ANNOTATING 200 MILLION WORDS: THE BANK OF ENGLISH PROJECT

Timo Järvinen
Research Unit for Computational Linguistics
University of Helsinki

Abstract

The Bank of English is an international English language project sponsored by Harper-Collins Publishers, Glasgow, and conducted by the COBUILD team at the University of Birmingham, UK. The text bank will comprise some 200 million words of both written and spoken English. The whole 200 million word corpus is being annotated morphologically and syntactically during 1993–94 at the Research Unit for Computational Linguistics (RUCL), University of Helsinki, using the English morphological analyser (ENGTWOL) and English Constraint Grammar (ENGCG) parser. The first half of the texts (103 million words) has already been processed in 1993. The project is lead by Prof. John Sinclair in Birmingham, and Prof. Fred Karlsson in Helsinki. The present author is responsible for conducting the annotation.

In the introduction of this paper the routines for dealing with large text corpora are presented and our analysing system outlined. Chapter 2 gives an overlook how the texts are preprocessed. Chapter 3 describes the lexicon updating, which is a preliminary step to the analysis. The last part presents the ENGCG parser and the ongoing development of its syntactic component.

1 INTRODUCTION

Each month the COBUILD team supplies an approximately 10 million word batch of markup coded running text (see Appendix A) in ASCII format. Every new batch is first scanned by the ENGTWOL lexical and morphological analyser [Koskenniemi, 1983] in filtering mode for the purpose of detecting words not included in the present lexicon. This is followed by a semi-automatic updating of the lexicon. After these adjustments, the whole system is used for annotating the data.

Our analysing system, which is presented in detail in [Karlsson, 1994], consists of the following successive stages:

- preprocessing
- ENGTWOL lexical analysis
- ENGCG morphological disambiguation
- ENGCG syntactic mapping and disambiguation

The main routines performed on the monthly data, including constant monitoring of both incoming texts and analysed output and management (documentation, backups) are closely linked to the updating of the preprocessing module and the ENGTWOL lexicon.

2 PREPROCESSOR

The preprocessing modules standardise the running text and tokenise it into a form suitable for the ENGTWOL lexical analyser.

ENGCG has been developed so that it takes into account various textual coding conventions [Karlsson, 1994]. We have developed preprocessing procedures further to cater for the different types of markup codes systematically. Since texts usually come from various sources, there may be undocumented idiosyncracies or systematic errors in some samples.

The information conveyed by the markup codes is utilised in the parsing process. Updating the preprocessing module to achieve the highest possible systematisation is therefore considered worthwhile. The present system can deal with any code properly if it is used unambiguously in either a sentence-delimiting function (e.g codes indicating headings, paragraph markers), sentence-internal function (e.g. font change codes) or word-internal (e.g. accent codes) function.

Since preprocessing is the first step before lexical filtering, it indicates the kinds of difficulties we are likely to encounter. If error messages are produced at this stage, I do the necessary adjustments to the preprocessor until it seems to produce the output smoothly. Errors in preprocessing may occasionally result in a truncation of lengthy passages of text or even a crash.

It is important for the utilisation of the corpus that no information is lost during standardisation. Therefore, we aim to mark all corrections made to the text. For example, the preprocessor inserts a code marking the correction when it separates strings such as *ofthe* and *andthe*.

Most errors are not corrected, such as confusion of sentence boundaries, truncation of sentences due to running headings or page numbers, misplacement or doubling of blocks of text, etc.

3 THE LEXICON

Filtering produces a list of all tokenised word-forms in the input text which are not included in the current ENGTWOL lexicon. The most common types are taken under closer scrutiny. It has to be decided whether these are genuine word forms or non-words (e.g. misspellings).

At the beginning, I used several days to update the lexical module for a new batch of text but experience and increased coverage of the lexicon have diminished the time needed for this task considerably. I have added words above a certain frequency routinely to the ENGTWOL lexicon. The frequency is not fixed but determined by practical considerations. For instance, when the data contain a great deal of duplication (as in the BBC material owing to the repetitive nature of daily broadcasting), simple token frequency is a poor indicator of what is a suitable item to add to the lexicon. However, sampling methods have not been developed to optimise the size of the lexicon, because it is not crucial for the present purpose.

