In the Heart of Semantic Technology

Paolo Lombardi

Expert System
Via Virgilio 56/Q 41100 Modena
Italy
plombardi@expertsystem.it

Abstract

The heart of semantic technology is a semantic network that is a lexical database in which terms are entered and grouped in nodes based on their meanings, that is to say the concepts they express. Leveraging our semantic network together our semantic engine (parser and a system of disambiguation), the semantic technology we developed allows a deep analysis of texts and in particular the automatic disambiguation of words with different meaning. The information of all the possible meanings of words and all of their grammatical details are fundamental in order to process text contents with high precision. That’s why each node or concept in our semantic network is characterized by grammatical information (verb, noun, etc.) and semantic information as well: each concept is linked to the others by semantic connections in a hierarchical and hereditary structure, in the form of a graph. Through a sequence of passages of analysis of the text (morphological and grammatical analysis, syntactical and logical analysis), in order to carry out word meaning disambiguation, our semantic technology looks up its semantic network to find all possible meanings and choose the right one.

What Is Semantic Technology?

Semantic technology in a software is a technology that “understands” the language in a way that's very similar to the way people do. Semantic technology collects all the structural and lexical text aspects in order to understand the meaning. It represents the knowledge contained in texts written in the everyday language. In other words, semantic technology comprehends natural language, understands the meanings, translates in different languages, and shares the knowledge, etc.

Semantic technology processing result is a cognitive and conceptual map, i.e. a structured representation of qualifying aspects of incoming unstructured data. The output structuring allows the automatic processing of the most relevant elements of the text. (Figure 1)

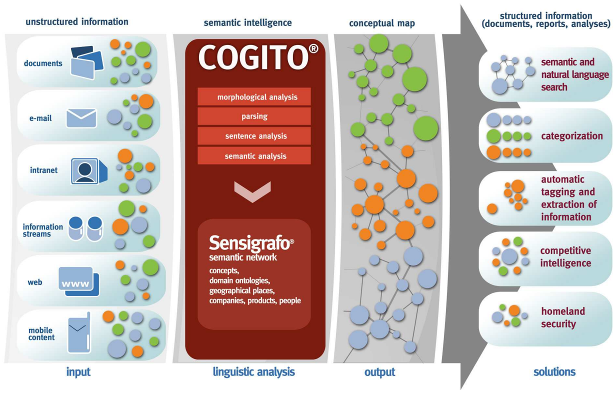


Figure 1
Expert System Semantic Technology COGITO
Functional Scheme

Semantic Technology Features

Semantic technology is composed of various modules dedicated to specific activities needed to disambiguate texts and process natural language which for example are essential for the automatic comprehension of questions formulated in the everyday language.
To understand automatically a text we need:
- a **semantic network**, which is the heart of semantic technology;
- a **parser** to trace each text back to its basic elements;
linguistic engines to query the semantic network (to link the basic elements of the texts with the semantic network of the meanings);
- a **system of disambiguation**.

Semantic Network

We’ve been developing a set of semantic networks firmly connected in the form of a graph that contains the conceptual representation of the language.
A semantic network is a lexical database in which terms are entered and grouped based on their meanings, that is to say the concepts they express. Therefore, they are not ordered alphabetically like in a standard dictionary but according to their meaning (this is why the network is called “semantic”) and to the various possible connections among these meanings (and this is why we talk about “semantic relations”.)
Each node in the semantic network is linked to the others by semantic connections in a hierarchical and hereditary structure, in the form of a graph.
Our network contains:
- information about connections among objects
- specifications about the lexical domains of each word
- information about the frequency of use

The richness of a semantic network is measured by both the quantity of words/concepts and of the semantic relationships. For example, concepts can be linked to each other in the following ways:
1) Subnomen (hyponymy) and supernomen (hypernymy). Hyponymy is the relation between a specific concept and a more general one. The supernomen is therefore the more generic term, a word having a general meaning in comparison with others representing specifications of that meaning: ex. “animal” is a supernomen of “cat”;

