expert.ai

DISAMBIGUATION

The Key to Contextualization

The Fundamentals of Disambiguation

Disambiguation is the process of removing confusion around terms that express more than one meaning and can lead to different interpretations of the same string of text. The goal of this process it to match each term to the author's expressed intent.

A human being possesses the competence and capabilities to understand and disambiguate both the spoken and the written language based on several elements (e.g., intellect, memory, culture, context).

On the contrary, **a machine does not have a reference system to support its understanding of the meaning of words, sentences and whole documents.** Therefore, machines need a system that provides access to background information — as similar as possible to human experience — which allows them to cope with text ambiguities and meaning recognition.

Such a system, combined with a computer's memory and computing capabilities, provides a powerful module capable of logical and comprehensive text understanding on a large scale.

The semantic disambiguator is the module of expert.ai technology that solves ambiguities and understands the meaning of each word in a text. This is enabled by a multi-level text analysis and the interaction of the disambiguator with expert.ai's knowledge graph. The disambiguator and knowledge graph constitute the core of the expert.ai's technology.

The multi-level text analysis consists of consecutive phases that lead to a cognitive and conceptual map of texts (i.e., the final output of the disambiguation process). In other words, the disambiguation process tends to represent a text in terms of its concepts, entities and the relationships that exist between them. This representation is called **disambiguation string**.

Since the disambiguation process is based on linguistic analysis of texts, the disambiguator is language dependent. For this reason, every language managed by expert.ai's platforms has its own disambiguator and knowledge graph.

The disambiguation process consists of four main phases:

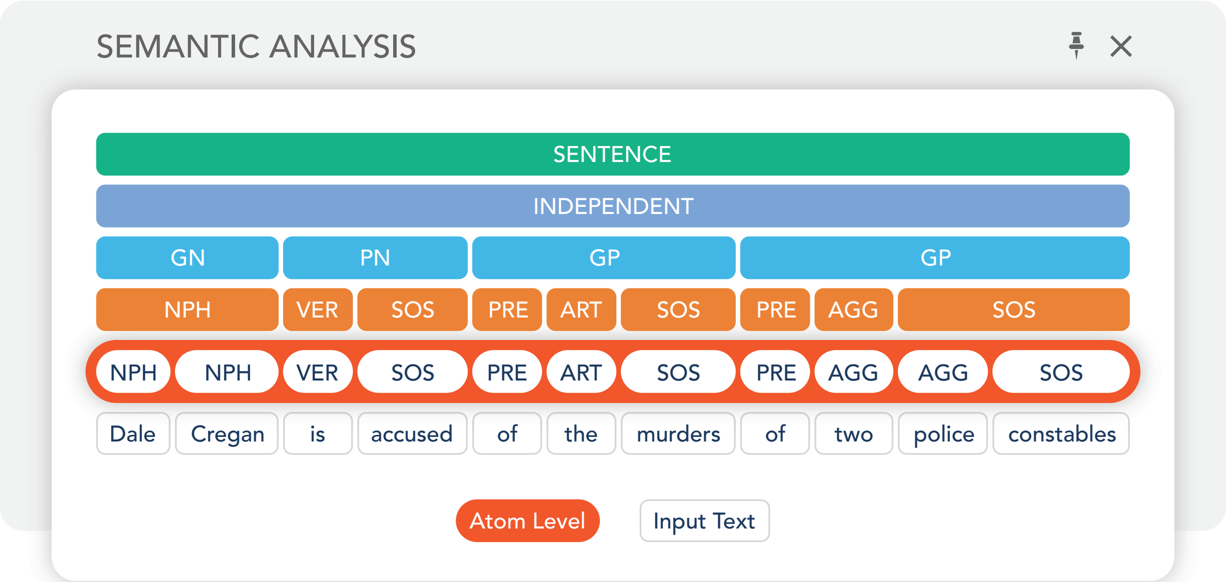
- **Lexical analysis**
- **Grammatical analysis**
- **Syntactical analysis**
- **Semantic analysis**

Lexical Analysis

This phase involves a first step during which the input text is "cleaned" in preparation for the lexical analysis (e.g., multiple spaces are removed, characters are made uniform, etc.). The real lexical analysis breaks up the stream of text into meaningful elements, called tokens. This process, called **tokenization**, classifies a word as a token.

Starting with a stream of characters unintelligible to the machine, the result of this elaboration step is a sequence of "atomic" or "indivisible" elements which is further elaborated in the next analysis phase.

