An Inheritance-based Lexicon for Message Understanding Systems

Lynne J. Cahill*
School of Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton BN1 9QH, UK
lynneca@cogs.susx.ac.uk

Introduction

POETIC (POrtable Extendable Traffic Information Collator)[1] (Gaizauskas et al., 1992) is a prototype system which analyses police reports of traffic incidents, builds a picture of the incident and broadcasts advisory messages automatically to motorists if necessary. The front end of the system can be viewed as a message understanding system, comprising two distinct components: a message analyser which is essentially a chart parser and which returns predicate calculus type semantic representations of fragmented parses of the input, and a discourse interpreter, which puts the fragmented parser output back together, and incorporates the new information into the knowledge it already has about the incident.

The message understanding part of the system was adapted to the domain of commercial joint ventures (henceforth JV) and entered for the fifth message understanding conference competition, sponsored by ARPA[2] (Gaizauskas et al., 1994). On the principal evaluation metric, the system fell in the third rank of seven statistically significant rankings with only three of the thirteen systems in its group performing significantly better, a pleasing result given the short time spent on the conversion to a completely different domain.

One of the main aims of the POETIC project was to develop an existing system (the TIC – Traffic Information Collator) to make it more readily portable to new police force sub domains, and increase extendability thus improving ease of maintenance. The level of success of this aim was tested by the conversion to the JV domain. The approach taken was to extract all domain specific knowledge into declarative knowledge bases and to develop these knowledge bases in such a way as to make them easily adaptable.

Naturally, one of the main areas of domain specific knowledge was the lexicon, which had to provide the, occasionally very specialised, words and expressions specific to the domain in question. In this paper, we discuss the lexicon system developed in POETIC and its conversion to use in the JV task.

The input to the POETIC system was verbatim police radio reports of traffic incidents, frequently in non-standard, ungrammatical or telegrammatic English, with extensive use of jargon and abbreviations. For the MUC-5 task, the input was "full" English newswire reports.

The parsing process

One of the novel aspects of the POETIC system is its overall approach to the parsing process. While a full parse of each input string is attempted, it is not required, or even expected. The parser returns fragmented analyses, which are then incorporated by a knowledge-based discourse interpreter into an overall picture of the incident being analysed. This means that the grammar is not required be able to cope with *all* possible input constructions, and that the lexicon does not have to have anything like total coverage. This was vital for the POETIC task since the input is frequently not in grammatical English, and spelling errors and typos, as well as new/unknown words are likely to occur, but much less likely to be needed.

The three-tier lexicon

In POETIC, a three tier lexicon system was used, in order to maximise modularity and minimise lookup in very large wide coverage lexicons. The first and smallest of the tiers consisted of the lexicon specific to an individual police force sublanguage. The language used by UK police forces is largely the same, but there are a few, often crucial, differences. For instance, the Sussex police force use the word 'black' to describe a fatal accident; the Metropolitan police force, in contrast, use the word 'black' to describe severe traffic congestion.

The second tier contained words which were specific to the traffic domain but shared across police forces, such as 'rta' (road traffic accident) and 'hgv' (heavy goods vehicle). These first two tiers were consulted in the first stage of parsing, and all possible analyses with these words were found. Only then was the third tier consulted, a general English lexicon containing basic syntax for around 7000 common English words. In order to prevent excessive consultation of this lexicon, those very common words in the data were included in the second tier.

* I would like to acknowledge the contribution to this work of my colleagues on the project, Roger Evans and Robert Gaizauskas.

[1]The POETIC project was funded jointly by the UK SERC and the DTI, under grant number IED4/1/1834, with Racal Research Ltd, the Automobile Association and NTL.

[2]Sussex participation supported by ARPA, the University of Sussex, Racal Research Ltd. and Integral Solutions Ltd.

Porting to a new police force domain therefore meant just altering the first tier of the lexicon, which contained around 100 words. The second tier contained around 1000 words. The total lexical coverage was relatively small, being around 8000, but this was because of the overall parsing strategy.

For the JV domain, the three tier lexicon structure was not needed, simply two-tiers: domain specific and general English. After a simple word frequency analysis of the test corpus (around 400,000 words), all those words which appeared more than 100 times were included in the lexicon. Subsequently important words which had not reached that threshold were added.