2) parsnomen (meronymy) and omninomen (holonymy), or the part-whole semantic relationship. A parsnomen is a noun that indicates a part of a whole (which is called the omninomen), such as for example the case of piece-cake (part=portion-whole=object) or plastic-bottle (part = material - whole = object);
3) relationships among nouns and verbs such as verb-subject or verb-object: given a noun and considering all possible “verb/subject” links, we obtain all the verbs normally (frequently) connected to that noun when it is the subject of the sentence: subject noun “food” verbs “to rot”, “to grow”, etc. The mechanism is the same when we consider the semantic relation “verb-object”: object noun “food” verb “to eat”, “to swallow”, “to grind”, ”to chew”…
4) Other kinds of connections, such as geographical links, are based on similar logic:
- each geographical element (not only countries, towns, rivers, valleys, etc. but also for example monuments) is connected to other geographical elements. Therefore, for example “St. Cloud” is linked to “Minnesota” which in turn is linked to “Midwest” which is linked to “USA”. Also, for example “Piccadilly” is linked to “London” which is linked to “England” which is linked to “Great Britain”, etc.

We’ve been developing an **Arabic semantic network** which contains more than 150 thousand concepts.
The concepts contain more than 130 thousand lemmas.
Thanks to the ability to manage derived and inflected forms, more than 450,000 different lemmas can be recognized.

We’ve developed semantic networks in other languages (English, Italian) focusing on the possibilities of a multilingual information retrieval, and cross-linguistic search activities.
Our semantic network is able to:
- manage all irregularities in the conjugation of verbs;
- manage all irregularities caused by inflected forms (feminine, dual, plural);
- recognize and process compound words;
- find the roots of words and possible prefixes and suffixes;
- identify the lexical function of prefixes and suffixes;
- analyse the form found in the text reducing it to its base form;
- individuate the correct semantic meaning;
- provide the description in English of the identified concept, presenting its Arabic vocalization.

Parser
The first step to take in order to understand the meaning of a sentence is to identify prefixes, suffixes and stems, and to determine the grammar role of each word. For example, in these sentences:

(a) “**سن** البرلمان الكردستاني قانونا جديدا“
(b) “ التربية في **سن** الطفولة المبكرة“

The word “*سن*” appears with two different grammar types: in sentence (a) the word is a verb, while in sentence (b) it's a noun.

According to traditional technology the two words are the same, while the semantic technologies assign different meaning to them.
Recognizing a word independently of its written form is equally important; nouns and verbs have several forms:

(a) “ انطلاق مهرجان «ألف **شاعر**.. لغة واحدة» أمس“
(b) “ كتاب "قصائد من **شعراء** جائزة نوبل“

In the sentences above, two forms (“**شاعر**”, “**شعراء**”) expressing the same concept are used. The parser performs a complete morphological, grammatical and syntactical analysis of the sentence, managing more than 3500 rules very quickly. Our parser uses an innovative and ad hoc methodology to query the semantic network, resulting in a significant improvement of the existing parsing. So semantic technology individuates gender - masculine/feminine - and number - singular/plural/dual – to recognize both words in the sentences above as forms of “acting person” associating all of them to their common meanings correctly, instead of individuating *n* different words as other systems do.

System of Disambiguation
For a human, the meaning is something obvious because of our capability to refer automatically to cultural elements that help us to understand the meaning of a word. The disambiguator of meanings included in our semantic technology thoroughly analyzes sentences or whole documents and distinguishes the right meaning for each element found, eliminating possible ambiguities.

The information of all the possible meanings of words are fundamental in order to process text contents with high precision. Being unable to detect different meaning brings a misleading understanding of the phrase, consider this example:

(a) يعتبر عقد *البيع* من أهم العقود وأكثرها انتشار*ا*
(b) المسبحة عبارة عن عقد مكون من مجموعة من الحبات
(c) أعلى المعدلات منذ أواخر عقد *1990*

The word “عقد” is ambiguous, because its meaning depends on context. In order to carry out word meaning disambiguation, our semantic technology looks up its lexicon to find all possible meanings. These lexicons are **semantic networks**. As already explained, semantic networks are not plain dictionaries, but resources that have been optimized for programmatical use, where word forms are knots linked to each other by multiple links denoting semantic or lexical relations. For example, the knots “secret agent” and “spy” are linked by a semantic relation named “synonymy” (they have similar meaning), while “angel” and “devil” are linked by “antonymy” (they indicate opposite concepts).

