Proceedings Paper:
Atwell, ES, Hughes, J and Souter, DC (1994) A unified multicorpus for training syntactic constraint models. In: Evett, L and Rose, T, (eds.) Workshop on Computational Linguistics for Speech and Handwriting Recognition. 1994 AISB Workshop on Computational Linguistics for Speech and Handwriting Recognition, 12 April 1994, University of Leeds, UK. AISB, 111 - 118.

A Unified MultiCorpus for training syntactic constraint models

Eric Atwell, John Hughes, Clive Souter
Centre for Computer Analysis of Language And Speech
Artificial Intelligence Division
School of Computer Studies
Leeds University
Leeds LS2 9JT, England, UK

eric@scs.leeds.ac.uk
john@scs.leeds.ac.uk
cs@scs.leeds.ac.uk

March 1994

1 Abstract

Tagged and parsed corpora (LOB, Brown, London-Lund, ICE, Lancaster-IBM, PoW, Nijmegen, UPenn, BNC, etc) are used as training data for statistical syntactic constraint models to improve recognition accuracy in speech and handwriting recognisers. However, linguists developing these linguistic resources have used quite different wordtagging and parse-tree labelling schemes in each of these annotated corpora. This restricts the accessibility of each corpus, making it impossible for speech and handwriting researchers to collate them into a single very large training set. This is particularly problematic as there is evidence that one of these parsed corpora on its own is too small for a general statistical model of higher-level syntactic structure, but the combined size of all the above annotated corpora should deliver a much more reliable model.

We are developing a set of mapping algorithms to map between the main tagsets and phrase structure grammar schemes used in the above corpora. We will develop a Multi-tagged Corpus and a MultiTreebank, a single text-set annotated with all the above tagging and parsing schemes. The text-set is the Spoken English Corpus; this is already annotated with two syntax schemes, and we plan to have added at least one more by the AISB Workshop. However, the main

deliverable to the speech and handwriting research community is not the SEC-based MultiTreebank, but the mapping suite used to produce it - this can be used to combine currently-incompatible syntactic training sets into a large unified multicorpus. Our development of the mapping algorithms aims to distinguish notational from substantive differences in the annotation schemes, and we will be able to evaluate tagging schemes in terms of how well they fit standard statistical language models such as n-pos (Markov) models.

2 Introduction

Several research projects around the world are building grammatically analysed corpora; that is, collections of text annotated with part-of-speech wordtags and syntax trees. However, projects have used quite different wordtagging and parsing schemes. In contrast to the Speech research community, which has reached broad agreement on an uncontentious set of labelling conventions for phonetic/phonemic analysis, there is no general consensus in the UK Natural Language research community on analogous conventions for grammatical analysis. Developers of corpora adhere to a variety of competing models or theories of grammar and parsing, with the effect of restricting the accessibility of their respective corpora, and the potential for collation into a single fully parsed corpus.

In view of this heterogeneity, we propose to investigate and develop methods of automatically mapping between the annotation schemes of the most widely known corpora, thus assessing their differences and improving the reusability of the corpora. Annotating a single corpus with the different schemes allows for comparisons, and will provide a rich test-bed for automatic parsers.

The most widely known tagged corpora for English are: the Lancaster-Oslo/Bergen (LOB) Corpus; the Brown Corpus; and the London-Lund Corpus. In addition, the International Corpus of English (ICE) should be included as its tagset has now been published ([souter+atwell93]). Parsed corpora for English include: the Lancaster-IBM Treebank; the Lancaster-Leeds Treebank; the Polytechnic of Wales (POW) Corpus; the Nijmegen Corpus and the University of Pennsylvania (UPenn) Treebank. We will include the British National Corpus (BNC) in the project when it becomes available. The tagging and parsing schemes with which we are most familiar at the outset are those used in LOB ([atwell81,82,83,84,87,88,91], [leech83a,b], [hughes+atwell93,94], [hughes94], [johannson86], [souter+atwell93]); POW ([atwell88,91], [souter+atwell88,92,93], [souter89a,b,90], [odonoghue91a,b], [hughes89,94]); Nijmegen ([souter92]); and ICE ([souter+atwell93]); by the end of this project we hope to have become unique 'tagging/parsing scheme polyglots'!

