Text Analytics in Ontotext

How we do it

This document is specially designed for clients, business partners, R&D engineers, software developers, annotators, and all those interested in how we do text analytics.

It is designed as a random-access reference tool, though some users may want to become familiar with the entire document.

Contents

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 text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modelling (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.

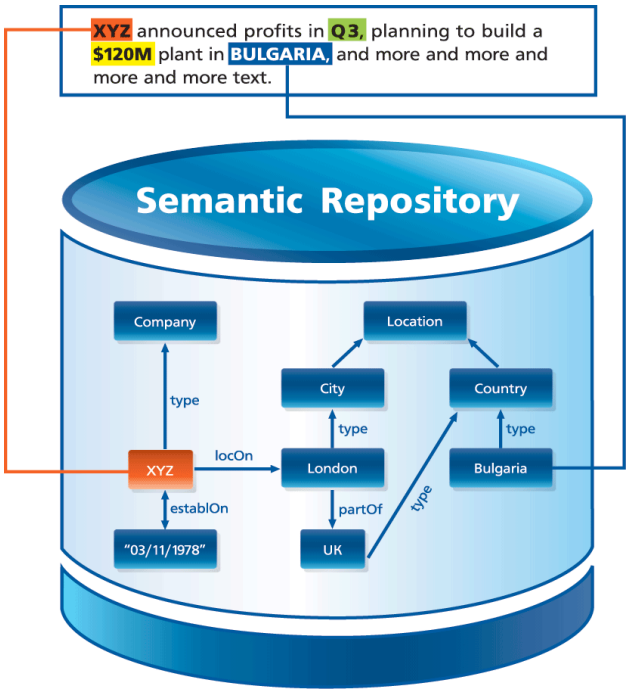


Figure 1 Semantic Annotation

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 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 Pepe Reina is a football player and lives in Liverpool, we know he plays for Liverpool F.C.
- Since Marie Curie worked at the Sorbonne, we know that she lived in France and not Poland.

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

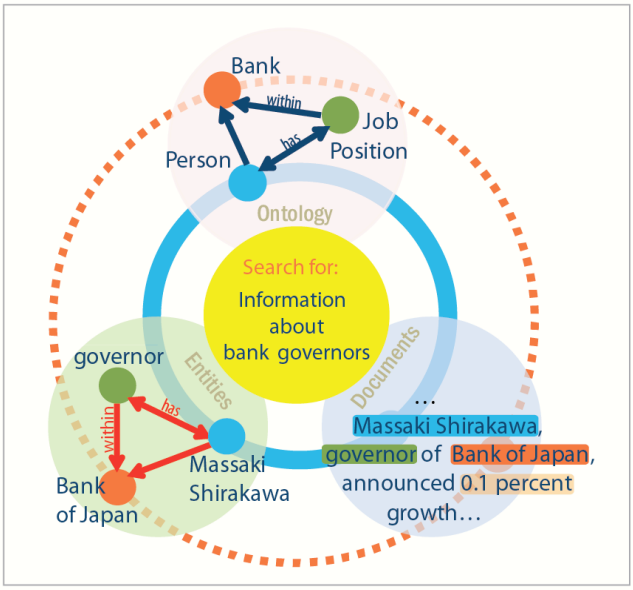


Figure 2 Linking Documents, Entities and Domain Models

How we do it

We attach concepts from the domain ontology to instances we found in the text. Then we disambiguate these instances, which means that from a list of candidates we choose the right instance, 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 labour-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. We use this 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 cost that vary depending on the task.

Methodology

Clear functional requirements

Building a solution based on semantics starts with the client's requirements - what they want to achieve with this solution, what their business objectives are, what business problems they want to solve.

Turning to text analytics usually aims to facilitate the management of big volumes of data and documents of a specific domain.

The first step of the solution development is to define the so-called functional requirements, or what the system is supposed to accomplish. Functional requirements are expressed in the form "system must do <requirement>" and they specify the particular results of the system. They drive the application architecture of the system. Functional requirements may be: specific smart search (faceted, FTS, etc.), content enrichment, documents feeds/data aggregation, etc.

For example, through semantic text analysis based enrichment of content, media content editors can dynamically create new content-based product offerings, while readers can benefit from adaptive content streams, personalised through their choices and behaviour.

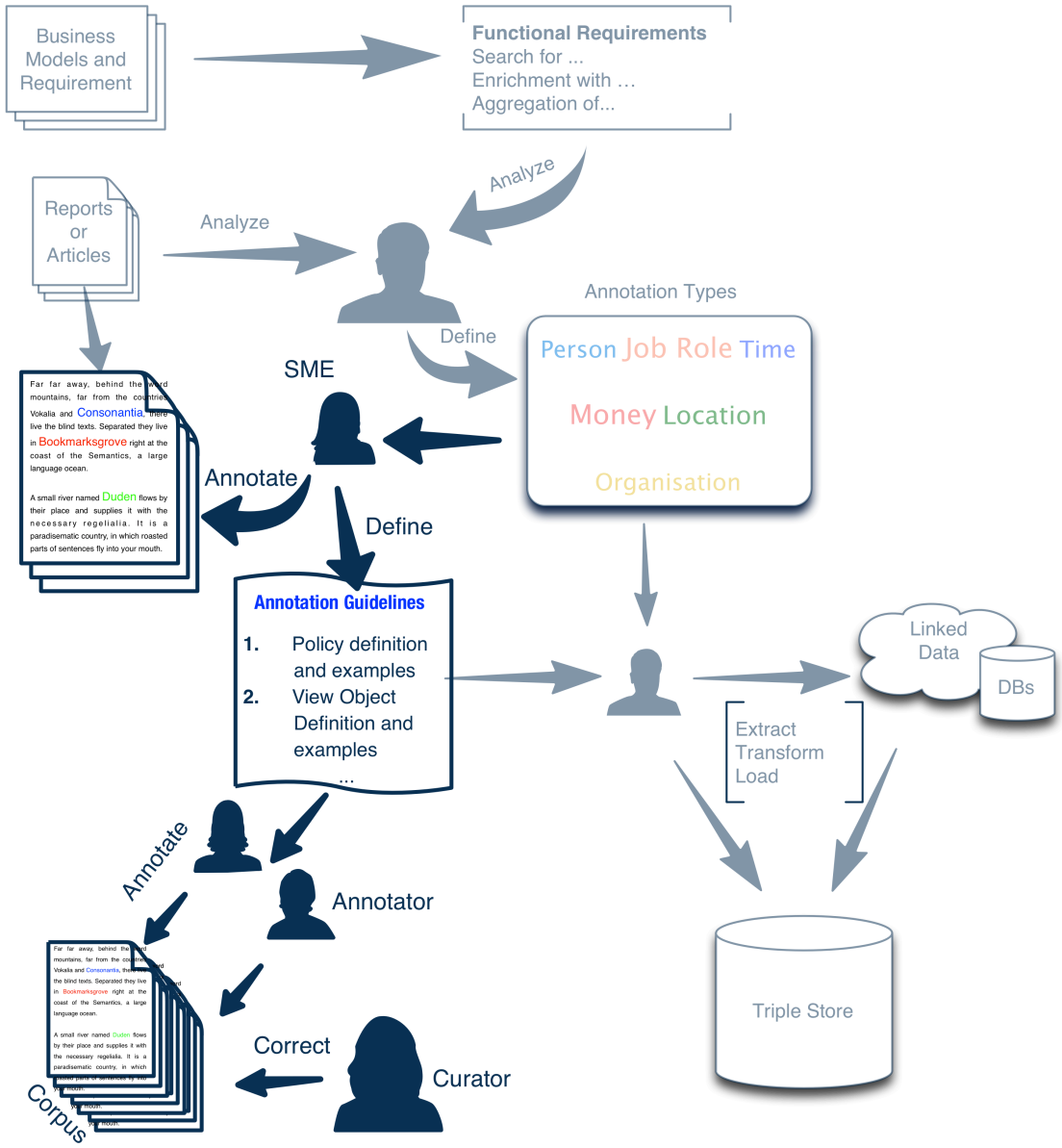


Figure 3 The Process of Building Semantic Annotation and Search Solution

Clear annotation types

Having clear annotation types is another important prerequisite for the annotation process. The Subject Matter Expert (SME) or someone who is very much familiar with the specificities of the domain defines the annotation types, based on empirical observations over the data and the documents.

The Annotation types (AT) are abstract descriptions of mentions, used for marking spans of text, i.e. recognising mentions of person, organisation, location, date, etc. in a document. An AT may have two parts:

- Type – the generic word that describes an entity;

• Feature – more specific sub categories of the type.

For example, an annotation labelled “address” (the type) can either be “street address”, “city”, or “country” (the features).

Initial corpus

The corpus is a collection of documents, which can be in different formats. The ones we presently support are XML, HTML, TXT, CSV, DOC, and PDF. Depending on the annotation task, these texts should be sampled to be representative and balanced. It means that the corpus should contain all types of texts (categories) present in that particular domain (e.g. for the news domain, it should contain texts about general news, social life, economy, finance, religion, sport, celebrities, etc.) and the proportion of the text types should be based on their share in real-life usage.

We usually start a new annotation task by creating an initial corpus of a small number of documents. In this way, we are able to see how well the annotation task and initial guidelines work and, if necessary, adjust the text analysis component/ guidelines/ text collection before adding more documents to our corpus.

There is no fixed number of how big the corpus needs to be in order to get good results, as this will depend largely on how complex the annotation task is. We usually use between 100 - 500 documents for evaluation and 700 - 2000 documents for ML training.

