Any Language Communications Inc.

High-Quality Machine Translation Using Relationship Analysis

Machine translation (MT) is different from other problems computers are asked to solve because natural language does not directly lend itself to either arithmetic or logical analyses. Virtually every aspect of language has been approached. Analyses have been done on the frequency of letters in words, on morphemes, on phrases, on every component of syntax, and on semantics. Statistical studies of word frequency coupled with context-free grammar rules attempt to create templates against which sentence fragments can be matched. None of these attempts has produced high-quality MT with an unrestricted vocabulary. The main problem with unrestricted MT systems has been, and continues to be, the development of a language-independent semantic interpreter. Researcher John McCarthy believes such an interpreter is the first step for artificial intelligence systems to be able to understand natural language.

While it's widely accepted that semantic analysis is required for true understanding, researchers have been stymied in attempts to do it. Two approaches are possible – either provide enough world knowledge for the computer to understand what is meant, or provide the analytical ability for the computer to interpret a sentence without understanding the world. Collecting world knowledge in a computer database has been largely deferred because of the massive amount of conflicting data that must be organized. Similarly, blind analysis has seemed intractable. Consequently, the most successful current systems are based on syntactic analysis with a semantic component, generally consisting of a keyword search coupled with a restricted vocabulary to resolve syntactic problems.

Any Language Communications (ALC) chose a semantics-oriented approach, called Relationship Analysis, to resolve the semantic analysis problem. In addition, while many commercial MT systems use semantic information to support their main syntactic analysis, Relationship Analysis turns this around by using syntactic information to support the semantic interpreter. Relationship Analysis has demonstrated high-quality translations for Arabic, English, French, German, Hindi, and Russian.

Major Features of the System

1. High-quality machine translations with an unrestricted vocabulary, between any of the languages supported on the system.
2. Contextual natural language understanding of words and phrases, including homonyms.
3. The availability of numeric language concept codes from the translations for more complex analyses by other computer programs.
4. Easy expandability to translate other languages. No language pair restrictions are required.

Relationship Analysis

Relationship Analysis is based on five concepts: The Theory of Universal Grammar, a weighted inheritance system, a specially developed number language, a type of genetic algorithm, and a data-driven

design. By merging these concepts, we've developed a system that's resolved the semantic analysis problem. In addition, by constructing parallel language databases and using numeric codes, Relationship Analysis can analyze semantics in any language supported by the system.

Theory of Universal Grammar: Proposed by Noam Chomsky at MIT in the 1950s, this theory posits that language is composed of two components: A surface structure and a deep structure. The surface structure consists of the word order, word endings, parts of speech, etc. (the syntax), and is specific for each language. So, English has its own syntax, Arabic has its own, and so forth. The deep structure is the actual meaning of the words (the semantics), and is universal across all languages. This means that the deep structure interpreted in English will have the same deep structure in French or in Chinese, or in any language. The ALC system includes surface structure components in the form of syntax analyzers for each source language and word arrangers for each target language, and a deep structure component in the form of a language-independent semantic interpreter.

Class/Category hierarchy: To implement deep structure analysis, we've developed a copyrighted class/category hierarchy. Containing five levels, in which each lower level inherits characteristics from the level above it, this hierarchy permits recognizing relationships between words. There are 16 classes and over 1000 categories that include all possible language concepts. In addition, weights are assigned to each word within the class/category structure, and these weights vary depending on relationships with other words in the sentence. This class/category structure has been found to be compliant with the world's major language families (Chinese, Germanic, Indic, Japanese, Malayo-Polynesian, Romance, Semitic, and Slavic).

Number language: Words have always been difficult for computers to evaluate. Consequently, each word and phrase entered in the system is transformed into a number that represents its relative place in the class/category organization. By forming pairs of the sentence words/idioms and comparing the values of these relative places (adjusted by the class/category weights), a value for the pair relationship can be obtained. Such valuations are calculated for all pairs in the sentence.

