



Semantic Multilingual Natural Language Understanding Using Relationship Analysis

Current approaches to natural language understanding involve statistical analyses to select meanings from a "world knowledge" database, to interpret the contextual meaning of messages. However, Any Language Communications has developed a novel system that uses the innate relationships of the words in a sensible message to determine the contextual meanings of all the words, without using world knowledge. The methodology is called "Relationship Analysis" and includes a weighted inheritance system, a number language word conversion, and a tailored genetic algorithm to select the best of the possible word meanings.

Relationship Analysis is a powerful language-independent method that's been tested with Arabic, English, French, German, Hindi, and Russian. When used with a k-nearest neighbor clustering method and a semantic taxonomy of language concepts, Relationship Analysis can do text analysis, data mining, high-quality machine translation, semantic metadata indexing, and semantic tagging for text in any language.

Background

Since the 1950s, as part of the field of artificial intelligence, there's been an attempt to collect information of all sorts into computer databases. However, organizing this massive amount of "world knowledge" has been difficult because the information is often contradictory. Natural language understanding (NLU) researchers hope to use this world knowledge to understand the semantics of messages for machine translation systems. Until world knowledge collection and organization are resolved, the process generally used is:

1. A human translator creates an excellent interpretation of a source language message.
2. This interpretation is stored as a semantic example.

3. New messages are analyzed and, if the new message is statistically similar to a saved example, that example is used as the semantic interpretation of the new message.

Language researchers throughout industry and academia use some combination of statistics, rules, and/or examples with these semantic interpretations to support their base syntactic analyses.

To avoid the manual effort required to create and interpret the many examples needed for example-based approaches, the Latent Semantic Indexing (LSI) method has been developed. Basically, a large collection of related documents are inspected for similarities among "significant" words, and a mapping is made of how closely related those words (concepts) seem to be. Better mappings are made by increasing the number of related documents inspected. A new message is compared to the map and, if a statistically close relationship exists, the message is considered to have similar concepts to the mapped messages. LSI can find messages containing concepts of interest to users in context, but it does have several weaknesses:

1. LSI requires training by the user from a pre-existing large collection of related documents.
2. LSI removes words it considers to have no significance and ignores grammar, making it unsuitable for machine translation.
3. LSI maps are created dynamically, so similar concepts may be mapped very differently.

A New Method

Any Language Communications (ALC) has developed a different method for the semantic analysis of free text that does not use either examples or LSI techniques to interpret natural language. The method includes a semantic taxonomy that uses a dynamic k-nearest neighbor clustering analysis with syntactic assistance. To properly analyze semantic concepts in any language, the taxonomy is composed of four dimensions and is called the Language Independent Semantic Taxonomy (LIST). The LIST can be considered a fixed mapping of language concepts that permits more intelligence into the meaning of each concept, and permits recognition of relationships among concepts. The method also keeps the message structure intact for possible translation into other languages. The method is essentially a mapping of language concepts from 4-space into clusters that describe the various interpretations of the



message, with the best interpretation being the one with the optimal mapping.

This method, used with Relationship Analysis, has been demonstrated for text analysis, data mining, and high-quality machine translation applications with Arabic, English, French, German, Hindi, and Russian messages. The method has also been used for designs of semantic metadata indexing and semantic text tagging applications. All interpretations are in context. Using this method, focused crawling of Web pages could be done on the content of the pages in context, rather than using URL extensions, hyperlink structure, or patterns of user behavior to recognize context.

Relationship Analysis

Human natural language understanding is generally considered from the listener's perspective, in which the world knowledge of the listener is used to disambiguate the various meanings of words s/he hears. Viewing communication from the listener's perspective leads to analyses using listener examples, i.e., example-based statistics.

Relationship Analysis views communication from the speaker's perspective, in which the speaker, in constructing a message to express a concept, has automatically (and invisibly) assigned only one meaning to each of the words. This can be illustrated by

I like to play squash

in which both "play" and "squash" have multiple possible meanings if expressed separately, but only one meaning in context. This context, assigned by the speaker and referred to as the "natural intelligence" of the message, requires no world knowledge from the listener. Using the Relationship Analysis approach permits the development of computer systems to extract this inherent natural intelligence from messages. Major Relationship Analysis components are:

Weighted Inheritance System. Word/phrase meanings are initially assigned a weight based on their common interpretation in the dictionary. For example, "hot" meaning "extremely warm" has a higher initial weight than "hot" meaning "radioactive". However, these weights are adjusted depending on relationships with other words in the message.