Initial Annotation Guidelines

We need to create initial annotation guidelines, which will be used as guidance for manually annotating the documents. Depending on the domain and complexity of the task, it can be done automatically or manually.

Automatic approach

Based on observation over the documents and the data, the text analysis expert creates the initial model of the phenomena (software text-analysis component - ML, rule-based) associated with the problem task we are trying to solve. In this way, the first annotation guidelines are automatically available. They describe the way the corpus should be annotated with the features in the model.

Manual approach

Based on observation over the documents and the data, and the cases in which entities appear in the text, or the context in which the mentions of AT appear, the MA experts create initial annotation guidelines. During the manual annotation process, they will enrich and refine them with specific use cases.

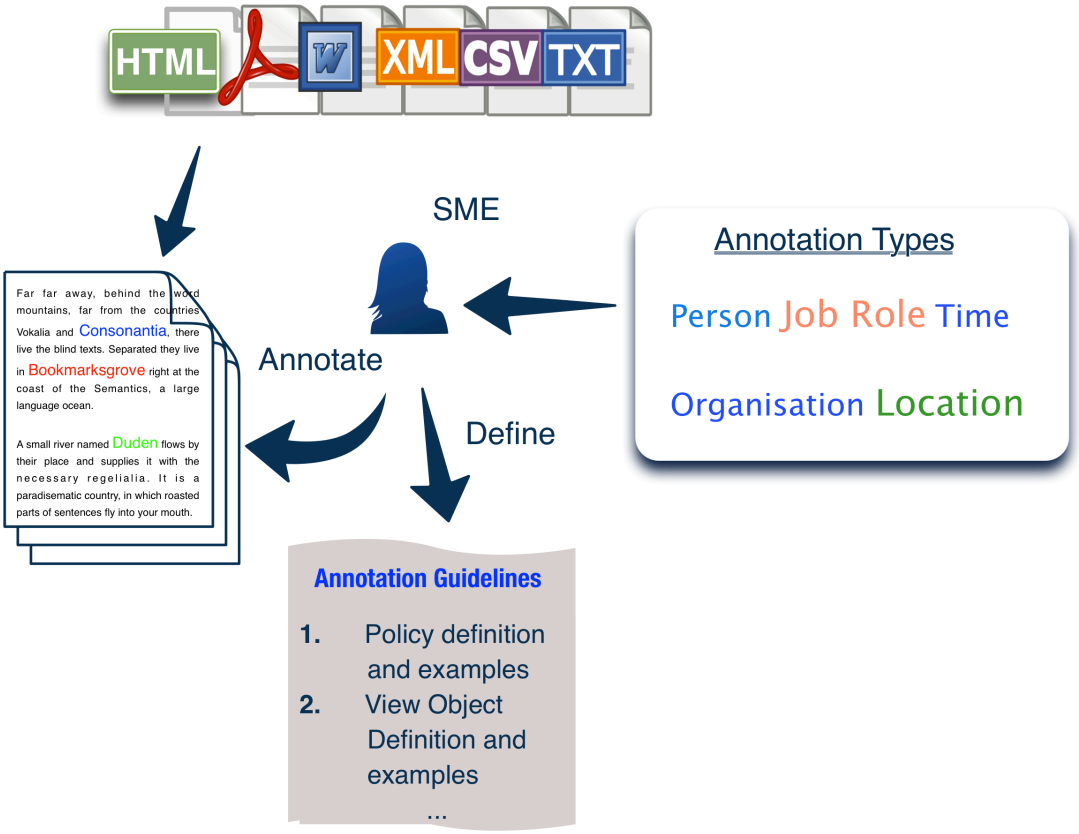


Figure 4: The Process of Creating Initial Annotation Guidelines

Semantic Annotation Cycle

The process of developing an annotated corpus is often cyclical, with changes made to the annotation types, annotation guidelines, and the tasks as the data is studied further.

In semi-automatic annotation, manual steps can come in several parts of the overall process. An initial manual step might identify a basic set of data or terms. These would be used to create a list of words that a computer could find in many documents (thousands or millions). Then a manual step would refine what the computer found and the results would be fed back into the automatic process to make it more precise. Later, a manual process might extend the automatic step for very specific uses, or to create new data as needs change.

Depending on what kind of approach we have decided to use for the automatic annotation - rule-based approach, ML, or both, the Gold Standard Corpora can significantly differ in number of documents. Let us have a look at both workflow cycles:

- ML approach
- Rule-based approach

ML approach

The semantic annotation cycle consists of the following steps:

Step 1: The initial set of documents is loaded and the project annotation schema (annotation types, features, values, etc.) is applied. It is critical to have a good annotation schema and accurate annotations for machine learning that relies on data outside of the text.

Step 2: Automatic annotation is performed, based on the initial model of the phenomena (software text-analysis component). This creates a pre-annotated corpus augmented with higher-level information from components such as tokenisers, sentence splitters, part of speech taggers, gazetteers, PER/ORG/LOC grammars, etc. Adding such information to a corpus allows the computer to find features that can make the defined task easier and more accurate.

Step 3: The pre-annotated corpus is then sent to MA experts for curation. A well-defined manual curation process is essential to ensure that all automatically pre-annotated entries are handled in a consistent manner. This process consists of several steps:

- All annotated entries are checked against the initial annotation guidelines.
- All erroneous entries are corrected, if possible, or deleted.
- Depending on the task or the project, omissions in the pre-processing stage are added as entries.

Step 4: Based on the observations on these pre-annotated documents and the data, and the cases in which entities appear in the text, or the context in which the mentions of annotation types appear, the MA experts revise the initial annotation guidelines and enrich them with specific use cases.

Step 5: A manually annotated corpus is created. It is further divided in two parts.

- One third or one forth of the documents is used to evaluate the performance of the model,

and depending on the achieved results, the model can be revised and the whole cycle repeated.

- The other part of the corpus, which is the biggest portion, is used for training and development of ML algorithms on the data:
    - The algorithms are trained and tested over the corpus
    - The results of training and testing are evaluated in order to see where the algorithms performed well and where they made mistakes.
    - The design of the model is revised and, if necessary, other annotation types are created.
    - The whole cycle or some parts of it are repeated.

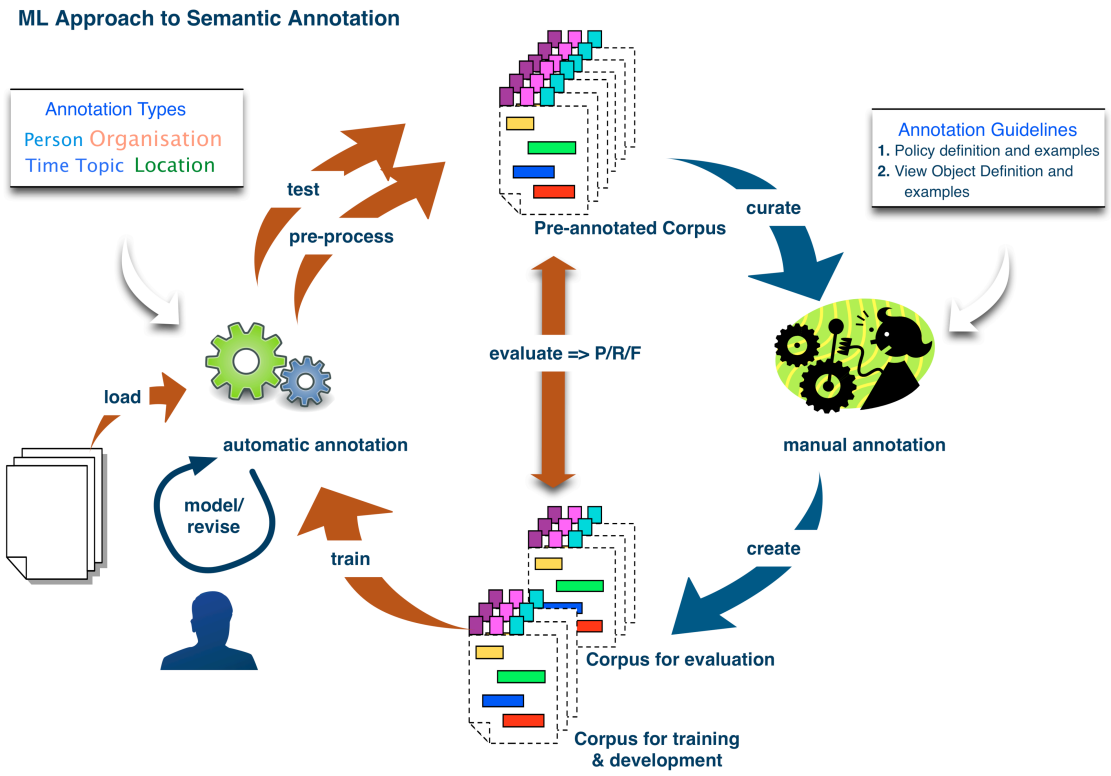


Figure 5: ML Approach to Semantic Annotation

Rule-based approach

The Rule-based approach consists of the following stages:

Step 1: The initial set of documents is loaded and the project annotation schema (annotation types, features, values, etc.) is applied.

Step 2: Automatic annotation is performed, based on the initial model of the phenomena (software text-analysis component). This creates a pre-annotated corpus augmented with higher-level information from components such as tokenisers, sentence splitters, part of speech taggers,

gazetteers, PER/ORG/LOC grammars, etc. Adding such information to a corpus allows the computer to find features that can make the defined task easier and more accurate.

