Any Language Communications Inc.

Multilingual Natural Language Understanding Using Relationship Analysis

Abstract. Current approaches to natural language understanding involve example-based statistical analyses or Latent Semantic Indexing 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 true contextual meaning of the message. This patented methodology is called "Relationship Analysis" and includes a class/category structure of language concepts, 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 has been tested using machine translations with English, French, and Arabic as source languages and English, French, German, Hindi, and Russian as target languages. A simplified form of Relationship Analysis does sophisticated text analyses, in which concepts in the text are recognized irrespective of the text language. Such analyses have been demonstrated using English and Arabic texts, with applications that include concept searches, email routing, semantic tagging, and semantic metadata indexing. In addition, a class/category data analysis provides machine-readable codes suitable for further computer system processing.

The Natural Intelligence of Language

Natural Language Understanding (NLU) is the analysis of communication between a "speaker" and a "listener", whether those individuals are communicating via literature, voice, or another medium. The listener interprets the speaker's intentions, picking the one meaning for each of the words/phrases that best matches the overall meaning of the message. Since people do this with apparent ease, the approach to computerizing NLU has been to mimic the human communications environment. That environment has been assumed to be based on the listener's "world knowledge", gleaned from a lifetime of experiences. For more than 10 years a major research effort has been undertaken to collect, categorize, and store this massive amount of often contradictory world knowledge information. The best analyses seem to rely on statistical methods, and nearly all NLU research in recent years has been to find the most successful statistical approach.

But what if human communication doesn't rely on the listener's analysis of the speaker's intentions at all? The speaker, in constructing a clear message, ensures that each word/phrase in the message has only one meaning. The listener extracts this "natural intelligence" from the message and recognizes the overall meaning from the collection of all the possible individual meanings. In other words, for sensible sentences the listener doesn't need world knowledge, and computers don't need it either.

Consider the following sentence:

They met at the bank.

This sentence is ambiguous and therefore can't be understood. For NLU (either human or machine) to be successful, the sentence must be further explained, such as:

They met at the bank to withdraw money.
They met at the bank where the fishing was best.
They met at the bank of spotlights.

As these examples illustrate, a message is ambiguous unless each of the words/phrases has a single distinct meaning. That meaning, that natural intelligence, makes sense - if the speaker didn't construct the message with distinct meanings for each word/phrase, the message would be ambiguous and neither human nor computer could understand it. Any Language Communications has found a way for computers to extract this natural intelligence through a method we call "Relationship Analysis."

Relationship Analysis

The basis for Relationship Analysis is our premise that the meaning of each word/phrase in a message can be determined from the possible meanings of the other words/phrases in the message. The relationships between these various possible meanings are clarified by using a copyrighted multilevel class/category structure of language concepts we've developed, called the Language Independent Semantic Taxonomy (LIST), containing 16 classes (such as Living Things, Human Society, Behavior & Ethics, etc.), and over 1000 categories (such as Sleeping, Eating, Medicine, etc.). A language in each of the "major" language families (comprising the first language of over 70% of the world's people) has been inspected for consistency with the LIST. Those language families are 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. All languages contain every category and no word/phrase has been found that doesn't fit some category.

Analysis of LIST relationships involves a weighted inheritance system for the words/phrases, a specially developed number language, and a type of tailored genetic algorithm. The underlying analysis uses a dynamic four-dimensional k-nearest neighbor clustering method.

It's clear that Relationship Analysis is a semantic analysis, but we've also found that including some language-specific syntax information (such as dependent/independent phrase parsing, part-of-speech clarification, possessive information, etc.) is necessary to produce accurate results, and syntax guidance distinctly improves system performance. We've also found that Relationship Analysis is language independent – messages in a variety of natural languages have been analyzed using the same software. The system has been tested with English, French, and Arabic as source languages and English, French, German, Hindi, and Russian as target languages. The following paragraphs describe major aspects of Relationship Analysis components.

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. We've also found that, occasionally, language processing requires additional analysis to link category relationships beyond those found in the hierarchy.

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 LIST organization. By forming pairs of the message words/phrases and comparing the values of their relative places (adjusted by the weights), a value for the pair relationship is obtained. Such valuations are calculated for all pairs and all 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 such as "The hot dog is ready to eat" results 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.

Language Independence and Machine Translation

We have constructed Relationship Analysis to be language independent. That means the same semantic interpreter can be (and is) used to interpret messages irrespective of the source language. While the correctness of interpretation results can be shown by the computer codes produced by the semantic interpreter, a more compelling demonstration is with machine translations (MT) of the codes. We've chosen several examples to illustrate semantic problems in sentence understanding. Note that none of these examples can be analyzed by purely syntactic processes, and none of them are properly interpreted by current example-based statistical systems.

