Multilingual Pattern Libraries for Question Answering: a Case Study for Definition Questions

Hristo Tanev[1], **Milen Kouylekov**[1,2], **Matteo Negri**[1], **Bonaventura Coppola**[1,2] **and Bernardo Magnini**[1]

[1] ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
Via Sommarive 18, 38050 Povo (TN), Italy
[2] Department of Information and Communication Technology, University of Trento
Via Sommarive 14, 38050 Povo (TN), Italy
{tanev | kouylekov | negri | coppolab | magnini}@itc.it

Abstract

In this paper we investigate the effectiveness of a novel resource for Multilingual Question Answering (QA). Such a resource consists of a set of multilingual pattern libraries for answer extraction and validation. In the spirit of the ongoing attempts to develop freely available resources for QA, we argue that the distribution and use of pattern libraries will contribute to make Multilingual QA a more feasible task.

1. Introduction

Current approaches to QA still fail to provide efficient and flexible strategies for managing different languages simultaneously. Moreover, most of the QA systems participating in the TREC competition frequently show complex architectures relying on tools and resources (*e.g.* syntactic parsers, WordNet, etc.) which are difficult to be adapted or ported to other languages. At the same time, multilinguality is gradually becoming a crucial issue with the increasing amount of multilingual on-line information.

In the spirit of the ongoing attempts to develop freely available resources for QA (*e.g.* Webclopedia[1]), this paper discusses the viability of a pattern-based approach as a more suitable alternative to address multilinguality, and presents a multilingual pattern library for definition questions as a novel resource for QA systems' developers.

The general effectiveness of patterns for answer extraction has been firstly shown in (Soubbotin 2001), which describes a QA system exploiting a collection of manually created surface text patterns to mine answers from the TREC target collection. A similar approach had also been adopted in (Joho & Sanderson 2000), which describes hand-crafted string level patterns for the extraction of definitions from text corpora. More recently, (Ravichandran & Hovy 2002) described an algorithm for the automatic acquisition of surface patterns, showing that even a limited number of simple surface patterns provides a viable solution for well defined question type categories. As an example, a typical high-precision pattern for questions about birth dates is:

```
<NAME> (<ANSWER>-
```

which matches specific text segments such as:

```
Napoleon (1769-
```

While surface level patterns described in (Ravichandran & Hovy 2002) show a certain lack of expressiveness, which is due to their weak generalization power, in this study we exploit syntactic and semantic features in order to broaden their coverage. In this framework, we introduce the notion of *linguistic regular expressions* as a novel kind of pattern structures which combine the expressive power of regular expressions and the abstraction provided by motivated linguistic generalizations. Moreover, we look at them as a first step towards the construction of multilingual QA systems based on pattern libraries as a key linguistic resource.

In order to extend string-level patterns to a multi-language environment, we will go through with two main steps. First, the initial surface pattern formalism (*i.e.* bare string patterns) is enhanced in order to improve expressiveness and to allow simple abstraction with respect to language-dependent features. For this purpose, we look at regular expressions as a general framework enabling the introduction of syntactic chunks and semantic-typed entities as components of the proposed formalism. As a second step, issues related to the construction of multilingual libraries are discussed and a specific approach is proposed. In order to show its feasibility, we applied it on the specific class of *definition questions.* Therefore, a library of multilingual patterns has been manually built and tested simultaneously for English, Italian and Bulgarian, three languages belonging to different Indo-European language groups, namely Germanic, Romance and Slavonic.

The paper is structured as follows. The general idea motivating the use of patterns for handling definition questions, and the main issues related to the construction

[1] http://www.isi.edu/natural-language/projects/webclopedia

of a multilingual pattern library for this question class are presented in Sections 2 and 3. Section 4 shows in detail the specific implemented approach. Sections 5 and 6 present and discuss empirical evaluation and results obtained in experiments carried out to verify the effectiveness of the general framework.

2. General Issues in Handling Definition Questions

We started the development of multilingual pattern libraries focusing on the specific class of *definition questions* such as *"Who is Aaron Copland?*" and "*What is a quasar?*". Such question type was first implicitly introduced in the 2002 edition of TREC competition, and then defined as a proper question class in 2003, where it represented the 10% of the whole question set.

