Finding short definitions of terms on Web pages

G. Lampouras, D. Galanis, I. Androutsopoulos

Natural Language Processing Group
Department of Informatics
Athens University of Economics and Business
http://nlp.cs.aueb.gr/

Finding definitions on the Web

- **QA systems**: find answers to natural language questions by examining documents (e.g., Web pages).
- We focus on **definition questions**. Very frequent and difficult. They cannot be answered by looking for particular types of named entities in sentences similar to the questions.
- Goal: **find on the Web short definitions** of terms **not covered by encyclopedias** and/or glossaries.
- E.g., “Who was Pythagoras?”

Snippets from Web pages:

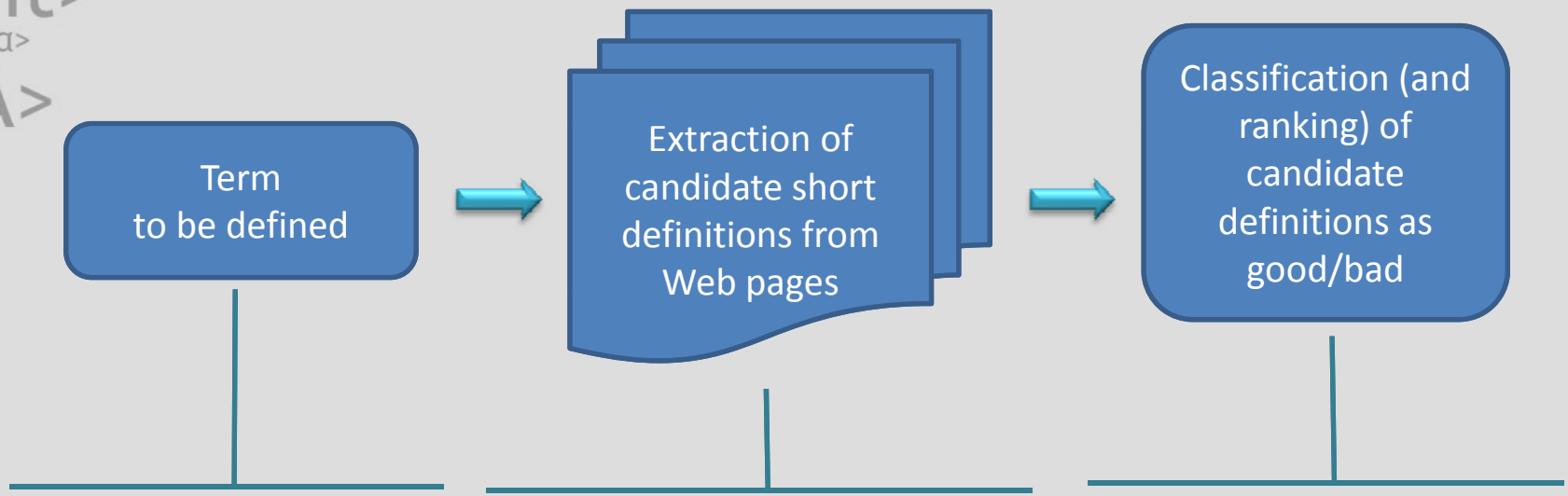
(…) not much is known about **Pythagoras**, other than that he was a mathematician and philosopher who founded a community in southern Italy (…)