Looking at the disambiguation output presented in the expert.ai Studio Semantic Analysis panel, the result of this subdivision is clearly visible in the diagram shown below (second line from the bottom). The bottom line represents the input text, the line above it represents what is called the "atom level." Every word of the input text has a corresponding "atom level" block on top of it that represents the tokenization process.



SEMANTIC ANALYSIS

SENTENCE										
INDEPENDENT										
GN		PN		GP			GP			
NPH		VER	SOS	PRE	ART	SOS	PRE	AGG	SOS	
NPH	NPH	VER	SOS	PRE	ART	SOS	PRE	AGG	AGG	SOS
Dale	Cregan	is	accused	of	the	murders	of	two	police	constables

Grammatical Analysis

During this phase, every token in the text is assigned a part of speech. The disambiguator, using the knowledge graph as a repository of known words, and applying a module that recognizes inflected forms, conjugations and the like, identifies nouns, proper nouns, verbs, adjectives, articles and so on. The interaction between the disambiguator and the knowledge graph a this phase also leads to:

- the recognition of known and unknown tokens (i.e., contained or not in the knowledge graph), and
- a first grouping of single tokens.

Among the known elements, the machine recognizes those groups of tokens that, when co-occurring in the language, denote a specific concept that is different from the concept represented by the single elements considered in isolation.

These particular groups of tokens are:

- compound words (white-collar, dog catcher);
- phrasal verbs (back up, check in);
- collocations (degree of purity, earth movements);
- idiomatic expressions (be in deep water, blood is thicker than water);
- proper nouns (Ministry of Foreign Affairs, William Butler Yeats).

Considering that:

- every token is identified as a part of speech;
- every token is characterized as known or unknown;
- significant groups of tokens have been located.

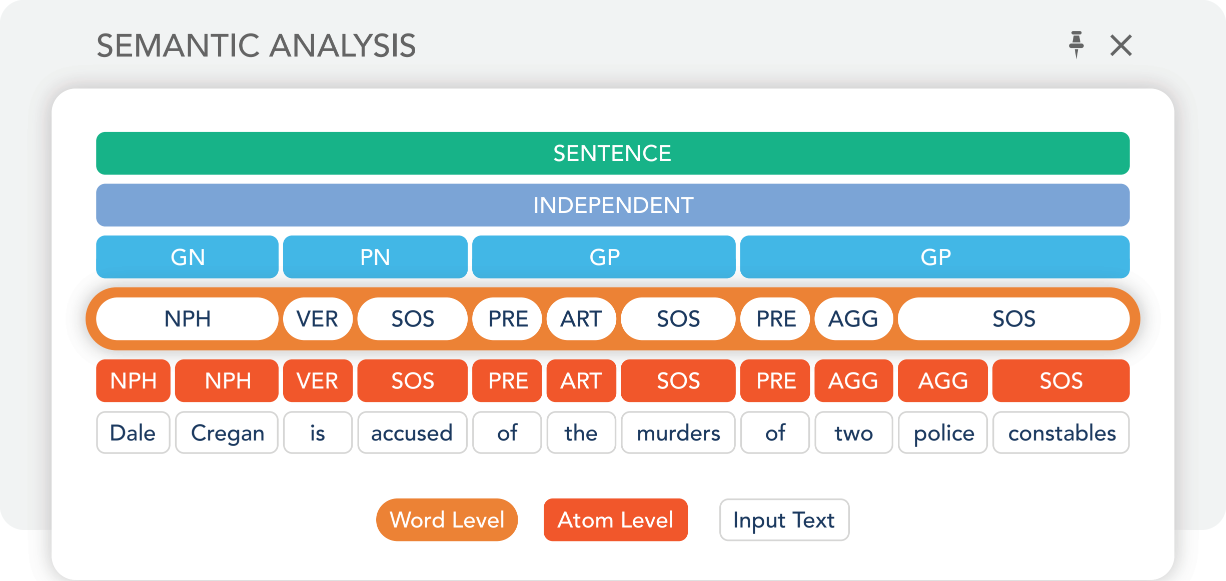
We can now assert that it is possible to recognize, in every known word or group of words, what is called a "lemma." A lemma is the base form of a word representing all its inflected forms.

For example: the noun **child** also represents **children**; the verb **go** also represents **went**, **goes**, **going**; the base form of adverbs and adjectives represent their comparative and superlative forms as well.

Starting from a mere sequence of tokens, what results from this elaboration stage is a sequence of elements, a large majority of which have been grammatically and lexically identified and classified.

Looking at the disambiguation output presented in expert.ai's Studio Semantic Analysis panel, the result of this elaboration is clearly visible in the diagram shown below (third line from the bottom). The bottom line represents the input text, the line above it represents the "atom level," then the next line represents the "word level."