Step 3: The pre-annotated corpus is then sent to MA experts for curation. A well-defined manual curation process is essential to ensure that all automatically pre-annotated entries are handled in a consistent manner. This process consists of several steps:

- All annotated entries are checked against the initial annotation guidelines.
- All erroneous entries are corrected, if possible, or deleted.
- Depending on the task or the project, omissions in the pre-processing stage are added as entries.

Step 4: Based on the observations on these pre-annotated documents and the data, and the cases in which entities appear in the text, or the context in which the mentions of annotation types appear, the MA experts revise the initial annotation guidelines and enrich them with specific use cases.

Step 5: A manually annotated corpus is created. It is used to evaluate the performance of the model, and depending on the achieved results, the model can be revised and the whole cycle repeated.

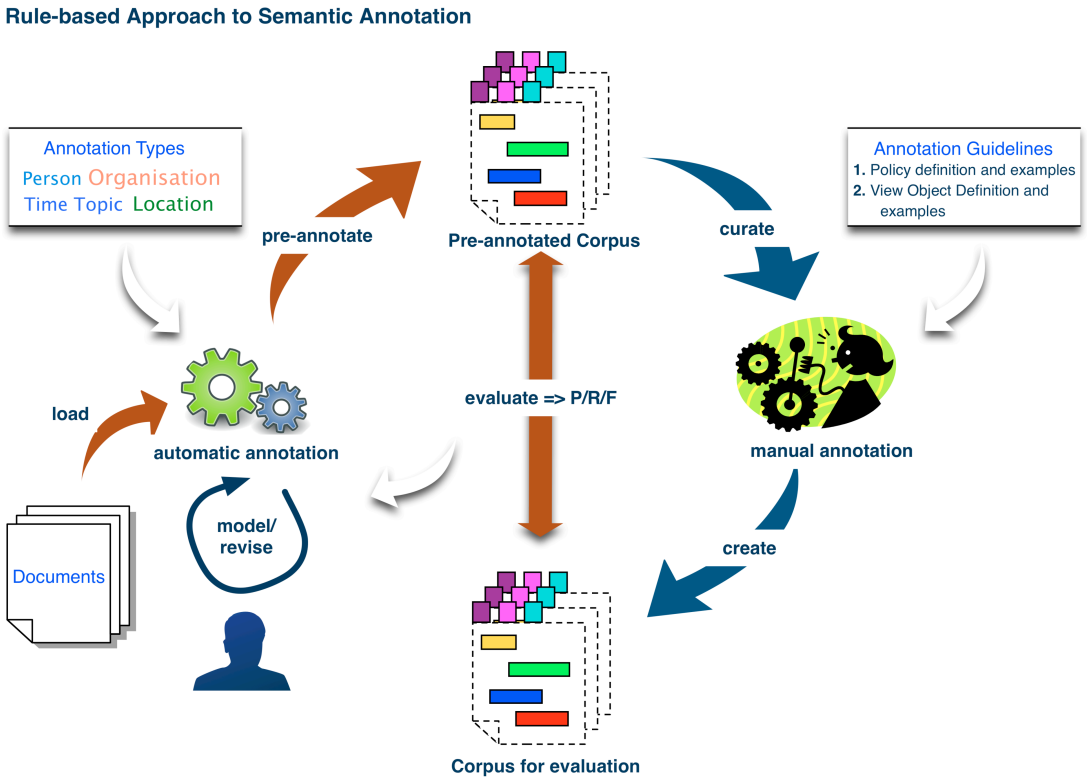


Figure 6: Rule-based Approach to Semantic Annotation

Semantic Extraction Service

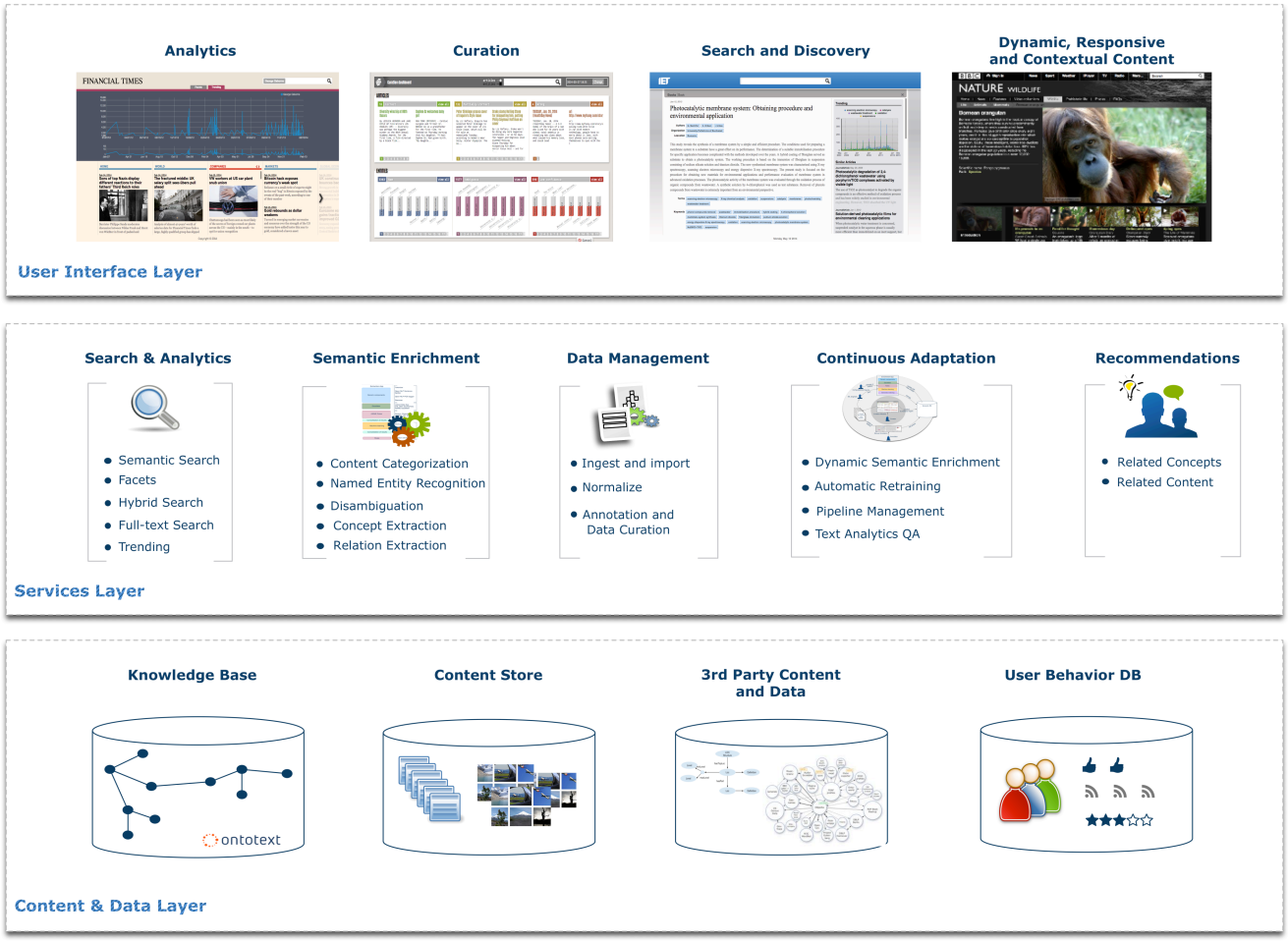


Figure 7 Semantic Extraction Services

The Semantic Extraction Service (SES) is a stand-alone service, which ensures life-cycle automation, i.e., development, evaluation, management and maintenance, of semantic extraction application (SEA). It uses the data in the OWLIM graph database as its dictionary of concepts (knowledge base) to recognise mentions of entities, as well as their relevance and algorithm's confidence, on the text. It comes with an out-of-the-box default pipeline, which is configured to produce media graph basic concept look-ups.

The service supports multiple pools of SEA and provides an API to manage them (start/stop/reload/configure), which is very convenient in cases when more sophisticated semantic annotation is required and also when different text analysis techniques are used to span across multiple data domains. Semantic annotation is covered by special queries in the SPARQL API, which retrieve the matched concepts and their features. Since the SEA are usually based on a mixture of machine learning routines and rules, SES also provides a Re-training API which allows both manual and automated re-training and evaluation of the underlying statistical models (for example, relevance and confidence criteria are updated automatically over time to adjust to new data).

Architecture features

As the Semantic Extraction Service is wired to OWLIM through its Plug-in API, its dictionaries are dynamically updated in a transactional fashion and are always in synchronization with the data in the semantic repository. It also provides high availability ensured by OWLIM's replication, load balancing and fail-over capabilities. Horizontal scaling also works in the same way as in OWLIM. Moreover, the pipelines running on a single cluster node could be pooled in the way so that multiple pipeline instances share the same dictionary and achieve more semantic annotation request throughput while optimizing memory usage. The pools are also capable of dynamically expanding with respect to the current load. Communication with the service is established over HTTP and the SPARQL protocol, which makes it easily integrable with various platforms written in different programming languages.