As a development and testing resource, we propose using the Lancaster-IBM Spoken English Corpus (SEC). The SEC is a collection of recordings of radio broadcasts with accompanying annotated transcriptions, collected by Lancaster

University and IBM UK as a public research resource. The SEC is available from the International Computer Archive of Modern English (ICAME) based at the Norwegian Computing Centre for the Humanities (in Bergen, Norway). The corpus is distributed in several forms: the digitised acoustic waveform; the graphemic transcription annotated with prosodic markings; and a part-of-speech analysis (using the LOB Corpus tagset). Skeletal parsing has been added to create the SEC Treebank, and this forms a subset of the Lancaster-IBM Treebank. Gerry Knowles (Lancaster) and Peter Roach (Leeds) are currently collaborating in an ESRC-funded project to set up a time-aligned database of recorded speech, accompanied by phonetic and graphemic transcriptions. Our proposal will produce, as a side-effect, several alternative tagged and parsed versions of the SEC which will be made available to the SEC database project collaborators. It will also be able to act as a test-bed for the comparison and evaluation of parsing schemes.

3 Objectives of the Project

The main objectives are as follows :

To design and implement algorithms for mapping between corpus annotation schemes; for both wordtag sets and phrase structure grammar schemes.

To empirically evaluate the accuracy and shortcomings of the developed mapping algorithms, by applying them to the tagged SEC and the SEC Treebank. The outcome of this evaluation will be to highlight the notational and substantive differences between the alternative tagging and parsing schemes.

To build a Multi-Tagged Corpus, by enhancing the Spoken English Corpus with different wordtagging schemes.

To build a Multi-Treebank, by enhancing the Spoken English Corpus with grammatical analyses according to several alternative grammatical theories.

To investigate the use of the Multi-Treebank as a benchmark for grammars and parsers.

We have chosen the SEC as a 'core' text for this project, because (i) the tagged SEC uses the same tagset as the LOB Corpus (widely considered to be the UK standard and our proposed primary tagset), (ii) the parsed SEC uses the same grammatical scheme as the Lancaster-IBM Treebank (our proposed primary parsing scheme); (iii) these are the annotation schemes which we have most experience of; and (iv) the text material, BBC radio broadcasts, are a neutral compromise between written and conversational spoken English genres.

Our aim is to develop bidirectional mappings for the above tagsets and grammar schemes, although we appreciate that for mapping from simple to delicate schemes this will not be possible, and that mappings will be imperfect. As mapping algorithms are developed and tested, and whilst building the Multi-Tagged Corpus and Multi-Treebank, we will compile "handbooks" of common errors and

their corrections. These will help future users of the developed mapping algorithms to straightforwardly post-edit their mapped corpora and treebanks, thus maximising resource reusability. To map between two tagsets other than LOB, two mappings will be necessary (via the primary tagset, our "interlingua" representation); similarly for non-terminal grammar schemes. We appreciate the danger of propogating incorrect mappings.

If there is sufficient time, we will go on to investigate mapping algorithms for other (more detailed) grammar schemes; for example the parsed POW Corpus (Systemic Functional Grammar), and the parsed Nijmegen Corpus (Extended Affix Grammar). The non-corpus-based Generalised Phrase Structure Grammar (GPSG) (as used in the Alvey Natural Language Toolkit ANLT) should also be included. Mapping from these to the Lancaster-IBM Treebank grammar scheme would only be uni-directional ie. from a detailed to a skeletal analysis.

The Multi-Treebank will be produced by applying the final version of each grammar scheme mapping algorithm to the SEC Treebank. Similarly, for the Multi-Tagged Corpus, the final version of each tagset mapping algorithm will be applied to the tagged SEC. The resulting annotations will then be intensively proofread and post-edited. This will require consultations with authorities in each of the tagsets and grammar schemes involved.

4 Phases of research

We are currently in the first of three phases. We plan to go on to map between phrase-structure parsing schemes; and to investigate applications of our multi-tagged corpus and multitreebank, as a benchmark.

