MULTI-LEVEL DISAMBIGUATION GRAMMAR INFERRED FROM ENGLISH CORPUS, TREEBANK, AND DICTIONARY

by Eric Atwell, Simon Arnfield, George Demetriou, Steve Hanlon, John Hughes, Uwe Jost, Rob Pocock, Clive Souter, and Joerg Ueberla

Summary. In this paper we will show that Grammatical Inference is applicable to Natural Language Processing. Given the wide and complex range of structures appearing in an unrestricted Natural Language like English, full Grammatical Inference, yielding a comprehensive syntactic and semantic definition of English, is too much to hope for at present. Instead, we focus on techniques for dealing with ambiguity resolution by probabilistic ranking; this does not require a full formal Chomskyan grammar. We give a short overview of the different levels and methods being investigated at CCALAS for probabilistic ranking of candidates in ambiguous English input.

Grammatical Inference from English corpora. An earlier title for this paper was "Overview of grammar acquisition research at CCALAS, Leeds University", but this was modified to avoid the impression of an incoherent set of research strands with no integrated, focussed common techniques or applications. The researchers in our group have no detailed development plan imposed 'from above', but are working on independent PhD programmes; however, there are common theoretical tennets, ideas, and potential applications linking individual projects. In fact, preparing for the Colloquia on Grammatical Inference has helped us to appreciate these overarching, linking themes, as we realised that the definitions stated in the Programme clearly applied to our own work at CCALAS:

'Grammatical Inference ... has suffered from the lack of a focused research community ... Simply stated, the grammatical inference problem is to learn an efficient description that captures the essence of a set of data. This description may be used subsequently to classify data, or to generate further examples of similar data.'

The data in our case is unrestricted English input, as exemplified by a Corpus or large collection of text samples. This renders a much harder challenge to Grammatical Inference than artificial languages, or selected examples of well-formed English sentences. The range of lexical items and grammatical constructs appearing in an unrestricted English Corpus is very large; and the problem is not just one of scale. The Corpus-based approach carries with it a blurring of the classical Chomskyan distinction between 'grammatical' and 'ungrammatical' English sentences. Indeed, [Sampson 87] went to the extreme of positing that there is NO boundary between grammatical and ungrammatical sentences in English; this might seem to imply that it is hopeless and even invalid to attempt to infer a grammar for English. Furthermore, the Corpus-based approach eschews the use of 'intuitively constructed' examples in training: a learning algorithm should be trained with 'real' sentences from a Corpus. It would seem to follow from this that we are also proscribed from artificially constructing negative counter-examples for our learning algorithms: we cannot guarantee that such counter-examples are truly illegal.

Tagged and parsed corpora and Machine Readable Dictionaries. On the other hand, Grammatical Inference research on Corpora can benefit from a wealth of human linguistic knowledge encoded in human-annotated corpora, and also human lexicographical knowledge encoded in general English dictionaries. At CCALAS we have several tagged and parsed Corpora, including: Brown, LOB (Lancaster-Oslo/Bergen), PoW (Polytechic of Wales), SEC (Lancaster/IBM Spoken English Corpus), SCRIBE (Spoken Corpus Recordings In British English), Nijmegen, SUSANNE. We also have a number of Machine Readable Dictionaries (MRDs) and lexical databases: LDOCE (Longman Dictionary of Contemporary English), OALDCE (Oxford Advanced Learner's Dictionary of Current English), CED (Collins English Dictionary), CELEX (Centre for Lexical Information, Nijmegen), MRC (Medical Research Council) Psycholinguistic Database, OED (Oxford English Dictionary). The Grammatical Inference problem for unrestricted English is transformed into the problem of utilising these resources, extracting the 'essence of English grammar' from them. The linguistic knowledge encoded in tagged and parsed corpora and machine readable dictionaries give us a bootstrap for Grammatical Inference, if only we can find a way of tapping it.

Centre for Computer Analysis of Language And Speech (CCALAS),
Artificial Intelligence Division, School of Computer Studies, The University of Leeds, Leeds LS2 9JT England

Grammar as Disambiguation Model. To simplify our task, we draw back from the Chomskyan definition of a grammar as a finite definition of the set of all and only sentences in a language [Chomsky 57], and have effectively been using the above looser definition (‘an efficient description that captures the essence of a Corpus of English text samples’). For some applications it is not necessary to decide whether a word-sequence is grammatical or not; our language description is used instead to disambiguate between alternative possible analyses of the input, by imposing an ordering on analyses rather than necessarily eliminating all but one.

