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

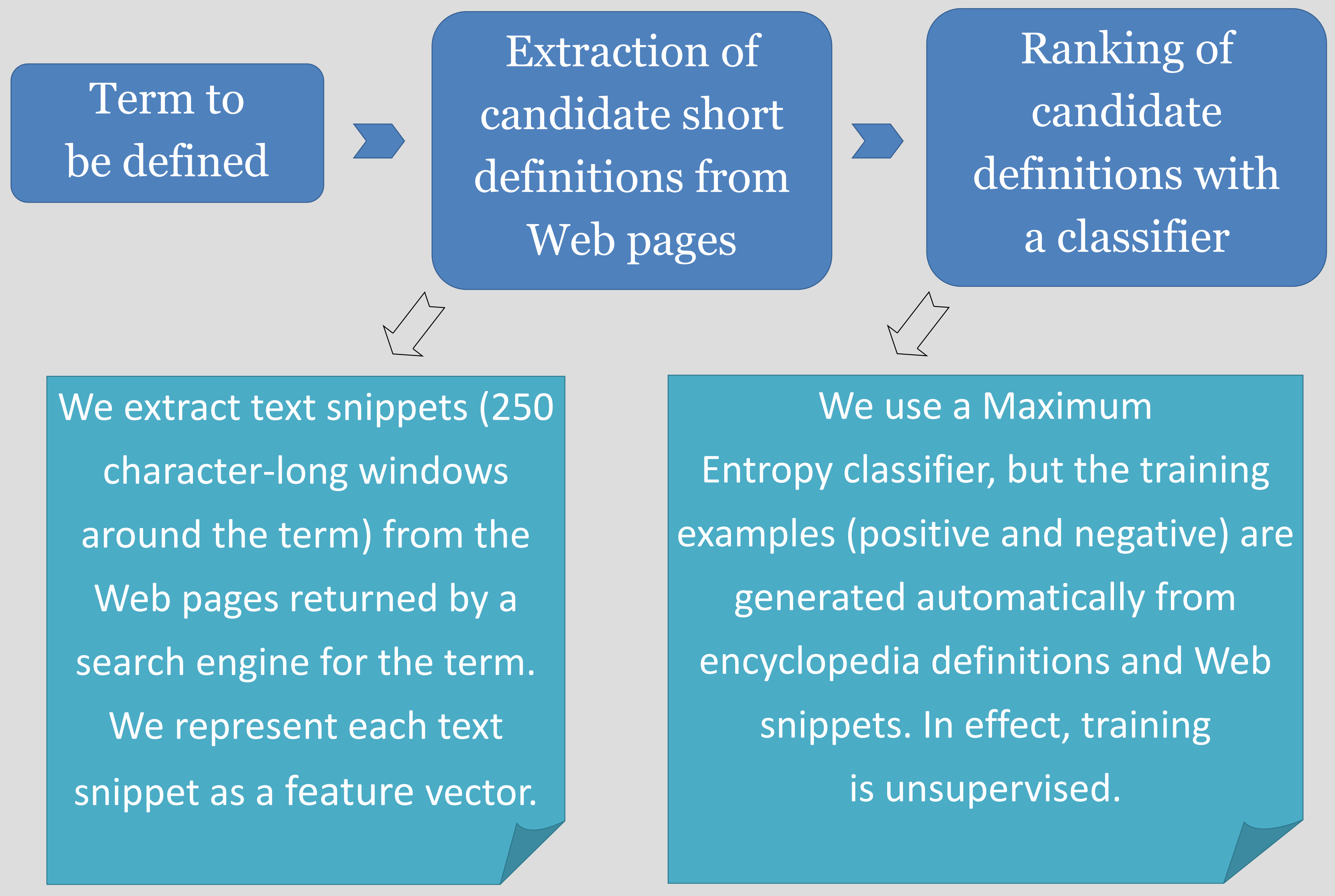
Why is this useful?

