TOPIC IDENTIFICATION TECHNIQUES FOR PREDICTIVE LANGUAGE ANALYSERS

J.I. Tait

University of Cambridge Computer Laboratory, Corn Exchange St., Cambridge CB2 3QG, England.

1. Introduction

The use of prediction as the basis for inferential analysis mechanisms for natural language has become increasingly popular in recent years. Examples of systems which use prediction are FRUMP (DeJong 79) and (Schank 75a). The property of interest here is that their basic mode of working is to determine whether an input text follows one of the systems pre-specified patterns; in other words they predict, to some extent, the form their input texts will take. A crucial problem for such systems is the selection of suitable sets of predictions, or patterns, to be applied to any particular text, and it is this problem I want to address in the paper.

I will assume that the predictions are organised into bundles according to the topic of the texts to which they apply. This is a generalisation of the script idea employed by (DeJong 79) and (Schank75a). I will call such bundles stereotypes.

The basis of the technique described here is a distinction between the process of suggesting possible topics of a section of text and the process of eliminating candidate topics (and associated predictions) which are not, in fact, appropriate for the text section. Those candidates which are not eliminated are then identified as the topics of the text

section. (There may only be one such candidate.) This approach allows the use of algorithms for suggesting possible topics which try to ensure that if the system possesses a suitable stereotype for a text section it is activated, even at the expense of activating large numbers of irrelevant stereotypes.

This technique has been tested in a computer system called Scrabble.

2. Suggesting Candidate Topics

The discovery of candidate topics for a text segment is driven by the association of a set of patterns of semantic primitives with each stereotype. (For the purposes of this paper it is assumed that the system has access to a lexicon containing entries whose semantic component is something like that used by (Wilks 77).) As a word is input to the system the senses of the word are examined to determine if any of them have a semantic description which contains a pattern associated with any of the system s stereotypes. If any do contain such a pattern the corresponding stereotypes are loaded into the active workspace of the system, unless they are already active.

3. Eliminating Irrelevant Candidates

In parallel with the suggestion process, the predictions of each stereotype in the active workspace are compared with the text. In Scrabble, the sentences of the text are first parsed into a variant of Conceptual Dependency (CD) representation (Schank 75b) by a program described in (Cater 80). The semantic representation scheme has been extended to include nominal descriptions similar in power to those used by (Wilks 77). The predictions are compared with the CD representation structures at the end of each sentence; but nothing in the scheme described in this paper could not be applied to a

system which integrated the process of parsing with that of determining whether or not a fragment of the text satisfies some prediction, as is done in (DeJong 79).

It is likely that stereotypes which are not relevant to the topic of the current text segment will have been loaded as a result of the suggestion process. Since the cost of the comparison of a prediction with the CD-representation of a sentence of the text is not trivial it is important that irrelevant stereotypes are removed from the active workspace as rapidly as possible. The primary algorithm used by Scrabble removes any stereotype which has failed to predict more of the propositions in incoming the text than it has successfully predicted. This simple algorithm has proved adequate in tests and its simplicity also ensures that the cost of removing irrelevant stereotpyes is minimised.

Further processing is subsequently done to separate stereotypes which were never appropriate for the text from stereotypes which were useful for the analysis of some part of the text, but are no longer useful.

4. An Example

Consider the following short text, adapted from (Charniak 78):

> Jack picked a can of tuna off the shelf. He put it in his basket. He paid for it and went home.

Assume that associated with the primitive pattern for food the system has stereotypes for eating in a restaurant, shopping at a supermarket, and preparing a meal in the kitchen. The lexicon entry for tuna (a large sea fish which is caught for food) will contain this pattern, and this will cause the loading of the above three stereotypes into the active workspace. The restaurant stereotype will not predict the first sentence, and so will immediately be unloaded. Both the supermarket and kitchen stereotypes expect sentences like

the first in the text. When the second sentence is read, the supermarket stereotype will be expecting it (since it expects purchases to be put into baskets), but the kitchen stereotype will not. However the kitchen stereotype will not be unloaded since, so far, it has predicted as many propositions as it has failed to predict. When the third sentence is read, again the supermarket stereotype has predicted propositions of this form, but the kitchen stereotype has not. Therefore the kitchen stereotype is removed from the active workspace, and the topic of text is firmly identified as a visit to the supermarket.

It should be noted that a completely realistic system would have to perform much more complex processing to analyse the above example. In such a system additional stereotypes would probably be activated by the occurrence of the primitive pattern for food, and it is likely that yet more stereotypes would be activated by different primitive patterns in the lexicon entries for the words in the input text.

5. Conclusions

The technique described in this paper for the identification of the topic of a text section has a number of advantages over previous schemes. First, its use of information which will probably already be stored in the natural language processing system's lexicon has obvious advantages over schemes which require large, separate data-structures purely for topic identification, as well as for making the predictions associated with a topic. In practice, Scrabble uses a slightly doctored lexicon to improve efficiency, but the necessary work could be done by an automatic preprocessing of the lexicon.

Second, the scheme described here can make use of nominals which suggest a candidate topic, and associated stereotypes, without complex manipulation of semantic information which is not useful for this purpose. The scheme of

(DeJong 79), for example, would perform complex operations on semantic representations associated with "pick" before it processed the more useful word "tuna" if it processed the above example text.

Third the use of semantic primitive patterns has greater generality than techniques which set up direct links between words and bundles of predictions, as appeared to be done in early versions of the SAM program (Schank 75a).

One final point. The technique for topic identification in this paper would not be practical either if it was very expensive to load stereotypes which turn out to be irrelevant, or if the cost of comparing the predictions of such stereotypes with the text representation was high. The Scrabble system, running under Cambridge LISP on an IBM 370/165 took 8770 milliseconds to analyse the example text above of which 756 milliseconds was used by loading and activating the two irrelevant stereotypes and 103 milliseconds was spent comparing their predictions with the CD-representation of the text. The system design is such that these figures would not increase dramatically if more stereotypes were considered whilst processing the example.

6. References

(Cater 80)
Cater, A.W.S. Analysing English Text: A Non-deterministic Approach with Limited Memory. AISB-80 Conference Proceedings. Society for the Study of Artificial Intelligence and the Simulation of Behaviour. July 1980.

(Charniak 78)
Charniak E. With Spoon in Hand this must be the Eating Frame. TINLAP-2. 1978.

(DeJong 79)
DeJong, G.F. Skimming Stories in Real Time: an Experiment in Integrated Understanding. Research Report No. 158. Yale University Department of Computer Science,

New Haven, Connecticut. May 1979.

(Schank 75a)

Schank, R.C. and the Yale A.I. Project. SAM -- A Story Understander. Research Report No. 43. Yale University Department of Computer Science, New Haven, Connecticut. 1975

(Schank 75b)

Schank R.C. Conceptual Information Processing. North-Holland, Amsterdam. 1975.

(Wilks 77)

Wilks, Y.A. Good and Bad Arguments about Semantic Primitives. Communication and Cognition, 10.1977.