Here, it is evident why some tokens have been grouped to form collocations (police constable) or proper nouns (Dale Cregan) and that every token or group of tokens is represented by a block stating its part of speech (NPH=Human Proper Noun, VER=Verb, NOU=Noun, ADJ=Adjective, etc.).



SEMANTIC ANALYSIS

SENTENCE										
INDEPENDENT										
GN		PN		GP			GP			
NPH		VER	SOS	PRE	ART	SOS	PRE	AGG	SOS	
NPH	NPH	VER	SOS	PRE	ART	SOS	PRE	AGG	AGG	SOS
Dale	Cregan	is	accused	of	the	murders	of	two	police	constables

Syntactical Analysis

During this phase, known and unknown tokens are arranged in a syntactical structure. In other words, the disambiguator operates several word grouping operations on different levels. It does this by reproducing the way in which:

- words are linked to each other to form phrases;
- phrases are linked to each other to form clauses;
- clauses are linked to each other to form sentences.

The first level of word groups is represented by phrases. A phrase is a group of words (sometimes a single word) that forms sort of a constituent acting as a single unit in the syntax of a sentence (e.g., noun phrases, verb phrases, preposition phrase, etc.).

The second level of word groups is represented by clauses. A clause is the smallest grammatical unit that can express a complete proposition. A clause can coincide with a sentence (**note: a sentence can be made of a single clause**), or more clauses can be combined to form a complex sentence.

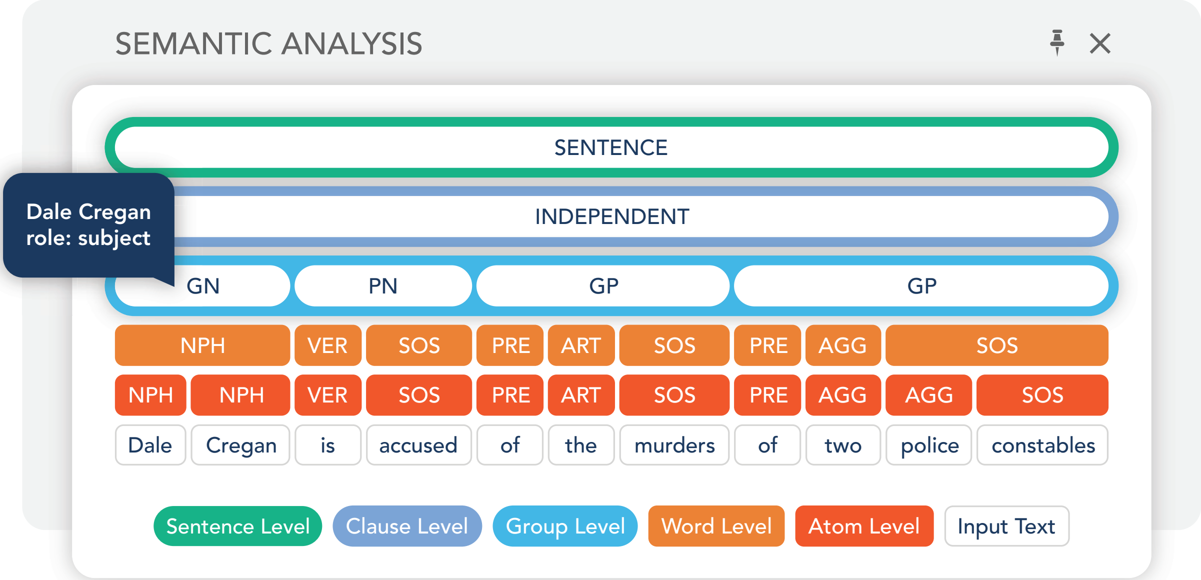
The third level of word groups is represented by sentences. A sentence is a grammatical unit that consists of one or several words linked to each other by a syntactic relation in order to convey meaning.

Phrases are further analyzed and elaborated to reach two goals:

- Attribute a logical role to each phrase (subject, object, verb, complement, etc.).
- Locate relationships between verbs, subjects and objects and between these and other complements whenever possible.

Looking at the disambiguation output presented in expert.ai's Studio Semantic Analysis panel, the result of this elaboration is clearly visible in the diagram shown below (reading the fourth, fifth and sixth lines from the bottom). The bottom line represents the input text, the line above it represents the "atom level," the third line represents the "word level" and the next three represent, respectively, the "group level," "clause level" and "sentence level."