In addition to these lexicons, there were databases of road and place names in POETIC and place and company names in the MUC-5 task. These had to be used with great care, due to their vastness and unreliability. Many important place names had several entries (e.g. Washington had 26) and some were the same as ordinary English words (e.g. 'Was', 'Of').

The inheritance based lexicon

The two domain specific tiers of the POETIC lexicon were written in DATR – an inheritance-based lexical representation language ((Evans and Gazdar, 1989a), (Evans and Gazdar, 1989b); for more about the development of the lexicon (Cahill and Evans, 1990), (Cahill, 1993)). The reasons for this were three-fold. First, one aim was to see how well suited the DATR language was to a relatively large-scale practical application. Secondly, it permitted the use of the two tiers without any implications for processing, since the two DATR theories could be compiled into a single lexicon for use at runtime. Thus, the domain specific part of the lexicon could be *maintained* separately, while being accessed as part of the main traffic lexicon. Finally, and most importantly, due to its hierarchical structure and inheritance mechanisms, the DATR language permitted much easier extension and adaptation of the lexicon, since changes affecting several entries could frequently be made at only one node at a high point in the hierarchy. Also, in a number of significant cases, it was possible to add a whole set of related entries very easily, only having to give minimal (sometimes even zero) individual information for each entry, all members of the set inheriting their main information from a common abstract node. Examples of this sort of thing in the POETIC domain are makes of car (e.g. 'Volvo'), all of which inherit all of their information from a single "CAR" node. In the JV domain, currencies inherit most of their information from a single "CURRENCY" node, with the individual currency name being the only piece of individual information.

Results

The lexicons used in the message understanding tasks described were both very small by most people's standards. The MUC-5 lexicon contained only 850 entries, while the POETIC lexicon contained just over 1000 entries. Even with the 7000-word general lexicon of English these are not large numbers by current thinking. The performance levels achieved with such small lexicons leads one to ask whether effort directed at constructing vast lexicons for NLP systems is genuinely worthwhile. Zipf's law states that, after a certain threshold, marginal cost (of increasing lexicon size) outweighs marginal utility (in terms of the frequency of occurrence of the additional entries). Although the 850 word lexicon for the MUC-5 task could undoubtedly be increased resulting in an improvement in performance of the system overall, the precise amount that it is worth increasing it by is debatable.

In the POETIC task, the question is even more glaring. There is a much broader range of information required in the MUC-5 task, and even though a very wide (possibly infinite) range of information may show up in the police logs, the range with which the POETIC system is expected to deal is strictly limited. It is extremely unlikely, therefore, that any significant improvement in the system as a whole would result from a great increase in the size of the lexicon.

What is clear is that far more important than the size and coverage of the lexicons used in such systems is the means of dealing with the cases of unrecognised words or phrases. The POETIC approach, fragmented parses pieced together by a knowledge driven discourse interpreter, can clearly be very effective. One advantage of such an approach is obvious – the time consuming and dreary task of adding thousands of lexical entries can be avoided. Even if automated lexical acquisition (which is not obviously feasible in many cases, such as the POETIC lexicon) can remove that problem, there is still the question of the efficiency of a system which must perform lexical lookup on a vast dictionary, followed by determining which of the possibly many analyses is the most appropriate.

References

Cahill, L. J. 1993. Some Reflections on the Conversion of the TIC lexicon to DATR. In Briscoe, de Paiva and Copestake (eds.) *Inheritance, Defaults and the Lexicon.*

Cahill, L. J. and R. Evans. 1990. An Application of DATR: The TIC Lexicon. In *Proceedings of the 9th European Conference on Artificial Intelligence*, pp. 120-125, Stockholm, 1990.

Evans, R. and G. Gazdar. 1989. Inference in DATR. In *Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics*, 1989.

Evans, R. and G. Gazdar. 1989. The semantics of DATR. In A. Cohn (ed.) *Proceedings of the Seventh Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, pp. 79-87, Pitman, London, 1989.

Gaizauskas, R., L. J. Cahill and R. Evans. 1994. Sussex University: Description of the Sussex System Used for MUC-5. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, Morgan Kaufmann, 1994.

Gaizauskas, R. J. and R. Evans and L. J. Cahill. 1992. POETIC: A System for Gathering and Disseminating Traffic Information. In *Proceedings of the International Conference on Artificial Intelligence Applications in Transportation Engineering, San Buenaventura, California*, June 1992.