The Disambiguation of the Texts

What does **disambiguating** mean?
It means receiving input texts and returning the same texts as output, where each term is marked with the concept it represents. From a computational point of view, it is a sequence of passages of analysis of the text and improvements in the interpretation of the concepts contained in it. This is because a program must be provided with a univocal representation of the world, creating a system of reference to represent the equivalent of the human experience of the world: a generic experience, of course, not an individual one.

The Steps of Disambiguation

To analyse the sentence:
“The salesclerk says that he can’t accept a credit card.”
a sequence of analyses is performed that can be described step by step.

First Step
Morphological and grammatical analysis:

definite article THE, singular
noun SALESCLERK, singular
verb TO SAY, third person singular, present indicative, transitive, active
conjunction THAT
adverb NOT
verb TO ACCEPT, third person singular, present indicative, transitive, active
indefinite article A, singular
noun CREDIT, singular
noun CARD, singular
The sentence is turned from a mere sequence of characters into an organized set of terms, each one with its grammatical value.

Second Step
Assembling of collocations/locutions (using the information of the semantic network):

a priority is set to order single words or groups of two or more words when these come to function as lexical and grammatical units. Card and credit are words with their own meanings, but in this case they must be joined in the expression credit card. Balance of the knowledge of the world and extensive linguistic knowledge base allows the generalization of this kind of analysis to the whole language.

Third Step
Syntactical and logical analysis:

the salesclerk (Subject) [The main clause starts]
says (verbal predicate) [The main clause ends]
that he [The subordinate clause starts]
can’t accept (verbal predicate)
a credit card (direct object) [The subordinate clause ends]

At this third step, all the existing systems of analysis stop and can not provide the single elements of the sentence with their exact meaning, thus making a true conceptual analysis impossible. To proceed with the disambiguation, the terms to be disambiguated are highlighted:
salesclerk
say
accept
credit card
Each one of the meanings (semantic network node or concept) related to each couple entry/grammatical role obtains a probability percentage according to:
- frequency of use
- domains (1)
- attributes of adjectives/nouns (2) and checks for semantic congruence
- attributes of the verbs (3)
- contextual information, fundamental to solve cases of complex ambiguity

In the end, the concept that obtains the higher percentage is considered as assigned.

Regarding the checks for **semantic congruence**, semantic software are able to answer correctly questions such as “Does the cat eat?” using rules of derivation generated automatically according to a representation of the world.

From the object “Animal” in the semantic network other objects are derived such as “Vertebrate”, “Invertebrate” and “Marine Animal”.
Rule number 1:
Each derived object inherits all the properties of the object from which it is derived (property of inheritance).
Let’s continue with the derivations:
“Animal” ---> “Terrestrial mammal” ---> ”Carnivore”--->”Felid”--->”Feline”--->”Cat”---> ”Persian”---->...

Elaborations:
Lives (Animal) = True
Lives (Mammal) = True
Eats (Animal) = True
Eats (Mammal) = True
Eats (Carnivore) = True

Nurses (Mammal) = True
Nurses (Felid) = True
Nurses (Feline) = True
Mews (Cat) = True

Answer:
Yes, the cat eats.

The disambiguation of meaning is one of the most complicated problems of semantics. To obtain a satisfying elaboration speed, the following are needed:
- a vast knowledge structured like an encyclopedia
- a set of disambiguation algorithms working perfectly

Disambiguating, in fact, is the true problem in the automatic interpretation of texts. In order to distinguish between
“The rust eats the tower.”