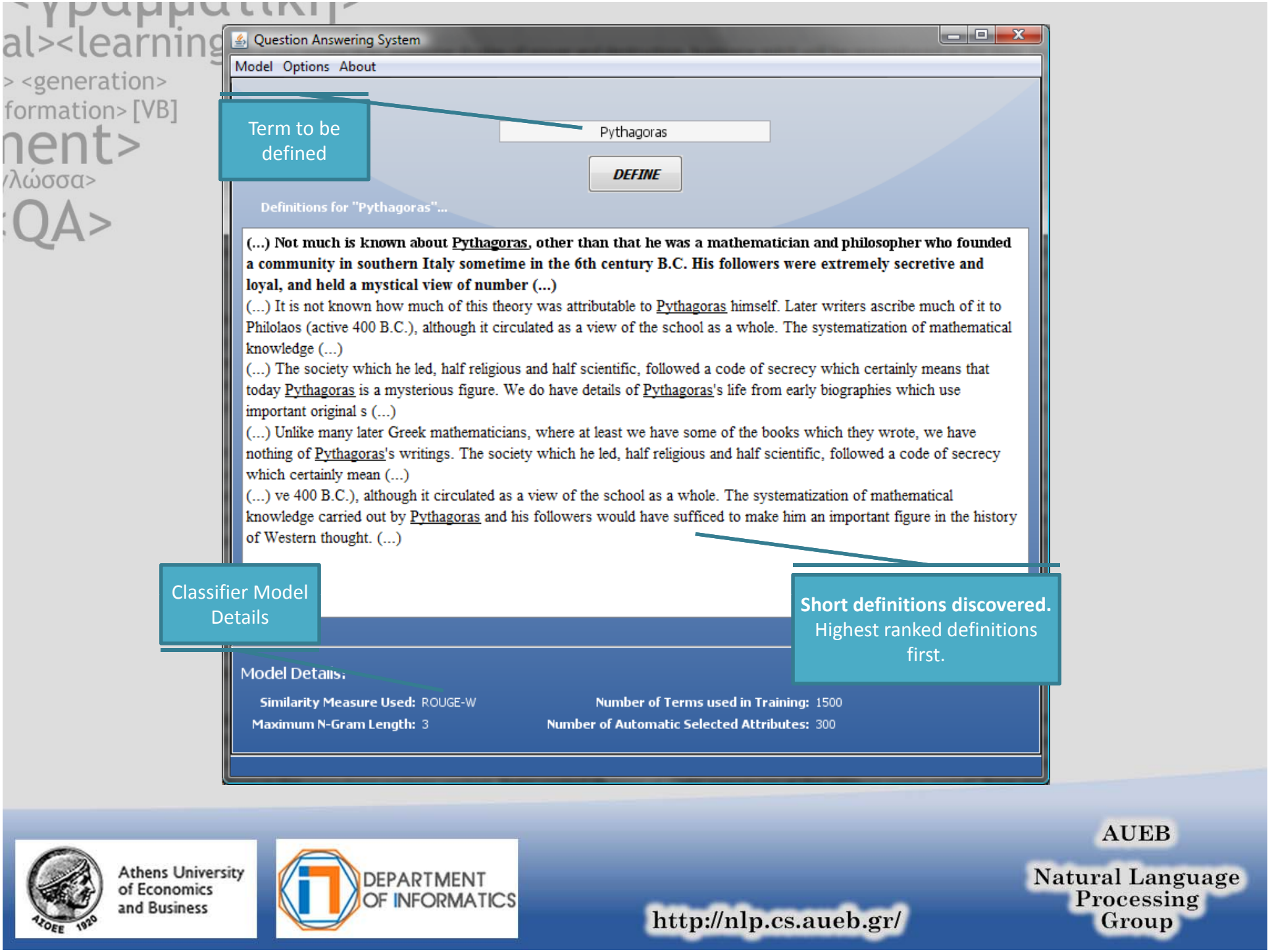
(…) it is not known how much of this theory was attributable to **Pythagoras** himself. Later writers ascribe much of it to Philolaos (active 400 B.C.), although it circulated (…)

(…) the society which he led, half religious and half scientific, followed a code of secrecy which certainly means that today **Pythagoras** is a mysterious figure (…)

(…) unlike many later Greek mathematicians, where at least we have some of the books which they wrote, we have nothing of **Pythagora's** writings (…)