While definition questions are among the most natural and frequent kinds of queries posed by humans, they raise specific issues in the development of a QA system. First, document retrieval for definition questions is a non trivial task (since few keywords are usually available to narrow the search space) and leads to a huge number of heterogeneous documents. Next, in many cases definitions appear in documents whose main topic is not related to the focus of the question; for this reason, documents containing definitions are not always ranked on the top by search engines.

As an example, querying Google with the simple question: "*What is an atom?*" (question n. 896 in the Text Retrieval Conference, query submitted on April 17th, 2003), most of the 5,290,000 retrieved documents did not contain an answer to the question. The first sensible answer, barely caught by browsing a huge portal ranked 15th (http://particleadventure.org), was: "*Moreover, experiments which looked into an atom using particle probes indicated that atoms had structure and were not just squishy balls. These experiments helped scientists determine that atoms have a tiny but dense, positive nucleus and a cloud of negative electrons (*e^-*)*". Experienced Web users may be able to circum-navigate the problem by resorting to the trick of predicting the ways in which the correct answer could appear within a document. For instance, querying Google with the exact string "*the atom is*", most of the 18,900 documents retrieved contain an acceptable answer. Similar results can be achieved by querying the search engine with: "*an atom is*" (32,900 hits), or "*an atom is defined as*" (299 hits).

In light of these considerations, different kinds of patterns, corresponding to different abstraction levels, could be excogitated. For instance, beyond the bare string-level patterns already mentioned in Section 1, a more structured pattern for definition questions is:

<NP> ((″who″ | ″which″) ″is″) ″called″ <FOCUS>

which will match text portions such as:

"*a negative particle called electron*"

In this pattern, the **<NP>** placeholder stands for a general noun phrase and **<FOCUS>** is the entity for which a definition is sought (e.g. the word *"electron"* in *"What is an electron?"*).

3. A Multilingual Pattern Library for Definition Questions

The ultimate purpose of a multilingual library for QA is to provide a fast and accurate way to extract and rank candidate answers to questions posed in different languages.

The problem of creating multilingual pattern libraries has to be faced considering three main dimensions: *multilinguality*, linguistic *abstraction level* (*e.g.* barely surface, syntactic, lexico semantic), and methods exploited for *acquisition*.

The multilinguality issue is related not only to the number of languages to which the libraries apply, but also to the degree of alignment between multilingual patterns. This means that patterns stored in a library are supposed to be aligned in multilingual structures which embody and synthesize variability and similarities between the considered languages.

The level of linguistic abstraction of a pattern and the acquisition methods adopted are issues related to each other. For instance, if we choose syntactic patterns, we are likely to need complex algorithms for learning structures. On the other hand, as shown in (Ravichandran & Hovy 2002), surface patterns can be acquired with relatively simple algorithms. Moreover, the actual construction of libraries for QA should necessarily take into account the existing trade-off between precision and recall when using patterns expressed at different abstraction levels. Surface patterns are in fact easier to acquire, but lack of representation power and provide high precision at the cost of a quite low coverage. On the other hand, syntactic patterns provide higher coverage with a reduced precision degree. For instance, the definition pattern

<FOCUS>, <NP>

will cover definitions like *"electron, a negative particle"* but will return also incorrect matches like *"electron, photon, neutrino"*.

In order to develop and test the effectivenes of the linguistic regular expressions here proposed (see Section 4), we opted for the manual development of a prototype library: a discussion about the optimal acquisition method for such expressions is intentionally postponed.

4. Pattern Formalism and Library Implementation

The library consists of two parts: *extraction patterns* and *validation patterns*. Extraction patterns, whose purpose is to extract candidate definitions from the text have low precision but high coverage. On the other hand, validation patterns are more accurate, since they are intended to measure the relation between the focus and the candidate definitions (see Magnini et al. 2002).

We propose a formalism based on regular expressions as a more expressive model than the string level patterns adopted by (Ravichandran & Hovy 2002) and (Joho & Sanderson 2000). Regular expressions allow for skipping word positions, introducing variants, forbidding some part of speech. Moreover, our regular expressions are parameterized with respect to the different languages considered.

