



Distinguo

Software for Text Analysis

Written by:
Dr. Laurent Debrauwer
Dr. Naouel Karam
Jean-Pierre Brulé

Semantic Searching and Text Analysis with Distinguo®

Automated analysis of natural language texts using mathematical models.

I. The Limitations of Statistic Search Methodologies

One of the greatest problems facing data-intensive industries today is the sheer volume of data that must be analyzed, sorted, and retrieved: organizations may deal with thousands or millions of documents in the form of e-mail, database records, text files, and more. Difficult enough when the data is located in structured data fields, the problem is exponentially more vexing when one is faced with vast quantities of natural language. Traditional search mechanisms focusing on statistics (i.e., the frequency of keywords) provide imperfect results: the keyword may be misspelled in some target documents; it may appear in a plural or conjugated form; it may be replaced by a synonym; it may have different meanings according to context. 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 syntax of texts, a determination of the meaning (semantic content) of a text, and a comparison of this meaning with the user's query.

Distinguo is a semantic search tool that provides two such solutions. **Distinguo Index** discerns the meaning of words, thereby allowing the user to broaden or narrow his search based on meaning rather than statistical frequency. **Distinguo Context** discerns the meaning not just of individual words, but of full sentences, allowing the user to search for texts containing similar meanings, although these texts express these meanings with different words. It can also compare entire texts with one another, resulting in a measurement of their similarity or difference.

II. Distinguo: Semantic Searches in Natural Language Texts

Distinguo 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

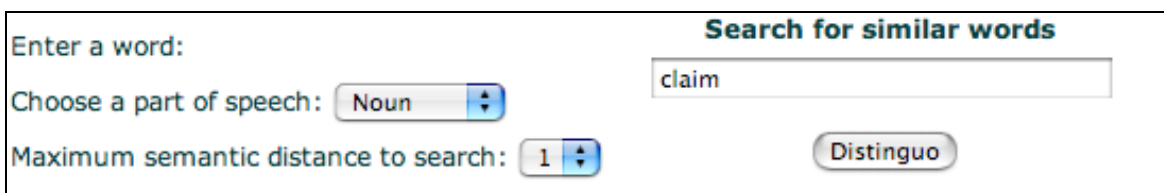
example, in an insurance report or a police blotter) will return documents containing the word "claim." But the search will overlook all documents that contain other, similar words: many tools will not even recognize alternate forms of the keyword (such as "claims" or "claimed") some can extend the search to include verbal synonyms (e.g. "affirm" and "assert") or related substantives (e.g., "allegation," "legal right," 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").

Moreover, 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 Distinguo Index can be tolerant of misspellings.

Distinguo Index is a revolutionary semantic search tool that works in just this way. The user enters a word:



The screenshot shows a web interface for the Distinguo Index. It has a title "Search for similar words" in blue. Below the title, there is a search input field containing the word "claim". To the left of the search field, there are two dropdown menus: "Choose a part of speech:" with "Noun" selected, and "Maximum semantic distance to search:" with "1" selected. A "Distinguo" button is located to the right of the search field.

And **Distinguo Index** returns the results (pictured here with a maximum degree of separation of "1"). Moreover, this semantic search tool 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.

Results for the noun "claim" thus appear as follows:

call	0
title	0
allegation	1
allegement	1
assertion	1
asseveration	1
avermment	1
demand	1
entitlement	1
insurance claim	1
legal right	1
own right	1
pretension	1
right	1

Distinguo Context

If a semantic search based on a single word results in more powerful results, imagine the potential of a search engine that understands entire *ideas*.

Distinguo 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, **Distinguo Context** can recognize similarities between texts *even if they do not include any of the same words*.

The contextual search methodology employed by **Distinguo 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; the result is a "translation" of the text's most basic meanings into a common meta-language. After applying the same process to other texts, these texts can be compared to 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 source text used for the query: "**Someone steals an electronic device.**" This query may be compared to several other texts, such as:

Text 1: "**Someone took** a wallet and a **cell phone** from a shopping 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.

III. Applications in Business

Businesses, administrations, or other organizations dealing with masses 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 lie in specialized uses. Here are just a few examples:

- An insurance company looking 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 charged with screening vast amounts of information (such as customs forms, communications, tax information).
- A company that requires more powerful ways of accessing information in a database or in its internal archive of documents.

Distinguo products rise to these challenges by:

- Parsing of texts in English or French (other languages available soon), to reduce texts to a simple XML output representing syntactic and semantic structures. **Distinguo** uses an ontological database of 500,000 terms/per language.
- Representing the text in the form of an ontology.
- Comparing the contents of a simple query to a large database of texts, locating texts containing similar meanings, and rating them according to a "similarity quotient" (rating of the closeness of the similarity). Technically, this is the calculation of the semantic similarity between two ontologies (i.e., between two texts).

IV. Summary

Nowadays, most consumer search engines sort their results according to number of criteria going from the number, proximity and location of terms matched, to 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.

Distinguo Index and **Distinguo Context** are based on cutting edge 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.

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.

Laurent Debrauwer's doctorate is in Computer Science, from the University of Lille (France)
Naouel Karam's doctorate is in Computer Science, from the University of Clermont-Ferrand (France)

Jean-Pierre Brulé, Polytechnician (École Polytechnique, France)

IV. Technical and Contact Information

Distinguo Index and **Distinguo Context** are based on algorithms for the parsing of language and the matching and ranking semantic results developed by Semantica Software of Luxembourg in conjunction with Ultralingua.

They are delivered as an Application Programming Interface (API), to be integrated into other solutions. **Distinguo** tools are supplied as fully documented C++ libraries, with an example of integration into an existing C++ program. It can be integrated into software solutions for its own features, or to supplement or refine statistical search methods.

The result of the syntactic analysis, as well as the format of the ontologies, 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.

Distinguo Index is available for English and French, and will be made available for other languages. **Distinguo Context** is available for English, but will also be expanded to other languages.

Distinguo Index and **Distinguo Context** are distributed by Ultralingua, Inc. For information, contact:

Chad Johnson
General Manager
Ultralingua, Inc.
1313 SE Fifth Street
Minneapolis, MN 55414
www.ultralingua.com
612-929-1400 (phone)
612-929-1401 (fax)