[*AMTA/SIG-IL First Workshop on Interlinguas*, San Diego, CA, October 28, 1997]

Simplification of Nomenclature leads to an Ideal IL for Human Language Communication

Young-Suk Lee
Clifford Weinstein
Dinesh Tummala
Linda Kukolich
MIT Lincoln Laboratory

Stephanie Seneff
Spoken Language Systems Group, MIT

In this talk, we advocate the view that simplification of nomenclature, stated in linguistically well-motivated concepts, leads to an ideal IL for human language communication systems including multi-lingual machine translation and spoken language dialogue systems. To advocate this view, we describe our method of constructing IL representations in developing our multi-lingual machine translation system, CCLINC (cf. Tummala et al. 1995, Weinstein et al. 1996, Lee et al. 1997), and give a demonstration of two-way English/Korean translation by CCLINC. At MIT Lincoln Laboratory, we have been developing a multi-lingual machine translation system, called CCLINC. The core of CCLINC consists of the language understanding system (TINA, cf. Seneff 1992) and the language generation system (GENESIS, cf. Glass et al. 1994). The system has been applied to English-to-French, English-to-Korean and Korean-to-English translations. In designing an IL representation, we have been following two developmental strategies: First, simplification of nomenclature. Second, preservation of the predicate/argument structure of the input sentence. These strategies are largely drawn from the experience of applying the language understanding/generation system to spoken language dialogue systems (cf. Zue et al 1996). Simplification of nomenclature provides a very general IL which can be used in various application areas including information access from a database, language tutoring, and conversational systems. Preservation of predicate/argument structure facilitates generation of multiple output languages, which are accurately ordered in each target language. The intermediate meaning representation, which we call a semantic frame, is derived from the parse tree of the input sentence. All major parse tree constituents (regardless of whether they are semantic or syntactic) are reduced into one of three major categories in the semantic frame, namely, clause, topic and predicate. All noun phrase expressions are mapped onto "topic." All adjectives, prepositional phrases, and verb phrases, are mapped onto "predicate." The frame hierarchy structure encodes dependencies among clauses, topics, and predicates. Thus the predicate/argument structure of the input sentence is preserved. Information regarding the sentence type such as declarative, and imperative, is encoded at the 'clause' level. The ontology of each category in the semantic frame is described in English for reasons of convenience, and the formalism has been incrementally enriched as we develop experience from working with diverse language classes.

From our experience of working with languages such as Korean, Japanese and Chinese, it has become clear that discourse understanding is critical in producing an IL which unambiguously captures the meaning of the input sentence. For instance, the (in)definiteness feature of a noun phrase, which is crucial in unambiguously picking out the entity referred to by the given noun phrase, can only be inferred from the discourse in these languages. Also these languages frequently allow argument (subjects/objects) drop, and the predicate/argument structure of the given input may be drawn from the discourse in such cases. Discourse module has already been utilized for conversational systems (cf. Seneff et al. 1996), and it is under development for CCLINC.

Finally, to illustrate how our view/strategy on IL has been implemented, we give a CCLINC system demonstration of two-way English/Korean translations. During the demonstration, we will discuss some compromises we had to make to produce a useful system within a short period of time.

References

James Glass, Joseph Polifroni and Stephanie Seneff, "Multilingual Language Generation Across Multiple Languages," International Conference on Spoken Language Processing, Yokohama, Japan 1994.

Young-Suk Lee, Clifford Weinstein, Stephanie Seneff and Dinesh Tummala, "Ambiguity Resolution for Machine Translation of Telegraphic Messages," Proceedings of 35th the Association for Computational Linguistics, Madrid, Spain 1997.

Stephanie Seneff, "TINA: A Natural Language System for Spoken Language Applications," Computational Linguistics, Vol.18, No.l, pp.61-86, 1992.

Stephanie Seneff, Dave Goddeau, Christine Pao, and Joe Polifroni, "Multimodal Discourse Modelling in a Multi-User Multi-Domain Environment," Proceedings of the Fourth International Conference on Spoken Language Processing, Philadelphia, USA 1996.

Dinesh Tummala, Stephanie Seneff, Douglas Paul, Clifford Weinstein, and Dennis Yang, "CCLINC: System Architecture and Concept Demonstration of Speech-to-Speech Translation for Limited-Domain Multilingual Applications," Proceedings of the 1995 ARPA Spoken Language Technology Workshop, Austin, Texas.

Clifford Weinstein, Young-Suk Lee, Dinesh Tummala and Stephanie Seneff, "Automatic English-to-Korean Text Translation of Telegraphic Messages in a Limited Domain," C-STAR II Proceedings of the Workshop, ATR International Workshop on Speech Translation, Kyoto, Japan 1996.

Victor Zue, Stephanie Seneff, Joe Polifroni, Helen Meng, James Glass, "Multilingual Human-computer Interactions: From Information Access to Language Learning," Proceedings of International Conference on Spoken Language Processing, Philadelphia, USA 1996.