Finding Short Definitions of Terms on Web Pages

Gerasimos Lampouras* and Ion Androutsopoulos**

* Department of Informatics, Athens University of Economics and Business, Greece
** Digital Curation Unit, Research Centre “Athena”, Athens, Greece

Why is this useful?

• Definition questions are very frequent in Web searches.
• Typical Question Answering systems have trouble with definition questions, because the answers are not named entities, and definitions can be phrased in many different ways.
• On-line encyclopedias and glossaries do not contain definitions for less known persons, products etc.

Training the classifier

• When training the system's classifier, we use terms for which many definitions exist in on-line encyclopedias.
• Web snippets for a term that are very similar to the corresponding encyclopedia definitions are taken to be positive training examples.
• Web snippets for a term that are very different from the corresponding encyclopedia definitions are taken to be positive training examples.
• Medium-similarity training snippets are discarded.
• Once the classifier has been trained, it can be used to classify snippets for which no encyclopedia definitions exist.

System Overview

1. Term to be defined
2. Extraction of candidate short definitions from Web pages
3. Ranking of candidate definitions with a classifier

Why is this useful?

• This system can be used as an add-on to search engines, to find short definitions (or lists of them) when no definitions are found in known encyclopedias and glossaries.
• The system does not use any named-entity recognizers, POS taggers, chunkers, parsers etc. It can be easily retrained for other languages.
• In INDIGO, users interact with robotic museum guides that generate texts from ontologies. The system we present can be used to answer definition questions, when the answers cannot be found in the ontologies.

Text Snippets

We extract text snippets (250 character-long windows around the term) from the Web pages returned by a search engine for the term.

We use a Maximum Entropy classifier, but the training examples (positive and negative) are generated automatically from encyclopedia definitions and Web snippets. In effect, training is unsupervised.

Training term

galaxy

(…)

(…)

(…)

(…)

(…)

A large aggregation of stars, bound together by gravity. There are three major classifications of galaxies—spiral, elliptical, and irregular.

A very large cluster of stars (tens of millions to trillions of stars) gravitationally bound together.

An organized system of many hundreds of millions of stars, often mixed with gas and dust. The universe contains billions of galaxies.

Evaluation results

50 250 500 1000 1500
52
50
50
51
48
41
Training terms

% of correctly answered test questions, 1 snippet allowed

% of correctly answered test questions, 5 snippets allowed

MRR

0.55
0.62
0.65
0.64
0.66

Mean Reciprocal Rank (MRR) calculated on the 5 snippets.

Allowing 1 snippet to be returned per test term. If it is an acceptable definition, we count the question as correctly answered.

Allowing 5 snippets to be returned per test term. If any of the five is an acceptable definition, we count the question as correctly answered.

The system is freely available from:
http://nlp.cs.aueb.gr/