4.1 Implementation of Algorithms For Mapping Between Tagsets

Mapping algorithms will be designed and implemented between the LOB Corpus tagset and each of: the tagged BROWN Corpus, the tagged ICE, the Lancaster-IBM Treebank, the UPenn Treebank, and the BNC tagset (when published). Each tagset will be considered in turn:

(1) Analysis of the notational and substantive differences between the LOB tagset and the 'current' tagset.

(2) Design and implementation of a mapping algorithm (two-way, where possible).

(3) Evaluate success of algorithm by applying it to the tagged SEC; incrementally improve in light of common errors and linguistic intuition.

A side-effect of this phase will be the production of a Multi-Tagged Corpus; the SEC automatically annotated with each tagset.

4.2 Implementation of Algorithms For Mapping Between Grammar Schemes

Initially mapping algorithms will be designed and implemented for between the Lancaster-IBM Treebank grammar scheme, and each of the UPenn Treebank and the Lancaster-Leeds Treebank. Each grammar scheme will be considered in turn:

(1) Analysis of the notational and substantive differences between the Lancaster-IBM grammar scheme and the current grammar scheme.

(2) Manually parse a subset of the SEC according to the current grammar scheme. This subset should be sufficient to allow a prototype mapping algorithm to be induced.

(3) Apply mapping algorithm to the parsed SEC; incrementally improve in light of common errors and linguistic intuition.

4.3 Assessment of the Multi-Treebank as a Benchmark for Grammars.

This requires analysis of the substantive differences between different parses of the SEC sentences; detailed analysis of how many and which constructs differ in the different language models. It may be possible to divide the sentences in the SEC into two subsets: a common core of "uncontentious" sentences which all or most theories analyse in much the same way; and a "troublesome" subset of sentences which linguists can concentrate their debate on.

4.4 Comparisons with other parsed corpora

Comparisons of the Multi-Treebank with other tagged and parsed corpora (LOB, UPenn, POW, Nijmegen, etc), to assess differences in the range and frequency distributions of grammatical constructs. The SEC consists of transcripts of radio broadcasts. Some Natural Language researchers may feel that the Spoken English dataset is thus inappropriate for their work, since the grammars and parsers they are developing are designed for a different type of language. It may be appropriate to augment the SEC dataset with additional material from alternative sources. On the other hand, it may be that the main differences are in vocabulary rather than syntax, and that the coverage of the SEC, though not complete or perfect, is adequate for most applications. We will try to find empirical evidence for or against the acceptability of Spoken English to the NL community.

4.5 Assessment of Multi-Treebank as a Benchmark for Parsers

This will involve attempting to parse the SEC text with other parsers, available from a variety of sources. To avoid the need for intensive manual proofreading or

checking of results, a (semi-)automatic assessment procedure will be developed.

5 Applications

The implemented mapping algorithms will be made widely available to the UK and international speech and language research community. They will allow research groups who are using corpus-based training data to make use of other corpora straightforwardly, without substantial modifications. Any current and future users of corpora will have a much expanded resource.

The Multi-Tagged Corpus and the Multi-Treebank will be distributed, along with the main Spoken English Corpus, through ICAME. They will also be available for incorporation into the SEC Speech Database currently being created by Gerry Knowles and Peter Roach, further enhancing the SEC as a general research resource.

Both the Multi-Treebank and the Multi-Tagged corpus will potentially be used by speech and language technology groups for many research and teaching purposes, including: training data for speech-recognisers, optical text recognisers, word processor text-critiquing systems, machine translation systems, natural language interfaces, and NLP applications generally; and for providing examples for English Language Teaching (ELT) grammar textbooks and training material. In addition, the Multi-Treebank may be used as a testbed and benchmark for parsers (explored in the workplan). It would also be a rich resource for grammar-learning experiments - a research topic of growing interest.

We envisage supplying the speech and handwriting research community with a valuable research resource, and the AISB Workshop will be an invaluable opportunity for us to survey potential customer requirements and preferences!

6 References

Atwell, Eric Steven. 1981. LOB Corpus Tagging Project: Manual Pre-edit Handbook. Departments of Computer Studies and Linguistics, Lancaster University.

Atwell, Eric Steven. 1982. LOB Corpus Tagging Project: Manual Post-edit Handbook. Departments of Computer Studies and Linguistics, Lancaster University.

