Mining Knowledge from Wikipedia for the Question Answering task

Davide Buscaldi*, **Paolo Rosso***

*Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Valencia, Spain
{dbuscaldi, prosso}@dsic.upv.es

Abstract

Although significant advances have been made recently in the Question Answering technology, more steps have to be undertaken in order to obtain better results. Moreover, the best systems at the CLEF and TREC evaluation exercises are very complex systems based on custom-built, expensive ontologies whose aim is to provide the systems with encyclopedic knowledge. In this paper we investigated the use of Wikipedia, the open domain encyclopedia, for the Question Answering task. Previous works considered Wikipedia as a resource where to look for the answers to the questions. We focused on some different aspects of the problem, such as the validation of the answers as returned by our Question Answering System and on the use of Wikipedia "categories" in order to determine a set of patterns that should fit with the expected answer. Validation consists in, given a possible answer, saying wether it is the right one or not. The possibility to exploit the categories of Wikipedia was not considered until now. We performed our experiments using the Spanish version of Wikipedia, with the set of questions of the last CLEF Spanish monolingual exercise. Results show that Wikipedia is a potentially useful resource for the Question Answering task.

1. Introduction

Encyclopedic knowledge is valuable for many Natural Language Processing (NLP) applications, and in particular for the Question Answering (QA) task. Recently, the availability of a large, open domain encyclopedia, such as the Wikipedia[1], has captured the attention of some researchers (Lita et al., 2004; Ahn et al., 2004) in the Question Answering field. Until now, the focus of these works was on the use of the encyclopedia in order to look for the answer to the questions. However, the results did not fulfill the expectations.

We investigated the use of Wikipedia in some slightly different aspects of the Question Answering task: answer validation and generation of answer patterns. In the first case, the problem consists in, given a possible answer, saying wether it is the right one or not. Previous work on answer validation has been carried out by exploiting the redundancy of the web (Magnini et al., 2002), giving good results. In the case of encyclopedias, redundancy is not an option, because usually each topic is covered by no more than one article. Therefore, the quality of the information extracted from the question is crucial to find the related article.

In the second case, the problem consists in building a regular expression pattern that (possibly) match the right answer. When a question pertains to a specific class, usually deduced by the structure of the question (for instance, the right answer for a question starting with the word *where* will be *at least* some kind of location) patterns can be built by hand, usually together with a custom-built ontology (Laurent et al., 2005; Amaral et al., 2005). However, when the question cannot be classified using a given taxonomy, the semantic class can be deduced by the question itself, such as in "Which *fruit* contains vitamine C?": in this case, the class is "fruit", and we want to find a suitable answer string for that class. In order to do that, we exploited the categorization of articles in Wikipedia. For instance, the article corresponding to the category *http://en.wikipedia.org/wiki/Category:Fruit* contains a list of fruits. It can be observed that the "category" entries constitute a sort of Wikipedia ontology, since some categories contain also subcategories.

2. Goal of the paper

The goal of this paper is to show some simple techniques to exploit the knowledge from Wikipedia. No in-depth analysis of the articles is needed, because we rely on the categorization of the articles that comes with Wikipedia and on simple rules in order to extract the useful information. The use of Wikipedia can overcome the issue of building a custom ontology, a task that is usually expensive both in time and money.

Moreover, an additional goal of the paper is to present some preliminary results obtained over the CLEF2005 Spanish monolingual test set which demonstrates that these techniques can be effectively used to improve the performance of our Question Answering system (named QUASAR), that already obtained good results in our last participation at the CLEF QA task (Gómez et al., 2005; Vallin et al., 2005).

3. Using Wikipedia for Validation

First of all, we need to briefly introduce the three-level question type taxonomy used by our Question Answering system. One or more classification patterns have been defined for each of the question types in Table 1. Any question that does not match any of the defined patterns is considered as a "GENERIC" question. Usually these questions start with the word "Which" or "What" ("Cuál", "Qué" in Spanish).

