SEMANTICA

www.semantica-software.com

Dr. Laurent Debrauwer (degree: University of Lille, France)
Dr. Naouel Karam (degree: University of Clermont-Ferrand, France)
Dr. Scott Carpenter (degree: University of Wisconsin, Madison WI, USA)
Jean-Pierre Brulé (degree: École Polytechnique, Paris, France)

Textual Analysis, Semantic Search and Text Summarization with Distingo®

I. Statistical vs. Semantic Search Methodologies

One of the greatest problems facing data-intensive industries today is the sheer volume of data requiring analysis, sorting and retrieval. Organizations may deal with thousands or millions of documents in the form of e-mail, database records, text files, and more. Difficult enough when data is held in structured data fields, the problem is considerably more vexing when one is confronted with complex instances of natural language. Traditional search mechanisms focusing on statistics (i.e., the frequency of keywords) provide imperfect results: the keyword may be misspelled in target documents; it may appear in a plural or conjugated form; it may not even be present (replaced by a synonym, or merely implied); it may have different meanings according to context. Moreover, statistical keyword searches do not allow one to recognize the logical relationships that may exist among keywords (for instance, when one word is the agent of an action and the other the recipient). In such cases traditional searches will typically return results that prove either too voluminous or too restricted to be helpful.

A useful supplement or replacement for statistical analysis would focus not on the number of occurrences of a word, but on the *meaning* of a word or a sentence. Such a solution requires sophisticated parsing of the linguistic syntax of texts, a determination of the meaning (semantic content) of a text, and a comparison of this meaning with the user's query.

II. Distingo: Semantic Search in Natural Language Texts

Distingo is a semantic search tool for English (French available soon) that provides two such solutions as fully documented C++ APIs:

- **Distingo Index** works with the meaning of individual words, thereby allowing the user to broaden or narrow his search based on meaning rather than statistical frequency. Distingo Index allows the user to search for terms according to relations of similarity, such as synonyms, hypernyms, hyponyms, holonyms and meronyms.

- **Distingo Context** parses and distills the meaning not just of individual words, but of full sentences. It can perform many different functions, including:

    - **Semantic comparison**: Distingo can compare texts with similar meanings, even when these meanings are expressed with different words. (The result is expressed as a measurement of semantic similarity.)
    - **Text summarization**: Distingo can distill a simplified version of a complex text, recording this summary as an XML output. The resulting output may have many uses, including an faster search (or comparison) of documents, or accelerated skimming by human readers.
    - **Keyword extraction**: Distingo can generate lists of keywords from natural language texts, reducing words to their root forms (singulars of plurals, infinitives of verbs, etc.). Keywords may be stored and used for statistical searches.
    - **Disambiguation**: When a polysemous word occurs in a text, Distingo can evaluate the surrounding context in order to infer which meaning of a word is intended. Once this meaning is determined, it may be linked to related meanings in Distingo's linguistic ontology.
    - **Error tolerance**: When target texts include spelling or grammar errors, Distingo can correct many such errors on the fly, thereby increasing the likelihood of successful results.

Both tools use a proprietary linguistic ontology (a hierarchical database of words and their meanings); **Distingo Context** also employs a proprietary syntax parser (a tool for determining the relationships of words in a sentence).

Distingo relies on leading research in meaning analysis, description logics, and mathematical search methods. Description logics ("DLs" — also called *terminological logics*) are a family of knowledge representation formalisms designed for representing and reasoning about terminological knowledge. In DLs, the conceptual knowledge of an application domain is represented in terms of concepts (unary predicates) that are interpreted as sets of individuals, and roles (binary predicates) that are interpreted as binary relations between individuals.

III. Details: Sample Inputs and Outputs

A: Distingo Index

Semantic search tools are programs designed to help broaden keyword searches based on the meaning of the keyword. In a typical search engine, a query for the word "claim" (for example, in an insurance report or a police blotter) will return documents containing the word "claim." But the search will overlook all documents that contain other, similar words: many tools will not even recognize alternate forms of the keyword (such as "claims" or "claimed") some can extend the search to include verbal synonyms (e.g. "affirm" and

"assert") or related substantives (e.g., "allegation," "assertion," or "entitlement").

A semantic search tool could broaden the search in just this way. By locating the query word within a vast linguistic ontology (a hierarchical network), a truly semantic search tool would identify the possible forms of this word, as well as synonyms (words with similar meanings), hyponyms (words that are lower in the hierarchy, as "cat" is lower than "mammal"), hypernyms (words that are higher in the hierarchy, as "mammal" is higher than "cat"), and meronyms (words pertaining to the same object, as "wheel" is related to "automobile").

Additionally, the user may choose how much to narrow or broaden the search, selecting only the most proximate meanings, only horizontal relationships (synonyms, meronyms), only vertical relationships (hyponyms, hypernyms), etc. The user may limit searches to a single degree of separation, or as many as three degrees of separation.
Finally, a semantic search tool like Distingo Index can be tolerant of misspellings.

Moreover, **Distingo Index** can compare two words and produce a measurement of their "semantic proximity." Thus one can measure the semantic distance between "steal" and "take," between "automobile" and "vehicle," etc.

B: Distingo Context

If a semantic search based on a single word results in more powerful search results, the potential of a search engine that understands entire sentences is even greater.