“The knight eats the tower.”
a program must be able to "reason". It must be taught that a language contains many ambiguities a man can solve without problems. But what a man knows because of education and experience, software must deduce from the text automatically, relying on coded knowledge and advanced technologies.
The research and development of automatic systems for the semantic disambiguation must solve a crucial problem: the administration of the number of existing combinations that can be generated when dealing with words and texts. These can be combined together in a very high number of ways, increasing exponentially.

A disambiguation system can work sentence by sentence or considering whole documents, according to the way it is configured. Distinguishing all the possible meanings of a text is just an additional, but extremely critical, step beyond the more common analyses: logical, grammatical, query of the semantic network, domain analysis.

There are many examples of interpretations of words that we humans can take for granted but a program can not, including expressions meant in a figurative sense.

Some examples of what semantic disambiguator can do?
Understanding univocally the following sentences:
Example 1:
*“He **has eaten** a chicken.”*
*“The sweater **was eaten** by the moths.”*
*“The rust **ate** the tower.”*
*“The slot machine **ate** his money in just one summer.”*
*“Your car **eats** to much oil.”*

Example 2:
*“We went out for a **row**.”*
*“The condemned is in a death **row**.”*
*“They’ve had a big **row**.”*
*“My **row** boat is the third in the **row**.”*

Arabic Semantic Network and Disambiguation

- تم سن القانون
- سن الطفل 5 اعوام
- سن السكين
- تسوس سن العقل سن

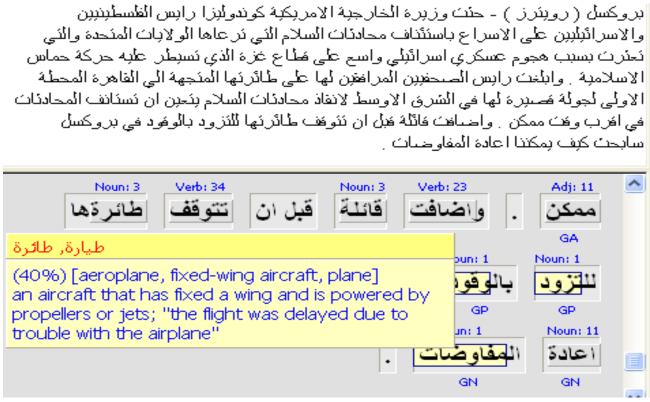
What does “سن” mean?

Thanks to a rich semantic network, it is possible to understand the different meanings of the word “سن” according to its context.

In the first sentence *“A new law was legislated”*, “سن” means to legislate.
In the second sentence *“The child is 5 years old”*, “سن” means old
In the third phrase *“The sharpening of the knife”*, “سن” means sharpening.
In the last phrase *“The wisdom tooth decayed”*, “سن” means wisdom tooth.

Our semantic network in Arabic includes more than 150 thousand word definitions or concepts and more of 130 lemmas. Thanks to the ability to manage derived and inflected forms, more than 450,000 different lemmas can be recognized. It includes
- a representation of all the significant geographic localities and the respective connections arranged in an intelligent way (geographic semantic network);
- a system that understands subjects and concepts included in a document and distinguishes automatically common names from brands (semantic network for companies, products, names and famous characters).

As described above, semantic technology performs a precise disambiguation of the meanings of the words, also leveraging the Arabic semantic network.
For the disambiguator process, an input text is provided. This text is analyzed (Figure 2) and the output is the same text as input, classified for the concept it represents (Figure 3).

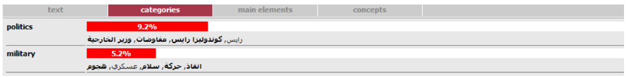


بروكسل (رويترز) - حثت وزيرة الخارجية الامريكية كوندوليزا رايس الفلسطينيين
والاسرائيليين على الاسراع باستئناف محادثات السلام التي ترعاها الولايات المتحدة والتي
تعثرت بسبب هجوم عسكري اسرائيلي واسع على قطاع غزة الذي تسيطر عليه حركة حماس
الاسلامية . وابلغت رايس الصحفيين المرافقين لها على طائرتها المتجهة الى القاهرة المحطة
الاولى لجولة قصيرة لها في الشرق الاوسط لانقاذ محادثات السلام يتعين ان تستانف المحادثات
في اقرب وقت ممكن . واضافت قائلة قبل ان تتوقف طائرتها للتزود بالوقود في بروكسل
سابحث كيف يمكننا اعادة المفاوضات .