We applied the Wikipedia-based answer validation technique to the following question types: "NAME" (includ-

[1] http://www.wikipedia.org

L0	L1	L2
NAME	ACRONYM	
	PERSON	
	TITLE	
	LOCATION	COUNTRY
		CITY
		GEOGRAPHICAL
DEFINITION		
DATE	DAY	
	MONTH	
	YEAR	
	WEEKDAY	
QUANTITY	MONEY	
	DIMENSION	
	AGE	

Table 1: Question type taxonomy used by QUASAR.

ing all subtypes but the “COUNTRY” one) and “DEFINITION”.

A Spanish snapshot of the Wikipedia (consisting of a single xml file containing about $75,000$ articles) was indexed using the well-known Lucene[2] search engine, splitting the data of each article into the following indices: *title*, *text*, *definition* and *category*. While *title* and *text* fields are obtained simply by picking the tagged data from the xml file (specific Wikipedia characters such as the double parenthesis are removed, too), for *definition* and *category* fields the data are obtained by means of simple patterns based on the analysis of the typical structure of the pages. Articles regarding discussions, user pages, pictures, metadata and, in general, unnecessary information were skipped. The “redirect” pages (i.e., pages that contain only a link to the page with the article) were filled with the text of the target page before indexing. All the text was indexed using the Snowball stemmer[3].

3.1. Validation of “Definition” Answers

Questions of this kind are easy to identify both in Spanish and English. They usually appear in the form `Who is <person name> ?` (*Quién es <person name>* in Spanish) or `What is <organization name> ?` (*Qué es <organization name>*). The validation of the answers of this kind is done in the following way:

1. Obtain the candidate answers from QUASAR;
2. For each candidate answer, perform the following search in Wikipedia:
   +title:*name* +definition:*candidate_ answer*
3. If at least a page is returned, then confirm the answer, else reject it.

The + operator is a Lucene operator used to force the returned page to contain that element. This guarantees that, if a page is retrieved, then it will contain the article about the person or organization we are looking for and that in its definition the same words which constitute the candidate answer will appear. Otherwise, if no page is retrieved, there may be two reasons for this: an article about the entity to be defined does not exist, or the definition we are searching is not correct. In the first case, we cannot perform the validation, while in the second one the result is that the returned definition should be labeled as a wrong answer. Therefore, a preliminary search is done in order to check wether an article about the named entity is present in Wikipedia or not. For instance, consider the following question:
Quién es Nelson Mandela ?
An article named *Nelson Mandela* is present in Wikipedia. We can move on to the next step. We obtain, from QUASAR, the three following candidate answers (ordered by their weight): *asistencia al encuentro del presidente de Sudáfrica*, *presidente de Sudáfrica* and *presidente de la República*.
The queries submitted to the search engine, are, respectively: [+title:*”Nelson Mandela”* +definition:*asistencia* +definition: *encuentro* +definition: *presidente* +definition: *Sudáfrica*], [+title:*”Nelson Mandela”* +definition: *presidente* +definition: *Sudáfrica*] and [+title:*”Nelson Mandela”* +definition: *presidente* +definition: *República*]. The first query returns no pages, while the other ones return correctly the page corresponding to the *Nelson Mandela* article. Therefore, the answer returned by the system after the validation process is the second one (because it has a greater weight than the third one).

3.2. Validation of “Name” Answers

In this case, a question can assume different forms. It can depend on what kind of name is being asked for, although there are many ways to formulate a “Name” question. However, the strategy can be considered as opposite to the one adopted for definition questions. In fact, the question contains a possible definition for a named entity we need to discover. The validation of the answers of this kind is performed in the following way:

1. Obtain the candidate answers from QUASAR;
2. For each candidate answer, perform the following search in Wikipedia:
   +title:*candidate_answer*
   +definition:*question_constraints*
3. If at least a page is returned, then confirm the answer, else reject it.

The *question_constraints* are selected by the QUASAR question analysis module, which uses some rules based on POS (Part-Of-Speech) labels, capitalization of words and their lemmas. In most cases the question constraints are represented by noun sequences, named entities, numbers, dates or quotations.
For instance, consider the following question: *Cuál es el nombre del secretario de las Naciones Unidas ?* (*What is the name of the secretary of the ONU?*). The question constraints in this case are represented by the word *secretario* (secretary) and the named entity *Naciones Unidas* (ONU).