System Overview



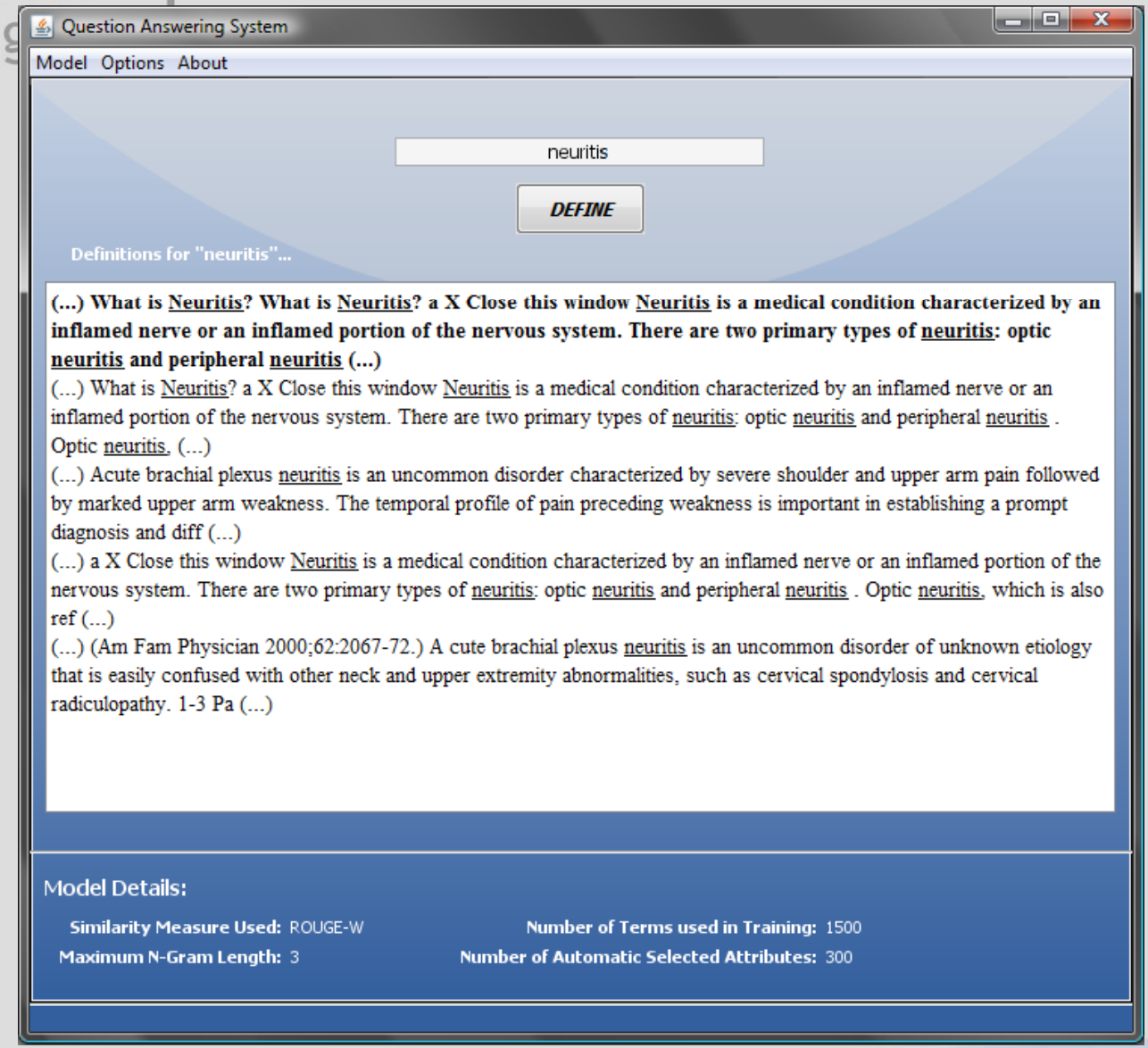


Athens University of Economics and Business

DEPARTMENT OF INFORMATICS

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

AUEB Natural Language Processing Group

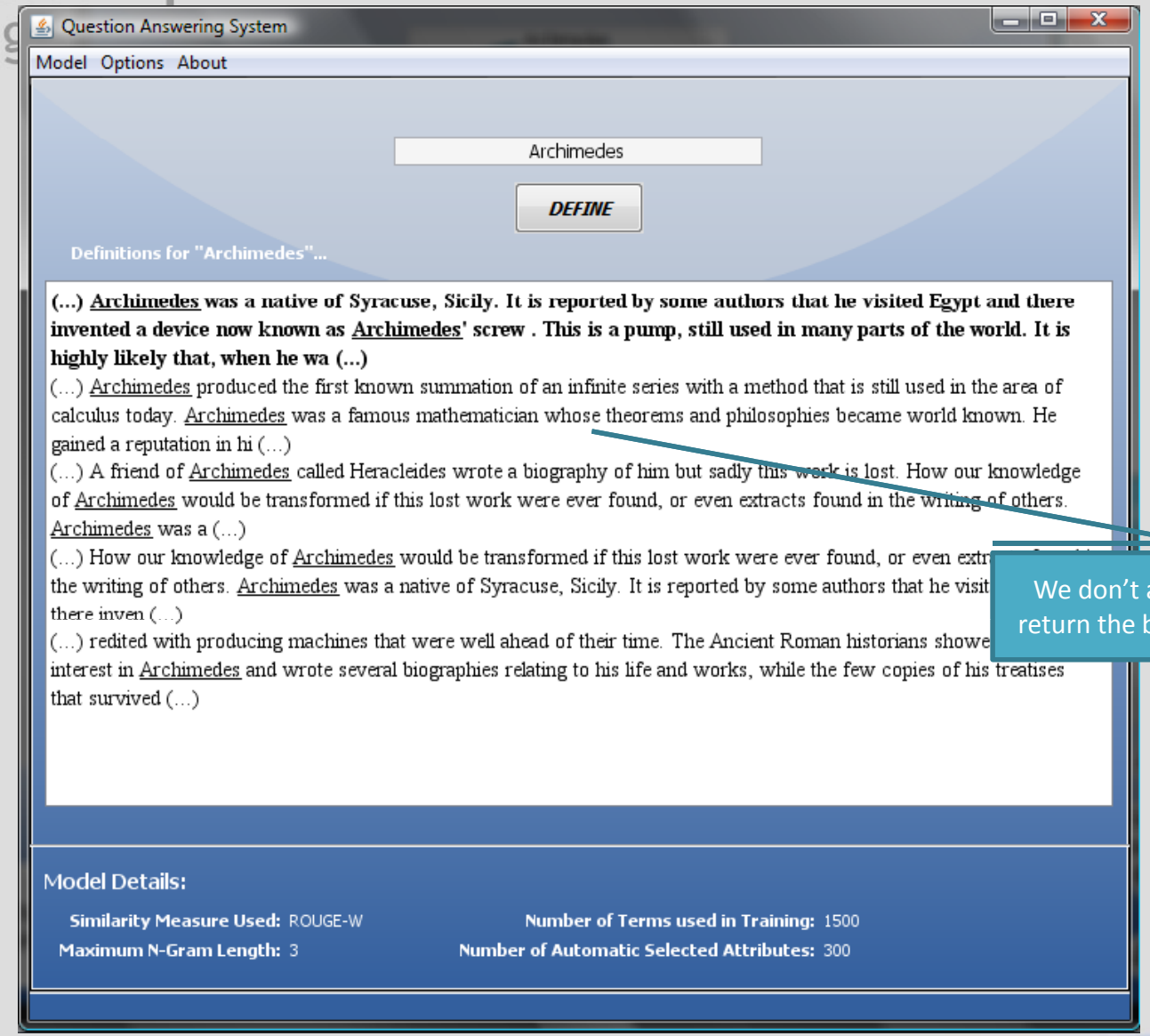


Athens University of Economics and Business

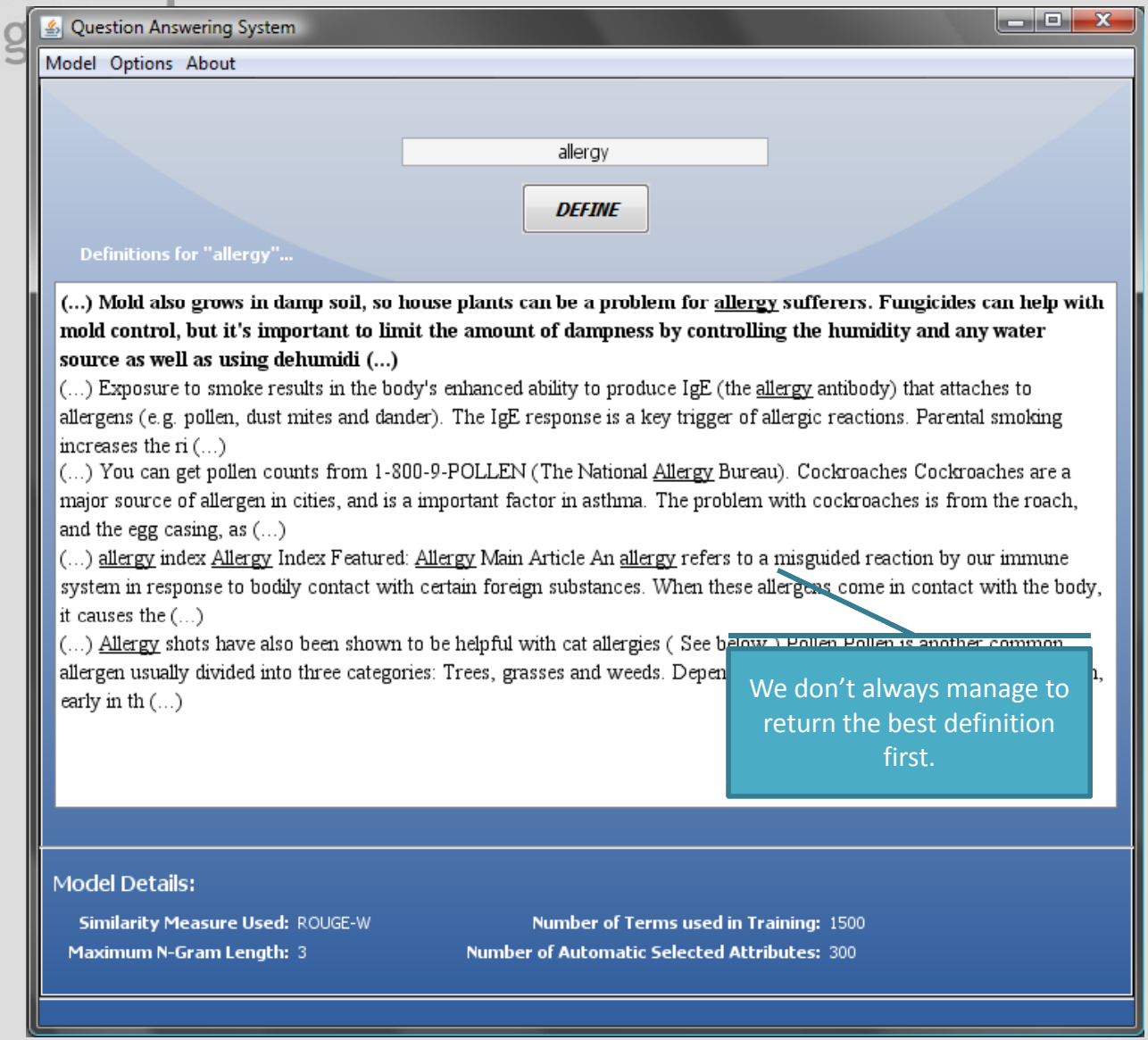
DEPARTMENT OF INFORMATICS

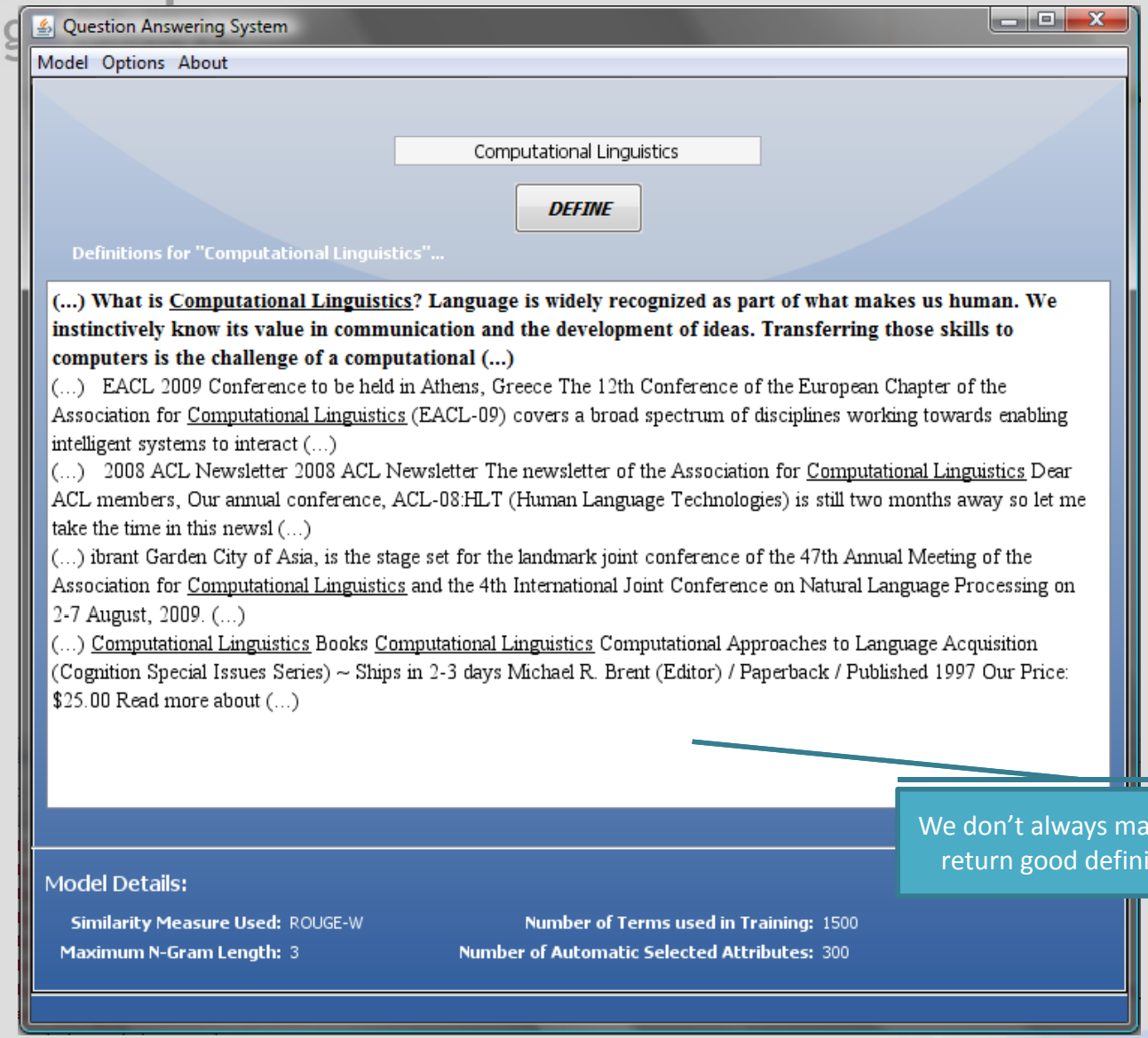
http://nlp.cs.aueb.gr/

AUEB Natural Language Processing Group



We don't always manage to return the best definition first.


Question Answering System
Model Options About
allergy
DEFINE
Definitions for "allergy"...
(...) Mold also grows in damp soil, so house plants can be a problem for allergy sufferers. Fungicides can help with mold control, but it's important to limit the amount of dampness by controlling the humidity and any water source as well as using dehumidi (...)
(...) Exposure to smoke results in the body's enhanced ability to produce IgE (the allergy antibody) that attaches to allergens (e.g. pollen, dust mites and dander). The IgE response is a key trigger of allergic reactions. Parental smoking increases the ri (...)