Any Language Communications Inc.

Number Language. Words have always been difficult for computers to evaluate. Consequently, each word/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 words/phrases and comparing their relationship values, values for all message elements are obtained. Such valuations are calculated for all the possible word/phrase meanings in the message.

Genetic Algorithm. The possible meanings for words quickly produce a massive number of possible message interpretations. Even a seven-word sentence can easily result in over 100,000 possible sentence 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 NLU 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 message solutions are compared with each other, with the "best" ones remaining while the others are cast off. Through multiple combinations and adjustments, the best message is developed. This may be the first use of genetic algorithm methods for natural language analysis.

Basically, Relationship Analysis is an approach to disambiguate natural language. It compares all the meanings of each word/phrase in a message with all the meanings of the other words/phrases to determine the closest relationships. Highly-related pairs of words/phrases are combined with other highly-related words/phrases until the best overall understanding of the message is selected. Relationship Analysis permits contextual natural language understanding with an unrestricted vocabulary, and uses the same semantic software irrespective of the natural language under analysis.

Because the analysis is in context, Relationship Analysis is able to disambiguate homonyms. For example, Relationship Analysis recognizes the different meanings of "running" in syntactically identical sentences such as



My refrigerator is running.
My candidate is running.
My nose is running.

k-Nearest Neighbor Clustering

While the k-nearest neighbor algorithm is most often used for classification, ALC has designed a variation of it that uses clusters for estimating relationships between words/phrases. This variation can be viewed as dynamic clusters of the meanings of the words/phrases in a message, with each cluster containing one complete set of relationships.

Word/phrase pairs are established for all meanings of all words/phrases. The "nearest neighbor" of each meaning of each pair is selected as the closest relationship and is assigned the highest numeric total, with more distant pairs considered as weaker relationships and assigned appropriately reduced numeric totals. The algorithm is modified by certain taxonomy characteristics and syntactic restrictions. Clusters are formed by adding word/phrase pairs that generate the highest combined relationship totals, with consistent meanings maintained throughout. For example, in "I like to play squash", consider the following hypothetical relationship totals:

play (sports):squash (game) = 1154
like (alike):squash (vegetable) = 437
like (loves):squash (game) = 296

The "like (loves)" would be selected as the meaning of "like" for this relationship, even though the total for "like (alike)" is higher. That's because the meaning of "squash" must be consistent within the extended relationship.

k-Nearest neighbor classically creates clusters with unique elements, with an element belonging to only one cluster. In the ALC variation, the same elements can belong to many clusters. Clusters are validated using the LIST. The system is dynamic in that clusters are initially chosen based on a first evaluation of the word/phrase relationships, but members of those clusters may change as the evaluation process continues. Cluster members are continually re-evaluated to improve the overall total of the message. As message length and complexity increases, cluster members change more frequently.

Any Language Communications Inc.

As Relationship Analysis proceeds, the weakest clusters are eliminated and the "cluster universe" is reduced to those requiring more detailed inspection. However, even in the final stages the same elements may be contained in multiple clusters.

Language Independent Semantic Taxonomy (LIST)

The LIST is an organization of the semantics of language and an implementation of the Deep Structure of the Theory of Universal Grammar. The LIST has been inspected for completeness against the following language families, comprising the first language of over 70% of the world's people: the Chinese family, the Germanic family, the Indic family, the Japanese family, the Malayo-Polynesian family, the Romance family, the Semitic family, and the Slavic family.

The LIST contains four dimensions:

1. Numeric semantic structure
2. Semantic weight
3. Antonyms
4. Distant attributes

Numeric Semantic Structure. ALC has organized language concepts into a hierarchical numeric semantic structure. There are 16 overall classes, detailed by numerous sub-classifications as needed for each concept (see Figure 1). Currently, nouns are described to 12 levels of detail and other parts-of-speech are described to 4 levels of detail. For example, the verb "like" (meaning "loves") is described as "11, 3, 7, 0". Words in other languages with the equivalent meaning to "like" (meaning "loves") would be described with the same numeric code. By using such parallel dictionaries, equivalent concepts in any language can be expressed with the same numbers, permitting the semantic analysis software to be language independent. The current full numeric semantic structure can be seen [here](#).