[2] http://lucene.apache.org

[3] http://snowball.tartarus.org

Suppose we obtain the following candidate answers: *Kofi Annan*, *Bill Clinton*.
Therefore, the queries submitted for the validation, are: [+title:*"Kofi Annan"* +definition:*secretario* +definition: *Naciones Unidas*], [+title:*"Bill Clinton"* +definition: *secretario* +definition: *Naciones Unidas*]. The second query returns no pages, while the first one returns the page corresponding to the *Kofi Annan* article. Therefore, *Kofi Annan* is validated as the correct answer.

4. Exploiting Wikipedia's Categories

A "Generic" question is a question that cannot be classified using the proposed taxonomy. This is usually due to the fact that the answer belongs to a category specified in the question itself. The major consequence of the impossibility to properly classify the question is that the system is not able to identify a candidate answer, due to the lack of a suitable pattern.
We noticed that the usual form of this kind of questions is `What/Which <category> <property>?` (*Qué <category> <property>* in Spanish), sometimes with a leading preposition. For instance: "*Which fruit contains vitamine C ?*", "*Which Russian president did attend to the G7 in Neaples ?*", "*In what team did Ayrton Senna begin his F1 career ?*" are all "Generic" questions, where the categories are respectively *fruit*, *Russian president* and *(F1) team*. Wikipedia categories can help in finding an answer to these questions. For instance, in the *category:Fruit* page[4] of the current English Wikipedia are listed 152 fruit names. Therefore, the pattern for a *good* candidate answer will correspond to one of these 152 names.
The resulting algorithm adopted to generate patterns using the Wikipedia categories is the following:

1. Extract the *category* from the question; this is done using the same algorithm used in order to identify the question constraints, but picking only the first of them.
2. Perform the following search in Wikipedia: +category:*<category>*;
3. From the best ranked page select the names of the listed articles and arrange them in a regular expression pattern; hand it over to QUASAR. If no page is returned, order QUASAR to generate a NIL answer.

Please note that a NIL answer corresponds to the situation in which the system is not able to find an answer to the question.
For instance, consider the question:
Qué fruta contiene vitamina C? (*Which fruit contains vitamine C ?*). The query submitted to Lucene is [+category:*"fruta"*]. Thanks to the fact that the pages have been indexed using a stemmer, the page corresponding to the category *Frutas* is returned. In the spanish version the listed fruits are 137. Therefore, a regular expression containing the fruit names separated by a | sign (*(?i)Ababaya|Abombo|Aceituna|Aguacate|...*)[5] is handed over to QUASAR, which will use that expression to identify the candidate answers, and select the best one according to their weights (Gómez et al., 2005).

5. Experiments

The set of questions used to test our approaches was the set of 200 Spanish monolingual questions from the CLEF 2005 Question Answering track (Vallin et al., 2005). Out of these 200 questions, *definition* ones are 50, 25 related to *organizations* and 25 to *persons*. The *name* questions are 42 (excluding the *location* questions), and *generic* questions are 25.
The results show that the use of Wikipedia allowed to obtain 9 right answers more with respect to the original results. This means a global improvement of $4,5\%$ in recall (i.e., number of right answers divided by the number of questions). Results grouped by question type are displayed in Table 2.

Question type	Answers (tot)	Rec. gain
All	9 (200)	4,5%
Definition	4 (50)	8,0%
Name	2 (42)	4,7%
Generic	3 (25)	12,0%

Table 2: Recall gain, grouped by question type.

Out of the 4 *definition* answers, 3 passed from being 'incorrect' (i.e., containing not only the right answer but also pieces of text semantically unrelated to the answer - they are labeled with an 'X' sign by CLEF evaluators) to right.
With the help of the error analysis, we found that in 15 questions of both *definition* and *name* types, the answer validation process failed due to the fact that QUASAR was not able to return good answer candidates, or did not return any at all. Similarly, 5 *generic* answers were actually present in Wikipedia but the passage retrieval module did not find passages where the answer was present. This means that with a "perfect" passage retrieval and answer extraction system, the potential improvement with the help of Wikipedia was of 29 questions (the 9 actually retrieved plus the 20 that failed due to the QA system), corresponding to a $14,5\%$ gain in recall. See Table 3 for details.

