Zhang Hanzhi

Semantics

22 January 2022

Code and Message: The Search for Meanings in Machine Translation

**I. Introduction**

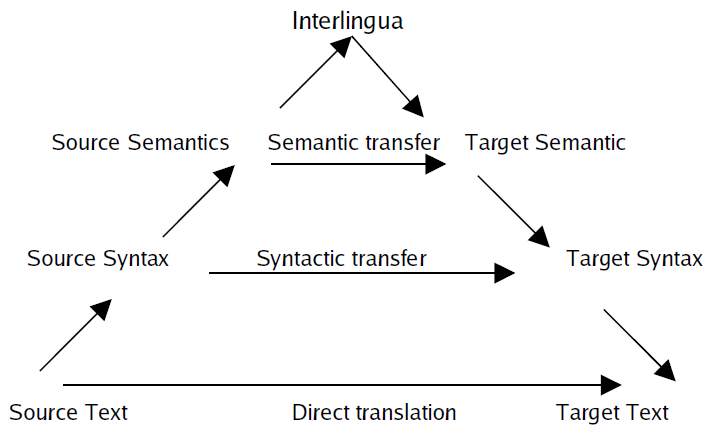
Back in 1962, when Machine Translation (MT) was a newly developed field, M. A. K. Halliday gave a visionary prediction on the contrast between the understanding of language from the linguistic and the MT perspective, as well as the possible obstacles MT would face due to this contradiction. In order to process natural language with the computer, the text must be treated as a fundamentally binary code, that is to say, code and message should be two distinct entities, so MT simply involves an decode-encode process from the source language to an intermediate representation of the message, and from the intermediate representation to the target language. This pattern contradicts the linguistics view of language, as Halliday poses, that the code and message in languages are one, with no means to separate them into independent entities. In the early stages of MT, such separation has been proven a formidable, if not impossible task.

Over the years, researchers in the fields of MT and linguistics have worked on their studies respectively. In traditional rule-based and example-based MT, researchers focus on the analysis and transfer of code, while message is largely ignored. In researches on linguistics, especially semantics, efforts are made to understand the meanings of words and sentences. With the development of statistical methods and Machine Learning, the possibility of building a completely statistical-based MT with little or no reference to linguistic and semantics knowledge comes into being. It seemed that MT and linguistics are developing on two parallel paths that seldom crosses.

Not all MT approaches are satisfied with working on language as a code without treating the message it carries. When traditional MT finds it hard to break through the bottleneck, and the demand for a more “universal” MT among various languages keeps growing, researchers picks up the idea of using an interlingua to represent the meaning of the text again. Such interlingua should be a metalanguage without ambiguity, so that it can be processed by a computer. It is better to be universal, so one set of interlingua may be shared by different languages using proper analysis and generation rules. These researchers are not alone. In the study of words and sentences meanings, linguists are also trying to construct a universal metalanguage that can be used to explain different languages. The essay aims to discuss relevant studies on MT and semantics, showing how a neutral or universal metalanguage is helpful in setting up a relation between code and message, providing Machine Translation more possibilities.

**II. Developing MT: Code without or with Message**

The graph below taken from Klyueva’s “Semantics in Machine Translation” (2007) shows different levels of machine translation. Among the models, interlingua approach is the only one that touches the meaning of the source text.



**Figure 1. MT Triangle (Klyueva, 2007)**

1. **Rule-based: Direct and Transfer Approach**

The Direct approach is developed from a lexicon-based, word-to-word translation model, where the program has a built-in bilingual dictionary for word substitution. This only produces satisfying result when the source and target language are closely related, that is to say, they have similar vocabulary and grammar. Hajič and colleagues’ Russian and Czech MT system built in 2003 is an example. For more distant languages, it may fail to match a word, or produce text with wrong sentence structures.

The Transfer approach goes beyond the lexical scope, parses the source text to build an intermediate syntactic structure, often in the form of a syntax tree. It then follows specified rules to construct a syntax tree in the target language, and generates lineal text from this structure. These three stages are called analysis-transfer-generation. The more distant two languages are, the more complicated the transfer rules would be. Thus, the Transfer approach produces better out outcome than the Direct approach, but still lacks accuracy when the languages are very different. Attempts are made to perform transfer on a deeper, semantic level, but the outcome did not improve significantly.