- 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



Term to be defined

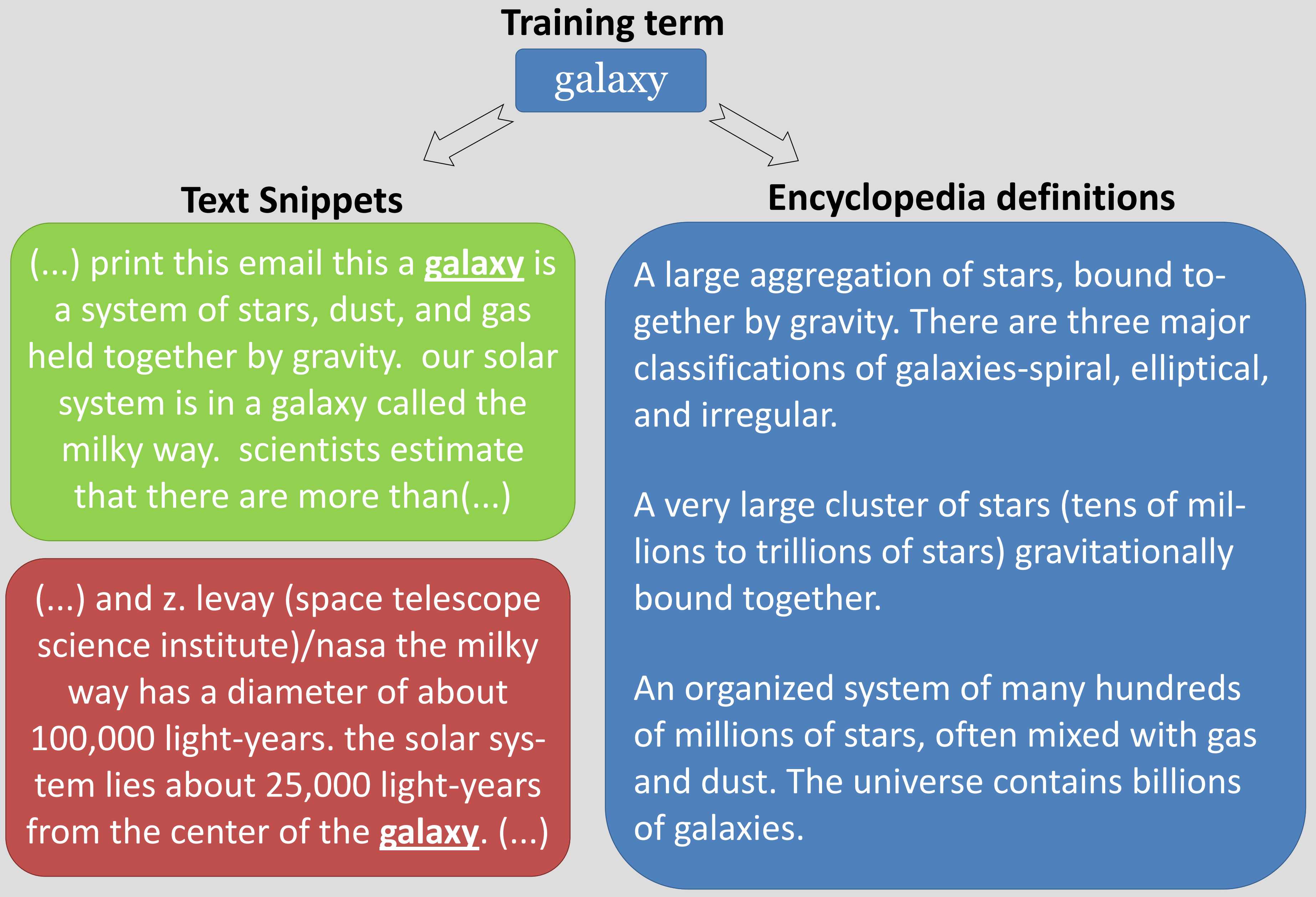
Extraction of candidate short definitions from Web pages

Ranking of candidate definitions with a classifier

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

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 example



Training term

galaxy

Text Snippets

(...) print this email this a **galaxy** is a system of stars, dust, and gas held together by gravity. our solar system is in a galaxy called the milky way. scientists estimate that there are more than(...)

(...) and z. levay (space telescope science institute)/nasa the milky way has a diameter of about 100,000 light-years. the solar system lies about 25,000 light-years from the center of the **galaxy**. (...)

Encyclopedia definitions

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.

Similarity measures (best results obtained with ROUGE) show if a snippet is very **similar** (**positive training example**) or very **dissimilar** (**negative training example**) to the encyclopedia definitions.

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.

Representing text snippets as vectors

- 22 manually-constructed features
- 300 automatically selected Boolean features.

  They show whether a particular word n-gram (n = 1, 2, 3) precedes or follows the term in the snippet.

  The n-grams are extracted from all the positive training snippets. We keep the 300 n-grams with the highest precision scores (the most reliable indicators) that exceed a frequency threshold.

(...) **ethanol** , also called (...)

(...) what is a ***galaxy*** (...)