(...) You can get pollen counts from 1-800-9-POLLEN (The National Allergy Bureau). Cockroaches Cockroaches are a major source of allergen in cities, and is a important factor in asthma. The problem with cockroaches is from the roach, and the egg casing, as (...)
(...) allergy index Allergy Index Featured: Allergy Main Article An allergy refers to a misguided reaction by our immune system in response to bodily contact with certain foreign substances. When these allergens come in contact with the body, it causes the (...)
(...) Allergy shots have also been shown to be helpful with cat allergies (See below) Pollen Pollen is another common allergen usually divided into three categories: Trees, grasses and weeds. Depen
early in th (...)
We don't always manage to return the best definition first.
Model Details:
Similarity Measure Used: ROUGE-W
Number of Terms used in Training: 1500
Maximum N-Gram Length: 3
Number of Automatic Selected Attributes: 300



We don't always manage to return good definitions.

Athens University of Economics and Business

DEPARTMENT OF INFORMATICS

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

AUEB Natural Language Processing Group

Why is this useful?

- Definition questions are **very frequent** (e.g., in logs of search engines).
- **Typical QA 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**.
- Our 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**.

Training the classifier

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

Training the classifier

Training Term

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.

a component of our Universe made up of gas and a large number (usually more than a million) of stars held together by gravity.

A large grouping of stars. Galaxies are found in a variety of sizes and shapes. Our own Milky Way galaxy is spiral in shape and contains several billion stars. Some galaxies are so distant that their light takes millions of years to reach the Earth.

Training the classifier

Training Term

galaxy

Web snippet

(...) 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(...)

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.

a component of our Universe made up of gas and a large number (usually more than a million) of stars held together by gravity.

A large grouping of stars. Galaxies are found in a variety of sizes and shapes. Our own Milky Way galaxy is spiral in shape and contains several billion stars. Some galaxies are so distant that their light takes millions of years to reach the Earth.

A positive training example

Training Term

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.

a component of our Universe made up of gas and a large number (usually more than a million) of stars held together by gravity.

A large grouping of stars. Galaxies are found in a variety of sizes and shapes. Our own Milky Way galaxy is spiral in shape and contains several billion stars. Some galaxies are so distant that their light takes millions of years to reach the Earth.

Web snippet

(...) 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(...)

Similarity measures (Rouge etc.) show the snippet is **very similar** to the encyclopedia definitions.

A negative training example

Training Term

galaxy

Web snippet

(...) (lowell observatory) 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.