Distingo Context shows how such a search is possible. This tool looks at the meaning of full sentences and documents. A contextual search tool, it recognizes the grammatical role played by words in the sentence (e.g., subject or object), and can detect the relationship between the parts of a sentence (objects, subjects, verbs, attributes, etc.), regardless of syntactic complexities such as the passive voice. Furthermore, **Distingo Context** can recognize similarities between texts *even if they do not include any of the same words.* The contextual search methodology employed by **Distingo Context** parses the sentences of a given text document in order to determine the syntactic composition of the phrases. It recognizes conjugated and plural forms of words, and it associates passive voice expressions with their active voice equivalents. Then it distills the meaning of the sentence by filtering it through a vast semantic ontology, resulting in a "translation" of the text's most basic meanings into a meta-language. After applying the same process to other texts, these texts can be compared with one another, resulting in the attribution of a coefficient — or a "proximity score" — that ranks the similarity of a text's ideas.

A sample comparison might begin with a query to be compared to several other texts. For example:

> **Query:** "**A person steals an electronic device**."
>
> **Text 1**: "**A pickpocket took** a wallet and a **cell phone** at a shopping center at the Winn- Dixie, 604 Crandon Blvd., at 9 p.m. Dec. 9. The victim has discovered items missing."
>
> **Text 2**: "A **burglar** broke into a truck and **stole** a purse. The truck was parked at

Northwest Fourth Avenue and Fourth Street. The incident was reported Dec. 8."

Text 3: "Two area **men** were arrested in September. They **abducted** a 17-year-old woman."

In this example, **Text 1** provides a perfect match: "Someone / took / cell phone." **Text 2** receives a lower matching score, because the object taken is not an electronic device. **Text 3** receives the lowest score, because "abduct" is more distant from the query verb ("steal") than "take", and "woman" is not an electronic device.

Note that a step of the comparison process is Distingo's distillation of the principal ideas of the text, which is stored in an XML schema. This intermediate output can be used for text summarization and/or keyword extraction.

IV. Problems and Solutions

Semantic search can be plagued by the difficulties associated with ambiguous meanings. Usually the difficulties are of the following sorts:

The definition of similarity: A statistical search for keywords simply looks for the repetition of certain words (i.e., *identical* instances), but Distingo looks for the repetition of ideas (i.e., *similar* instances). How similar does a match have to be before it is useful? Low similarity may produce too many results, many of which may be false positives. High similarity may produce too few results. Distingo thus allows users to calibrate the desired level of similarity with the user's needs and data. (Semantica personnel can help with this calibration.)

Ambiguity of search: Most words have several different meanings, and if Distingo does not know which meanings matter to the user, it may return results that lead in the wrong direction. Thus it is important that users can select the meaning or meanings they wish the search to pursue. Moreover, users can direct Distingo to infer the meaning of certain words based on their semantic context.

Discipline-specific language: Distingo's standard linguistic ontology includes many but not all specialized words and definitions. It is thus possible for users to add their own terms.

V. Applications in Business

Businesses, administrations, or other organizations dealing with large quantities of information need filters for retrieving pertinent information. While semantic and contextual search tools could someday be integrated into large, public search engines to great benefit, the more immediate applications typically lie in specialized uses. Here are just a few examples:

- An insurance company seeks to compare new claims to similar claims filed in the past; a single determination of whether to honor a claim or not could result in hundreds of thousands of dollars of savings or of extra cost.

- A government agency needs to screen vast amounts of information (such as customs forms, communications, tax information).

- A law office needs to extract keywords for the indexing of documents.

- A search engine needs to find keywords that may appear in inflected forms (e.g., plurals or conjugated forms), or even words that are spelled entirely differently but have similar meanings.

- A company needs to generate simplified versions of texts for skimming by human readers.

VI. Summary

Currently most consumer search engines sort their results according to various criteria, such as the number, proximity and location of terms matched; or page-related factors, such as the number of links made to a page or the number of times a page is accessed from a results list.

The ranking algorithms used by the search engines are not published and we know only a little about their ranking criteria. The novelty of the approaches described in this paper is that they **allow the user to express his query as a natural language description**. The criteria used when sorting the retrieved documents is semantic relevancy with respect to the query.

A number of similarity measures for ontological structures have been proposed in different domains like databases, artificial intelligence and semantic web. Some work extends the comparison to semantic structures (set of super and sub-concepts of a concept) and relations between the concepts. We believe that a semantic and contextual approach in search terminologies is more complete because it operates in semantic descriptions expressed in description logics rather than structures. In addition, it involves both name and complex description matching.

References:

R. Küsters, "Non-Standard Inferences in Description Logics," *2100 of Lecture Notes in Artificial Intelligence*. Springer-Verlag 2001.
Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge: University Press, 2003.
G. Bisson. Learning in FOL with a Similarity Measure. In *10th National Conference on Artificial Intelligence*. Morgan Kaufmann, 1992.
A. Budanitsky. *Semantic Distance in WordNet: An Experimental, Application- oriented Evaluation of Five Measures*, 2001.

VII. Technical and Contact Information

Distingo Index and **Distingo Context** are delivered as Application Programming Interfaces (API), to be integrated into other solutions. **Distingo** tools are supplied as fully documented C++ libraries, with an example of integration into an existing C++ program.

The result of Distingo's syntactic analysis is represented in XML. The calculation of semantic similarities may be in the form of a numerical coefficient, or an ontology showing the information present in the first ontology and missing in the second.

The format of the texts and of the XML is a string of characters in the programming language C.

Distingo Index and **Context** are available for English and will soon be available for French (somewhat limited feature set in French).

Distingo products are distributed by Semantica and Ultralingua, Inc. For information and pricing, contact:

business (at) semantica-software.com

contact (at) ultralingua.com