For example, 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.

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.

The following groups show that Relationship Analysis also disambiguates syntactically identical sentences with a homonym in French.
C'est la crème pour le café.
That is the cream for the coffee.

C'est la table pour le café.
That is the table for the café.

C'est la crème de la promotion.
That is the top of the class.

With Arabic as the source language and English as the target language:

The following two groups show that Relationship Analysis also disambiguates syntactically identical sentences with a homonym in Arabic (the first Arabic word is either "dealt with" or "ate", depending on the context).
تناول الأستاذ الموضوع في الدرس
The professor dealt with the subject during the lesson.

تناول الأستاذ الطعام في الدرس
The professor ate the food during the lesson.

The following two groups show that Relationship Analysis can recognize "animal" sounds from "human" sounds (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.

يصيح الأستاذ على تلاميذه
The professor yells at his students.

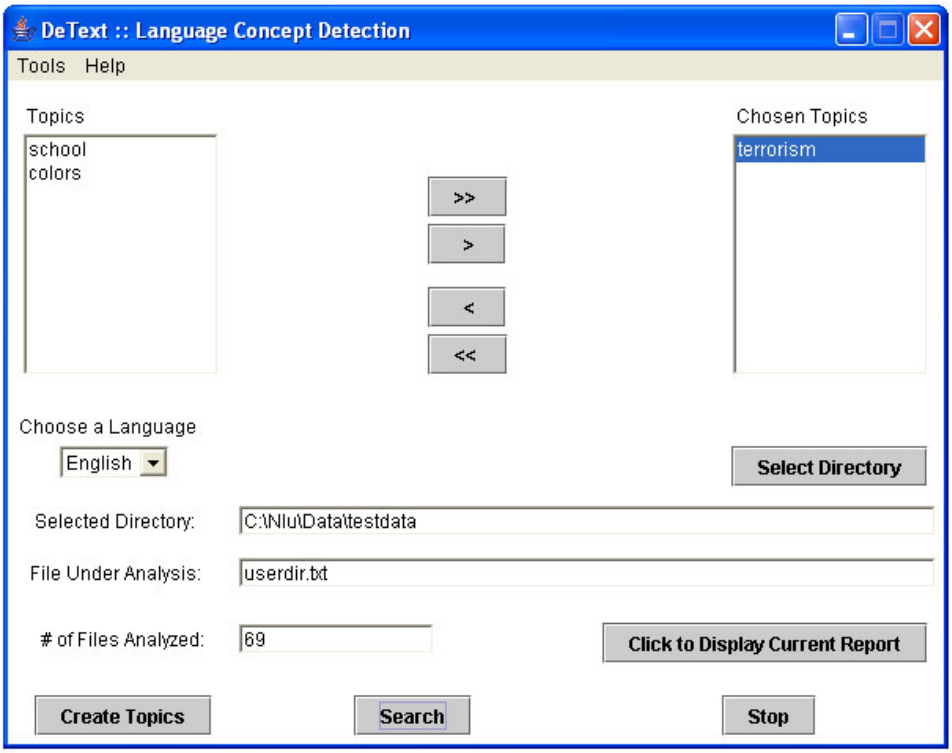
The following two groups show that Relationship Analysis uses all the sentence information to determine the best overall meaning in Arabic as well as it did in English. Note that the two sentences are identical except that a word has been added to the second sentence, changing the meaning of "Jeddah" to "grandmother". In addition, note that Relationship Analysis allows identification of "places" or "people" in the target language, permitting proper 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.

Text Analysis

While MT requires detailed Relationship Analysis to select the proper meaning of each word/phrase, a more general text analysis to determine overall concepts can be done using a simplified form of Relationship Analysis. Such concept determination could be used commercially for message routing or as a sophisticated search to find text referring to topics of interest, for further consideration by a human analyst. We've developed a system called DeText to do this kind of software analysis.



DeText permits users to describe Topics (such as terrorism, global warming, etc.) in general, and then use those general descriptions to find references to those topics in new text. For example, if "banks" was the Topic, the system needs to know if financial institutions, or aircraft movements, or river structures, etc. is the topic of interest. Users describe a topic by choosing five words that remind him/her of the topic. For our example, "terrorism" is the Topic, which was described with "bombs", "fight", "guns", "explosions", and "war" as clarifying words.

Scenario 1 – Tricky terrorists. Let's say computers have been seized from possible terrorists. These computers contain hundreds of files, each hundreds of pages long, with file names such as "antiques", "sports", "recipes", etc. However, the terrorists know to go to page 127 of a particular file to find a single paragraph detailing their plans. If the police had to manually read all those files, the specific paragraph may be missed or may not be discovered until after the terrorist activity has happened. Relationship Analysis (through DeText) to the rescue!

