006). Evaluating the accuracy of an unlexicalized statistical parser on the PARC depbank. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 41–48.

[5] MSTParser scores did not fit in the table, but largely correspond in distributional characteristics, and are available upon request.

[6] Not with multiple coders yet, but percentage agreement for dependency annotation typically varies from 93-98%.

Table 2: Performance scores of the parsing systems for major GRs. Some of the relations could not be reliably be mapped, and are absent for the CCG parser.

	RASP			CCG parser			MEGRASP		
relation	*Prec*	*Rec*	F_1	*Prec*	*Rec*	F_1	*Prec*	*Rec*	F_1
aux	89.13	69.87	78.33	90.81	62.21	73.84	98.13	96.21	97.16
com	67.80	6.12	11.23	-	-	-	93.15	88.52	90.78
comp	22.73	64.18	33.57	24.53	53.66	33.67	80.00	84.72	82.29
coord	70.42	64.31	67.23	82.50	30.62	44.66	75.07	83.93	79.26
cpzr	74.67	20.97	32.75	-	-	-	90.16	85.77	87.91
det	90.34	89.38	89.86	60.88	82.54	70.07	96.38	97.27	96.82
jct	57.85	56.68	57.26	54.71	5.16	9.42	85.14	83.05	84.08
mod	63.04	76.93	69.29	16.89	47.43	24.91	90.00	90.63	90.32
obj	73.34	75.50	74.40	46.09	69.25	55.34	91.93	91.10	91.52
obj2	32.81	55.13	41.13	53.37	39.16	45.18	83.33	74.14	78.47
pobj	88.11	75.51	81.33	-	-	-	91.94	93.05	92.49
pred	54.77	48.94	51.69	64.60	15.55	25.07	90.21	91.08	90.65
quant	55.87	68.87	61.69	-	-	-	83.10	91.46	87.08
subj	74.53	67.58	70.89	66.94	66.11	66.52	94.68	95.01	94.84
xcomp	52.17	64.97	57.87	1.62	3.35	2.19	92.11	87.13	89.55
xmod	12.93	15.32	14.02	2.60	24.19	4.69	56.64	65.32	60.67

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80.

Buttery, P. and Korhonen, A. (2007). I will shoot your shopping down and you can shoot all my tins—automatic lexical acquisition from the CHILDES database. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 33–40.

Carroll, J., Briscoe, T., and Sanfilippo, A. (1998). Parser evaluation: a survey and a new proposal. In *Proceedings of the 1st LREC*, pages 447–454.

Charniak, E. and Johnson, M. (2001). Edit detection and parsing for transcribed speech. In *Proceedings of NAACL*, pages 118–126.

Clark, S. and Curran, J. R. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Dredze, M., Blitzer, J., Pratim Talukdar, P., Ganchev, K., Graca, J. a., and Pereira, F. (2007). Frustratingly hard domain adaptation for dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1051–1055.

Geertzen, J. (2009). Semantic interpretation of Dutch spoken dialogue. In *Proceedings of the Eight IWCS*, pages 286–290.

Lease, M. and Johnson, M. (2006). Early deletion of fillers in processing conversational speech. In *Proceedings of the HLT-NAACL*, pages 73–76.

Lin, D. (1998). A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering*, 4(2):97–114.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, third edition.

McDonald, R., Crammer, K., and Pereira, F. (2005). Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting on ACL*, pages 91–98.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., and Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the ACL-2007 workshop on Cognitive Aspects of Computational Language Acquisition*.

Sagae, K., MacWhinney, B., and Lavie, A. (2004). Adding syntactic annotations to transcripts of parent-child dialogs. In *In Proceedings of the Fourth LREC*, pages 1815–1818.