Continuous Adaptation

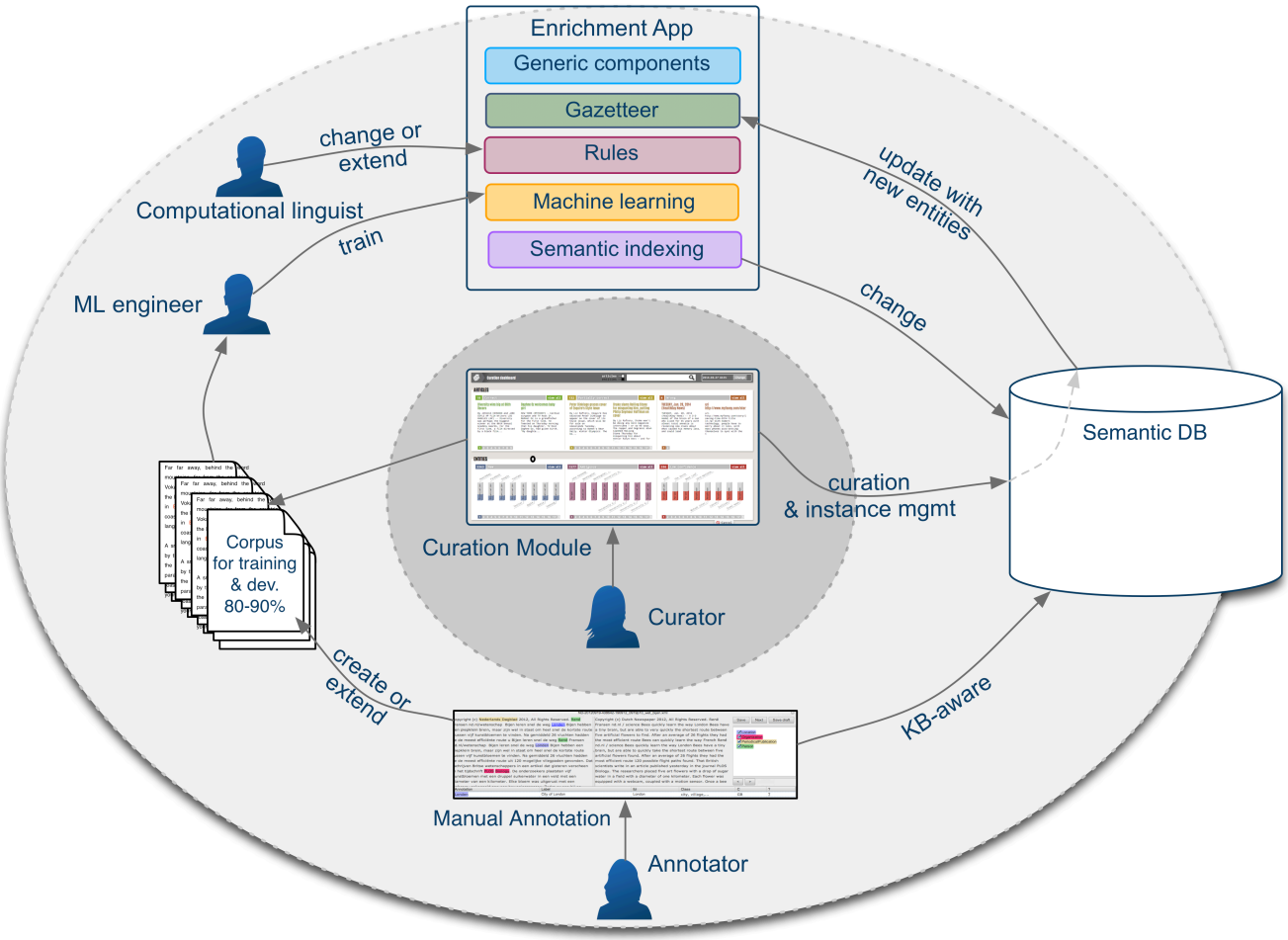


Figure 8 Continuous Adaptation

Concept Extraction Pipeline

The following is a high-level description of the phases (sub-pipelines) involved in an example concept extraction pipeline.

Pre-processing phase

During this phase, generic pre-processing takes place:

- Tokenisation (ANNIE English Tokeniser PR);
- Sentence splitting (ANNIE Sentence Splitter PR);
- Stemming (Snowball Stemmer PR);
- POS-tagging (OpenNLP English POS Tagger PR);
- Chunking (OpenNLP English Chunker PR);
- Lemmatization (Gate Morpholocial Analyser PR).

Most of the processing resources used during this phase are part of the GATE distribution. Several additional rules that improve the output and facilitate the subsequent tasks are also provided (e.g. rules that insert additional splits on newline characters and document elements available through the "Original markups" annotation set; rules that modify noun phrase chunks in order to improve the extraction of keyphrase candidates; rules that generate canonical forms for such noun phrases).

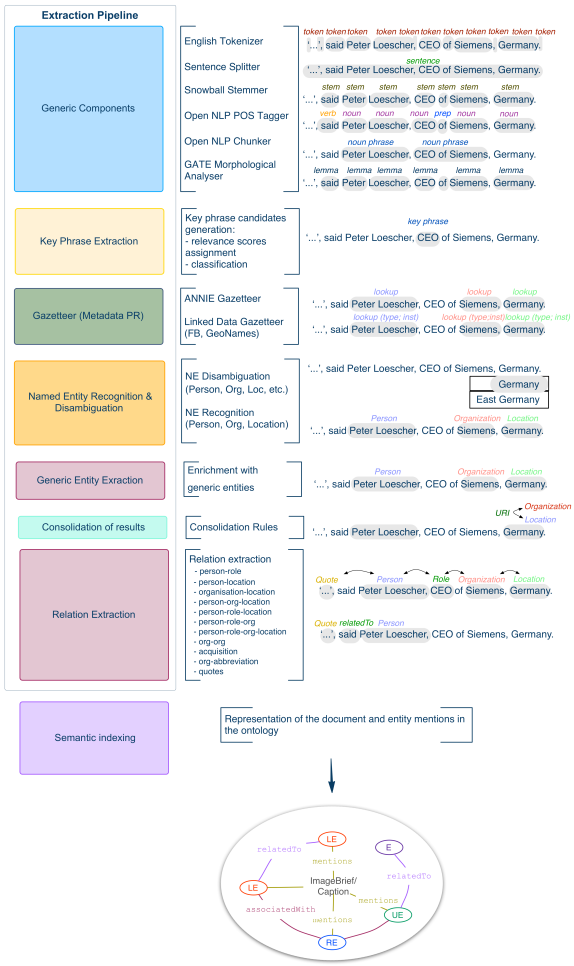


Figure 9 Example Concept Extraction Pipeline

Keyphrase extraction phase

This phase consists of logic that generates keyphrase candidates, assigns relevance scores to the candidates, and classifies them into positive or negative instances via a specialized processing resource for supervised classification.

At the completion of this phase, positive keyphrase instances are stored in a separate set for further reference.

Gazetteer-based enrichment phase

At this stage, document content is enriched by means of semantic gazetteers that annotate various types of entities. These gazetteer lookups are not part of the extraction results, but they provide features required by the statistical models and rules that produce the final set of entities.

All sub-pipelines are responsible for the execution of a single gazetteer (ANNIE Gazetteer, LD Gazetteer), and the consequent transfer of features from the gazetteer-generated annotations to other annotation types involved in the named entity recognition and disambiguation phases.

The “LD Gazetteer” component's cache is populated with instances from the following sources: DBpedia, Freebase, Geonames.

Named entity recognition and disambiguation phase

Named entity disambiguation

Using the named entity candidates discovered by the LD Gazetteer, and given the specific article context, this phase makes use of a specialized classifier to assign a "positive" or "negative" label to each candidate. As a result, the ambiguity associated with the complete set of gazetteer lookups is eliminated – at most one named entity remains per document offset, and the redundant ("negative") named entity candidates are removed.

The disambiguation mechanism relies on a set of Lucene indexes that store:

- a short textual description of each candidate (based on DBpedia and Freebase abstracts);
- a set of URIs representing the entities that appear in the full DBpedia article for each candidate.

Various similarity scores are computed by a specialized processing resource that accesses the indices and evaluates the correspondence between the candidate and the context (the article content and the content stored in the aforementioned indices). The final classification is conducted by a separate processing resource, based on these pre-computed scores and some additional features.

Currently, the component supports disambiguation of named entities belonging to either of the following classes: “Person”, “Location”, “Organization”, "PeriodicalPublication", "Event", "RecurringEvent", "Activity", "AnatomicalStructure", "Award", "CelestialBody", "Color", "Currency", "Device", "Disease", "Drug", "Food", "GovernmentType", "Holiday", "Ideology", "Language", "MeanOfTransportation", "MusicGenre", "ProgrammingLanguage", "Project", "Species", "Work", "Thing".

Discovery of novel named entities

Named entities not available in the “LD Gazetteer” component's cache are not recognized, and therefore not handled by the above-described disambiguation mechanism. The recognition of novel entities belonging to the “Person”, “Location” and “Organization” classes is handled by the PLO

Tagger processing resource, which compensates for the lack of perfect coverage by the classifier-based tagging approach.

The results extracted during phases 4.1 and 4.2 are combined in a way that eliminates the overlapping among annotations produced by the disambiguation classifier and the PLO tagger components. The implemented logic guarantees that the disambiguated entities that have a meaningful URI are preferred to the anonymous entities discovered by the PLO tagger.

Generic entity extraction phase

This phase implements a rule-based enrichment with entities of generic type. Currently, these include:

- dates (normalization logic is provided as well)
- numbers
- money
- percentages
- measurements

Result consolidation phase

This phase contains rules that take into account all entity types discovered during the preceding phases in order to refine the extraction results. At its end, instance URIs are generated and assigned to entities that have no such identifiers.