On the group level, tokens such as "of," "the" and "murders" have been put together because they form a prepositional phrase (GP), whereas "is accused" is recognized to be a predicate nominal (PN). The clause and sentence levels coincide in this example because the sentence is made of a single INDEPENDENT clause. The word balloon shows the role attributed to the noun phrase "Dale Cregan," which is recognized as the subject of the sentence.



SEMANTIC ANALYSIS

Dale Cregan
role: subject

SENTENCE										
INDEPENDENT										
GN		PN		GP			GP			
NPH		VER	SOS	PRE	ART	SOS	PRE	AGG	SOS	
NPH	NPH	VER	SOS	PRE	ART	SOS	PRE	AGG	AGG	SOS
Dale	Cregan	is	accused	of	the	murders	of	two	police	constables

Semantic Analysis

During the final and most complex phase, the tokens recognized during grammatical analysis are associated with a meaning thanks to a new interaction cycle between the disambiguator and the knowledge graph. Excluding function words such as articles, prepositions and conjunctions (which bear little lexical meaning), every token is associated with several concepts (syncons) within the knowledge graph, all of which are potential candidates to represent the token.

The choice is ultimately made in consideration of the base form of each token with respect to its part of speech. For example, **trained** can be a verb or an adjective (in the first case, the base form is **train**, in the second case, it is **trained**).

When ambiguity arises (e.g., if several syncons exist with the same part of speech and contain the same base form), the list of candidates that could be associated with a token are ranked according to:

- frequency of use;
- domain(s);
- attributes;
- semantic consistency.

At this point, the disambiguator skims the list of candidates for each token by applying each of the following principles:

- the score of each candidate;
- the grammatical and syntactical characteristics of the token;
- the position of the token in the sentence;
- the relation of the token with the syntactical elements surrounding it.

To put it simply, these principles consider the context in which each token appears to determine its meaning. **By means of successive refining stages applied to every element in the process, the disambiguator eliminates all candidate syncons for each token except one, that which is definitively associated with the token.** This process occurs simultaneously for all lexical elements in the input text, thus leading to the final text disambiguation and to the complete comprehension of its meaning.

Since a standard knowledge graph is an extraordinarily rich yet not totally comprehensive database, some terms or meanings present in a text may not be available during the disambiguation process. These are concepts that usually belong to special knowledge fields (technical terms) or unknown entities (e.g., proper names of companies, people, institutions, products, etc.) which, of course, cannot all be entered in the semantic network.

When the disambiguator encounters such elements, these are always PoS tagged and syntactically disambiguated by context. What happens with sense disambiguation depends on the unknown element itself and the information available to the disambiguator.

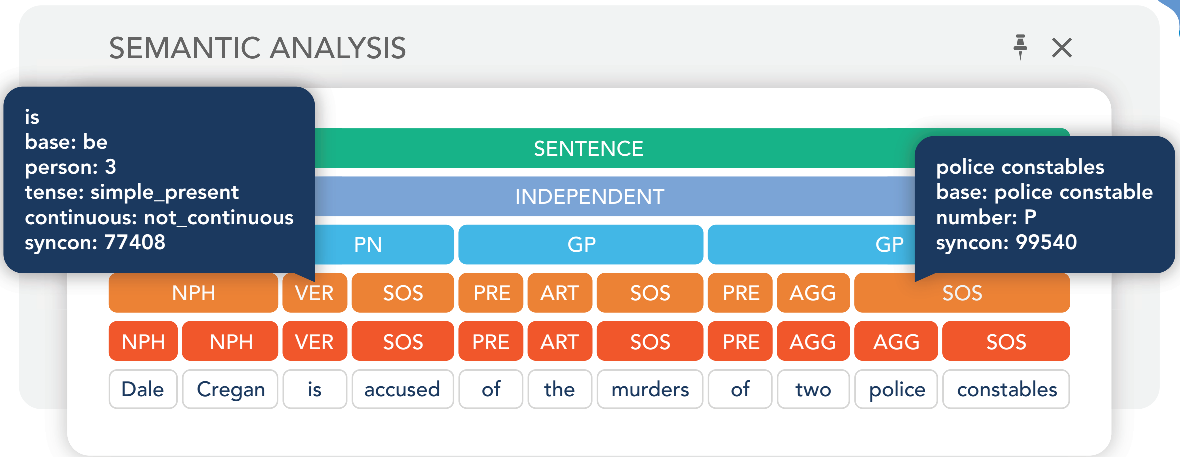
Whenever possible, the disambiguator tries to infer, from the context, the nature of the unknown element, virtually linking it to a known concept. In other words, **when the disambiguator comes across an unknown element in a text, it applies complex algorithms and heuristic rules to use the known words surrounding the unknown element to tag it as virtual "child" (type) of a key syncon.** A typical example of unknown elements which are systematically associated to a concept contained in knowledge graph are human proper names, which are linked to the concept of person.