طيارة, طائرة

(40%) [aeroplane, fixed-wing aircraft, plane]
an aircraft that has fixed a wing and is powered by propellers or jets; "the flight was delayed due to trouble with the airplane"

Figure 2
The Disambiguation Process
The meaning of the highlighted word is correctly recognized as “airplane”



text	categories	main elements	concepts
politics	9.2%		
	رايس, كوندوليزا رايس, مفاوضات, وزير الخارجية		
military	5.2%		
	انقاذ, حركة, سلام, عسكري, هجوم		

Figure 3
The text above was focused only on politics and military issues.

Using rules related to the specific language, the text content is at this point categorized by concepts, providing us a form of synthesis (this step is called semantic synthesis). The text above, for example, originally contains something about conversations, Brussels and airplane (Figure 4). The semantic synthesis includes the identification of the main elements of the text.

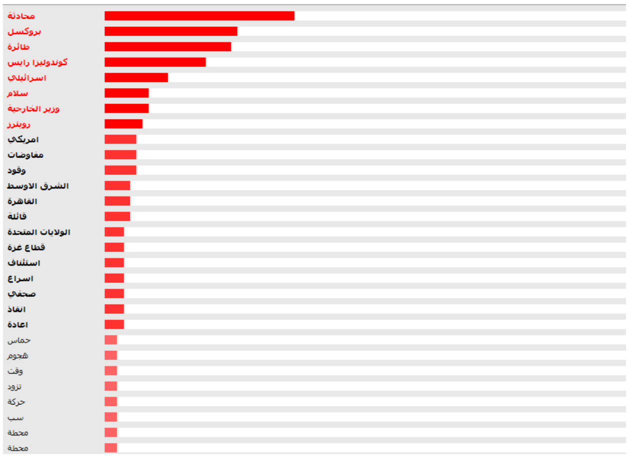


Figure 4
Semantic Synthesis and Main Elements

The next step is based on verbs, nouns and adjectives analysis. In this way, the system is able to recognize not only the entities, but also the action in which these entities are involved. (Figure 5)

