Any Language Communications Inc.

Search RAE - Multilingual Semantic Concept Detection Using Relationship Analysis

Recognizing language concepts in documents has been difficult because of the inherent necessity to understand the context in which the words are written. For example, "bank" will refer to different concepts in sentences like, "They met at the bank to withdraw money", "They met at the bank where the fishing was best", or "They met at the bank of spotlights." It is evident that syntactic analysis is inadequate – semantic analysis must be done. Relationship Analysis, an interlingual semantic natural language understanding method, provides such capabilities. With parallel dictionaries and a taxonomy of over 14,000 language concepts, the Relationship Analysis Engine (RAE) has been used to develop Search RAE, a product that scans documents for specific topics of interest. Information that relates to the topics is harvested, in whatever languages are supported by the system. Since Relationship Analysis includes an inheritance-oriented hierarchical arrangement of concepts, Search RAE can analyze text at different classification levels of detail. For example, users can find information on "spaniels", or "canines", or "mammals", or "animals", or "living things", depending on their needs. Applications for Search RAE include security scans of suspicious computers and email routing for large organizations.

Multilingual Semantic Concepts

The history of semantic natural language understanding (NLU) has been one of great promise and few results. While significant advances have been made in the morphological, lexical, and syntactic analysis of language, semantic analysis remains elusive. Our approach was to use the old idea of "semantic concept groups", expanded and combined with recent views of various language structures, to develop a multilingual semantic text analysis system.

Grouping words and phrases into distinct concepts permits more meanings of the words to be understood. However, defining the concept groups is both labor intensive and somewhat arbitrary, and most NLU systems use fewer than 50 groups. Not surprisingly, vocabulary restrictions must be imposed to generate proper semantic interpretations.

To develop systems with an unrestricted vocabulary, we use our Language Independent Semantic Taxonomy (LIST) structure with more than 14,000 semantic concept groups. As part of our multilingual NLU and machine translation research and development, we've tested these groups for Arabic, English, French, German, Hindi, and Russian and have found them to be consistent across language boundaries. By ordering these semantic groups in an inheritance-oriented hierarchical structure, by assigning numeric codes to them, and by comparing the likelihood of each definition of each word with the definitions of all other words in a sentence phrase, we've been able to select the correct interpretation of all the words in a message. We call this process Relationship Analysis, and a simple application of it is used to scan texts for language concepts. The application, called Search RAE, works irrespective of the language in which the text was written, for all languages supported by the system.

Any Language Communications Inc.

Relationship Analysis for Semantic Scans

Search RAE is a software product designed to scan computer files for topics of interest. Other systems rely on user-selected keyword searches that require that the keywords be in the text for the text to be selected. Search RAE uses a simplified version of Relationship Analysis to generate semantic concept searches. The following example illustrates the process:

Search RAE requests a user to choose a Topic, then to identify five words that describe the topic. This clarification is needed to help the system understand the concepts for which to search. 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. For our example, "terrorism" is the Topic, with "bombs", "fight", "guns", "explosions", and "war" as clarifying words that came immediately to mind (others could just as likely have been chosen). A Relationship Analysis is then done on those six words to determine the semantic concepts. Let's consider two scenarios:

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 identifiers 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.

In doing a Search RAE scan, the following paragraph was discovered:

"Our plans are the following - the killing Monday of a worker from Bulgaria, who must be shot by a sniper outside the West Bank city of Jenin, about 45 miles north of Jerusalem. We have support from a renegade branch of al-Aqsa Martyrs Brigades, the militant wing of Arafat's and Abbas's mainstream Fatah movement, who will assert responsibility for the killing."

Note that this paragraph doesn't contain any of the Topic or Clarification words (terrorism, bombs, fight, guns, explosions, war). Search RAE uses those words only to identify the concepts to search for, and finds information that matches those concepts.

Scenario 2 – Trickier terrorists. But what if the hundreds of files, each containing hundreds of pages, are in Arabic? Since Relationship Analysis is numerically (rather than language) based and is concerned only with concepts, the text language doesn't matter, as long as the language is supported by the system. The same scan can be done, which 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 can be in English, even if the text language is different.

Relationship Analysis for 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. To emphasize the problem, a Gartner Group report predicts a 50% yearly increase in such email.

An embedded version of Search RAE 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.

Conclusions

Search RAE has demonstrated that multilingual interpretation of text is possible, and that semantic concepts can be isolated for a variety of languages. The practical applications are "topic of interest" scans of computer files and email routing, based on the intent of the message.

Search RAE identifies, but does not translate. It can be used as the back end of a speech recognition system and as the front end of a machine translation system (such as our tRAEnslate), providing a complete natural language processing system. In addition, it can be used as part of a message analysis system, to recognize the concepts to which the system should respond.