Starting from a structured set of complex yet senseless elements, the end result of this elaboration stage is a fully analyzed text that has a known and usable meaning (as it is or for further elaboration). Using the knowledge graph as a repository of words, meanings, and grammatical information, and methodically evaluating the context in which words appear, the disambiguator discovers and formalizes the syntactical and semantic principles that underly a text and reproduce the disambiguation process of the human mind.

Looking at the disambiguation output presented in expert.ai's Studio Semantic Analysis panel, the result of this elaboration is clearly visible when you drag the cursor over the different blocks on the "word level" of analysis. In fact, both word balloons show all the grammatical and semantic information generated for the tokens is and police constables.

The first is a verb (VER): its base form is "be" and in the text it appears conjugated to the third-person singular. The tense is simple present, and it is associated to syncon 77408, identifying in the knowledge graph the concept of the verb be.

The second is a noun: its base form is police constables and the token in the text represents the plural (number: P) form of the concept. The syncon to which it is attributed is 99540 identifying in the knowledge graph the concept of police constable.



SEMANTIC ANALYSIS

is
base: be
person: 3
tense: simple_present
continuous: not_continuous
syncon: 77408

police constables
base: police constable
number: P
syncon: 99540

SENTENCE

INDEPENDENT

PN GP GP

NPH VER SOS PRE ART SOS PRE AGG SOS

NPH NPH VER SOS PRE ART SOS PRE AGG AGG SOS

Dale Cregan is accused of the murders of two police constables

In computational linguistics, word-sense disambiguation is a task that is addressed using a variety of techniques. These can be divided into three main families:

- keyword/statistic approach
- shallow linguistics
- deep semantic analysis

Keyword/Statistical Approach

This approach to natural language processing usually involves tokenization of a stream of text, as well as sentence splitting. Probabilistic and statistical methods are added to try and solve some language ambiguities. For this system, a document is only made of character strings that appear several times in the text.

The result of this process is an alphabetical index of elements, generally used for search purposes. This approach has its power in the indexing phase, which is simple and quick, but its limitations are evident in the advanced search for information, which reveals overload and underload problems.

Shallow Linguistic Approach

This approach (also known as machine learning) adds part-of-speech (PoS) tagging, lemmatization and, to a certain degree, logical grouping of words to a keyword/statistical approach. However, these approaches only consider context in which words appear to determine, probabilistically, what sense a word has in a text.

For example, if the word **bass** is near the words **sea** or **fishing** it is probably used in the "fish" sense. However, if **bass** is near **music** or song it is probably used in the "music" sense. These rules are usually automatically generated by the computer, using a training corpus of words tagged with their senses. But no attempt is made to understand the document.

Deep Semantic Analysis

Contrasting the other approaches, semantic analysis builds up an elaborated representation of the document in order to reach a deep understanding of the text. In fact, this method best exploits the potential of linguistics applied to computer science and carries out a complete multi-level text analysis (lexical, grammatical, syntactical and semantic analysis) interacting with a proprietary semantic network, which represents a repository of the possible meanings for a word.

In this repository, all words are not only listed, but they are organized in groups representing the same concept. The principle behind this is that working with "concepts," rather than simple keywords, makes it possible to fully understand the meaning of single words, sentences and whole documents.

Using the most out-of-date technologies, disambiguation accuracy is between 45% and 55% — for a few systems, it may even reach 73-75% in closed contexts. With deep semantic analysis, users can achieve accuracy levels higher than 90% in totally open contexts.

Deep semantic analysis also solves the problem of overload and underload. In fact, when faced with advanced information search, this approach maintains both high precision (retrieving only accurate results that are relevant to the search) and high recall (retrieving a high percentage of relevant documents).

Get **Started**

Interested in learning how NLU AI will transform your company? Get started here.

See what expert.ai can do for you!

expert.ai

About us
Expert.ai is the premier artificial intelligence platform for language understanding that augments business operations, scales data science capabilities, simplifies AI adoption and provides the insight required to improve decision making throughout organizations. The expert.ai brand is owned by Expert System (EXSY:MIL), that has cemented itself at the forefront of AI-based natural language solutions across Insurance, Banking, Publishing, Defence & Intelligence, Life Science & Pharma, Oil Gas & Energy, and more.

www.expert.ai **info@expert.ai**