The Direct and Transfer model are both rule-based, apart from difficulties in writing rules for distinct languages, it also takes long terms to test, adjust and modify the rules over and over again. In order to cut down the time consuming, inefficient rule-making process, researchers move on to an example-based approach.

1. **Example-Based: Statistical Approach**

Açıkgöz and Sert describes Example-based MT as a “translation by analogy”. This model is built upon the theory that only natural language itself could be the “absolute source for analyzing any sort of language produced by its users” (2006). The theory is a step forward from Halliday’s proposition, that not only “message” independent from the natural language code does not exist, but any man-made rules about language is not fully credible. Thus, Example-based MT returns to language itself, relying on a gigantic parallel corpus from which the program should be able to find exact or similar cases for any given piece of text. With the development of artificial intelligences, computers are gaining the power to maintain a giant database, pick out the most appropriate example judging from the frequencies of words and word combinations.

With the Example-based MT, it seems that the computer is translating in a way more similar to human translators, that is “learning” through the experience. It can even take the context into account, as in the case of “She sat by the river, looking to the opposite bank”, a well-trained, “smart” enough statistical model will be able to establish a link between the word “bank” and “river” in its context, picking out the word in the target language for “slope beside a body of water” rather than a financial institution. Note that the computer does not have to “understand” the meaning of river or bank to create this link: the process is no different from a picture recognition system: the data is stored in the computer, a “logic” that only the computer itself can recognize and make use of, to human observers they are just meaningless numerals.

No matter how smart the computer is trained, the model has two fatal flaws. Among the sea of languages on earth, very few pairs of them have parallel corpus large enough to “teach” the computer language patterns. As for the model training itself, corpuses in a specific field produce better results than one containing all kinds of randomly selected text, for the computer can easily be “confused”. Melby suggests that such MT models work better when they less ambitious, focusing on a narrow topic in a “rather dry and monotonous style” (1996), like law documents or drug instructions. When it comes to broad topics, literature or casual spoken words, the model is not effective. Another issue is whether to trust the computing power – is a statistically calculated pattern bound to be superior to one summarized from human’s knowledge and experience?

1. **Interlingua-Based: The Search for Meanings**

With the flaws of MT models that focus on designing rules or finding patterns with little if no exploration of meaning, the debate centered around a question posed by John Searle in his Chinese Room Argument (1980). Searle sets up the plot in which a man with no knowledge of Chinese is locked in a room with batches of Chinese and English script and instructions, so that he could follow the rules to “translate” between the two languages as a computer does. The question is whether he can cheat the people outside to believe there is a Chinese speaker in the room. If so, then a computer can appear to talk with human without actual understanding the conversation, hence the Turing Test would be invalid. For researchers in MT, it does not matter whether the computer really understands or not, as long as it produces high-quality translation.

Can he cheat people outside? Is understanding the language necessary for producing good translation? If the computer is to understand the language, the way it interprets or compiles programming languages can act as an inspiration. Programming languages are man-made languages with many common features with natural languages, like syntax and semantics. When programs are compiled, the lexical, syntax and semantics analysis are the same steps MT systems perform. Then a compiler generates assembly language code, and then machine code. Suppose that machine code is only “understandable” to the computer, and programming languages like C or Python is only readable to human programmers, then the assembly language code is the “intermediate representation” as the interlingua in MT triangle. Hence it can be inferred that the transfer between source language and interlingua should be a reduction process, and that the interlingua should bear similarities with the assembly language: simple, unambiguous and readable. MT researchers then find themselves in a search for a representation of language meanings, presumably in the form of a set of universal lexemes.

The idea of a universal lexeme is tempting, for it not only suggests possibilities for more accurate translation, but also reduces the MT among N languages from finding N × N pairs of transfer rules or patterns, to creating N pairs of reduction and generation strategies for the same interlingua. Despite these theoretical prospects, the approach is not passionately welcomed by all. Some like the MT pioneer Weaver simply believes it is unnecessary. Take his words in 1949: given a text Russian, which he does not know, he pretends it is written in English and “strips off the code” to “retrieve the information contained in the text”. Such are the supporters of the theory that understanding is not a must, and would not make such big improvement. Açıkgöz and Sert see the setbacks of the model from another perspective, that compressing natural languages into a universal set of symbols “denies the very nature and cultural value of language”. This is to say, even if a set of symbols is picked to represent the meanings of all languages, it would only reach the purpose through cutting the nature and culture in languages, resulting in a small intersection that is really not a “universal” lexeme.