Genetic algorithm: The number of possible meanings for words quickly produces a massive number of possible sentence interpretations. Even a seven-word sentence can easily result in over 100,000 possible interpretations. While only a few of the sentences will be "sensible", the computer has no way of knowing which are and which aren't, and the combinatorial explosion of the analysis can overrun the processing capabilities of most computers. In fact, this was one of the major reasons for failure in early MT attempts. Recently, a mathematical technique called "genetic algorithms" was developed to address these "hard" problems, and has been applied to weather forecasting, pipeline analysis, traveling salesman problems, etc. Conceptually similar to the way body cells produce DNA, the most viable products survive to combine with other viable products to produce the "fittest" final product. In Relationship Analysis, partial sentence solutions are compared with each other, with the "best" ones remaining while the others are cast off. These partial sentence comparisons are done with numeric totals, using the class/category hierarchy and the number language, from the word pair Relationship Analysis. The most reasonable word pairs tend to combine to produce the highest totals. Through multiple combinations and adjustments, the best sentence is developed. This may be the first use of genetic algorithm methods for natural language analysis.

Data-driven design: One of the key features of Relationship Analysis is that the same semantic interpretation software can be used to understand any natural language. This is possible because the information necessary to understand a language is contained in parallel language dictionary databases, not in the code. All that's required to interpret a particular language is to point the system to the dictionary for that language.

Because Relationship Analysis physically separates the language-specific syntactic components from the language-independent semantic component, complete translation flexibility is available. Our syntactic components are a syntax analyzer for each source language and a word arranger for each target language. So any language that has a syntax analyzer can be translated into any language that has a word arranger. The "language pair" restriction common in commercial MT systems is not a restriction with Relationship Analysis. As long as a word arranger has been written for English, any language may be translated into English as long as that language has a syntax analyzer and a dictionary of words/phrases. Similarly, to translate all source languages into Spanish, all that would be needed is a word arranger for Spanish and a Spanish dictionary. Figure 1 illustrates this flexibility for various languages.

Translation Examples

The following examples show the power and flexibility of Relationship Analysis for natural language understanding and machine translation. Note that this also demonstrates that the system can be used for many topics, as an indication of the unrestricted vocabulary capability of the system. None of these examples can be correctly analyzed by syntactic systems.

With English as the source language and French and German as target languages:

The following two groups show that Relationship Analysis can disambiguate syntactically identical sentences with a homonym.

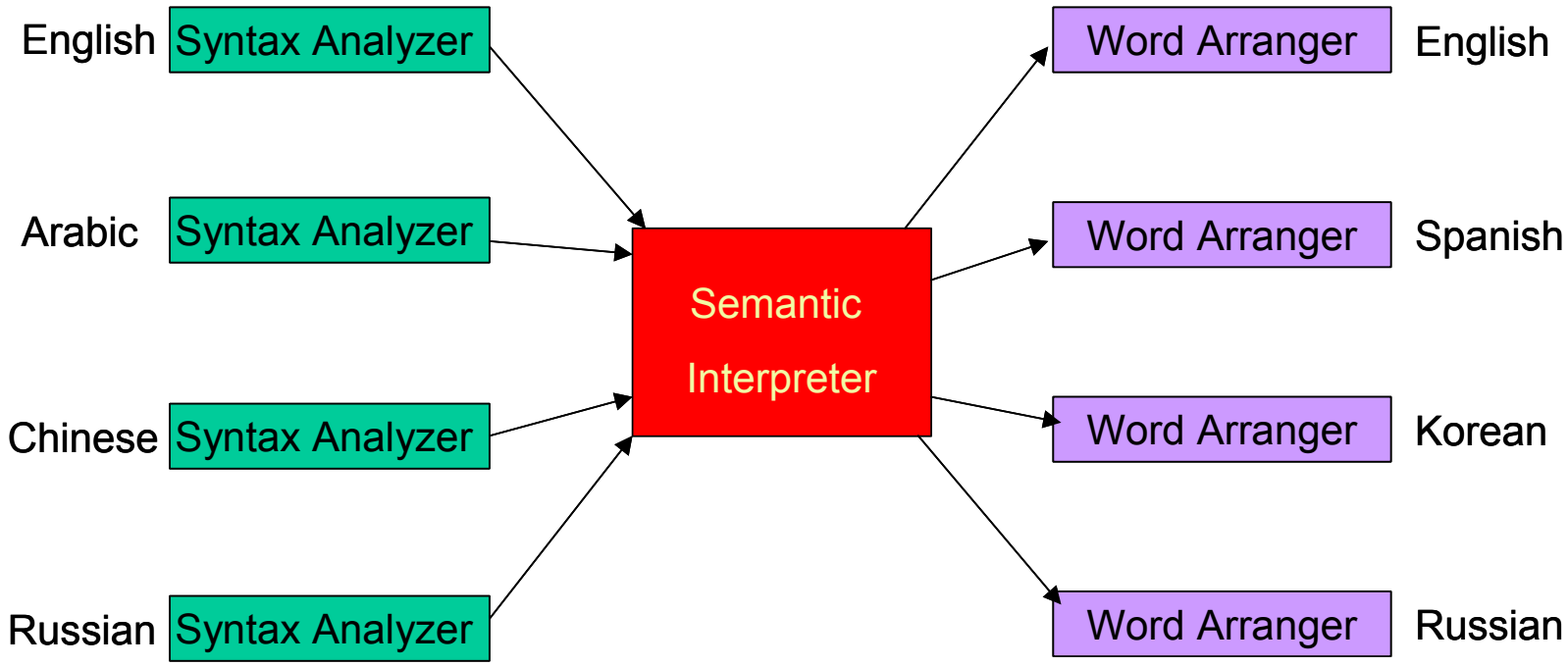
My refrigerator is running and my nose is running.
Mon réfrigérateur fonctionne et mon nez coule.
Mein Kühlschrank läuft und meine Nase rinnt.