In doing the DeText scan, the following paragraph was discovered:

"Have snipers from our militant branch hide near the main entrance to Parliament. When the Prime Minister comes out, he will be shot by at least two snipers from different directions at the same time. This will cause initial confusion by the security forces that should permit our men to escape. One of the snipers will leave evidence that casts blame on a renegade branch of the opposition party."

Note that this paragraph doesn't contain any of the "topic" or "clarification" words (terrorism, bombs, fight, guns, explosions, war). DeText uses those words only to identify the concept to search for, and finds information that relates to that concept.

Scenario 2 – Trickier terrorists. But what if the hundreds of files, each containing hundreds of pages, are in Arabic? Since Relationship Analysis is concerned only with concepts, the text language doesn't matter. By selecting Arabic as the text language, the same scan can be done which, using our test files, yields the following paragraph:

بلير ينتقد الخطة الفرنسية الألمانية بنزع سلاح العراق ويقول إنه من العبث الاعتقاد بأن مفتشي الأسلحة قد يعثرون على أسلحة الدمار الشامل دون تعاون كامل من بغداد. البرلمان التركي يبحث نشر آلاف الجنود الأميركان. بوش يعتبر أن إصدار مجلس الأمن لقرار جديد يجيز العمل العسكري ليس أمرا ضروريا لواشنطن.

Note that the Topic and Clarification words could be in English, even if the text language is not English.

Email Routing

The text analysis application of Relationship Analysis also has a direct use for commercial email routing. Large companies get hundreds of emails from customers every day asking for product information, to register a compliment or a complaint, to ask for service, etc. Currently, these emails are manually routed to the appropriate customer service area.

An embedded version of DeText can be used as part of an email routing system. Emails can be forwarded for semantic interpretation and automatically routed, providing faster and less expensive customer service. Using the interlingual nature of Relationship Analysis, user emails can be written in any language supported by the system and still be semantically analyzed and routed to the correct area.

Semantic Tags

Another use for Relationship Analysis is to set machine-readable semantic tags in text, which can be referenced by analysis software in other computer systems. The LIST data structure defines numeric codes that uniquely identify semantic concepts to an arbitrarily low level. The following example is the last internal file generated by the NLU software for MT purposes:

```
sentences(1,3,[1],"the","",126,1,50,"adj",'N')
sentences(1,3,[2],"hot","",238,1,78,"adj",'N')
sentences(1,3,[3],"dog","",333,2,214,"subject",'N')
sentences(1,3,[4],"is","",730,1,107,"verb",'N')
sentences(1,3,[5,6],"ready_to","",138,1,19,"prep",'N')
sentences(1,3,[7],"bark","",784,2,1,"verb",'N')
sentences(1,3,[8],".","",0,1,0,"",'N')
```

The numbers (shown in color here) permit the MT software to select the specific target language word/phrase to correspond with the source language concept. However, they can also be written as hidden fields with the text or used in an index file to function as semantic tags for sophisticated search and analysis software. Note that these numeric tags are language independent – the same numbers are used irrespective of the source language. This permits analyses across natural language boundaries.

Semantic Metadata Indexing

An extension to semantic tags is semantic metadata indexing, in which the concepts contained in text are identified and categorized in an index to permit faster recognition of related items. Two approaches are to index the locations of embedded semantic tags or to index the text concepts without tagging them. While Relationship Analysis can do either approach, we prefer the non-tag index because

1. Tagging requires the additional step to tag text within documents.
2. Tagging requires tag consistency among organizations and for all languages.
3. Tagging restricts the index to tagged text.
4. Tagging may be inaccurate if the text changes.

The LIST organization provides a natural index structure and Relationship Analysis handles the semantic interpretation, for text in any language supported by the system, for Web pages and non-Web pages in various formats (.TXT, .DOC, .RTF, etc.). Our system design also recognizes changed text pages, and automatically re-indexes them.

Semantic metadata indexing permits users to quickly search large text repositories with complex queries. For example, a user may request, "Give me all information on arms purchases by Iran that were not small arms." Only the information directly relevant to the request will be returned.

Summary

Relationship Analysis is a powerful semantics-oriented analysis technique that has produced contextually correct interpretations of words and phrases, and linguistically accurate machine translations of those words, in a variety of natural languages. We have also found that Relationship Analysis can recognize nuances in messages not possible in purely syntactic approaches. Identified potential applications for Relationship Analysis are natural language understanding, machine translation, text analysis, email routing, semantic tags, and semantic metadata indexing. Research continues on the relationship between syntax and semantics, and development continues to test and expand the system to other languages.