The disambiguation problem manifests at several different levels. For example, consider an input string of words such as:

stephen left school last year .

A word may belong to more than one category according to context; for example:

- the Part of Speech (POS) of the word ‘school’ can be Noun or Verb;

- the word ‘school’ can have several different meanings or senses: in the word sense classification used in LDOCE, ‘school’ can mean (1) a place of education, (2) a course of learing, (3) a body of students, (4) a specialised training establishment, (5) a university department, (6) a school of thought, (7) a University (in USA), (8) to teach, (9) a group of sea animals;

- in a topic or subject taxonomy, the uses of ‘school’ can be broadly divided into three subject fields: Education, Art, Zoology;

- a classification inference algorithm can provide alternative ways of categorising words according to other contextual cues, such as the general position witih a sentence that the word tends to appear in;

- in spoken English, words can have different pitch and stress associated with them according to context - a major stress on ‘school’ would indicate a contrast with other things stephen might have left;

- there is also structural ambiguity, for example in the parse-tree annotation scheme used in the Spoken English Corpus (SEC) Treebank, ‘left school’ could constitute a noun phrase (labelled ‘[N left school N]’), or alternatively could be a verb followed by a noun phrase object/complement (labelled ‘[V left V][N school N]’).

Grammatical Inference can supply efficient descriptions or models of contextual constraints and preferences at each of these levels. These models do not eliminate ‘illegal’ (ungrammatical) analyses, but rather provide a numeric evaluation and rank ordering of possible analyses. Whether some or all but one analysis are subsequently eliminated (eg by thresholding) depends on the application.

An application: Speech, Handwriting, and Optical Character Recogition. Ambiguity resolution is a central issue in correct transcription of input from speech, handwriting and Optical Character Recognisers. Output from an English recognition system (be it speech, handwriting or optical character) is typically a sequence of candidate sets, referred to as a recognition lattice. For example, on ‘hearing’ the sentence "Stephen left school last year", an English speech recognition system may produce the following lattice of candidates (in order of decreasing similarity to input speech signal):

stephen	stiffen	stiffens
left	lift	loft
school	scowl	scull
lest	last	least
yearn	your	year

To disambiguate such a lattice, a standard technique is to use a language model to constrain the possible choices, so that the chosen sequence of words is the most linguistically plausible. Most language models for lattice disambiguation provide only a limited coverage of the linguistic knowledge available - restricted to word- or wordtag n-grams [Jelinek 90]. Analysis of recognition lattices involves traversing a much larger search space than when analysing sentences, and the necessity of real-time computability acts as a constraint on language model complexity. Because of this, sophisticated language analysis systems have not been successful in disambiguating recognition lattices. [Keenan 92] found that the ANLT parser was too powerful for such a task, requiring long computation times to discover a very large number of ambiguous analyses of even simple sentences. There is a clear need for a language model incorporating a broader range of linguistic knowledge than word and wordtag n-grams, while remaining computationally feasible.

Probabilistic coocurrence preference models as grammars. N-grams or Markov Models are a conceptually simple mathematical means for representing an observable, real-world sequence of events or symbols. They are equivalent to a non-deterministic finite automata, where the transition from the current symbol to the next is determined by probability, based upon a small fixed-size ’window’ of previous symbols. Markov theory is computationally efficient and provides simple but very general and powerful models for applications throughout science. In NLP, common applications are in speech-processing (symbols are "acoustic chunks" such as phonemes) and in grammatical tagging (symbols are part-of-speech wordtags). However, Markov models are not readily applicable to higher levels of linguistic analysis (eg semantics) involving links between units beyond a small fixed-size window. Collocations are a variation on Markov models or n-grams, more appropriate to language modelling. An n-gram model records ALL n-length symbol-sequences in the training set; for example, a word bigram model records all pairs of words in the training set (and their frequencies of occurrence), while a word trigram model records all word-triples. A word collocation model records combinations which occur together significantly more frequently than predicted by their probabilities in isolation (using some application-specific measure of significance); as only significant combinations (and their frequencies) are recorded, a much larger ’window’ can be used than for a strict n-gram model of equivalent size. In semantic analysis, coocurrence constraints and preferences are rarely confined to immediate neighbours: a sense of one word can ‘match up’ with a sense on any other word in the sentence. Looser generalisations beyond pure Markov models allow us to explore this.