This phase involves the execution of the “Orthomatcher” processing resource, which deals with the discovery of orthographical variations of the labels and aliases of people, location and organization entities. Subsequently, the trusted URIs of disambiguated entities are propagated to novel aliases annotated by the PLO tagger, based on linkage done by the “Orthomatcher” component. At the end of the phase, URIs are generated for the entities that have not been assigned a valid URI during the above-described phases.

Relation extraction phase

This phase conducts rule-based extraction of various relationships between the atomic entities discovered at the preceding stages. Currently, the following types of relations are supported:

“Person – Role”
“Person – Location”
“Organization – Location”
“Person – Role - Organization”
“Person – Role - Location”
“Person – Organization – Location”
“Person – Role - Organization – Location”
“Organization – Organization”

"Acquisition"
"Organization – Abbreviation"
"Quotation"

Clean-up phase

During this phase, a final clean up takes place, through which redundant intermediate annotations are removed, the document readability is improved, and the annotation sets are reorganized in order to assume the structure expected by the components that process the documents after the completion of the concept extraction pipeline.

Types of Extracted Information

Text can be analysed at multiple levels including:

- Categorisation to a specific categories such as blog post, political news, sport news, etc.;
- Topic extraction - recognizing important words and phrases in the text;
- Named entity recognition (NER) - extracting people, organization, location, time, amounts of money, etc.;
- Term definitions organized in hierarchies or thesauri;
- Concept extraction - extracting well-defined rich entities in a database;
- Relation extraction between all concepts.

Categorisation

News stories are classified into one or more of the following news categories:

- National;
- International;
- Economy;
- Politics;
- Sports;
- Media/culture;
- Science;
- Religion.

We are in the process of collecting documents from categories that have no or less than useful stories such as Science, Religion, Techs and Sports.

Topic extraction

Topics are important phrases or simply keywords that are explicitly mentioned in the document. These phrases, however, are not bound to a specific ontology or knowledge and are extracted on the fly from the stories. As such, they represent the news in terms of coverage better but do not bring along rich background information the way concepts do.

For instance, a news story about Mark Rutte mentions that he is the leader of the "People's Party for Freedom and Democracy (VVD)", however, the abbreviation "VVD" is not part of the available names for this party (as "ANP", for example). Therefore, this will not be extracted by the concept extraction algorithm. Instead, it will be extracted by the topic extraction routine. Abbreviations are used in particular in informal chats and as search words. That is why having topics is very valuable for

the search and navigation of content. The extraction of topics also adds dynamics and increases the coverage of the system.

Term extraction

In Term extraction, algorithms automatically link chunks of text to terms of a thesaurus or other representations. For instance, a scientific paper abstract is linked to terms from the INSPEC thesaurus.

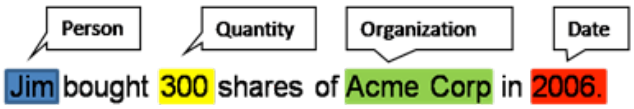
Named Entity Recognition

In the Named Entity Recognition task, an algorithm seeks to locate and classify chunks of text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

NER systems take an unannotated block of text, such as this one:

Jim bought 300 shares of Acme Corp in 2006.

and produce this one:



Concept extraction

Concept extraction is a task that aims to link a chunk of text to a rich concept representation of an object in a database. These objects, which we call concepts, have additional links to other concepts and thus represent a rich graph structure. For instance, Barack Obama is linked to the Democratic Party (his political orientation), other Democrats such as Joe Biden, a place of birth, a spouse, children, education, etc.

The main problem in this task is disambiguating between multiple candidates for a single mention of their name in the text.

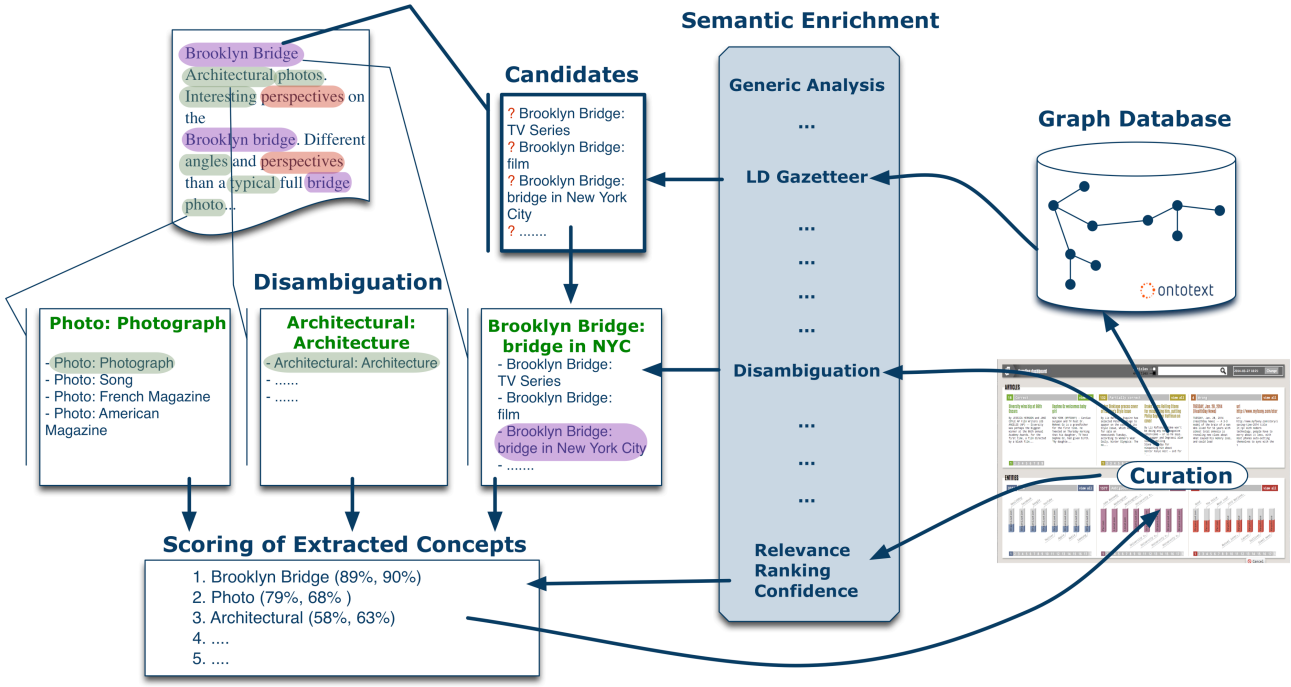


Figure 10 Disambiguation Process

To illustrate the levels of ambiguity and the importance of handling them, below we present the level of ambiguity based on 354 manually curated documents.

Definitions:

Concept: an instance in the knowledge base, with a unique ID and multiple names/labels

Mention: an actual appearance of an entity name in the text

A mention of a concept is a match for any of the concept names (alternative names). The number of mentions with more than two competing concepts in the gold standard is 9304. The mention with most ambiguities is “Paul Smith”, for which there are 70 concepts. The average number of candidates for a single mention is six. We may consider this as a representative number of concepts between which our algorithm disambiguates on average.

All mentions are 16401 and the percentage of ambiguous mentions is 57%.

Below we show the first 10 ambiguous mentions.

Paul Smith: 70 candidates
John: 64 candidates
William: 62 candidates
Premier League: 59 candidates
De Hoop: 55 candidates
Alex: 54 candidates
Alexander: 52 candidates
Washington: 49 candidates
John Roberts: 48 candidates
Michael: 44 candidates

Bellow we represent only the first 10 alternative concepts for “Paul Smith”, identified by their URI.

- http://dbpedia.org/resource/Paul_Smith_(cricketer)http://dbpedia.org/resource/Paul_Birch_(actor)
- http://dbpedia.org/resource/Paul_Smith_(academic)http://dbpedia.org/resource/Paul_Smith_(musician)
- http://dbpedia.org/resource/Paul_Smith_(fullback)http://rdf.freebase.com/ns/m.05x4g5p
- http://dbpedia.org/resource/Paul_Girard_Smithhttp://dbpedia.org/resource/Paul_Smith_(composer)
- http://rdf.freebase.com/ns/m.066849fhttp://dbpedia.org/resource/Paul_M._Smith_(photographer)

In the second sheet, we represent the top 100 most frequently appearing ambiguous concepts in the gold standard set. It can be seen that the Netherlands and, in general, locations like Europe are mentioned many times throughout this set of documents and all of them carry a certain ambiguity.

Relation extraction

Relations are references between one or more concepts and things that are extracted only from the text of the story. These should not be mistaken with relations that already exist in the knowledge base. Relations add new knowledge and connections between concepts in the knowledge base and can be of the following types:

Generic Relations

Generic Relations attempt to connect at least one recognized concept to other concepts or topics without predefining relationship types.

PersonCareer relations

PersonCareer relation expresses a relation between a person and: (i) a company where he has a position (e.g. John Smith from General Electrics); (ii) his occupation (President Obama, general director Mark Thomas); (iii) his position within a location (Mayor of London); (iv) his position within an organization active in a specific location (Ben Bernanke is the chairman of the Federal Reserve, the central bank of the United States).

Company relations

Company relations express relation between a company and: (i) its location (Siemens of Germany); (ii) its daughter company (VOX Global, a subsidiary of Omnicom Group Inc (OMC); (iii) its mother company; (iv) competitor companies (The Toulouse, France-based aircraft maker is outselling rival Bombardier Inc.); (v) company type (bank, telecom, etc.); (vi) customer companies; (vii) collaborator companies (Bloomberg Finance L.P., a Delaware limited partnership); (viii) another

company such as merger/acquisition (Amgen Inc. agreed to acquire Onyx Pharmaceuticals Inc.); (ix) company's abbreviation; (x) and a quotation made by the same company ("...", an allegation denied by Herbalife.).

