MSc project—A (related to Evolutext)

PROPOSED STUDY : Alignment of thematic corpus and thesaurus with relevant web resources

Aims

1—assessing the relevancy of automatized entity linking of an ancient text through DBpedia Spotlight tool

2—linking the concepts of an existing thesaurus with DBpedia data

3—enriching the thesaurus with extracted concepts linked to DBpedia

DBpedia is a knowledge base built by transforming Wikipedia pages into a machine-readable format called RDF (Resource Description Framework). A side project, DBpedia Spotlight, automates the annotation of natural language in order to extract words or groups of words and link them with their corresponding DBpedia entity.
The first task of this project will be to use DBpedia Spotlight to process an English text (translation of an ancient zoological Greek text) and assess the quality of the results (precision and recall of the entity linking process). Complementary filter could be implemented to refine the linking process (by exclusion of not pertinent information, such as too general information on unspecialized terms). The second task will to be to connect this annotation with an existing thesaurus on ancient zoology. One way to do that could be to link the concepts of the thesaurus with DBpedia entities, so that both the text and the thesaurus would be linked with the same pivot knowledge reference. Beside DBpedia data, an exploratory research will be conducted on other knowledge bases relevant to the topic, that could be aligned with the concepts of the thesaurus. An additional (but optional) task would be to design a regulated process of feed-back from annotated text to thesaurus, enriching the so far uncomplete thesaurus with new concepts.

The thesaurus will be provided with full access, along with the English text to operate on (in xml format). Other software and knowledge bases used have to be in open access.

Pre-requisite : Semantic Web technologies (RDF, SPARQL)

MSc project—B (related to Evolutext)

PROPOSED STUDY : Semantic annotation tool relying on an in progress specialized thesaurus

Aim

Designing a fair and ergonomic tool to annotate text in connection with an existing thesaurus (providing suggestion of labels)

Many annotators (and web annotators) already exist, but with complex environment or limited reference dictionaries. The tool must allow tagging with new labels that will be automatically added to the thesaurus. This interaction (go-to-and-fro) from thesaurus to text will be an added value to existing tools.

MSc project—C (related to Evolutext)

PROPOSED STUDY : Automatized semantic annotation using a thematic thesaurus

Aim

1. **Semantic annotation of text segments with zoonyms based on the Thezoo thesaurus (zoonyms and few general zool. sub-domains based on their representation by sets of semantically related terms) Improve knowledge extraction from texts**
2. **Reusing state-of-the-art NLP methods and supervised learning algorithms and libraries for basic categorization of text segments**

The issue of this project is to build a tool designed to automatically annotate a text using the vocabulary of an existing thesaurus specialized in ancient zoology (built with opentheso). The vocabulary is around 1000 words rich, not all being relevant to the targeted text.

MSc project—B+C (related to Evolutext)

PROPOSED STUDY : Automatized semantic annotation using a thematic zoology-related thesaurus

Aim: Designing a tool to annotate text in connection with an existing thesaurus

1. **Investigate state-of-the-art NLP methods and supervised learning algorithms and libraries for basic categorization of text segments**
2. **Improve knowledge extraction from texts by implementing semantic annotation of text segments with zoonyms based on the Thezoo thesaurus (zoonyms and few general zoological sub-domains based on their representation by sets of semantically related terms)**

The issue of this project is to build a tool designed to automatically annotate a text using the vocabulary of an existing thesaurus specialized in ancient zoology (built with opentheso). The object of the study is an English text, which is a translation of an ancient zoological Greek text. The thesaurus' vocabulary is around 1000 words rich, not all being relevant to the targeted text.

Many annotators (and web annotators) already exist, but with complex environment or limited reference dictionaries. The tool must allow tagging with new labels that will be automatically added to the thesaurus. This interaction (go-to-and-fro) from thesaurus to text will be an added value to existing tools.

Contacts : isabelle.thery@cepam.cnrs.fr , arnaud.zucker@cepam.cnrs.fr