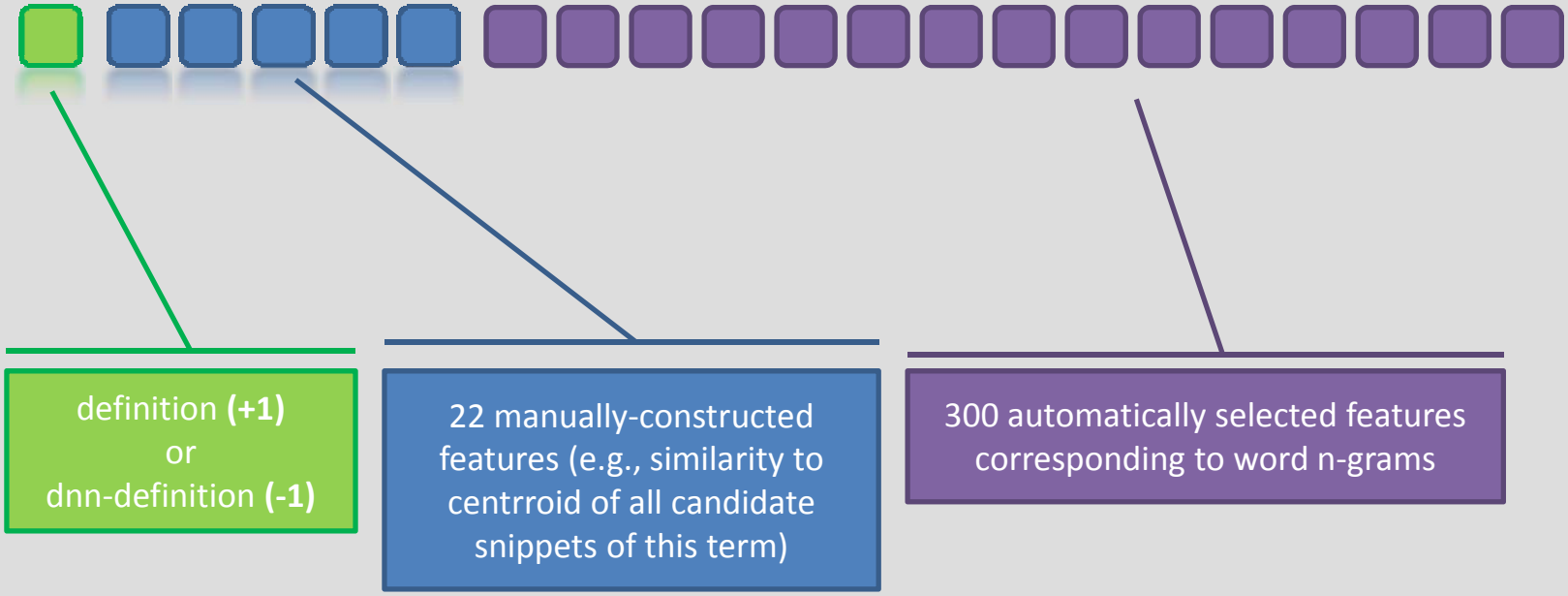
an organized system of many hundreds of millions of stars, often mixed with gas and dust. The universe contains billions of galaxies.

a component of our Universe made up of gas and a large number (usually more than a million) of stars held together by gravity.

A large grouping of stars. Galaxies are found in a variety of sizes and shapes. Our own Milky Way galaxy is spiral in shape and contains several billion stars. Some galaxies are so distant that their light takes millions of years to reach the Earth.

Similarity measures (Rouge etc.) show the snippet is **very different** than the encyclopedia definitions.

Representing text snippets as vectors



Automatically selected features

300 automatically selected numeric features.

Each shows 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.

The value of the feature is the ROUGE-W score between a pattern and the left or right context of the term.

(...) email this *a* ***galaxy*** *is a system* (...)
(...) **ethanol** , also called (...)

- are found
- is a
- history of
- this a
- also called
- also

Evaluation results

Snippets from encyclopedias/glossaries ignored during testing.