Other relation types

- Relation between two locations such as sub-region of, part of, located in, etc. (Tegucigalpa, Honduras; Buddhist kingdom of Mustang, northwest of Kathmandu, Nepal)
- Relation between a person and quotation ("There's a perception that doctors are meant to heal wounds, not bleed them," Mr Kamara said.)

Curation

Usually automatic annotation is not of sufficient quality to enable focused search and retrieval: either too many or too few terms are semantically annotated. Therefore, in the semi-automatic approach to annotation, after we run the corpus through the automatic pre-processing stage, we need to train human annotators to manually mark up (annotate) or check (curate) the annotated dataset. Manual annotation or curation is a very important part in the development of semantic annotation and search solutions.

Here, we associate specific spans of text with specific labels following strict guidelines that describe what spans to annotate and how to label them. The basic annotation unit within a project corpus is the so-called "entity". Entities refer to real-world objects. The span of text that refers to an entity is called a "mention" of that entity. An entity may appear several times in the same document and different mentions may refer to the same entity.

Just as the entity is the basic unit of annotation, so marking up entities and mentions is the basic sub-task of the annotation process. In this sub-task, stretches of text are marked as being mentions of an entity of a particular annotation type.

Types of annotation projects

There are many types of annotation projects, some of which do not fit neatly into categories. Below are four examples that describe the most common types. The general guidelines cover all of them but the specific guideline processes may be quite different.

Gold Standard

The goal of the Gold Standard is to create a top threshold for measuring computer performance through manual annotations applied to the same text by more than one person. This is often used at the outset of a project to see how often two or more annotators agree on a particular annotation or type of annotation. When the independent annotators reach some level of agreement, may be 85% of the time, this becomes the target for a computer-generated annotation. In other contexts, we might create a very precise manual annotation of 100% accuracy in order to train an automatic process (or judge the effectiveness of an automatic process).

Quality Assurance

A quality assurance project often looks for errors in previously processed documents. The errors may either be corrected, or simply collected and analysed, depending on the need. A QA process may also create its own gold standard on a subset of all documents — a mini corpus - to target specific problems. QA processes often seek 100% accuracy, unlike other processes, but on small, manageable sets of documents. Thus, the guidelines for annotators normally will be stricter and the process somewhat different.

Processing

A processing project may seek to add annotations that would be used in a processing step and then could be removed. For example, a processing project might highlight errors in a document and provide corrective labels, which could be used to make automatic corrections in the original and then stripped from the text itself. Alternatively, a processing annotation could be used in automatic routing of a document in an information management system. There are many types of processing projects. Typically, the labels will be quite specific with no possibility for variation.

Social

Social annotation (or "social tagging") is becoming increasingly popular both on the Web and in organisations. Often, social tagging is an attempt to arrive at a new nomenclature or to accommodate bottom-up terminology (AKA "folksonomy") with a more formal system of names and labels. Social annotation might uncover new popular trends, help with navigation, or generate civil discussion of various topics. In general, there is no attempt to constrain the annotations or how they are applied but to harvest the various inputs to see if/when patterns emerge.

Here in Ontotext, we use manual annotation for creating Gold Standard Corpora that are used for ML training or system evaluation.

Curation Guidelines

The Annotation Guidelines are a very important part of manual annotation and curation as they provide a common understanding of what defines an annotation and detail the process by focusing on the annotation types used (depending on the project type). The guidelines document describes precisely and with clear examples, for each entity type, how to map it from the text to the annotation, including:

- Which pieces of text should be annotated?
- How should spans of text be mapped to mentions? Which text should be included/excluded?
- How should special cases be dealt with?
- What information should be recorded for different entities?

In manual annotation and curation there is a fine line between having an annotation that is the most precise (has high informativity), and having an annotation that is not too difficult for annotators to complete accurately (which results in high levels of correctness). If the Annotation Guidelines provide too many rules and exceptions, it would be difficult for the annotators to follow them, which will in turn hamper the ML task.

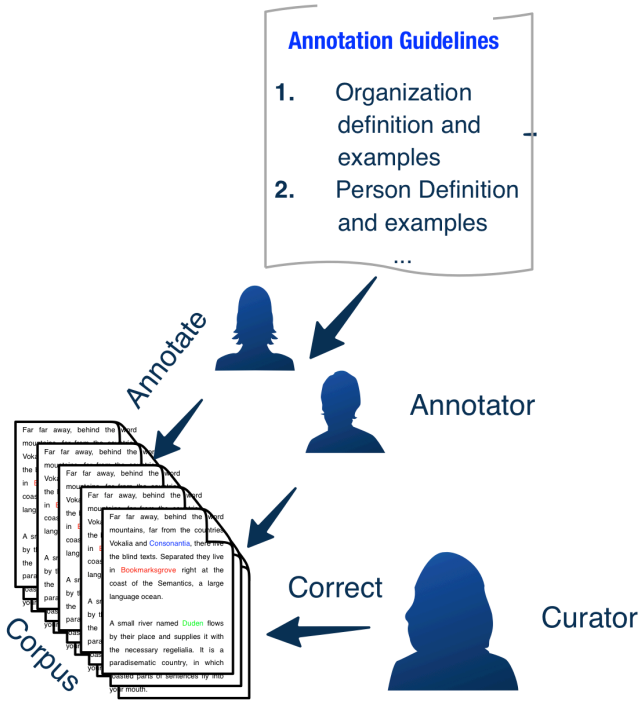


Figure 11: The Annotation/Curation Process

The following exemplifies why annotators need a common understanding of what spans to annotate and how to label them. Here, the different ways in which the Organisation "QBC Productions Inc. of East Anglia" could be tagged are illustrated below:

- [QBC Productions] Inc. of East Anglia
- [QBC Productions Inc.] of East Anglia
- [QBC Productions Inc. of East Anglia]

Each of these might look correct to an individual annotator, but only one actually corresponds to the correct mark-up in the Annotation Guidelines.

Here are some techniques/templates for preparing Annotation Guidelines, which show how to provide general rules, a clear definition for each annotation type, a brief list of positive annotation examples, a brief list of exceptions, etc.

Accuracy and Performance

Gold Standard Corpus

A central challenge of semantic annotation is the lack of objective standards for assessing the success of the process of converting data to higher-level information.

Once the initial corpus is annotated by at least two people (more is preferable, but not always practical), the gold standard corpus can be created. It is the final version of the annotated data. It uses the most up-to-date specifications that were created during the annotation process, and it has everything tagged correctly according to the most recent guidelines.

Note that the gold standard does not assess whether the system accurately classifies phenomena, but the extent to which the system agrees with the human annotators where those classifications are concerned. Here it is important to decide what level of agreement is acceptable knowing that humans are bound to make mistakes.

Evaluation Metrics

The evaluation metric mathematically defines how to measure the system's performance against the manually annotated gold standard:

- Precision - what percentage of the annotations proposed by the system are correct (when compared to the gold standard);
- Recall - what percentage of the annotations in the gold standard were correctly identified by the system;
- F- score - an average that rewards precision and recall values that are close together. (For many annotation tasks, we are interested in obtaining high levels of precision in conjunction with high, or at least reasonable, levels of recall).

Sometimes traditional methods for IE are not sufficient for ontology-based IE as the distinction between right and wrong is less obvious. For example, recognising a Person as a Location is clearly wrong, but recognising a Research Assistant as a Lecturer is not so wrong.

Annex A: Manual Annotation Guidelines – Example 1

Introduction

This document provides instructions for annotators and curators, who will be involved in the annotation exercise. The goal of this exercise is to create a gold standard corpus based on .... *(insert document type)*. ....*(Project name)* gold standard annotations mark semantic units in .... *(document type)* such as .... *(insert list of entities)*.

This document provides a common understanding of what defines an annotation, and details the process of manual annotation focusing on annotation types, which are used to mark the semantic units.

Terminology

This section describes the terms used to discuss annotations. For the purposes of annotating the .... *(project name)* gold standard, a single terminology will be adopted. It makes use of the following example:

"Example sentence or two from the document here"

Annotation Terminology

Table 1: Annotation Terminology

Term	Description	Example
Entity	An entity is a thing in the world. It has an existence independent of the text: it is not a piece of text. It may be concrete or abstract. Explicit entities are mentioned in the text. Implicit entities are not mentioned, but their existence may be inferred from the text. We are not interested in annotating implied entities, only those that are explicitly mentioned in the text.	Insert an example of an entity here.
Mention	A mention is the textual realization of an entity. A single explicit entity may have more than one mention. An implicit entity has no mentions.	Insert an example of a mention here.

What is an Annotation?

An annotation is a piece of information attached to some text, usually describing the text in some way. An annotation may be:

- Attached to a particular region of a document, such as to a word or group of words.
- Attached to the document as a whole, and independent of a particular span of text.

In *(project name)*, we are interested in *(insert explanation of entities selected)*. Annotating a document is the task of marking the mentions of these entities in a document, describing their type. Typically, Annotation Editor will display annotation by highlighting the text to which it is attached with some colour.

This section gives descriptions of the *(project name)* gold standard annotations for entities, and what things they refer to. It tells you:

- What *(project name)* annotations stand for?
- How to make sense of annotations in a ready-annotated abstract.
- What an annotation means if you find it attached to a piece of text in a document.
- It does not tell you:
- How to add annotations to an un-annotated document.
- The detail of the mapping between surface text and either entities or relationships.
- It does not say how to decide whether a piece of surface text should have an annotation.