N-grams, collocational models, and other probabilistic coocurrence preference models of that ilk all have the advantage of being automatically extractable from appropriate training data. Such a model can be automatically inferred using a variety of large-scale linguistic resources (tagged corpus, treebank and machine readable dictionary). This is in contrast to many other NLP systems, where linguistic knowledge is supplied from intuition. We recognise that each individual Model is theoretically and practically inadequate as a model of linguistic knowledge, but believe that taken as a combination they will provide a holistic model of constraints sufficient for at least some applications. For English speech or handwriting recognition, the optimal analysis is not required to be fully correct at all levels; its purpose is to indicate the correct words.

Multi-level grammar for disambiguation and recognition. Several researchers from CCALAS are presenting posters at the Grammatical Inference colloquium giving further details of their individual contributions to the problem of English disambiguation and recognition. The following is an overview of the components. The levels of ‘grammatical description’ being inferred from Corpus, Treebank, and Machine Readable Dictionary are:

1. LDOCE Semantic Primitives:
All word sense-definitions in the Longman Dictionary Of Contemporary English (LDOCE) are written in terms of the "Longman Defining Vocabulary" (LDV). This is a closed set of approximately 2000 words, which effectively constitute semantic primitives. [Demetriou 93] shows that the definition of a word can be used as the basis for semantic constraints, by maximising overlap between definitions. When a word has more than one sense, we order the candidate senses according to the number of LDV-primitives which overlap or also occur in senses of other words in the sentence.

2. Semantic Subject Field Markers:
LDOCE also has a set of Subject Field Markers (SFMs) which provide a taxonomic semantics, at a higher level of abstraction than the sense-definitions. Each word is associated with a small number of Subject Field Markers, and [Jost 93], [Jost and Atwell 93] show that these can be used as semantic tags in a disambiguation algorithm based on a probabilistic cooccurrence preference model. The model is embodied in a table of pairs of SFMs which cooccur in sentences in the training Corpora (LOB Corpus and New Scientist Corpus, together nearly two million words). This is NOT a Markov table, as a pair is NOT required to cooccur as immediate neighbours: semantic preferences can span a whole sentence. This is a grammar in a very loose sense, as it does not capture linear precedence rules or neighbour

constraints; however it does meet the Grammatical Inference Cooloquium definition of being "an efficient description that captures the essence of a set of data, and can be used subsequently to classify data": the model can successfully select the correct sense of a word according to context.

3. Syntactic Phrase Structure Boundaries

Given a parsed Corpus, a straightforward way to infer a grammar is to extract every non-terminal node label and its sequence of immediate daughters, and store these as lhs and rhs respectively of a context-free rule. However, [Atwell 88b] showed that this resulted in a huge, unwieldy context-free grammar: even after deletion of multiple copies of rules (corresponding to repeated occurrence of the same grammatical construct in the Treebank), a Treebank of c50,000 words yielded c8,500 context-free rules, too many for the parser-generator investigated. So, we turned to a simpler, smaller model of English Treebank syntax. A two-dimensional phrase-structure syntax tree can be transformed into a one-dimensional sequence of labelled brackets; this can be captured in a simpler Markovian model. A Markovian parser derived from the Spoken English Corpus (SEC) Treebank has been developed at Leeds [Atwell 83, Pocock 91, Atwell & Pocock 92a,b,93, Pocock and Atwell 93], for the M.O.D funded Speech-Oriented Probabilistic Parsing (SOPP) project. The model used is a variant of standard Markov theory, in that both the training set and desired output are required to be an alternating sequence of wordtags and labelled bracket combinations. The parser implementation uses this adapted model for a "bracket-insertion" procedure, augmented with a collocational "tree-closing" procedure to ensure parse trees are correctly balanced. With experiments in parsing lattices using equivalent sized training sets, the Markov Model based parser is much faster and more robust than a probabilistic chart parser developed as part of the SOPP project. Its optimal parsetree is unlikely to be structurally perfect, but it dominates the correct word-sequence, which is adequate for lattice disambiguation.