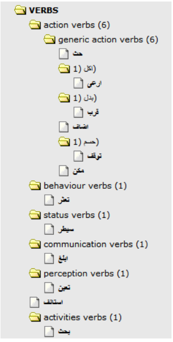
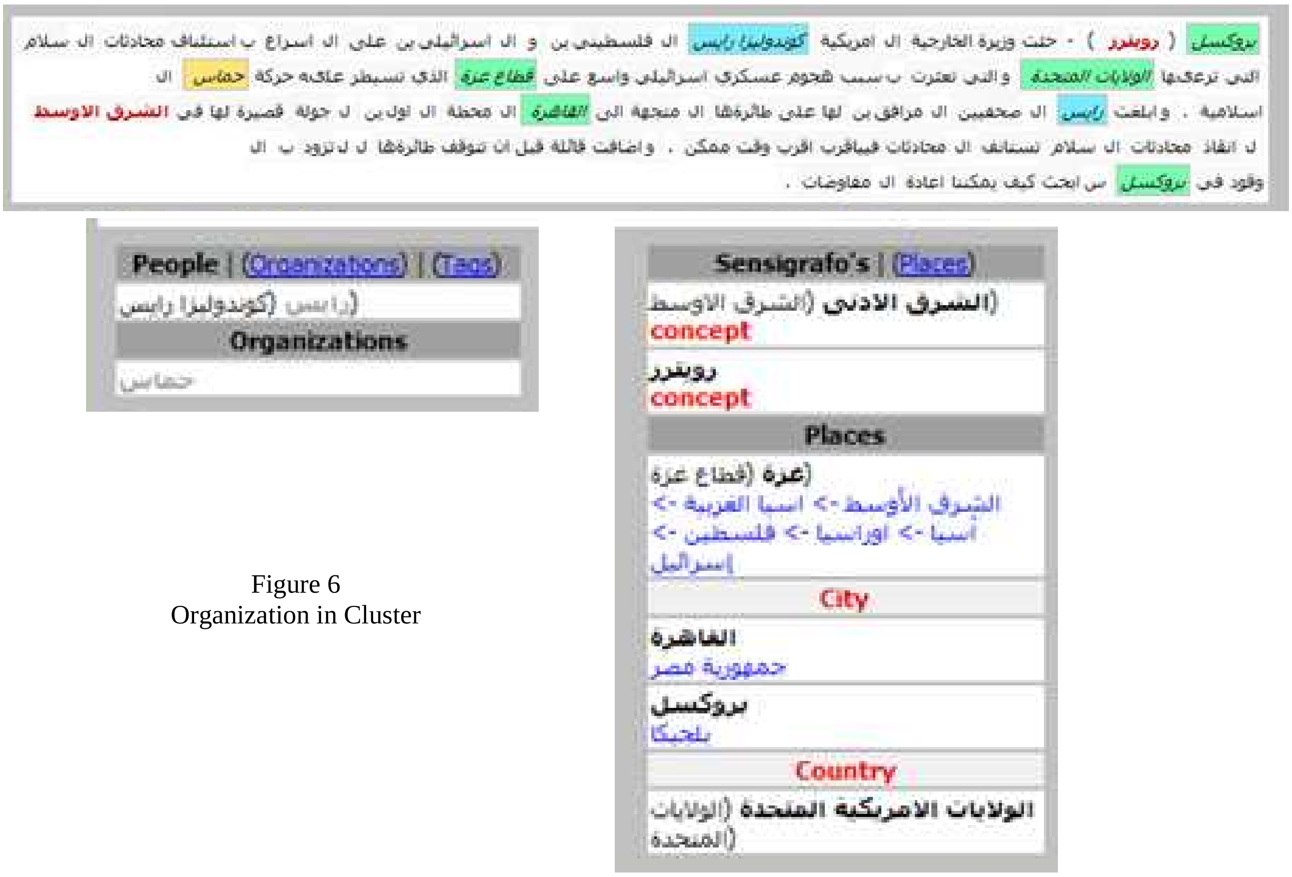


Figure 5
Entity Hierarchy, Verbs

In the last step, data are organized in cluster. Analyzing a single piece of information is not a reliable way to make an analysis. A better way consists of considering a concept. For example, a "president" could be a "person with institutional title", like a "politician" that could be a "head of state". But "president" could be only a generic title.

Having a knowledge organized in a semantic net allows one to think new means of analysis. We can ask the system to provide information about "titled people", obtaining a result where the substantive "president" is indicated.

A major hurdle is that proper nouns must be clusterized and identified.
As can be seen in the Figure 6 below, some proper nouns are recognized, like persons, cities, countries and so on.
But, if it might be easy to recognize a country noun, most of the time it's difficult to recognize people.



بروكسل (رويترز) - حثت وزيرة الخارجية ال امريكية كوندوليزا رايس ال فلسطيني ين و ال اسرائيلي ين على ال اسراع ب استئناف محادثات ال سلام
التي ترعاها الولايات المتحدة والتي تعثرت ب سبب هجوم عسكري اسرائيلي واسع على قطاع غزة الذي تسيطر عليه حركة حماس ال
اسلامية . وابلغت رايس ال صحفيين ال مرافق ين لها على طائرتها ال متجهة الى القاهرة ال محطة ال اول ين ل جولة قصيرة لها في الشرق الاوسط
ل انقاذ محادثات ال سلام تستانف ال محادثات فياقرب اقرب وقت ممكن . واضافت قائلة قبل ان توقف طائرتها ل ل تزود ب ال
وقود في بروكسل س ابحث كيف يمكننا اعادة ال مفاوضات .

Figure 6
Organization in Cluster