For these things, refer to the more detailed description found in the “Annotating Entities in Text” section.

Annotating Entities in Text

The basic annotation unit within the *(Project Name)* corpus is the entity. Entities refer to real-world objects: *(insert entities specific to Project Name)*. Entities are grounded in the text. The span of text that refers to an entity is a mention of that entity. An entity may appear several times in the same document. Different mentions may refer to the same entity: “example here”. Just as the entity is the basic unit of annotation, so marking up entities and mentions is the basic sub-task of the annotation process. In this sub-task, stretches of text are marked as being mentions of an entity of a particular type. This section describes, for each of the entity types, how annotators should map from the surface text to annotation:

- Which pieces of text should be annotated?
- How should spans of text be mapped to mentions? Which text should be included/excluded?
- How should special cases be dealt with?
- What information should be recorded for different entities?

Entity Types

This table summarizes all of the entities to be annotated and any features that they have. It gives a brief description and examples.

Table 2 Descriptions of Entity Types

Type	Features	Description	Examples
Entity X	Features of Entity X	Description of Entity X	Examples of Entity X

It should be noted that this table would contain information pertaining to all the entities relevant to the specific annotation project.

…. *(ENTITY X)*

- What is *(Entity X)*? (General Definition)
- Insert a brief list of examples of what would be considered *(Entity X)*.
- Insert any other pertinent information for determining *(Entity X)*.
- What is not *(Entity X)*?
- Insert a brief list of examples of what would not be considered *(Entity X)*.
- Insert any other pertinent information for determining what is not considered *(Entity X)*.
- Other Rules.
- Insert any other information that is relevant to determining what is or is not *(Entity X)*
- Insert any specific circumstances that will need further instruction, i.e. – modification by other words.

Annex B: Manual Annotation Guidelines – Example 2

The corpora contains:	About the project:
Number of documents:	**Name:**
Domain:	**Type:**
Crawled URLs:	**Client:**
Language:	**Project Lead:**
Used for (training, evaluation):	**Confidentiality:**

These guidelines provide instructions for manual curation of previously processed documents from the general news domain. They will be used as a gold standard corpus for automatic text analysis evaluation, as well as for training the ML model. The level of granularity for the creation of Annotation Types and their Features is based on more concrete classes for **Person**, **Organisation**, **Location**, **Event**, **RecurringEvent**, **Project**, **ProductModel**, **PeriodicalPublication**, **Religion** of the PROTON Ontology, with extension of classes and properties from DbPedia, Freebase and GeoNames.

The document curation was divided into two consecutive jobs:

1. Instance disambiguation - possible true candidates are verified by assigning the correct URL.
2. Tagging (only for **Person**, **Organisation**, **Location**) - the labels **Person**, **Organisation**, or **Location** assigned during the automated annotation stage are verified and new labels are added to specific words or phrases in the text.

General rules

For the instance disambiguation job:

Verify the correct instance candidate by opening the URL of "Lookup" labels in a web browser.
Example: "Tijdens de slotbijeenkomst van de vrijgemaakt-gereformeerde Schooldag in de Broederkerk in [Kampen] zingen honderden bezoekers Psalm 119"
There are 3 instance annotation options for Kampen with the following URLs:

- http://www.geonames.org/2753106/kampen.html - positive
- http://www.geonames.org/2753105/kampen.html - negative
- http://www.geonames.org/2753107/kampen.html - negative

  In this example, the text is about Broederkerk, which is a church in the city of Kampen, Overijssel. This helps to select the first instance annotation of Kampen as correct. The main label for this instance annotation is Kampen.

- The name of the recognized Entity coincides with the Main label or any of the instance labels.

- Mentions cannot have two assigned instances at the same time.
- When two instance candidates are correct but one references DBpedia and the other one - freebase, the one referencing DBpedia is assigned.
- When all available instances reference wrong URLs then none of them is assigned.
- When the only instance candidate references freebase where all the available information for it is that it is a person or an organisation, it is not assigned.
- When there are several instance candidates, which are valid locations but the text does not provide enough information to verify which one is correct, then the instance candidate of the bigger populated place is assigned.
  Example: the text is about church movements in different cities one of which is Zwolle and there are two valid locations in the Netherlands with that name - one in Overijssel (bigger) and another one in Gelderland (smaller).
- When a mention in the text references a specific part of a location such as East, West, South, etc. and the location candidate references the general location, it is assigned as correct.
  Ex: "Zuid [Soedan]", "Oost [Europa]", "Westelijke [Sahara], etc.

For the tagging job:

- Wrong assignment of annotation type label should be deleted or converted to the right one.
- Mentions that have been omitted during the automated annotation stage must be labelled with the respective annotation type label.
- When the same mention occurs several times in a document but refers to different annotation type labels, each mention is tagged with the correct annotation type label depending on the context.

Example when "Rijksmuseum Twenthe" refers to the building where the Rijksmuseum Twenthe is situated, the mention is annotated as location and when it refers to the private organisation in the form of a foundation then it is annotated as organisation.

- When valid annotations are fused together with other words in the text due to bad formatting, they are not assigned as correct.

  Ex: "ESMRatificatieNoodfonds", "Manon UphoffGeboren", etc.
- When during the instance disambiguation job a partially correct location candidate was assigned (a mention in the text references a specific part of a location such as East, West, South, etc. and the location candidate references the general region), then in the tagging job the whole location is annotated.

  Ex: "[Zuid Soedan]", [Oost Europa]
- When several mentions consisting of two words are listed, both mentions are tagged as one location/org.

  Ex: [West en Oost Europa], [Tweede en Eerste Kamer], etc.
- When several names of persons are listed, each name is tagged separately.

Ex: "[Alexandra], [Judith] en [Kitty Leschan]", etc.

- When the mention references a person of royal origin who is best known with his/her title, both title and name are tagged

  Ex: "[Prins Jean], "[Koningin Beatrix]", etc.

Annotation Types

Person

Definition:

An individual referred to by short name, first name, family name, and full name

Instance is true if it is:

- One name/two name/full name is associated with any of: profession, organisation, birth date, location, or anything else the document talks about.

  Ex: the document is about "politics" and the recognized person is "Mark Rutte" - in the KB, there is a person with the name "Mark Rutte" associated with "politics", so we assume he is the same person.

- A Person is of royal descent in which case there is usually only one name after the royal title.

  Ex: "koningin [Beatrix]", "Prins [Carlos]", etc."

- Historical figures associated with a religion such as [Jesus], [Ezechiël], [Jesaja], koning [Saul], etc.

  Ex: "Waren [Abraham], [Mozes] en [Samuël] wellicht rijp voor behandeling?"

- A Person is mentioned at least twice in the text, once with a short and once with a full name.

  Ex: "En dan waren er nog de toespraken, ditmaal tijdens de morgensamenkomst in dezelfde Broederkerk, uitgesproken door de Kamper onderzoeker dr. [Hans Schaeffer] en dr. [Henk Geertsema]... [Schaeffer] en [Geertsema]" legden...

Instance is not true if it is:

- The name of the recognized Entity coincides with the Main label or any of the instance labels but it references a person associated with a different profession, organisation, location, birth date, or anything else the document talks about.

  Ex: the document is about "banking" and it mentions the "investment bank Nomura". The recognized person is "Jens Søndergaard" - in the KB, there is a person with the name "Jens Søndergaard" but this annotation instance is about a Danish expressionist painter, so it is not the same person.

- The person candidate is split into several fragments pointing to irrelevant persons.

Ex: "[Eberhard] [van der Laan]"

- a Person is mentioned at least twice in the text, once with a a full and once with short name but the short name mention references a different instance.

  Ex: "Dus hielden respectievelijk [Mark Rutte] en [Diederik Samsom] hun mond over punten die in de formatie een rol gaan spelen. Na de bijdragen van [Rutte] en [Samsom] bloedde het debat vrijwel dood."

- A fictional character in a book, movie, theater play, etc.

  Ex: "[Orfeo] ed [Euridice]", "[Jesus Christ] Superstar", etc.

  Ex: "1670 van Jan Steen waarop [Cleopatra] wedijvert met [Marcus Antonius] wie het rijkste is."

- The name is used as a metonym.

  Ex: "St Paul" refers to the St. Paul Church and not the saint.

  Ex: "Het Kröller-Müller Museum in Otterlo heeft maar liefst dertien [Maillols]."

Organisation

Definition:

A group with a particular skill set, strategy, resources, or priorities, such as commercial, international, religious organisations, executive, legislative, judicial government, etc.

Instance is true if it is:

- A commercial organisation (organisation that buys or sells goods or services for a profit. It may also be a business or it may merely be a sub-organisation of a business entity such as airlines, banks, sports clubs, insurance companies, news agencies, media companies, telecom, etc.).

  Ex: [Thomas Cook Airlines], [Federal Reserve Bank of Boston], [Dick, Kerr's Ladies F.C.], [Football Association] or [(FA)], [Reuters], etc.

- An international organisation (an international organisation is an organisation with an international membership, scope, or presence.).

  Ex: international organisations such as [Internationaal Monetair Fonds], [Internationaal Olympisch Comité] or [(IOC)], etc.

  Ex: multinational companies such as [General Electric], [Heineken], [Philips], etc.

  Ex: international governmental organisations such as [Europese Commissie], [Europese Unie] or [(EU)], [VN-vluchtelingenorganisatie Unhcr], etc.

- An educational organisation (educational organisation refers to nonprofit organisations providing educational services such as universities, schools, etc.).

  Ex: [Harvard University], [Hogeschool Zuyd], "[Vrije Universiteit] in Amsterdam", etc.