Supporters of the Interlingua approach poses the argument that MT must go beyond pure linguistic information (syntax and semantics) and “understand” the content of texts (Hutchins and Somers, 1992). The faith in an abstract, universal, language-independent representation of meaning (the interlingua) is shared by some researchers in the field, that Halliday is wrong, code and message can be separated. Others like Volk admits the impracticability constructing a perfect ideal model of interlingua as well as the theory that language cannot be separated from its entities either culturally or linguistically, but suggests that approximate models might be good enough for production (1998).

The next part will introduce linguists and MT researchers’ recent works on universal lexeme and interlingua, and how they solve the problems along the way.

**III. Finding a Universal Metalanguage**

Linguists’ search for a universal symbol system to interpret language aims to find a neutral tool to study language with. Before this, most linguists use their own mother language as a metalanguage to study foreign languages or the language itself. Given the linguistic relativity as a premise, a natural language is bound to carry its own cultural and linguistic background, and therefore cannot be absolutely neutral.

The search is directed in two ways, reduction through paraphrasing with a selected vocabulary from natural languages, or reduction into formal/logical forms. Both ways have been adopted by interlingua MT, separately or combined.

1. **Natural Semantic Metalanguage**

Ever since the 1980s, Anna Wierzbicka and colleagues have been working on the extraction of an ultimate core vocabulary through reductive paraphrase. The vocabulary, as Goddard and Wierzbicka purposed, should consist of all and only the simple, basic and universal concepts of languages, often referred to as “semantic primes”, also “the simplest lexis of paraphrase and explanation” (McCarthy 1990). The notion of semantic primes relies on that some meanings are simpler than others, and that simpler meanings can be used explain more complex ones. (Goddard and Wierzbicka, 2007)

The purpose for the Natural Semantic Metalanguage (NSM) research is despite that English is filling the role for a “Esperanto” in international communication, it is, at all levels from vocabulary, grammar to pragmatic norms, far from neutral and value-free. NSM English, English version of the neutral, universal NSM, is a culture-free “nuclear English” with no literary, aesthetic or emotional aspirations, thus it is fair and neutral. The original NSM is in English, but all language-specific NSM’s are isomorphic, so if something can be said in one NSM, it can be said in any NSM.

Although the extensive corpus of NSM semantic studies has shown that NSM is sufficient for studying languages, it is so not sufficient for more practical tasks like MT and international communication, where speed or compleness is as important as fairness and accuracy. Sgall and Panevová (1987) summarizes the requirements on linguistic theory for MT as adequacy, testability, economy and modularity. NSM serves especially well in adequacy, that the “open-endedness” of language requires that the completeness of a language description must be reduced to an estimated core to fit into a computer program, NSM has reduced enough. But it is not economy, for the theory is too complex to be applicable. As for global communication, expressiveness is also a problem. An optimal semantic metalanguage must consist only of the meanings present in all natural languages, hence we again face the problem of content loss in the reduction inside one language, as well as the intersection between languages. Goddard and Wierzbicka admit that a truly nuclear, culture-free subset cannot fulfil the role for communication, while a subset rich and large enough for the job would fail to be “culture-free”.

The dilemma is more disappointing to MT researchers than linguists, for despite the lack of completeness and efficiency in NSM, which prevents it from being applied in production, it is sufficient for academic studies. The ideal semantic metalanguage is to be as simple, regular, logical, rich, and creative as possible, while NSM is a carefully balanced compromise among these features. Allan doubts that any artificial semantic formalism is really “a degenerate form of a natural language” (1986), however either for research or production purpose, the degeneration has to be tolerated.

1. **Formal Semantics and Logical Metalanguage**

Another reductive approach is formal semantics and logical representation. Related works are based on truth-conditional symbols and rules, so that the system functions in a way very similar to compilers. The Principle of Compositionality suggests that the syntactic rule system and the semantic rule system are linked, so that a recursive and infinite syntax produces recursive and infinite semantic meanings, the same way natural language words combine to form phrases, sentences or paragraphs.