4. Wordtag n-grams:

These are widely used in handwriting, speech and optical character recognition, e.g [Jelinek 90], [Keenan 92], [Hanlon and Boyle 91]. They have also been successfully for error-recognition in Word Processor input, by artificially "ambiguating" the input into a lattice and then selecting the best candidate according to context, in much the same way as for a speech or handwriting input lattice [Atwell 87a], [Atwell and Elliott 87]. Wordtag n-gram models were originally used for the automatic part-of-speech tagging of corpora [Atwell 83, Owen 87]. [Leech et al 83, Atwell et al 84] describes the CLAWS system for tagging the LOB Corpus [Johansson et al 86]. CLAWS was the first NLP system to go beyond a Markov model to wider collocations: the "augmented first-order model" [Atwell 83] added only 'significant' trigrams to the core bigram model, avoiding the size and computational problems of a full trigram model.

5. Word-Collocational Preferences:

Lexicographers have long known that word-collocations are an alternative source of lexical semantic patterns or constraints [Sinclair 87]. [Rose & Evitt 92] use word-collocations as a readily trainable surrogate for traditional NLP semantics in disambiguation of handwriting lattices. Word collocations are recognised within English Language Teaching (ELT) and applied linguistics as indicators of native speaker naturalness; [Howarth 93] applied statistical significance measures of coocurrence to extract word collocations from the LOB Corpus, but found that grammatical inference has to be supplemented with linguist's intuition in finding 'natural' collocations in English texts.

6. The relationship between prosody and syntax:

Prosody is the pattern of stress and intonation in spoken language. There is a complex interrelationship between prosody and syntax in spoken English, and a computational model of this can be applied in both speech recognition and generation. The Spoken English Corpus (SEC) is a transcribed corpus of BBC radio broadcasts which contains both prosodic and syntactic annotations, added by expert linguists. [Arnfield and Atwell 93] presents a grammar of coocurrences of wordtags and prosodic markers, inferred from the SEC; and shows that the placement of stresses on words in an utterance may be largely predicted from syntactic wordclasses. This leads us to the idea of prosodically-oriented wordtags. Current tagsets used in tagged corpora, such as the LOB corpus tagset [Johansson et al 86], are large and finely-grained, with well over a hundred wordclasses; this causes computational problems, and also means very large training sets are required for Grammatical Inference, to ensure adequate representation of all wordtags. For research on spoken English, it might be advantageous to group together wordtags on grounds of prosodic similarity, to derive a smaller more manageable tagset.

7. Wordtag Inference:
Research on grammatical inference of wordtag n-gram models, mentioned above, relies on intuitively-defined wordtag sets; but in practice linguists' intuitions rarely agree, and there are many different competing word classification sets (see, for example, [Sampson 87]). Furthermore, as word-classes are defined on intuitive grounds rather than in terms of some formal model, grammatical inference based on n-gram models can only ever provide an approximation of a 'correct' grammar. An alternative is to seek a classification of words purely based on computable coocurrence and collocational measures; this tagset can then be used in a 'purer' wordtag n-gram model. Research on inference of word-classification systems, such as [Atwell 87, Atwell and Drakos 87, Atwell and Elliot 87, Hughes and Atwell 93, Finch 93] might seem to be orthogonal to the aims of mainstream Grammatical Inference researchers; but we foresee that paradigmatic and syntagmatic relationships can be combined in a fully inferrable n-gram model which does not require a pre-tagged training corpus.

Future Work We plan to implement a generic system for disambiguating English recognition system lattices, using the above probabilistic language model components to provide linguistic constraints. This will traverse a sentence-lattice using dynamic programming: for each candidate-set, the algorithm will find relative probabilities for each candidate word-hypothesis with respect to the given language model. The different linguistic constraint levels will then be integrated in a holistic lattice disambiguation model. As the interface between each component and the lattice is via assignment of independent probabilities to word-hypotheses, there is no need for sophisticated blackboard or inter-process communication. Dynamic programming lattice-traversal modules for each level will execute independently (conceptually in parallel, with separate windows on the same section of the lattice); each word-hypothesis will be annotated with a set of probabilities (one with respect to each level); these probabilites are then combined by multiplication, or some more sophisticated function to be refined empirically.