Question type	Answers (tot)	Pot. Rec. gain
All	29 (200)	14,5%
Definition	8 (50)	16,0%
Name	7 (42)	16,6%
Generic	5 (25)	20,0%

Table 3: Potential recall gain (i.e., questions where Wikipedia could be useful but was not possible to use its information), grouped by question type.

Another interesting feature discovered by error analysis is that when Wikipedia proved to be useless, it is usually due to one of the following reasons:

- The question is about facts unrelated with the Spanish world. That is, the answer could be present in another

[4] http://en.wikipedia.org/wiki/Category:Fruit

[5] in Java regular expressions the $(?i)$ sign is used to ignore the capitalization

localization of Wikipedia. For instance, the answer to the question *Who is Giulio Andreotti?* could be find in the Italian or English versions of Wikipedia.

- The question is about facts too specific to be taken into account into the Wikipedia. For instance, *Who discovered the galleon San Diego?*, or *Who is Rolf Ekeus?*.

More less significative failure reasons were the ambiguity of some categories (for instance, “Qué *plataforma* estaba acampada en el Paseo de la Castellana de Madrid ?” - “Which platform was camped at Paseo de la Castellana in Madrid?”), or the fact that the category was imaginary (for instance, “Para qué *periódico* trabajaba Clark Kent?”, “For which newspaper does Clark Kent work?”: the category *newspapers* (periódicos) exists, but does not contain the *Daily Planet*).

6. Conclusions

We presented some methods to exploit the Wikipedia open source encyclopedia for the Question Answering task. Although the results obtained showed that Wikipedia can be actually used to improve the performance of our Question Answering system, especially for “Generic” questions, they are well below the potential. This is due mainly to the following three reasons: the performance of passage retrieval and answer extraction systems, the localization of Wikipedia editions, and the fact that knowledge related to small-scale events or less known people usually is not included into the Wikipedia. In this last case, no action can be taken, since it is a feature of a massive distributed project like Wikipedia; however, we can work to improve the passage retrieval system and answer extraction subsystem, obtaining better passages and candidate answers. Another interesting work direction should be a multilingual approach that could take into account the various localizations of Wikipedia in the other languages, preferably those containing many articles.

Acknowledgements

We would like to thank R2D2 CICYT (TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects for partially funding this work. A special acknowledgement to Piedachu Peris García for her support in reviewing and correcting this work.

7. References

David Ahn, Valentin Jijkoun, Gilad Mishne, Karin Mller, Maarten de Rijke, and Stefan Schlobach. 2004. Using wikipedia at the trec qa track. In *TREC 2004 Proceedings*.

Carlos Amaral, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, and Cláudia Pinto. 2005. Priberams question answering system for portuguese. In *CLEF 2005 Working notes*.

José Manuel Gómez, Davide Buscaldi, Empar Bisbal, Paolo Rosso, and Emilio Sanchis. 2005. Quasar: The question answering system of the universidad politécnica de valencia. In Springer, editor, *CLEF 2005 Proceedings, LNCS*, Vienna.

Dominique Laurent, Patrick Séguéla, and Sophie Nègre. 2005. Cross lingual question answering using qristal for clef 2005. In *CLEF 2005 Working notes*.

Lucian Vlad Lita, Warren A.Hunt, and Eric Nyberg. 2004. Resource analysis for question answering. In *ACL 2004 Proceedings*. Association of Computational Linguistics, July.

Bernardo Magnini, Matteo Negri, Roberto Prevete, and Hristo Tanev. 2002. Is it the right answer? exploiting web redundancy for answer validation. In *ACL*, pages 425–432.

Alessandro Vallin, Danilo Giampiccolo, Lili Aunimo, Christelle Ayache, Petya Osenova, Anselmo Peas, Maarten de Rijke, Bogdan Sacaleanu, Diana Santos, and Richard Sutcliffe. 2005. Overview of the clef 2005 multilingual question answering track. In *CLEF 2005 Working notes*.