My lexical practices differ somewhat from the updating procedure documented in [Voutilainen, 1994]. If our aim is to supply every word in running text with all proper morphological and syntactic readings, we cannot deprive frequent nonstandard words (e.g. *larn, veggie, wanna*) of their obvious morphological readings because this might cause the whole sentence to be misanalysed. Since prescriptive considerations were not taken into account in the design of ENGTWOL, many entries marked as informal' or lang' in conventional dictionaries were added to the lexicon. I have also included highly domain-specific entries into the lexicon if they were frequent enough in certain types of data, especially when heuristics might produce erroneous or incomplete analyses for the word in question (e.g. species of fish which have the same form in singular and plural: *brill, chub, garfish*)[1].

One advantage of including all frequent graphical words to the lexicon is that ENGTWOL filtering of incoming texts produces output which can be more reliably dealt with by automatic means. When all frequent nonstandard and even foreign words are listed in the lexicon, the output can be used in a straightforward way for generating new entries.

The procedure of adding new entries to the lexicon goes as follows: first, all words are classified according to the part-of-speech they belong to. Second, new entries in the ENGTWOL format are generated automatically from these word-lists using ready-made tools presented in [Voutilainen, 1994]. Lists of new entries are carefully checked up, and additional features (such as transitivity and complementation features for verbs) are supplied manually. In describing the items, I have relied mainly on Collins COBUILD Dictionary (1987) and Collins English Dictionary (1991) which have been available for us in electronic form. But when the usage and distribution seems to be unclear, I have generated an on-line concordance directly from the corpus. Since I have dealt with words which have a frequency of, say, at least 10 tokens in the corpus, this method seems to be quite reliable.

We cannot detect errors in the lexicon during the initial filtering phase. Once a certain string has had one or more entries in the lexicon, it is not present in the output of the filtering, and other potential uses might not be added to the lexicon[2]. And frequent errors tend to get corrected since all incorrect analyses detected during the manual inspection are corrected directly in the lexicon.

The ENGTWOL lexicon which is used in the Bank analyses contains approximately 75,000 entries. Morphological analysis caters for all inflected forms of the lexical items. The coverage of the lexicon before updating is between 97% – 98% of all word-form tokens in running text. Appendix A presents the number of additional lexical entries generated from each batch of data. The cumulative trend shows that a very small number of new entries is needed when analysing the latter half of the corpus.

Morphological heuristics is applied after ENGTWOL analysis as a separate module (by Voutilainen, Tapanainen). It assigns reliable analyses to words which were not included in the lexicon. This also contributes to the fact that lexicon updating will be a minor task in the future.

4 ENGCG DISAMBIGUATION AND SYNTAX

English Constraint Grammar is a rule-based morphological and dependency-oriented surface syntactic analyser of running English text.

Morphological disambiguation of multiple part-of-speech and other inflectional tags is carried out before syntactic analysis. Morphological disambiguation reached a mature level well before the beginning of this project (see evaluation in [Voutilainen, 1992]).

The morphological disambiguation rules (some 1100 in the present grammar) were written by Atro Voutilainen. The Bank data is analysed using both grammar-based' and heuristic' disambiguation rules. This leaves less morphological ambiguity (below 3%), although the error rate is still extremely low (below 0.5%).

[1] The default category of morphological heuristics is a singular noun. In the case of a potential plural form (s-ending), an underspecified tag SG/PL is given.

[2] Although missing entries are possible to find indirectly, e.g. *-ing* and *-ed* forms in the filtering output indicates that the base form is not described in the lexicon as a verb

4.1 Current state of ENGCG syntax

The first version of ENGCG syntax was written by Arto Anttila [Anttila, 1994]. At the beginning of the Bank project, new Constraint Grammar Parser implementations for syntactic mapping and disambiguation were written by Pasi Tapanainen. These have been tested during the first months of this project. Some adjustment to the syntax was needed to cater for new specifications, e.g. in rule application order.