It may be very costly to compute all levels in parallel at run-time. An altrnative possible solution, explored by [Ueberla 93a,b], is to use a technique called Classification And Regression Trees (CART) to combine knowledge sources more efficiently. This requires a systematic study of the influence of different types of knowledge sources on the quality of a language model. The CART technique will yeild an efficient combination of language models: at each point during analysis of new input, CART chooses the specific submodel that will most help the prediction of the next word and minimise perplexity.

As a resource for evaluating the success of the implemented lattice disambiguation system, we propose to collect together recognition lattices (along with the correct sequence of words for each lattice) from the NLP / Pattern Recognition research community. As the lattices will be gathered from many different sources, we plan to devise a standard format, to which all lattices will be converted. We will consult with the research community via SALT (UK) and ELSNET (European) Networks of Excellence, both in the gathering of lattice data and in formulation of formatting standards; these must also conform to Text Encoding Initiative (TEI) Guidelines. This will also have the side effect of enabling us to build links with researchers who may be potential customers for our work. The Lattice Corpus will be the first of its kind. To be reasonably representative a large sample is required. Initially we aim for a number of recognition lattices equivalent to 50,000 words, which is comparable in size to current parsed corpora eg. Polytechnic of Wales (POW), Spoken English Corpus (SEC) and Lancaster-Leeds Treebanks. The Corpus may become a standard test resource, distributed it through text archives and file servers including ICAME at Bergen, and Oxford. Such a Lattice Corpus will also be a rich training data source for further Grammatical Inference research, with direct practical applications in speech, handwriting and optical character recognition.

References

Arnfield, Simon and Eric Atwell. 1993. A syntax based grammar of stress sequences. to appear in Grammatical Inference: theory, applications, and alternatives: Colloquium Digest, Institution of Electrical Engineers.

Atwell, Eric Steven. 1983. Constituent Likelihood Grammar. ICAME Journal 7: 34-67.

Atwell, Eric Steven, Geoffrey Leech and Roger Garside. 1984. Analysis of the LOB Corpus: progress and prospects. In Corpus Linguistics: Proceedings of the ICAME 4th International Conference. Jan Aarts and Willem Meijs (eds.). 40-52, Amsterdam: Rodopi.

Atwell, Eric Steven. 1987a. Constituent Likelihood Grammar. In (Garside et al. 1987): 57-65.

Atwell, Eric Steven. 1987b. A parsing expert system which learns from corpus analysis. In Corpus Linguistics and Beyond: Proceedings of the ICAME 7th International Conference. Willem Meijs (ed.). 227-235. Amsterdam: Rodopi.

Atwell, Eric Steven and Stephen Elliot. 1987. Dealing with ill-formed English text. in (Garside et al. 1987): 120-138.

Atwell, Eric Steven Atwell and Nikos Drakos. 1987. Pattern Recognition Applied to the Acquisition of a Grammatical Classification System from Unrestricted English Text. In Bente Maegaard (ed), Proceedings of the Third Conference of European Chapter of the Association for Computational Linguistics, pp56-63, New Jersey, Association for Computational Linguistics.

Atwell, Eric Steven. 1987. How to detect grammatical errors in a text without parsing it. In Bente Maegaard (ed), Proceedings of the Third Conference of European Chapter of the Association for Computational Linguistics, pp38-45, New Jersey, Association for Computational Linguistics.

Atwell, Eric Steven. 1988a. Grammatical analysis of English by statistical pattern recognition. In Josef Kittler (ed), Pattern Recognition: Proceedings of the 4th International Conference, Cambridge: 626-635. Berlin: Springer-Verlag.

Atwell, Eric Steven. 1988b. Transforming a Parsed Corpus into a Corpus Parser. In Merja Kyto, Ossi Ihalainen and Matti Risanen (eds), Corpus Linguistics, hard and soft: Proceedings of the ICAME 8th International Conference. 61-70. Amsterdam: Rodopi.

Atwell, Eric Steven, David Hogg, and Robert Pocock. 1992. Speech-Oriented Probabilistic Parser Project: Interim Reports 1&2 A.I. Division, School of Computer Studies, Leeds University.

Atwell, Eric Steven, and Robert Pocock. 1992. Speech-Oriented Probabilistic Parser Project. Presented to ICAME 13th International Conference, Nijmegen.

