Semantic Annotation

Huge part of the available information on the internet is unstructured - online news, emails, blogs, tweets, comments, various companies' documents, clinical trials, etc. This makes it difficult for companies to dig all relevant information and extract the knowledge they need. Here comes the text analytics, which tries to bridge this gap. By using text analytics methods, one can easily summarize similar information from different sources, derive the important conceptual elements from the texts, structure them and provide the more thorough and high-quality analysis of this information.

What is semantic annotation

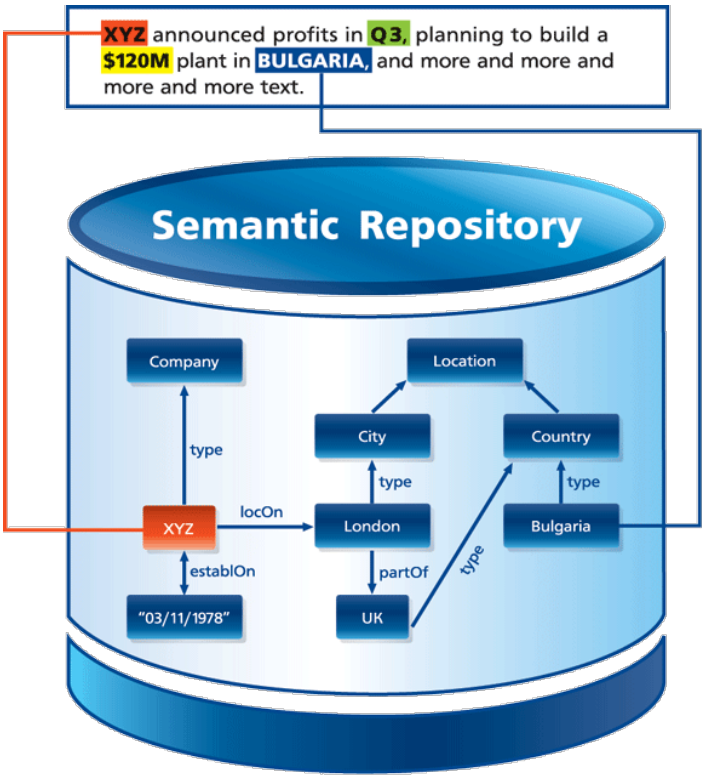
Typical semantic annotation tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

Annotation, or tagging, is about attaching names, attributes, comments, descriptions, etc., to a whole document, document snippets, phrases or words. It provides additional information (meta-data) about an existing piece of text. Compared to tagging, which adds relevance and precision to the retrieved information, semantic annotation goes one level deeper:

- It enriches the unstructured or semi-structured data with a context that is further linked to the domain structured knowledge.
- It allows results that are not explicitly related to the original search.

Semantic Annotation helps to bridge the ambiguity of the natural language when expressing notions and their computational representation in a formal language. By telling a computer how data items are related and how these relations can be evaluated automatically, it becomes possible to process complex filter and search operations.

We call Semantic Annotation the meta-data, as well as the process of adding it to specific ranges of text within a document.

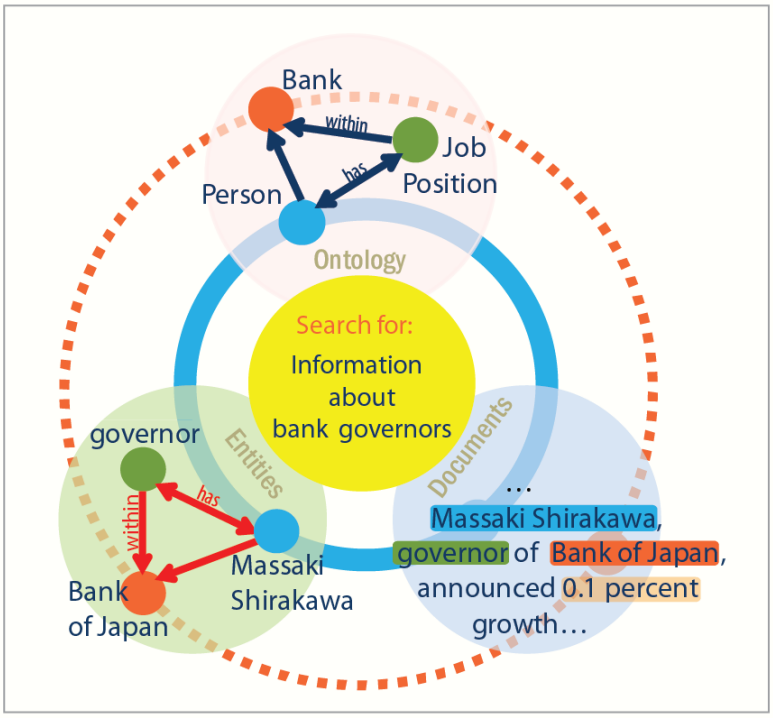


How we use it

We can make inferences about all kinds of things once we have the annotations linked to an ontology and the background knowledge. We know that:

- Cities are located in countries, organisations are located in the cities, people work for organisations, etc.
- Since Marie Curie worked at the Sorbonne, we know that at that time she lived in France and not Poland.

Semantic annotation links mentions in unstructured or semi-structured texts to an abstract model of the relevant domain and to their respective instances in the background knowledge.



How we do it

We attach concepts from the domain ontology to instances we have found in the text. Then, we disambiguate these instances. This means that we choose the right instance from a list of candidates, according to the context in which it appears. For example, London, GB vs London, Canada.

There are three basic approaches for adding semantic annotations to texts: automatic, manual and semi-automatic. Within each of these general approaches there exists a range of techniques to handle different types of annotation tasks, each with its own set of advantages and disadvantages over the available alternatives.

- Automatic annotation - learning algorithms search for patterns in text and require no external input. It is less precise but can operate with considerable speed and over many more documents than humans can reasonably address.
- Manual annotation - humans do all of the annotation. It is more precise and reliable, but very labor-intensive and is often used to train a machine to perform automatic annotation.
- Semi-automatic annotation - learning algorithms are trained via a text corpus that has been manually annotated to replicate the human's annotation decisions. This is the most used method in Ontotext. It is more precise as well as cost and labour effective.

Ultimately, each manual, automated, or semi-automated method for analysing textual data has its own set of benefits and costs that vary depending on the task at hand.