I have tested all constraints extensively with different types of text from the Bank. I have revised almost all syntactic rules and written new ones. The current ENGCG parser uses 282 syntactic mapping rules, 492 syntactic constraints and 204 heuristic syntactic constraints. The mapping rules should be the most reliable, since they attach all possible syntactic alternatives to the morphologically disambiguated output. Syntactic rules prune contextually inappropriate syntactic tags, or accept just one contextually appropriate tag. Syntactic and heuristic rule components are formally similar but they differ in reliability. It is possible not to use heuristic rules at all if one aims at maximally error-free output, but the cost is an increase in ambiguity.

During the project, the quality of syntax has improved considerably. The current error rate, when parsing new unrestricted running text, is approximately 2%, i.e., 2 words out of 100 get the wrong syntactic code. But the ambiguity rate is still fairly high, 16.4% in a 0.5m word sample, which means that 16 words out of 100 still have more than one morphological or syntactic alternative. Much of the remaining ambiguity is of the prepositional attachment type. This particular type of ambiguity accounts for approximately 20% of all remaining ambiguity. More heuristic rules are needed for pruning the remaining ambiguities. Of course, many of the remaining ambiguities (especially PP attachment) are genuine and should be retained.

The speed of the whole system used in morphological and syntactic annotation is about 400 words per second on a SUN SparcStation 10/30.

4.2 Developing the syntax

Facilities for the fast compilation of a parser with a new rule file and the speed of the analysis makes a very good environment for the linguist to test new constraints.

A special debugging version of the parser can be used for testing purposes. The debugging version takes fully disambiguated ENGCG texts as input. Ideally, every rule is tested against a representative sample from a corpus. This would set the requirement that the test corpus should be made of large random samples. However, it is time-consuming to prepare manually large amounts of corrected and disambiguated data, even from ENGCG output. Therefore, a very large test corpus is beyond the scope of this project.

The current syntactic test corpus contains approximately 30,000 words. It is large enough for testing reliable syntactic rules, but if we want to rate the acceptability of heuristic syntactic rules, a larger syntactic corpus would be necessary. The test corpus consists of 16 individual text samples from the Bank of English data. The texts have been chosen so that they take text type variation into account. All samples but one are continuous, unedited subparts of the corpus.

It seems worthwhile to continue preparing a disambiguated corpus from selected pieces of text. Once new data is received, it is expedient to add a representative sample from it to the test corpus. A manually disambiguated test corpus constitutes a very straightforward documentation of the applied parsing scheme (as described in [Sampson, 1987]).

5 CONCLUSION

The analysing system has reached a mature stage, where all technical problems seem to be solved. We have developed methods dealing with the data with a considerable degree of automatisation. ENGCG has proved to be a fast and accurate rule-based system for analysing unrestricted text.

Writing and documenting ENGCG syntax will be the main concern during the following months. Our part of the project will be completed by March, 1995.

It is possible that the whole 200-million corpus will be analysed afresh near the end of the project. This would put to use all the improvements made during the two-year period and would guarantee a maximal degree of uniformity and the overall accuracy of the annotated corpus.

6 ACKNOWLEDGEMENTS

Special thanks are due to Harper Collins Publishers, Glasgow, for permission to use both Collins COBUILD and Collins English Dictionary in electronic form. Personally, I am greatly indebted to Pasi Tapanainen for solutions to an incalculable number of technical problems and to Atro Voutilainen for guidance and supervision during this project. I wish to thank also Prof. Fred Karlsson, Juha Heikkilä, Kari Pitkänen and Sari Salmisuo for reviewing earlier drafts of this paper.