Atwell, Eric Steven, 1983. Constituent-likelihood grammar. ICAME Journal 7: 34-66.

Atwell, Eric Steven, Leech, Geoffrey and Garside, Roger. 1984. Analysis of the LOB Corpus: progress and prospects. In Jan Aarts and Willem Meijs (eds.). Corpus Linguistics. 40-52. Amsterdam: Rodopi.

Atwell, Eric Steven. 1987. A parsing expert system which learns from corpus analysis. In Willem Meijs (ed.). Corpus Linguistics and Beyond. 227-235,

Amsterdam: Rodopi.

Atwell, Eric Steven. 1988. Transforming a Parsed Corpus into a Corpus Parser. In Merja Kyto, Ossi Ihalainen and Matti Rissanen (eds.). Corpus Linguistics, hard and soft. 61-70. Amsterdam: Rodopi.

Atwell, Eric, Clive Souter and Tim O'Donoghue. 1988. Prototype Parser 1. COMMUNAL Research Report No. 17. School of Computer Studies, The University of Leeds.

Atwell, Eric, Tim O'Donoghue and Clive Souter. 1991. Training Parsers with Parsed Corpora. Research Report 91.20. School of Computer Studies, University of Leeds.

John Hughes. 1989. A learning interface to the realistic annealing parser. Technical report, School of Computer Studies, University of Leeds.

John Hughes. 1994. Automatically Acquiring a Classification of Words. PhD Thesis, School of Computer Studies, University of Leeds.

John Hughes and Eric Atwell. 1993. Acquiring and Evaluating a Classification of Words. Proceedings of IEE Grammatical Inference Colloquium. University of Essex, Colchester.

John Hughes and Eric Atwell. 1994. The Automated Evaluation of Inferred Word Classifications. Proccedings of 11th European Conference on Artificial Intelligence. Amsterdam, The Netherlands.

Johannson, Stig, Eric Atwell, Roger Garside, and Geoffrey Leech. 1986. The Tagged LOB Corpus. University of Bergen, Norway: Norwegian Computing Centre for the Humanities.

Leech, Geoffrey, Roger Garside and Eric Steven Atwell. 1983a. The Automatic Grammatical Tagging of the LOB Corpus. ICAME Journal 7: 13-33.

Leech, Geoffrey, Roger Garside and Eric Steven Atwell. 1983b. Recent developments in the use of computer corpora in English language research. Transactions of the Philological Society 1983: 23-40.

O'Donoghue, Tim F. 1991a. EPOW: The Edited Polytechnic of Wales Corpus. Research Report 91.11. School of Computer Studies, University of Leeds. Also appeared in Proceedings of the 5th International Conference on Symbolic and Logical Computing. Dakota State University, USA, April 1991.

O'Donoghue, Tim F. 1991b. Taking a parsed corpus to the cleaners: the EPOW corpus. ICAME Journal 15: 55-62.

Souter, Clive and Eric Atwell. 1988. Constraints on Legal Syntactic Configurations. COMMUNAL Research Report No. 14. School of Computer Studies, The University of Leeds.

Souter, Clive. 1989a. The COMMUNAL Project: Extracting a grammar from the Polytechnic of Wales corpus. ICAME Journal 13: 20-27.

Souter, Clive. 1989b. A Short Handbook to the Polytechnic of Wales Corpus. ICAME, Norwegian Computing Centre for the Humanities, P.O. Box 53, Bergen University, N-5027 Bergen, Norway.

Souter, Clive. 1990. Systemic Functional Grammars and Corpora. In J. Aarts and W. Meijs eds. Theory and Practice in Corpus Linguistics. 179-211, Rodopi Press, Amsterdam.

Souter, Clive. 1992. The Nijmegen Linguistic Database program. ICAME Journal 16: 70-79.

Souter, Clive and Eric Atwell. 1992. A Richly Annotated Corpus for Probabilistic Parsing. Research Report 92.13, School of Computer Studies, University of Leeds. Also appeared in Proceedings of AAAI workshop on Statistically-Based NLP Techniques, San Jose, California, July 12-17, 1992.

Souter, Clive and Eric Atwell eds. 1993. Corpus-Based Computational Linguistics. Amsterdam: Rodopi Press.