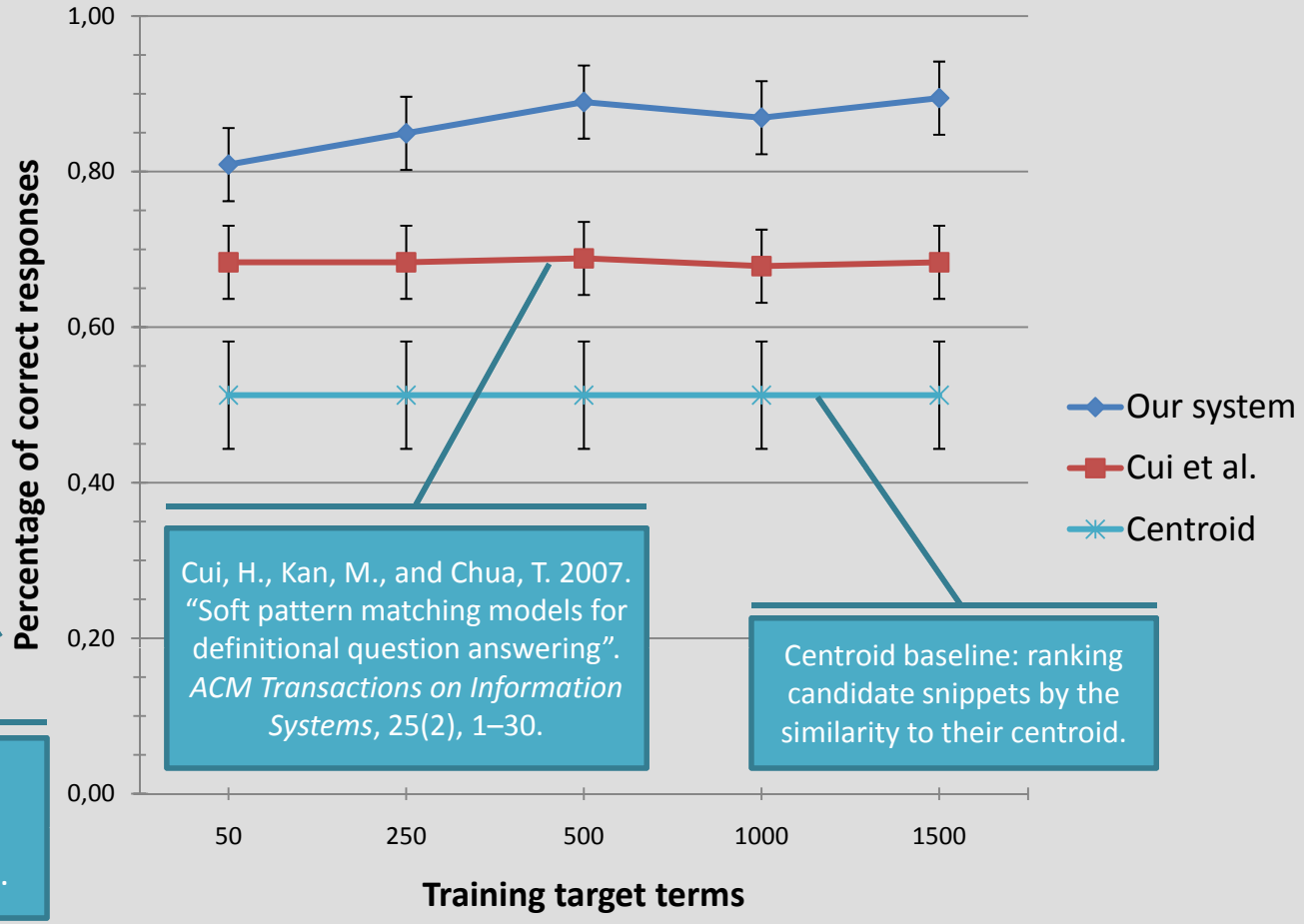
Allowing 5 snippets to be returned per term. If any of the five are acceptable (by human judges), we count the definition question as correctly answered.

Allowing 1 snippet to be returned per term. If it is acceptable (by human judges), we count the definition question as correctly answered.

	50	250	500	1000	1500
1 snippet per term	0,41	0,48	0,51	0,50	0,52
5 snippet per term	0,81	0,85	0,89	0,87	0,90
MRR	0,55	0,62	0,65	0,64	0,66

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

Comparing to other systems



Allowing again 5 snippets to be returned per term.

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.

Centroid baseline: ranking candidate snippets by the similarity to their centroid.

For more information

- G. Lampouras and I. Androutsopoulos, “Finding Short Definitions of Terms on Web Pages”. Proceedings of the *2009 Conference on Empirical Methods on Natural Language Processing* (EMNLP 2009 at ACL/IJCNLP 2009), Suntec, Singapore, 2009.
- G. Lampouras, "Methods to Automatically Detect Definitions in Document Collections", MSc thesis, Department of Informatics, Athens University of Economics and Business, 2008 (in Greek).
- I. Androutsopoulos and D. Galanis, “A Practically Unsupervised Learning Method to Identify Single-Snippet Answers to Definition Questions on the Web”. Proceedings of *HLT/EMNLP 2005*, Vancouver, Canada, pp. 323-330.
- S. Miliaraki and I. Androutsopoulos, “Learning to Identify Single-Snippet Answers to Definition Questions”. Proceedings of *COLING 2004*, Geneva, Switzerland, pp. 1360-1366.