The “translations” of sentences can be specified into a logic with model theory and proof theory, so that formal semantics can be applied to natural language the same way it is applied to formal languages. Below are some examples of lambda calculus and beta reduction, taken from Peter Selinger and Ted Briscoe’s lecture notes:

1. λ x.(x+1)

(λ x.(x+1)) 3 -> 3 + 1

(b) λ x [snore1(x)]

λ x [snore1(x)] (max1) -> snore1(max1)

The semantic reduction (a) produces 3 + 1 (which can be “translated” to 4 through universal mathematical rules), while that of (b) produces snore1(max1), in Briscoe’s design means “Max snores”. The problem is that mathematical rules is universal, but if language users wishes to do inference in a model-theoretic fashion, they would need to carry around the “whole actual world” (and all the other possible variations). The verb snore may be bound with its meaning (the action or a habit of snoring), while the nouns, especially names like “Max”, faces the linguistic problem of reference/denotation. It is hard to imagine that a MT should maintain a whole set of denotations for each context given in its source text. With even higher accuracy than NSM provided by the nature of formal and logic expressions, the practicality of the model is worse.

**C. Application of the Interlingua**

AlAnsary examined a series of interlingua-based MT systems in 2014, including DLT, UNITRAN, KANT. Among the four systems, UNITRAN and KANT shares the idea of the natural metalanguage approach, with UNITRAN using a Lexical Conceptual Structure (LCS) for meaning representation, and KANT a “well-defined subset of the source language”. DLT and UNL adopts the formal semantics approach, with DLT using a logical language and UNL a formal artificial language. Note that the interlingua-based models are different from the traditional rule-based and example-based MT in that they use interlingua as an intermediate code for meaning; transfer rules or statistical analysis may still be used in these models, along with more recent neural network methods.

An interlingua describes the core structure embedded in and shared by languages, in which meanings are composed into text. AlAnsary concludes that such an approach with regularity and predictability is supposed to have economy, modularity, localization, and back-translation possibility, as well as the potential in other NLP areas apart from MT, such as summarization, rephrasing and question answering. Despite the supposed advantages, the interlingua approach faces huge obstacles when it comes to mapping natural language words to an unambiguous conceptual metalanguage and designing a grammar for this language. Because of this challenge caused by the infinite and variable nature of languages, most interlingua-based system never makes it beyond the research phase. As for those who manages to put their systems into production, they circumvent the problem the same way their pioneers in traditional MT do, either by “limiting input texts to specific domains (such as Kant)” or “controlling input language vocabulary and structures (such as Kant and DLT’s prototype)” (AlAnsary, 2014).

A second issue in interlingua-based MT is the problem of ambiguity. The ambiguity of languages exists on each level of parsing from lexical to semantic, not separately by as a whole. A programming language is a language without ambiguity right from the design stage, but natural languages are “born” to be ambiguous. In the MT triangle, the interlingua that represents meaning is extracted from the semantic level, but in order to produce all the possible meanings of a sentence, it must go back to as far as the lexical parsing stage. Take the following example in Chinese:

1. lexical scope: 你 好 好学 。

syntax scope: n. (v. +) adv. adj.

you are very hardworking

semantics scope: You are very hardworking.

1. lexical scope: 你 好好 学 。

syntax scope: n. adv. v.

you well study

semantics scope: You (should) study well (study hard).

For this case the ambiguity in lexical parsing, the ambiguity is carried all the way into the semantic stage, producing two interpretations that are both grammatically and logically acceptable. In such and similar cases in interlingua-based MT, the lexical and syntax parser would first produce all possible alternative syntax trees regardless of their semantic probability. The trees with wrong grammar structures or meanings that do not make sense will be eliminated in the semantics level, either by the program or its user. If the ambiguity is carried to the interlingua level, many MT models would generate all these possible cases into target language for the user to choose from.

The final issue is how “universal” the interlingua could be. Despite the possibility to reduce the complexity of MT among N languages from N2 to 2N, the problem is that the complexity of adding new languages to an interlingua-based system is not lineal. Linguists’ study on NSM shows that a truly universal lexeme would be an extremely small intersection of semantic primes. The smaller the subset is, the less expressive it would be, and the more complicated are the analysis and generation rules. Adding new languages to an interlingua-based MT that is grammatically distinct from the already existing ones may force changes in the interlingua, and thus every existing language’s rules should be modified. For such reasons, very few interlingua-based MT systems can intermediate between more than ten languages. However, supporting ten languages is already a big step forward compared to traditional systems.