- Health and residential care institutions such as hospitals, dentistries, pharmacies, nursing homes, hospices, etc.

  Ex: [Universitair Medisch Centrum Groningen], etc.

- Religious organisation (religion-supporting organisations).

  Ex: [Operatie Mobilisatie], [Baptisten Gemeente Amersfoort], [Rooms-Katholieke Kerk], [Ordine Constantiniano], [Youth for Christ], etc.

- Executive government (the branch of the government that is responsible for carrying out the laws such as the Cabinet, the different Departments).

  Ex: [Raad van State], [Financiën], [Sociaal Economische Raad], etc.

- Judicial government (the branch of the government responsible for the administration of justice.

  Ex: [Hoge Raad der Nederlanden], [Staten-Generaal], etc.

- Legislative government.

  Ex: [Tweede Kamer], [Eerste Kamer], [Senaat], etc.

- Non-government organisation (non-governmental organisation, or NGO is a legally constituted organisation created by natural or legal persons that operates independently from any government. NGO's can be: Trusts, charities and foundations.).

  Ex: [Foundation4Life], [HelpAge International], [Partners in Health], etc.

- Political entity (a unit with political responsibilities) such as Parliament, a Political party.

  Ex: [PvdA], [GroenLinks], [ChristenUnie], [Muslim Brotherhood], etc.

- Military organisations and army structures such as Army, Navy, Air Force, Police, Firefighters, Strategic Missile Force, Coast Guard, Special Forces, Marines, Border Patrol, etc.

  Ex: Koninklijke Landmacht, etc.

- National sports teams.

  Ex: [Oranje] or [Nederlands voetbalelftal], etc.

Instance is not true if it is:

- A location although there is an organisation with the same name.

  Ex: [Manchester], [Liverpool], etc.

- Part of the name of an event.

  Ex: "Bernardus [Porsche] Golf Cup", "de [Coca Cola] Games in 1984", etc.

Location

Definition:

Location is a particular place such as a geographical region, a natural location, a public or commercial place, roads, structures, buildings, etc.

Instance is true if it is:

- a region such as dioceses, parishes, villages, towns, cities, provinces, states, countries, continents.

  Ex: [Europa], [Philadelphia], [Netherlands], [Noord-Brabant], [Den Haag], [De Lage Vuursche], etc.

- a natural location such as mountains, mountain ranges, woods, rivers, wells, fields, valleys, gardens, nature reserves, allotments, beaches, national parks.

  Ex: [Rhein], [Noordzeekanaal], [Mount Everest], [Vondelpark], [Te Urewera National Park], [Stille Oceaan], etc.

- a public place such as squares, opera houses, museums, markets, airports, stations, swimming pools, sports facilities, youth centers, parks, town halls, theaters, cinemas, galleries, camping grounds, NASA launch pads, club houses, libraries, parking lots, playgrounds, cemeteries.

  Ex: [Paleis Soestdijk], [Schiphol], [Rijksmuseum Twenthe], [Covent Garden], [Centre Pompidou], [Station Amsterdam Centraal], etc.

  Ex: " In de [Kunsthal] is een proefopstelling gemaakt."

- A commercial place such as chemists, pubs, restaurants, depots, hostels, hotels, industrial parks, nightclubs, music venues.

  Ex: [Ventana Inn & Spa], [Dean&DeLuca], etc.

- Roads (streets, motorways).

  Ex: [Champs-Élysées], [Damrak], etc.

- Structures (bridges, ports, dams).

  Ex: [London Bridge], etc.

- Assorted buildings (houses, monasteries, creches, mills, army barracks, castles, retirement homes, towers, halls, rooms, vicarages, courtyards)

  Ex: [Haarlemmerpoort], etc.

- Before or after a postal code (part of an address)

  Ex: "[London] SW1A 2HB [United Kingdom]"

- Preceded by prepositions for place (in/on/at/by/next to/beside/near/between/behind/in front of/over/above/under/below) and prepositions for direction

(from/to/into/toward/through/onto/across).

Ex: "in het [Van Gogh Museum], [Tate Modern] in [Londen]"

- When annotation is partially correct - the location candidate references the general location and not the more specific North/South/East/West part of the location.

  Ex: "Zuid [Soedan]", Oost [Europa], etc.

Instance is not true if it is:

- The location candidate is split into several fragments pointing to irrelevant locations.

  Ex: "De [Zwarte] [Ruiter]", "New [Mexico]", etc.

- Denotes an organisation's branch.

  Ex: "IKEA [Eindhoven]", "Porsche [Netherlands]", "Baptisten Gemeente [Amersfoort]", "Ondernemersvereniging Centrum [Utrecht]", etc.

  Exception: When the organisation does not have branches in other locations even though the location is not introduced by a preposition for place.

  Ex: "AMC [Amsterdam]", "Sint Antonius Ziekenhuis [Nieuwegein]", etc.

- Is a highway, etc.

  Ex: "...rijkswegen [A 15] en [A 20]..."

- Is part of the name of an organisation.

  Ex: "Community of Protestant Churches in [Europe] (CPCE)", "[Malmö] Aviation", [Nederlands Fotomuseum], etc.

- is part of a JobPosition/Position in society.

  Ex: "President van de [Verenigde Staten]", "Minister-president van [Aruba]", etc.

  Ex: "Prins Jean van [Bourbon-Parma] van [Nassau] van [Luxemburg]", etc.

  Exception:

  Ex: "De presidenten van Egypte en Iran, respectievelijk, Mohammed Morsi en Mahmoud Ahmadinejad.."

  Ex: "het regime van president Assad van Syrië"

- is part of a show/program/tv channel etc:

  Ex: "Ze komen voor een nieuw tv-programma op [Nederland] 3 waarvan zij de presentatie in handen hebben."

  Ex: "Ik hou van [Holland]"

  Ex: "[South Park]"

- is part of the name of an event.

  Ex: [Munich] Oktoberfest, "[Amsterdam] Fashion Week", "[Kyoto] protocol", etc.

- is part of the name of an organisation.

  Ex: "Universiteit van [Amsterdam]" (or the UvA), etc.

- is part of the name of a product/model.

  Ex: "De [Ibiza] was weliswaar een schot in de roos, maar de [Altea] en de [Toledo] hadden niet echt een richting."

- is part of or the name of a sport club.

  Ex: [Manchester] United, etc.

- is part of a title of a book, movie, play, course, etc.

  Ex: "[Paris], [Texas]", etc.

- is part of a title of a periodical publication, etc.

  Ex: "[New York] Daily News

- is used as a metonym.

  Ex: [WallStreet] = as the financial market of the US

  Ex: [Hollywood] = the US film industry in general

- general directions such as East, West, North, South.

  Ex: "Zeker het christendom is gauw het mikpunt, denkt de godsdienstsocioloog, omdat zij lange tijd dominant was in het Western."

Religion

Definition:

An organized collection of belief systems, cultural systems, and worldviews such as Islam, Protestantism, Buddhism, etc.

Instance is true if it:

- references a world religion.

  Ex: "De pers viel hem scherp aan over al dan niet correct weergegeven standpunten over de [islam] en moslims."

  Ex: "Een bekend voorbeeld zijn de beelden van heiligen die tijdens de [Reformatie] uit de kerken werden verwijderd."

Instance is not true if it:

- references an adherent of a religion such as Muslim, Protestant, Buddhist, Jew, etc.

  Ex: "Als een [protestant] aan het eind van zijn leven is gekomen..."

- references the religious institution.

  Ex: [Protestantse Kerk in Nederland], etc.

- references the title of a book, movie, play, course, etc.

  Ex: "De Alpha-cursus, later [Christianity] Explored, totdat we gewoon begonnen met Bijbellezen..."

ProductModel

Definition:

Different consumer products, or product models for example: iPhone, iPad, SUBARU Legacy etc.

Instance is true if it is:

- the recognized product/model coincides with the Main label or any of the instance labels.

  Ex: [Boeing 747], [Opel Mokka], [Peugeot 5008], [Ford Galaxy], etc.

- the recognized product/make partially coincides with the Main label or any of the instance labels because it references the basic product/model.

  Ex: "[Citroën C4] Grand Picasso", "[Hyundai i20] First Edition", etc.

Instance is not true if it is:

- the mention references the company making the product/model.

  Ex: "De nieuwe samenwerking tussen [Mercedes] en [Renault] moet gaan zorgen voor een nieuwe limousine van [Renault]."

- is part of the name of an organisation.

  Ex: "[Bugatti] Club", etc.

Event

Definition:

Major events, diseases, natural disasters

Instance is true if it is:

- a valid event, decease, natural disaster, etc.

  Ex: [Eerste Wereldoorlog], [9/11], [HIV], [depression], [Bijlmerramp], [Koude Oorlog], [Holocaust], etc.

  Exception: When the name of a location is used as a name of the event.

Ex: [Tsjernobyl] - when mentioned in "kernramp van Tsjernobyl"

Instance is not true if it is:

- a location associated with an event.

  Ex: [Oostfront], etc.

RecurringEvent

Definition:

Recurring events such as the Olympics, film festivals, conventions, etc

Instance is true if it is:

- a recurring event such as the Olympics, film festivals, conventions.

  Ex: [Oktoberfest], [Bernardus Porsche Golf Cup], etc.

PeriodicalPublication

Definition:

Regular publications such as newspapers, journals and magazines

Instance is true if it is:

- a regular publication.

  Ex: [Nederlands Dagblad], [De Journalist], [Kathimerini], [Autoweek], [Bild], etc.