Each pattern expression is a sequence of elements, representing a class of words or a syntactic chunk (in current implementation only NP chunks are considered). Each class can be either a part of speech, or a lexical or a semantic class. Our regular expressions allow for the following syntactic forms (where $e_1, e_2, \ldots, e_n$ are regular expressions themselves):

- Alternatives: **[**e_1 | e_2 | **...** | e_n**],** stating that one of the expressions $e_1, e_2, \ldots, e_n$ should appear in the text.

- Language alternatives: **[eng:**e_1 | **ita:**e_2 | **bul:**e_3**],** where **eng**, **ita**, and **bul** bind e_1, e_2, and e_3 respectively to English, Italian, and Bulgarian.

- Negation: **[**~ e_1 e_2 ...e_n**],** stating that e_1 e_2 ...e_n should not match.

- Repetition specificator: e**(**m**),** stating that the expression e should appear at most m times.

The following notation are further used to denote certain word or chunk classes:

"s" **-** string which should be matched in the text

lemma: *s* **-** the lemma of the word should be *s*

<**NP**> **-** noun phrase

<**PERSON**> **-** matches any noun which designates person (e.g. "author", "writer", "philosopher", etc.)

<**HYPERNYM**> **-** matches any noun which is a hypernym of the question focus in MultiWordNet.

<**FOCUS**> **-** stands for the question focus

Prep, **Noun**, **Verb**, etc. are used to denote words which belong to the corresponding part of speech

<**WX**> **-** matches any word

For example, considering definition questions, the following multilingual pattern with a variable component hold for English, Italian and Bulgarian (in Bulgarian Cyrillic alphabet is used, but we use Latin transcripts in our examples):

```
[~ Prep ] <FOCUS> [~ Noun](1) [eng:
lemma:be | ita: lemma:essere| bul:
lemma:sam] [~ Prep Verb Conj](3) Noun
```

This pattern captures the following sequence of words: a word which is not a preposition (**[~ Prep]**); the focus (**<FOCUS>**); one word which is not a noun may follow (**[~ Noun](1)**); the auxiliary verb "*be*" appearing in one of the considered languages; at most 3 words (**[~ Prep Verb Conj](3)**), none of which is a preposition, a finite verb form or a conjunction; a noun **(Noun)**.

Such an expression will capture a broad range of fragments like: "*Socrates is the greatest philosopher*", "*Socrates may be the greatest philosopher, but…*", "*Socrates is considered the greatest philosopher*". It will not wrongly capture "*The followers of Socrates are…*", since no preposition is allowed before the focus.

Another frequently used matching pattern is:

```
[eng: "called" | ita: "detto" |bul:
"narechen" ] [~ Punct](4) <FOCUS>
```

It captures fragments like: *"…a philosopher called Socrates*" and its translation in Italian and Bulgarian.

A certain fragment can be matched by more than one pattern. For example the fragment *"… a philosopher called Socrates*" is also matched by the pattern:

```
<PERSON>[~ Noun FiniteVerbForm](3)
<FOCUS>
```

Our aligned representation makes feasible the use of multilingual pattern matching algorithms, and creates an appropriate framework for transferring knowledge acquired from one language to the others. Linguistic regular expressions can be represented in XML; for instance:

```
<pattern>
     <multilingual>
          <ENG> called</ENG>
          <ITA> detto </ITA>
          <BUL> narechen </BUL>
     </multilingual>
     <repeat max="4">
          <no><part-of-speech"Punct"/></no>
     </repeat>
     <FOCUS/>
</pattern>
```

5. Experiments and Results

Our evaluation was performed using the described library in the context of the QA task. Two Web based QA systems (one for both English and Italian, and one for Bulgarian) were used to support the experiments. These systems used the multilingual library to extract definitions. As a semantic resource for English and Italian we used MultiWordNet (Pianta et. al 2002). For Bulgarian we used gazetteer lists of words which frequently appear in definitions (*e.g. "author"*, *"astronomer"*, *"philosopher"*, etc.). These resources were used when semantic elements of the patterns were matched in the text.