The most recent breakthrough is through constructing language-aware interlingua for multilingual neural machine translation (NMT). This new type of interlingua does not seek to be universal, but captures the diversity and specificity of different languages, giving inferior performance compared with models with one interlingua to fit all. One example is the language-aware interlingua NMT with an Encoder-Decoder architecture developed by Alibaba Group’s Machine Intelligence Technology Lab in 2020, with its highly elaborated training model, the test shows statistically significant improvements in both many-to-one and one-to-many translation directions on WMT data.

**V. Conclusion**

Both linguistics and Machine Translation had come a long way since Halliday’s proposition that natural language code and message are inseparable. Linguists tried to find a neutral metalanguage to study the meaning and semantics of languages, either in the form of paraphrasing with a subset vocabulary, or binding words and meanings to formal and logic expressions. One the other hand, MT researchers tries to import an interlingua to achieve better accuracy and efficiency compared to traditional rule-based and example-based MT systems: the interlingua is a reduction of languages, so it shall express meanings “understandable” to both human and computers; the interlingua is designed to be as universal as possible, so it may support translation between a wider range of language pairs. Despite the problems with complicated rules, loss of language features, and ambiguity, the obstacles are partly overcome by limiting the source text to a specified field, balancing the number of semantic primes, and allowing ambiguity in earlier stages of parsing to preserve the possibilities of meanings. Great progress has been made in the field of interlingua-based MT, with advancing artificial intelligence and neural network that seem to promise smarter models and a brighter future.

**Works Cited**

Açıkgöz, Fırat & Sert, Olcay: Interlingual machine translation: prospects and setbacks. *Translation Journal* 10 (3), July 2006.

Alansary, Sameh: Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas. *The Egyptian Journal of Language Engineering*. 2014.

Allan, Keith: *Linguistic Semantics*. London: Routledge & Kegan Paul. 1986.

Briscoe, Ted: Introduction to Formal Semantics for Natural Language. Lecture notes of *CS 135: Computational Semantics* at University of Cambridge, 2011.

Goddard, Cliff & Wierzbicka, Anna: Semantic primes and cultural scripts in language learning and intercultural communication. *Cultural Linguistics: Implications for Second Language Learning and Intercultural Communication*. 2007.

Goddard, Cliff: Natural Semantic Metalanguage: The state of the art. 2008.

Hajič, J., Homola, P. and Kuboň, V. : A simple multilingual machine translation system. *In Proceedings of the MT Summit IX*, New Orleans, 2003.

Halliday, Μ. Α. Κ. : Linguistics and Machine Translation. *STUF - Language Typology and Universals* 15. April 1, 1962.

Hutchins, W. J. and Somers, H. L. : An Introduction to Machine Translation, (chapter1, 4, 17) (chapter 1 p.8) London Academic Press Limited, 1992.

Klyueva, N. : Semantics in Machine Translation. *WDS'07 Proceedings of Contributed Papers: Part I - Mathematics and Computer Sciences* (eds. J. Safrankova and J. Pavlu), Prague, Matfyzpress, 2007.

McCarthy, Michael: *Vocabulary*. Oxford: Oxford University Press. 1990.

Melby, A. K. : *The Possibility of Language*. Warner John Benjamins Publishing. 1996.

Selinger, Peter: Lecture Notes on the Lambda Calculus. Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada.

Sgall, Petr & Panevová, Jarmila: Machine Translation, Linguistics, and Interlingua. *Third Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Copenhagen. 1987.

Volk, M. : The Automatic Translation of Idioms - Machine Translation vs. Translation Memory Systems. In: Nico Weber (ed.): *Machine Translation: Theory, Applications, and Evaluation. An assessment of the state of the art*. St. Augustin: gardez-Verlag. 1998. <http://www.ling.su.se/DaLi/volk/publications.html>

Weaver, W. : Translation. Repr. in: Locke, W.N. and Booth, A.D. (eds.) *Machine translation of languages: fourteen essays*. Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology, 1955, pp. 15-23.

Zhu, Changfeng et al: Language-aware Interlingua for Multilingual Neural Machine Translation, Machine Intelligence Technology Lab, Alibaba Group. ACL 2020.