Atwell, Eric Steven. 1992. Overview of Grammar Acquisition Research. In Workshop on sublanguage grammar and lexicon acquisition for speech and language: proceedings. Henry Thompson (ed.). 65-70, HCRC, Edinburgh University.

Atwell, Eric. 1993. Corpus-Based Statistical Modelling of English Grammar. in (Souter & Atwell 93): 195-215.

Chomsky, Noam. 1957. Syntactic Structures. The hague: Mouton.

Demetriou, George. 1993. Lexical Disambiguation Using CHIP (Constraint Handling In Prolog). In Proceedings of 1993 European conference of the Association for Computational Linguistics, Utrecht (forthcoming).

Finch, Steve. 1993. Finding structure in language. Unpublished PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Garside, Roger, Geoffrey Leech and Geoffrey Sampson (eds.). 1987 The Computational Analysis of English: a corpus-based approach. London and New York: Longman.

Hanlon, Stephen, and Roger Boyle. 1991. Syntactic knowledge in word-level text recognition. Research Report, School of Computer Studies, Leeds University.

Howarth, Peter. 1993. Automatic analysis of conventional language in written English. Seminar paper, CCALAS, Leeds University.

Hughes, John and Eric Atwell. 1993. Automatically acquiring and evaluating a classification of words. to appear in Grammatical Inference: theory, applications, and alternatives: Colloquium Digest, Institution of Electrical Engineers.

Guthrie, Louise. 1993. A Note on Lexical Disambiguation. in (Souter & Atwell 93): 217-239.

Jelinek, Fred. 1990. Self-organized language modeling for speech recognition. In Alex Waibel and Kai-Fu Lee (eds). Readings in Speech Recognition: 450-506. Morgan Kaufmann.

Johansson, Stig, Eric Steven Atwell, Roger Garside and Geoffrey Leech. 1986. The Tagged LOB Corpus. Norwegian Computing Centre for the Humanities, Bergen University.

Jost, Uwe. 1993. A coocurrence model of LDOCE subject fields for semantic disambiguation. Undergraduate project report, School of Computer Studies, Leeds University.

Jost, Uwe and Eric Atwell. 1993. Deriving a probabilistic grammar of semantic markers from unrestricted English text. to appear in Grammatical Inference: theory, applications, and alternatives: Colloquium Digest, Institution of Electrical Engineers.

Keenan, Francis. 1992. Large vocabulary syntactic analysis for text recognition. Unpublished PhD thesis, Dept of Computing, Nottingham Trent University.

Leech, Geoffrey, Roger Garside, and Eric Steven Atwell. 1983. The automatic grammatical tagging of the LOB corpus. ICAME Journal 7: 13-33.

Owen, Marion. 1987. Evaluating automatic grammatical tagging of text. ICAME Journal 11: 18-26.

Pocock, Robert. 1991. The construction and evaluation of a Markov Model based parser Undergraduate project report, School of Computer Studies, Leeds University.

Pocock, Robert and Eric Atwell. 1993. Probabilistic Syntax Models For Treebank - Trained Lattice Parsing. submitted to 1993 Association for Computational Linguistics conference.

Rose, Tony, and Lindsay Evitt. 1992. A large vocabulary semantic analyser for handwriting recognition. AISB Quarterly. 80: 34-39.

Sampson, Geoffrey. 1987. Evidence against the "Grammatical"/"Ungrammatical" Distinction. In Corpus Linguistics and Beyond: Proceedings of the ICAME 7th International Conference. Willem Meijs (ed.). 219-226. Amsterdam: Rodopi.

Sinclair, John (ed). 1987. Looking Up: an account of the COBUILD project in lexical computing. London: Collins.

Souter, Clive, & Eric Atwell (eds). 1993. Corpus-Based Computational Linguistics. Amsterdam: Rodopi.

Ueberla, Joerg. 1993. State Language Models for Speech Recognition. Research Report, School of Computer Studies, Leeds University; also published as Technical Report, School of Computing Science, Simon Fraser University.

Ueberla, Joerg. 1993. Analysis of a simple bipos language model - attempt at a strategy to improve language models for speech recognition. to appear in Grammatical Inference: theory, applications, and alternatives: Colloquium Digest, Institution of Electrical Engineers.