PROPOSALS FOR A HIERARCHY OF FORMAL TRANSLATION MODELS

Klaus-Jürgen Engelberg

Universität Konstanz, Philosophische Fakultät, Fachgruppe Sprachwissenschaft, BRD

The present deplorable state-of-the-art in the field of machine translation seems greatly due to a fundamental lack of formal translation models needed in natural language processing.

From the methodological point of view it appears difficult to delineate a borderline between translation theory and modern theoretical linguistics (availing itself of model theoretical semantics) or full natural language understanding systems as developed in Artificial Intelligence research. It seems plausible to postulate that any prospective translation theory should draw on ideas from both fields. Unfortunately, problems discussed in painstaking detail in linguistics like differences in quantifier scope appear to be of lesser concern to a translator (since these ambiguities may well remain present in the target language), neither seems a full or deep understanding necessary in many cases, standard syntactic phrasing may suffice. More specifically, we regard the problems of disambiguation, mandatory insertion of lexical items not conventionally implied in the source language and coreference/anaphora resolution as the crucial problem areas of machine translation.

In this paper, we will endeavour - in this preliminary draft only in a very sketchy manner - to set up a hierarchy of formal translation models ordered according to their increas-

ing systematic disambiguation power for certain types of texts.

Quite analogous to complexity considerations in mathematics, the power of a translation system is assumed to be measured by the amount of storage needed for the lexical component (AI-people might call this long-term-memory) and/or for the transient or dynamic data (short-term-memory) built up during the interpreting process of a particular text. Any model will be capable to translate only certain restricted types of texts in a systematic manner and with satisfactory results, but the idea is that any model will also contain components of lower levels of complexity. This is to make sure that in cases in which disambiguation on purely syntactic grounds is possible no such process via 'deep' semantic representations will be attempted for this particular case. The rationale, of course, will be to utilize ever larger portions of contextual (or rather co-textual) information for these ends. As the reader will notice, powerfull translation system have to incorporate more and more knowledge-of-the world into the database, as becomes apparent from the famous example:

The soldiers shot the women. They fell down. ⇒
Les soldats abbatirent les femmes. Ils/elles? tomberent.

Syntactic methods

Level Syn1: Word-to-word translation
Is out for apparent reasons! (although a full bilingual dictionary would require a considerable amount of storage space in a computer)

Level Syn2: Constituent preserving translation
These models utilize the immediate syntactical context (e.g. valency of verbs) for disambiguation purposes. In such a system a rule may look like

x sich erinnern ⇒ x remember, but, x erinnern y ⇒ x remind y

At any rate, a valency oriented lexicon would be helpful in the following models, too. The search strategy would be longest match first.

Level Syn3: Tree-to-tree translation

Unbouded translations allow for reordering of arbitrarily long portions of a sentence. We think it reasonable to assume that a quarter-century of Generative Grammar research in Linguistics will have produced enough theoretical and practical apparatus to deal with any type of tree-restructuring that may be needed in direct syntactic translations between natural languages (also cf. the French system GETA).

Semantic methods

Level Sem1: Case.grammar oriented translations

There are several MT systems that impose heavy restrictions on the possible arguments of verbs by encoding semantic features in the lexicon (e.g. METEO in Canada). By this, of course, disambiguation can take place only within the limits of a single sentence or clause.

Level Sem2: Translations using coherence relations

The basis of this approach is the assumption that there exist finitely many determined and computable coherence relations between two subsequent sentences and/or clauses in certain types of texts. (sometime called the cohesive-ties-approach). They may be even indications of these relations at the surface level of the discours e.g. 'whereas' suggesting CONTRAST or 'then' suggesting TIME-SEQUENCE, other relations may be ELABORATION, EFFECT, CAUSE (Hirst /1981/). Processing of these texts could be done by semantic finite state automata that would accept only highly constrained discourses in which no abrupt shifts of focus would be allowed. At last at this level of complexity it seems necessary to assume that the vocabulary should be organized - in addition to the usual lexicographic

order - as a sort of semantic network containing all types of sense relations like super-subset relation, antonymy, converseness, time-sequence - existing even between several places verbs.

Level Sem3: Translations using story trees

These models dynamically build up a tree-like macrostructure for a text in which arbitrary deep embeddings of themes and sub-themes are represented. In this approach, coherence relations between entire portions of text or paragraphs could be established - thus allowing for coreferen across long distances in a text (vide Rumelhart /1975/). This process may be facilitated by what Y. Wilks chose to call 'paraplates' in the database.

Level Sem4: Translations using semantic networks

This model is designed for not so orderly texts as assumed in the previous levels. A semantic network as the dynamic macrostructure of a text would allow for multiple views or thematic structures associated with a portion of a text. To make this effective, a very rich fabric of various types of associative links would be needed in the database.

Level Sem5: Frame-based translations

'Frames' or 'scripts' have been widely discussed in the AI community in the past 10 years or so. The idea seems to bee to aggregate all sorts of information object-centred linked with a particular 'stereotypical situation' into a structured entity - called 'frame'. This approach would, in principle, allow one - by default reasoning - to recover information not explicitly mentioned in the text. In particular, this may be helpful when translating into a western language from Russian, in which the definite/indefinite or known/unknown distinction in nouns is lacking. Consider the translation problems in the following example (drawing on Schank's favourite script):

Petr posel v restoran. Oficiant podal emu menju. ⇒
Peter went to a restaurant. The waiter handed him the menu.

Scripts could account for associations induced by 'spatial- - temporal contiguities' as present in this example.

Doubts as to the feasibility of MT based on frames - except possibly in very restricted areas of discourse - have come from various quarters. First, the coding effort could turn out to be enormous. Second, a intricate problem seems to be how to find out which script is relevant to the current portion of text.