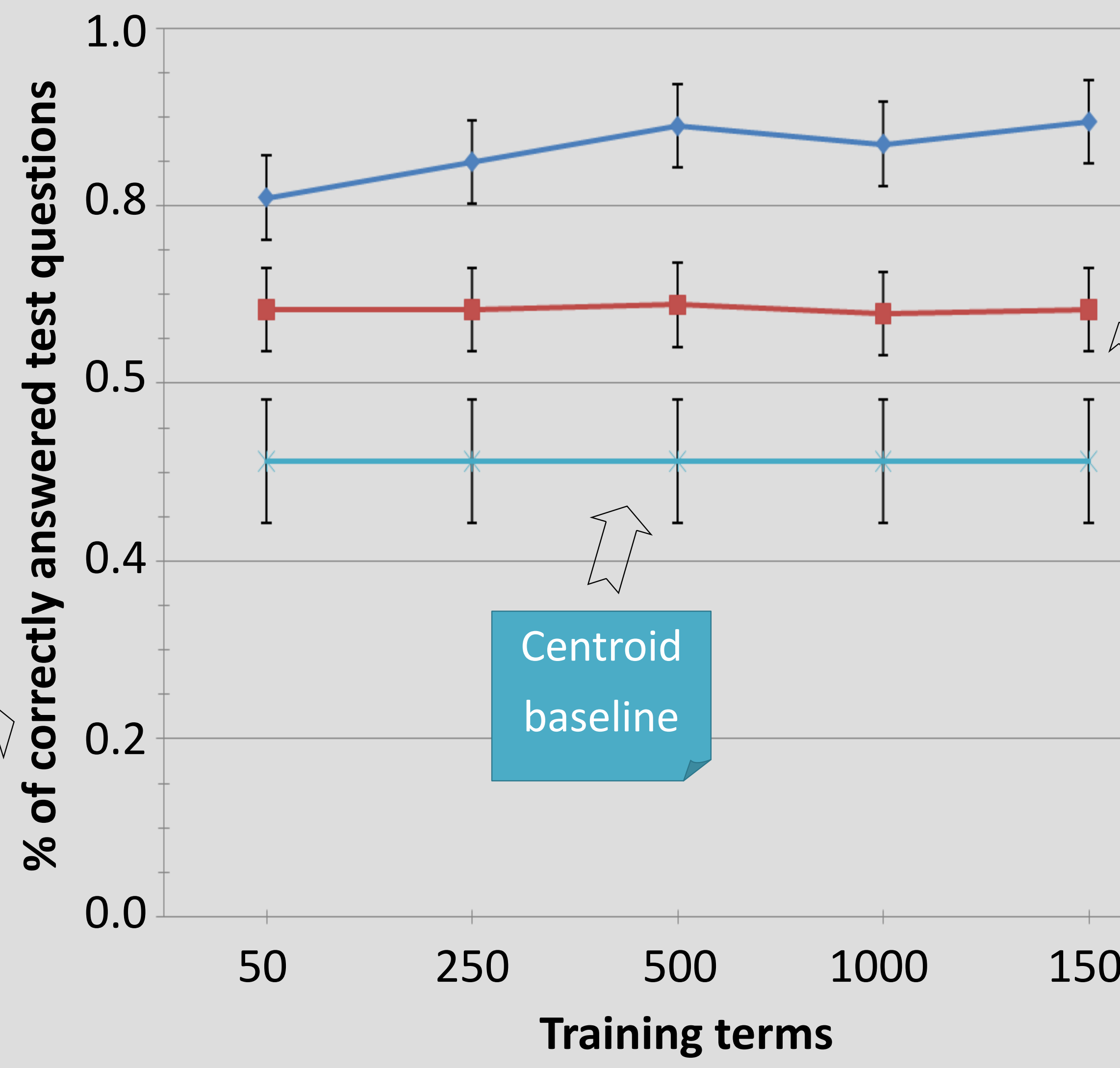
Evaluation results

50	250	500	1000	1500	Training terms
41	48	51	50	52	% of correctly answered test questions, 1 snippet allowed
81	85	89	87	90	% of correctly answered test questions, 5 snippets allowed
0.55	0.62	0.65	0.64	0.66	MRR

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.



Our system

Cui, H., Kan, M., and Chua, T. 2007. “Soft pattern matching models for definitional question answering”. *ACM Transactions on Information Systems*, 25(2), 1–30.

Athens University of Economics and Business

DEPARTMENT OF INFORMATICS

The system is freely available from:

http://nlp.cs.aueb.gr/

INDIGO Interaction with Personality and Dialogue Enabled Robots