

Project 2: Ontology Learning

This project is about mapping given named entities to a small tourism ontology using textual tourism data.

You have learnt in your first exercise how to use parsers to identify some basic semantic information from sentences.

Here, you have to do the best possible source(names entities) to target (ontology hierarchy) mapping using a textual corpus and text mining, NLP approaches. You are given a small ontology ([taxonomy21.jpg](#)) and some named entities ([t21.nelist](#)) to resolve to this ontology. This task was a sub-task in the PASCAL Ontology Learning Challenge (2006) (<http://olc.ijs.si/>) (Task 2.1: Ontology Population)

Background:

Ontologies are formal, explicit specifications of shared conceptualizations, representing concepts and their relations that are relevant to a given domain of discourse. Currently, ontologies are mostly developed as well as used through a manual process, which is very subjective and may cause major barriers to their large-scale use in such areas as Knowledge Discovery and Semantic Web. As human language is a primary mode of knowledge transfer, linguistic and statistical analysis of relevant documents for this purpose seems a viable option. More precisely, automation of ontology construction (ontology learning) and use (ontology population through knowledge markup) can be implemented by a combined use of linguistic analysis and machine learning approaches for text mining.

Task: Ontology population

You are provided with a list of named entities to be assigned to concepts stemming from a tourism taxonomy (Lonely Planet Dataset). In particular, for this purpose a subset of 96 concepts of the ontology — consisting of 681 concepts overall — will be used. The corresponding corpus from which these named entities originate will also be provided in a linguistically preprocessed form. The corpus consists of 1801 descriptions of destinations from LonelyPlanet. You are asked to generate assignments for the named-entities to one out of 96 concepts. The results should be provided in a file where each line contains a named-entity identifier followed by the concept identifier it is related to.

The following data is to be used for this task:

- 1) one file containing the structure of underlying concept hierarchy
- 2) one file containing the 1106 named entities to assign to their corresponding concepts
- 3) one file containing the 96 target concepts
- 4) the preprocessed corpus

The file with the target concepts is as follows:

country
river
hotel
region
mountain_range
...

And the list of instances to tag could look as follows:

France
Germany
Rhein
Seine
Alps
...

As result of the task we will expect a file as follows:

France country
Germany country
Rhein river
Seine river
Alps mountain_range
...

Data Sets (<http://olc.ijs.si/lpReadme.html>)

t21.tax (<http://olc.ijs.si/lpTxt/t21.tax>)

Taxonomy definition for the task T2.1. The file contains 103 lines in the form of is_a(A,B)

which means A is a type of B. All concepts are derived from the “root” concept and are formed using lower-case ASCII letters and the character '_'.

t21.nelist (<http://olc.ijs.si/lpTxt/t21.nelist>)

Contains the list of 1106 named entities to be used for task T2.1, one per line.

t21.colist (<http://olc.ijs.si/lpTxt/t21.colist>)

List of 96 target concepts for task T2.1, one per line.

flatcorpus.txt (<http://olc.ijs.si/lpTxt/flatcorpus.txt>)

Contains the entire corpus of 1801 documents, one per line, starting with “filename: ” and followed by a clear-text version of the document without newlines.

corpus.zip (<http://olc.ijs.si/lpTxt/corpus.zip>)

An archive of the text-only version of the corpus, one file per document.

taxonomy21.jpg (<http://olc.ijs.si/lpTxt/taxonomy21.jpg>)

An informative picture of the taxonomy used in task T2.1.

Evaluation

You have to generate the required output file to be automatically evaluated from here:

<http://olc.ijs.si/eval.html>

Suggestions:

1. OpenNLP has a sentence detector, tokenizer and a named entity recognizer (besides the shallow parser that some of you used in your first exercise.) The setup is the same as the shallow parser and the api is neat.
2. <http://secondstring.sourceforge.net/> - A suite of very efficient string matching tools. You may be interested in looking at this toolkit for matching your found out named entities with the given named entities here.
3. **Extra Credit (20 points):** You may want to export the tourism ontology to Protege ontology editor (this will help you to reason over different ontology concepts), use one of their text corpus linguistic annotation plugins and write wrapper classes to populate instances automatically.
4. **Extra Credit (15 points):** The named entities may match to more than one concept in the ontology. You should have a statistical learning component to take care of these cases (based on corpus statistics)

PLEASE NOTE: START EARLY! You may collaborate to figure out how to use these systems to process data. Your actual solution should be independently motivated. Please reach the TA (gtalk: sauravsahay) in case you have questions.

Due date: Sunday, October 5th.

Deliverables:

1. Diagrams of your problem solving methodology, and pseudo-code of your approach. (25 points)
2. Evaluation Results (20 points, extra credit for very good results :-)
3. A analysis of your results and how you could improve the system. (25 points)
4. Code (30 points)