[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

Rapportage from the discussion group 'Discourse and Linguistics for MT'

Doug ARNOLD
Department of Language and Linguistics
University of Essex
Wivenhoe Park
Colchester C04 3SQ, UK

July, 1990

There were three presentations in this session:

1. Karen Sparck Jones (Cambridge): 'Large Scale Discourse Structure and MT' (KSJ);
2. Bill Black (UMIST): 'Running a Robust MT Dialogue System' (BB);
3. Danny Jones (UMIST) 'High Quality Translation for Monolinguals' (DJ).

1 Karen Sparck Jones (Cambridge): 'Large Scale Discourse Structure and MT' (KSJ)

KSJ's talk, derived from ongoing work on text summarization being carried out in Cambridge, provided an overview of current approaches to the description of large scale text structure, and raised the question of whether, or how far this sort of analysis would be necessary for MT. Part of the talk (and much of the following discussion) was concerned with a (constructed) example discourse, exemplifying a number of interesting and problematic phenomena.

It is clear that 'large scale' analysis (i.e. going beyond sentence level) is essential in NLP in general (e.g. for anaphora; in text generation or paraphrase, for maintaining coherent presentation of ideas, and cohesive flow of style).

KSJ noted that analyses based simply on commonalities across text-types, or genres of texts are far too general to be of much use, and that a distinction between linguistic and non-linguistic structure is useful. For each, one can approach the discovery of structure in a 'top down' (TD) or a 'bottom up' (BU) fashion (the usual caveats about this distinction apply).

A typical example of the BU approach to non-linguistic structure is to look for content relations between propositions, e.g. a hierarchy linking propositions expressed in the text to the other propositions which they entail (such a structure is potentially quite independent of any presentational structure the text may have). Examples of this approach can be found in the work of Kintsch and van Dijk. The typical TD approach to non-linguistic structure involves some kind of Script (e.g. à la Schank et al).

A typical BU approach to linguistic structure is Grosz's work involving identification of focus spaces (the structure assigned is clearly a straightforward reflection of the presentational structure). TD approaches include McKeown's Rhetorical Schemata, rhetorical grammars (Mann), and work that tries to identify causality relations or component parts of arguments (Reichman).

The talk was based on work done as part of a SERC funded project on General Techniques for Automatic Summarizing being carried out in Cambridge.

2 Bill Black (UMIST): 'Running a Robust MT Dialogue System' (BB)

BB's talk raised the question of what knowledge sources are necessary to maintain robust dialogue. The particular context is dialogues about the contents of 'Yellow Pages' lists of services, written in French.

A wide range of knowledge sources is necessary for dialogue management, including at least the following: syntactic, semantic and lexical knowledge, including knowledge of idioms; knowledge about the particular application (e.g. about the relation between suppliers of plumbing materials, plumbers, and equipment hire firms: any of which may be appropriate suggestions for someone who wants a new sink); 'real world' or commonsense knowledge (e.g. for anaphora, and for knowing that cutting prototypically involves lawnmowers in 'cut grass', whereas in 'cut cake' it prototypically involves knives); knowledge about the expressivity/limitations of knowledgebase language, and the query language; 'pragmatic' knowledge about conversational implicature and presupposition.

BB suggested that the last of these can be dealt with by means of dialogue grammars, which decompose a dialogue into a sequence of exchanges (of different kinds) and further into a variety of moves. This is also helpful in handling ellipsis, and various kinds of ambiguity (e.g. a grasp of what stage of a dialogue on is at can allow a decision about whether 'okay' signals understanding, or agreement, or marks a topic boundary). Some kind of dialogue grammar is also essential if a system is to be able to engage in 'meta-dialogue' (e.g. recovering from misunderstandings).

BB discussed how these kinds of knowledge should be organised, noting that the desire for re-usability motivated separating the different kinds of knowledge (which can be problematic, since the boundaries are not always clear).

3 Danny Jones (UMIST) 'High Quality Translation for Monolinguals' (DJ)

This talk reported work on a system for translating business letters from English to Japanese, for users who have no knowledge of the target language (this is unusual: most MT systems are intended either for bi-linguals, such as translators, or at least assume revision (post-editing) by competent speakers of the target language). In contrast to the mainstream 'rule based' approach to MT, this system is 'example based'. The system operates by searching a network of equivalences between source and target texts, or text fragments, based on a description of a linguistic pattern, and a context. In the limiting case, a single such equivalence might produce the translation of an entire letter (for example, a letter of complaint about non-delivery of goods). (Notice that because the output of the translation is an actual piece of more or less complete target text, correctness can be guaranteed, and the system should be usable by monolinguals).

In most cases, however, business letters cannot be complete canned texts, there will be at least some gaps to be filled in by the user - minimally, gaps for dates, maximally entire paragraphs. Where the gaps are not very simple, there are a number of possibilities:

- the user may be prepared to accept that the letter says less than they would like to say, in order to ensure that at least something (and something correct) is said;
- the user may be able to interact with the system to find some approximation to his/her original intention which the system can handle.

However, the chief problem with the approach is the inflexibility inherent in dealing with relatively large chunks of text. But dealing with smaller chunks of text leads, in the end to what is essentially a rule based system, with the associated worries about the correctness of the output that results from combining translations of small chunks. So the problem is to find:

- the right size of 'chunk' for the particular domain - this is essentially an application of discourse grammar for the text type; and

- the right level of description for the chunks in the source language - e.g. rather than a literal match, or a pattern match, one might look for a match at some abstract level (e.g. the level of discourse function).

More information about this project can be found in: D. Jones and J. Tsujii (1990) 'Interactive High Quality Translation for Monolinguals'. Third International Conference on Theoretical and Methodological Issues in Machine Translation 11-13 June 1990, Linguistics Research Center, Austin, Texas.

4 Discussion

The discussion session that followed lunch was wide ranging, and included further quite detailed discussion of KSJ's example dialogue, and a short description by Abid Khan (Cranfield) of his work on using anaphoric relations in discovering discourse structure. Other questions addressed concerned whether it is necessary to look beyond simple discourse grammars to the more advanced techniques outlined by KSJ, and how far a complete a complete discourse representation is necessary for MT.

There was some attempt to assess the state of the art of NLP in the area of discourse. It seemed clear that though there are many fundamental questions to be answered, and that the generality of current techniques for analyzing discourse is doubtful, enough is known for practically useful descriptions of discourse structure to be made in restricted types of text and subject matter (e.g. business letters - cf. DJ's talk).