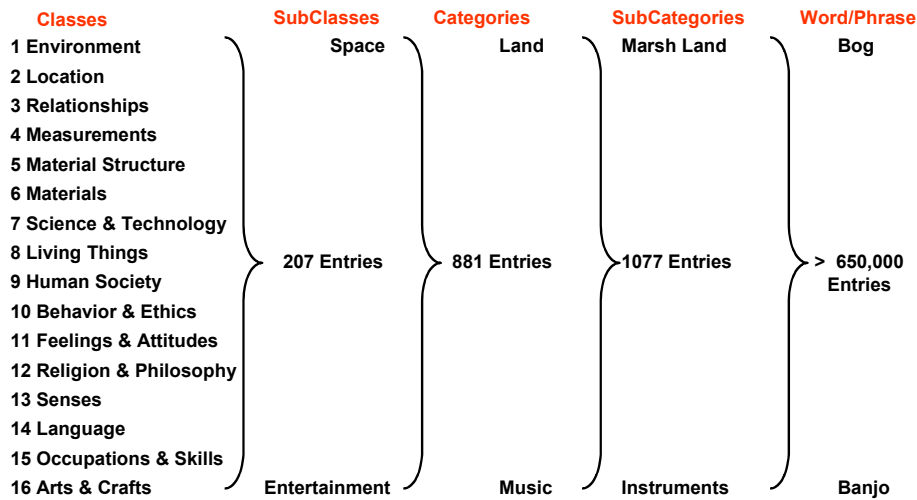


Figure 1 - Numeric Semantic Structure

Semantic Weight. Each meaning of each word/phrase is assigned a "weight", depending on how that meaning may be evaluated within the message context. Current weights are Default, Global, Local, and Ignore. For example, "hot" can have 18 meanings, including "heated", "being excited", and "radioactive". "Heated" is a Default weight, because people generally think of "heated" when they see "hot". "Being excited" is a Global weight, because that meaning is used in diverse contexts of "hot". "Radioactive" is a Local weight, because that meaning is restricted to a focused domain. Commonly used words (such as "is" or "the") are assigned a weight of "Ignore", which eliminates their influence as part of the Relationship Analysis calculations.

Antonyms. The LIST is arranged for relationship recognition among language concepts. Since the analysis is done via k-nearest neighbor and antonyms are similar concepts but opposite in meaning, they must be recognized and handled. For example, "Being Attractive" and "Being Ugly" are both concepts of physical attraction and are consequently close neighbors, but analysis of a message like, "That's a pretty ugly tie" must recognize those opposite meanings. The Antonym dimension of the LIST performs this task.



Distant Attributes. Sometimes an important attribute of a concept is distant in the LIST from that concept and would, in the course of Relationship Analysis, generate a weak relationship calculation. For example, "writing" is an attribute of humans, but not of all animals. If it was placed as a near-neighbor to humans in the LIST, then relationships with any animal for writing would also be strong. The Distant Attributes dimension is available for such close relationships between specific concepts for the analysis.

Summary

Relationship Analysis is a powerful semantics-oriented analysis technique that produces contextually correct interpretations of words and phrases, and linguistically accurate machine translations, in a variety of natural languages. Relationship Analysis can also recognize nuances in messages not possible in purely syntactic approaches. For example, in sentences such as "They met at the bank to withdraw money", "They met at the bank where the fishing was best", and "They met at the bank of spotlights", semantic analysis correctly interprets "bank" in every case. When used with a structured taxonomy in which the various meanings of words/phrases are organized, the most-related interpretation is selected as the most likely meaning of the message. This technique allows on-the-fly analyses of individual messages, in context, without the need for example-based statistics or latent semantic indexing methods.

By using the structured taxonomy to create parallel dictionaries, the method becomes language-independent and messages in any language can be interpreted using the same semantic code. The method generates a numeric interpretation for each word/phrase in the message, which the system can translate into words for human users or can forward to other software for further analysis. Using this method has facilitated the development of text analysis, data mining, and high-quality machine translation applications, and has been the basis semantic metadata indexing and semantic tagging system designs.