We used as a test set the 50 definition questions from the 2003 edition of TREC, which were translated in Bulgarian and Italian.

When a definition question is posed to the QA system, it first downloads documents from the Web, and next extracts and ranks definitions according to the matching patterns. In our evaluation framework weights were given to the different patterns. Those weights were defined manually for the scope of the experiments, though in general more refined statistical techniques can be used. The same definition extraction strategy is adopted for all of the three languages:

- For every definition question we retrieve the 100 top ranked text snippets returned by Google.
- For every snippet in which the focus of the definition question appears, we extract its left and right context as a candidate definition.
- All the patterns are matched with candidate definitions and the sum of their weighs is assigned as a *primary score* of the candidate definition.
- From each candidate definition the system selects the noun phrase which is closest to the focus. Then it checks the co-occurrence of this *definition core* with the question focus in a validation pattern. Additional Web queries are generated to collect this frequency information. For example: if the question is "*Who is Aaron Copland?*" and the closest noun in the candidate definition is "*composer*" we count the frequency of the validation pattern "*Aaron Copland is a composer*" on the Web (we use AltaVista). Next, the Pointwise Mutual Information (PMI) between "*Aaron Copland*" and "*composer*" will be calculated. PMI is assigned as a *secondary score* to the candidate definition. The same techniques is used for Italian. Conversely, for Bulgarian we collect frequency information from the snippets already obtained, since AltaVista does not provide reliable support for this language.
- The final score of every candidate definition is obtained by multiplying its primary and secondary score. Finally, candidate definitions are sorted according to their score.

For every question we considered for evaluation up to five definitions and calculated the Mean Reciprocal Rank (MRR) for every language. MRR was standard measure for evaluation of the performance of the QA systems in TREC 9 and TREC 10. It is calculated as the mean of the *reciprocal ranks* of the different questions. Given an ordered list of answers to a question, the reciprocal rank is calculated as *1/r* where *r* is the position of the first correct answer. For example, if the first answer in the list is correct, the reciprocal rank of the question will be 1, if the second is correct and the first incorrect, then the reciprocal rank will be .5, etc.

We also computed the percentage of questions answered correctly in the top five definitions returned. The following table shows the results for the three languages.

	MRR	Correctly answered
English	0,54	78%
Italian	0,37	60%
Bulgarian	0,55	62%

6. Conclusions

We presented a prototype of multilingual pattern library for the extraction of answers to definition questions. The coverage of such a library has been tested in a Web based QA scenario. Evaluation shows that we answer 78% of the definition questions in English while giving a right answer to 60% and 62% for Bulgarian and English questions respectively. This gap between languages can be easily explained by the fact that much more Web pages exist for English rather than for Italian and Bulgarian. The fact that for Bulgarian we have the highest MRR is somehow surprising, since on the Web this language is represented with much smaller number of pages with respect to English and Italian. However, most of the pages in Bulgarian come from official sources, such as news agencies and for this reason the average quality of the Web pages is higher than for the English and Italian languages.

Our results can be regarded as a clue to the viability of the general approach of using pattern libraries for QA. We intend to improve our prototype and to extend it for question types other than definition questions.

References

Magnini, B., Negri, M., Prevete, R., Tanev, H. (2002). Is it the Right Answer? Exploiting Wed Redundancy for Answer Validation. In Proceedings of the 40th ACL Conference. University of Pennsylvania, Philadelphia.

Pianta E., Bentivogli L., Girardi C. (2002). MultiWordNet: Developing an Aligned Multilingual Database. In Proceedings of the 1st International Global WordNet Conference, Mysore, India.

Ravichandran D., Hovy, E. (2002). Learning Surface Text Patterns for a Question Answering System. In Proceedings of the 40th ACL Conference. University of Pennsylvania, Philadelphia.

Joho H., Sanderson M. (2000). Retrieving Descriptive Phrases from Large Amounts of Free Text. In Proceeding of 9th International Conference in Information and Knowledge management.

Soubbotin M. M. (2001). Patterns of Potential Answer Expressions as Clues to the Right Answers. In Proceedings of the 10th Text Retrieval Conference. Gaithersburgh, MD.