

Issues on Interlingua Machine Translation Systems

Sameh Alansary

Sameh.alansary@bibalex.org

Head of the Arabic UNL Language Centre,
Bibliotheca Alexandrina,
P.O. Box 138, 21526, El Shatby,
Alexandria, Egypt.

Department of Phonetics and Linguistics
Faculty of Arts
Alexandria University
El Shatby, Alexandria, Egypt.

Extended Abstract

This paper presents Interlingua machine translation focusing on three systems, namely, KANT, UNITRANS and UNL. It explains the basic properties of each system to come up with a concrete and objective evaluation of Interlingua MT.

The first ideas about machine translation (MT) came up even before the first computer was invented, and has developed over many years of research. As an automated system that facilitates communication and helps overcome language barriers, MT has now become a very vital part of everyone's life. Machine translation has passed through different levels of progress since the 17th century with the goal of analyzing Source Language (SL) texts to produce equivalent texts in the Target Language (TL), ideally without human intervention. Different approaches have been adopted throughout this progress. However, researchers still try to design more accurate, fast processing approaches to avoid the problems of older ones; the conflicting rules and restrictions of the Rule-based approach, the word selection problems of the statistical approach and the direct approach's problems where words are translated directly without passing through additional analyses, a fact that yielded inaccurate results. It was then that the first ideas about interlingua-based machine translation arose and can be traced back to Descartes and Leibniz, who are considered the God fathers of the first interlingua-based MT. Interlingua-based machine translation relies on an artificial intermediate language which can be used as a common, formal representation into which source natural language (SL) may be translated, and from which target natural language (TL) can be generated. It is an instance of rule-based machine translation approach; however, it excels in that it provides a SINGLE underlying representation of meaning for both (SL) and (TL). The source language text, or the text to be translated written in one language (SL), is transformed into an interlingua; i.e. an abstract language-independent representation that can be used as a pivot representation of text meaning; subsequently, the target natural language (TL) text is generated out of this intermediate representation.

Compared with other machine translation approaches, interlingua-based machine translation has two main advantages; first, as each language has an independent mapping to and from the interlingua, any number of source and target languages can be connected, without the need to define explicit rules for each language pair in each translation direction. Second, the intermediate language-independent representation of meaning should be able to provide a neutral basis for the comparison of equivalent texts that differ syntactically, but share the same meaning.

Nevertheless, for an interlingua to be indeed efficient, it must satisfy two primary conditions: a) accurate translation from the source natural language into the interlingua must be easier than direct translation into another natural language; b) accurate translation from the interlingua into the target language must be easy. However, many properties of language usage act to impede such accurate and simple transformation because of two reasons. First, the form of text often under-determines the content, which causes problems when attempting to analyze the source text (e.g., in order to convert it to some interlingua). Second, the content often under-determines the form, which causes problems when trying to synthesize the target text. These and other issues in performing machine translation mean that defining an interlingua that can unambiguously capture the appropriate meaning from any language, and explicitly preserve the appropriate semantic, pragmatic and other contextual information is a hugely formidable task.

KANT is an Interlingua which has been primarily targeted towards the translation of technical text in controlled sub-domains. It uses controlled vocabulary and grammar for each source language, and explicit, yet focused, semantic models for each technical domain to achieve very high accuracy in translation. Designed for multilingual document production, KANT has been applied to the domains of electric power utility management, heavy equipment technical documentation, medical records, car manuals, and TV captions. Some of the limitations of such a system are: a) the need for consistency in source grammar and terminology, such that the use of a domain sublanguage is justified beyond its use for accurate MT; b) the texts to be translated focus on an area of technical information (such as computers or motor vehicles), such that it is feasible to design a domain-specific approach to writing (terminology and grammar).

UNITRAN is another Interlingua MT system that translates Spanish, English, and German bi-directionally. It is a Principle-Based approach to machine translation. It operates cross-linguistically but still accounting for knowledge that is specific to each language.

UNL, the Universal Networking Language, is another Interlingua MT system that adopts a tridimensional theory of meaning whose components are: concepts, concept relations, and concept predicates. It is a language for the representation of document knowledge in a language-independent format, aiming at creating an Interlingua for all human languages and supporting the exchange of textual information in multilingual environments. It supports a multitude of languages without any restrictions on specific domain. Whenever there is a need to represent knowledge in a domain-independent manner, researchers turn back to natural language to explore the semantic atoms used by natural languages for expressing knowledge. UNL follows this philosophy by providing an interlingua equivalent of these semantic atoms. UNL can be proposed as a firm document knowledge representation language because; first, the set of necessary relations existing between concepts is already standardized and have a strong linguistic basis such as the logical, temporal, spatial and causative relations which are widely employed in semantic analysis as well as in knowledge representation. Second, the set of attributes that modify concepts and relations is fixed and well defined, guaranteeing a precise definition of contextual information. Third, in UNL, semantic atoms, or Universal Words (UWs), are not simply concepts but word senses, extracted mainly from the English lexicon and are organized hierarchically. Accordingly, we can claim that UNL syntax and semantics are well defined.

This paper argues that the UNL is indeed a suitable interlingua for automated translation, ranging from fully automatic MT to interactive MT for four reasons. First, UNL has achieved the principle of universality in interlingua-based machine translation. Second, Due to the precise nature of UNL, human non-specialists can improve a UNL representation interactively, a posteriori, from any UNL-related language, and on demand. Third, in many contexts other than translation, an interlingua; i.e., a semantic-oriented representation, like UNL is actually the best solution because applications related to information processing in multilingual contexts do not need a very precise representation of the form of the information, what they need is a precise enough representation of the semantic content of this information. Third, applications such as information retrieval and abstracting have already been prototyped successfully using UNL. Fourth, it is far easier to generate SQL or SQL-like queries and answers from a UNL representation than from natural language text in many languages.

The paper concludes that Interlingua-based Machine Translation systems other than UNL have been built under limited conditions; they mirror one-to-many (languages) or many (languages)- to-one approaches, often involving English at the "one" end. In addition, their communication is restricted to the exchange of basic information ignoring the richness and flexibility of the human mind. On the other hand, the Universal Networking Language (UNL) adopts a different approach to interlingua-based MT. It produces a representation of the semantic content of texts, removing away the details of the source language -a fact that qualifies it to be a language-independent representation- while keeping enough linguistic information to make feasible text generation in a multilingual environment that includes more than a dozen languages.