Extraction of Definitions for Bulgarian

Hristo Tanev
Institute for the Protection and the Security of the Citizen
Joint Research Center
via E.Fermi 1
21020 Ispra, Italy
Hristo.Tanev@jrc.it

Abstract

We participated at the Monolingual Bulgarian QA task at CLEF-2006 with a definition extraction system based on linguistic templates and keywords. Our system uses a partial syntactic parser for Bulgarian to detect noun phrases as candidates for definitions. Our system answered correctly to 28% of the definition questions.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids

1 Introduction

This year we participated at the Monolingual Bulgarian QA task with a system which answers definition questions. Our work was inspired by the online Bulgarian QA system "Socrates" [1].

We think that automatic extraction of definitions is important for several reasons:

First, albeit the number of online encyclopaedic resources in English increases in quality and range (for example Wikipedia (http://www.wikipedia.org) provides over 1 million English articles), for many languages like Bulgarian the quantity and quality of such resources are not sufficient. As a result, no encyclopaedic entries can be found for many topics on the Bulgarian Web. For example, question number 9 from the Bulgarian QA test set of CLEF 2006 is "Kakvo e OneNote?" ("What is OneNote?"). The Bulgarian version of Wikipedia provides no article for OneNote (though the English version does). If we search for OneNote with Google (http://www.google.bg) in the Bulgarian-language pages we can hardly find good descriptions of OneNote. On the other hand, if we search for definitions on the Bulgarian Web using the automatic definition extraction service of "Socrates" (http://tanev.dir.bg/Socrat.htm), we find that OneNote is an application which has functions of a notebook and following the link returned we can see a relevant description of OneNote.

Second, the encyclopaedic resources usually give high-quality well-structured descriptions of a term, but more information can be captured by scanning free texts. Such information can be

more subjective, controversial, or incomplete with respect to the encyclopaedias, but nevertheless it can be useful. For example for question 126 "Kakvo e Evrovizia?" ("What is Eurovision?") the Bulgarian version of Wikipedia returns a short definition and a list of winners; on the other hand, one of the definitions returned by "Socrates" on-line definition extraction is that "Evrovizia e nay-golemiat skandal na godinata" ("Eurovision is the biggest scandal of the year"). Following the link returned we can get interesting information considering a scandal around the Bulgarian participation in this song contest.

Third, if a definition pattern like "*TERM is DEFINITION*" is present in a document, even if the definition extracted is not informative enough, the pattern itself means that *TERM* has an important role in the article. In this way, identification of definition templates can be used to rank better the results from a search engine.

Fourth, but not last in importance is the fact that automatic definition extraction can help to the people who build dictionaries and encyclopaedic resources like Wikipedia by providing them with relevant textual fragments.

Our definition extraction system uses linguistic templates and clues similar to the ones described in [1] and [2].

In this paper we will give an overview of the linguistic templates and rules used by our system, as well as our participation at CLEF 2006.

2 Definition Extraction Patterns

Definition questions ask for a definition of a person or a term (e.g. "Who is Galileo?" - answer:"Italian astronomer"). Techniques which rely on Named Entity recognition are not useful for this type of questions. On the other hand, templates provide a reliable instrument for definition extraction. For example, the approach described in [3] used only superficial patterns of the type: "*a TERM is DEFINITION*", "*TERM, DEFINITION*". However, such approaches are error prone, since similar patterns can be encountered in non definition contexts. For example, "The charge of a positron is about..." is not a definition of the positron, though the pattern "a positron is" is present as a substring. We use linguistic constraints and rules to avoid or mitigate the effect of similar errors.

For each definition question, our system first tries to match one of its templates on the linguistically pre-processed text. We used the LINGUA language engine [4] to perform text segmentation, part-of-speech tagging and parsing.

The phrases which match one of these patterns are considered candidate definitions. For every candidate a set of linguistic constraints are applied. For example, if the template "*TERM is DEFINITION*" is found in the text, *DEFINITION* should be parsed by the parser as a noun phrase and has to agree by gender and number with *TERM*.

3 Linguistic and Lexical Clues

Our experiments demonstrated that for high-quality definition extraction it is not enough to capture a fragment which matches certain patterns. It is necessary also to analyze the content of the phrase and its context. Each phrase - a candidate for a definition is evaluated using a set of evaluation rules which consider its syntactic context and lexical content.

Here we are going to give some examples for evaluation rules:

If a phrase matches a pattern like "*TERM is DEFINITION*" or "*TERM, DEFINITION*" we give lower weight to the matches where this pattern is preceded by a preposition: "*Prep TERM is DEFINITION*" or "*Prep TERM, DEFINITION*". In most such constructions the definition does not refer to *TERM* but to another phrase which contains it.

If a phrase matches a pattern like "*TERM is DEFINITION*", we give lower weight to the matches where the *TERM* is part of a bigger noun phrase like in "svobodniat elektron e valna"

("the free electron is a wave"). In this case the definition refers to another term ("the free electron") rather than to "electron" itself.

If we have the pattern "*TERM, DEFINITION*", but it is a part of a comma separated list, then the candidate for definition is most probably not a definition. Therefore its weight is decreased.

If a candidate definition for a person contains one of the keywords designating occupation, social role, or other words used for famous people, like "shampion" ("champion"), "golemiat"("the great"), etc., higher weight is given to this definition.

Longer definitions obtain higher weight.

After the application of all the rules, each candidate definition phrase obtains a weight; phrases are sorted according to this weight and the best one is chosen.

4 Experiments and Future Directions

We participated in the Bulgarian Monolingual QA task at CLEF 2006. We run our system only on the definition questions. The accuracy we achieved on this question subset was moderate - about 28%.

There is a lot of space for improvement in our definition extraction system: First of all, we may enlarge the lexicon with "interesting" words when evaluating definitions of people. We may learn automatically syntactic and lexical clues for the definitions context and structure. The Bulgarian section of Wikipedia can be used as a training corpus. Finally, we may estimate the informativeness of a definition by considering the Inverse Document Frequency of its words.

Our definition extraction system may have a broad range of applications, especially in the context of Internet. It may be used to build profiles of people and oragnizations and extract relations between them, to classify automatically terms, to populate ontologies, etc.

References

[1] Tanev, H. "Socrates - a Question Answering prototype for Bulgarian" In RANLP-2003 Proceedings, Borovets - Bulgaria, September, 2003

[2] Tanev, H., Kouylekov M., Negri M., Coppola B., and Magnini B. "Multilingual Pattern Libraries for Question Answering: a Case Study for Definition Questions" In LREC 2004 Proceedings, Lisbon, Portugal, 2004

[3] Ravichandran D., Hovy E. "Learning Surface Text Patterns for a Question Answering System" In Proceedings of ACL 2002, Philadelphia, 2002

[4] Tanev, H. and Mitkov R. "Shallow Language Processing Architecture for Bulgarian" In Proceedings of COLING 2002, Taiwan, 2002