My candidate is running.
Mon candidat se présente aux élections.
Mein Kandidat stellt sich der Wahl.

The following group shows that Relationship Analysis uses all the sentence information to determine the best overall meaning (compare with the previous sentence, "My candidate is running").

My candidate is running a temperature.
Mon candidat fait une fièvre.
Mein Kandidat hat ein Fieber.

Figure 1. Relationship Analysis Translation Flexibility



The following two groups show that Relationship Analysis is sensitive to changes in sentence meaning caused by changing a single word.

The hot dog is ready to eat.
Le hot-dog est prêt à manger.
Die Frankfurter Wurst is fertig zu essen.

The hot dog is ready to bark.
Le chien chaud est prêt à aboyer.
Der heisse Hund ist fertig zu bellen.

With Arabic as the source language and English, French and German as target languages:

The following two groups show that Relationship Analysis also disambiguates syntactically identical sentences with a homonym in Arabic (note that the first word means either "dealt with" or "ate", depending on the context).

تناول الأستاذ الموضوع في الدرس
The professor dealt with the subject during the lesson.
Le professeur a traité le sujet pendant la leçon.
Der Professor hat der Gegenstand während des Unterricht behandelt.

تناول الأستاذ الطعام في الدرس
The professor ate the food during the lesson.
Le professeur mange la nourriture pendant la leçon.
Der Professor aβ die Speise während des Unterricht.

The following two groups show that Relationship Analysis can recognize "animal" sounds from "human" sounds (note that the first Arabic word is identical in both groups, but changes meaning depending if the sound is animal or human).

يصيح الديك
The rooster crows.
Le coq chante.
Der Hahn kräht.

يصيح الأستاذ على تلاميذه
The professor yells at his students.
Le professor hurle à ses étudiants.
Der Professor schimpft bei seine Studenten.

The following two groups shows that Relationship Analysis uses all the sentence information to determine the best overall meaning in Arabic as well as it did in English. Note that the two sentences are identical except that an additional word has been added to the second sentence, changing its meaning. In addition, note that Relationship Analysis allows identification of "places" or "people" in the target language, permitting the proper words to be selected in the French and German translations.

أريد أن أزور جدة
I want to visit Jeddah.
Je veux visiter Jeddah.
Ich will nach Jeddah reisen.

أريد أن أزور جدة زوجتي
I want to visit my wife's grandmother.
Je veux rendre visite à la grandmère de ma femme.
Ich will die Grossmutter meiner Frau besuchen.

Summary

Relationship Analysis is a powerful semantics-oriented analysis technique that produces contextually correct interpretations of words and phrases, and linguistically accurate machine translations of those words, in a variety of natural languages. Relationship Analysis can also recognize nuances in messages not possible in purely syntactic approaches, permitting more fluent translations into target languages. In addition, Relationship Analysis produces numeric codes for the semantic interpretation of the words, permitting further computer analysis of the message.