A List of annotated Bank of English data

data	size in words	additional lexical entries
Today	10,019,195	6,540
Times	10,090,991	1,837
BBC	18,076,124	3,379
The Economist, WSJ	11,195,100	455
British Books 1	9,232,527	1,488
British Books 2	13,925,852	1,961
Independent, Magazines	10,199,542	1,143
Magazines	10,365,173	1,059
American books	10,532,267	972
Total:	103,636,771	18,834

The table above shows the size of the 11 batches annotated so far in words and the number of new lexical entries[3] derived from them.

B An Example of the ENGCG analysed sentence (from the American Books data)

The original text:

```
<t>
The situation at Stanford, to be examined
in more detail later, is hardly unique.
```

Annotated text:

```
<t>
"<The>"
  "the" <*> <Def> DET CENTRAL ART SG/PL @DN>
"<situation>"
  "situation" N NOM SG @SUBJ
"<at>"
  "at" PREP @<NOM
"<Stanford>"
   "stanford" <*> <Proper> N NOM SG @<P
"<$,>"
"<to>"
  "to" INFMARK> @INFMARK>
"<be>"
  "be" <SV> <SVC/N> <SVC/A> V INF @-FAUXV
"<examined>"
  "examine" <SVO> <P/in> PCP2 @-FMAINV
"<in>"
  "in" PREP @ADVL
"<more>"
  "much" <Quant> DET POST CMP SG @QN>
"<detail>"
  "detail" N NOM SG @<P
"<later>"
  "late" ADV CMP @ADVL
"<$,>"
"<is>"
  "be" <SV> <SVC/A> V PRES SG3 VFIN @+FMAINV
"<hardly>"
  "hardly" ADV @ADVL @AD-A>
"<unique>"
  "unique" A ABS @PCOMPL-S
"<$.>"
```

Syntactic tags, listed in [Tapanainen, 1994; Voutilainen, 1992] are marked with an at-sign (@). The shallow syntax distinguishes four verb chain labels and nominal head and modifier functions. Modifier functions have a pointer (> or <) to the head to the right or to the left, respectively. PP and adverbial attachment is solved when it can be done reliably.

[3]The same WSJ material from ACL has been used in updating the ENGTWOL lexicon before this project

References

[Anttila, 1994] Arto Anttila. 1994. How to recognise subjects in English. In *Karlsson & al* 1994.

[Garside, 1987] Roger Garside, Geoffrey Leech and Geoffrey Sampson. 1987. The Computational Analysis of English - A Corpus-Based Approach. London: Longman.

[Karlsson, 1994] Fred Karlsson. 1994. Robust parsing of unconstrained text. In *Nelleke Oostdijk and Pieter de Haan (eds.), Corpus-based Research Into Language.*, pp. 121–142, Rodopi, Amsterdam-Atlanta.

[Karlsson, 1994] Fred Karlsson. 1994. The formalism and Environment of CG Parsing. In *Karlsson & al* 1994.

[Karlsson, 1994] Fred Karlsson, Atro Voutilainen, Juha Heikkilä and Arto Anttila (eds.). 1994. Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text. Berlin/New York:Mouton de Gruyter.

[Koskenniemi, 1983] Kimmo Koskenniemi. 1983. Two-level morphology: a general computational model for word-form recognition and production. Publications nro. 11. Dept. of General Linguistics, University of Helsinki. 1983.

[Sampson, 1987] Geoffrey Sampson. 1987. The grammatical database and parsing scheme. In *Garside* 1987, pp. 82–96.

[Tapanainen, 1994] Pasi Tapanainen and Timo Järvinen. Syntactic analysis of natural language using linguistic rules and corpus-based patterns. In *proceedings of COLING-94*. Kyoto, 1994.

[Voutilainen, 1992] Atro Voutilainen, Juha Heikkilä and Arto Anttila. 1992. Constraint grammar of English. A Performance-Oriented Introduction. Publications No. 21, Department of General Linguistics, University of Helsinki.

[Voutilainen, 1994] Atro Voutilainen and Juha Heikkilä. 1994. Compiling and testing the lexicon. In *Karlsson & al* 1994.