Towards a Multilingual QA System based on the Web Data Redundancy[*]

Rita Aceves-Pérez[1], **Luis Villaseñor-Pineda**[1] **and Manuel Montes-y-Gómez**[1,2]

[1] National Institute of Astrophysics, Optics and Electronics, Mexico.
{rmaceves, villasen, mmontesg}@inaoep.mx
[2] Polytechnic University of Valencia, Spain.
{mmontes}@dsic.upv.es

Abstract. This paper explores the feasibility of a multilingual question answering approach based on the Web redundancy. The paper introduces a system prototype that combines a translation machine with a statistical QA method. The main advantage of this proposal is its small dependence to a given language. The experimental results demonstrated the great potential of the approach and gave interesting insights about the Web redundancy and the online translators.

1 Introduction

The documents accessible from the Web may satisfy almost every information need. However, without the appropriate access mechanisms all these documents are practically useless. In order to solve this dilemma several text processing approaches have emerged. For instance: information retrieval and question answering (QA).

The goal of a QA system is to retrieve answers to questions rather than full documents to general queries [4]. For example, given a question like: “where is the Amparo Museum located?”, a QA system must respond “Puebla” instaed of just returning a list of documents related to the Amparo Museum.

In recent years, due to the Web growth, there has been an explosive demand for better multilingual information access approaches. Multilingual QA systems are one example [3]. These systems allow answering a question based on a set of documents from several languages.

In this paper we present our first experiment on multilingual question answering on the Web. This experiment considers answering English questions using Spanish Web documents and vice versa. The system architecture that we propose is different from the traditional approaches [4]. It is based on the use of online translation machines and simple pattern matching methods, rather than on sophisticated linguistic analyses of both questions and documents. In some degree, the purpose of this paper is to analyze the performance of automatic online translators, as well as, to study the impact of the incomplete and wrong translations over the answer precision.

The rest of the paper is organized as follows. Section 2 introduces a statistical QA system for the Web. Section 3 proposes a general architecture for a multilingual QA

[*] This work was partially financed by the CONACYT Project 43990A-1 and scholarship 184663. We also thank to the Secretaría de Estado de Educación y Universidades, España.

system. Sections 4 and 5 show the experimental results. Finally, section 6 draws our conclusions and future work.

2 Statistical QA on the web

This section describes the general architecture of a statistical QA system that allows finding answers to factual questions from the Web. It consists of three major modules: (i) query reformulation, (ii) snippets recollection, and (iii) answer extraction.

The architecture is supported on the idea that the questions and their answers are commonly expressed using the same words, and that the probability of finding a simple (lexical) matching between them increases with the redundancy of the target collection [1]. It was originally adapted to Spanish [2], however it is general enough to work with questions in other languages that share some morpho-syntactic characteristics of the Spanish, such as: English, Italian, French, Portuguese and Catalan.

2.1 Query reformulation

Given a question, this module generates a set of query reformulations. These reformulations are expressions that were probably used to write down the expected answer. We performed several experiments in order to determine the most general and useful reformulations. The following paragraphs present those with the best results. All the cases are illustrated for the question: Who received the Nobel Peace Prize in 1992?

First reformulation: "bag of words"
This reformulation is the set of non stop-words of the question. For instance, *received Nobel Peace Prize 1992.*

Second reformulation: "verb movement"
One of our first observations after checking a list of factual questions was that the verb is frequently right after the wh-word. We also know that in order to transform an interrogative sentence into a declarative one is necessary to eliminate the verb or to move it to the final position of the sentence. The resultant sentence is expected to be more abundant in the Web that the original one.

In order to take advantage of this phenomenon, but without using any kind of linguistic resource, we propose to build a set of query reformulations eliminating or moving to the end of the sentence the first and second words of the question.

Two examples of these kinds of reformulations are: "*the Nobel Peace Prize in 1992 received"* and *"Nobel Peace Prize in 1992"*.

Third reformulation: "components"
In this case the question will be divided in components. A component is an expression delimited by prepositions. Therefore, a question Q having m prepositions will be represented by a set of components $C = \{c_1, c_2,\ldots, c_{m+1}\}$. Each component c_i is a sub string of the original query. New reformulations will be defined combining them.

Some examples of this kind of query reformulations are: "*received the Nobel Prize*" "*of Peace*" "*in 1992*", and "*in 1992 received the Nobel Peace Prize*".

2.2 Snippets recollection

Once the set of reformulations has been generated, this module sends them to a search engine (currently we are using Google), and then it collects the returned snippets.

2.3 Answer extraction

From the snippets collected from the Web we compute all the n-grams (from unigrams to pentagrams) as possible answers to the given question. Then, using some statistical criteria the n-grams are ranked by decreasing likelihood of being the correct answer. The top five are presented to the user.
The method for the n-gram extraction and ranking is as follows:

1. Extract the twenty most frequent unigrams satisfying a predefined typographic criterion (capitalized proper nouns, numbers and names of months).
2. Determine all the n-grams, from bi-grams to pentagrams, just containing the frequent unigrams.
3. Rank the n-grams based on their compensated relative frequency[1]:

$$P_{x(n)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n-i+1}\frac{f_{\hat{x}_j(i)}}{\sum_{y\in G_i} f_{y(i)}}$$

4. Show to the user the top five n-grams as possible answers.

Applying this method we obtained the following answers to the example question: *Rigoberta Menchu*, *Rigoberta Menchu Tum*, *Menchu*, *Rigoberta Menchu Recibio*, *Rigoberta*.

3 Multilingual QA Prototype

A multilingual QA system enables the users to formulate a question in a language different from the reference corpus. Most common multilingual systems work with two languages, one for the question and another for the target collection. However, a full multilingual QA system would allow searching for answers on documents from several languages. The figure 1 shows our proposal for a multilingual QA system for the Web. This architecture consists of two main modules: (i) a translation machine, and (ii) a language-independent statistical QA system.

The main advantage of this prototype is it small dependence to the target language, which allows using the system in a full multilingual scenario. On the other hand, one of the main disadvantages of this architecture is it high dependence to the quality of question translations, and also to the Web redundancy in the target language.

[1] We introduce the notation x(i) for the sake of simplicity. In this case x(i) indicates the i-gram x, G_i is the set of all i-grams, and $\hat{x}_j(k)$ represents the k-gram x contained in n-gram x(i) at the position j.

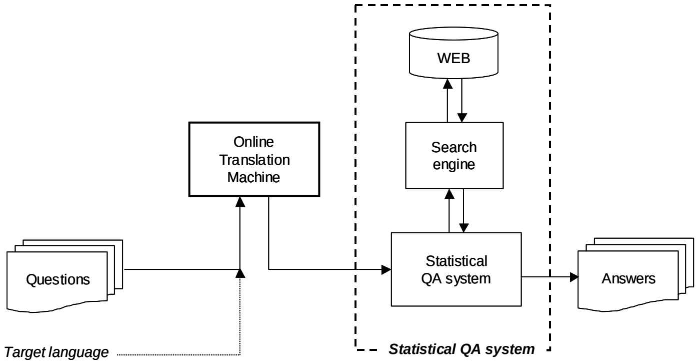

Figure 1. Multilingual QA prototype

4 Experiments

The experimental evaluation considered a set of 165 factual questions. These questions were taken from the English training corpus of CLEF 2003[2]. The answers for these questions were of four types: names of persons, organizations, locations and dates.

We probed with 3 different online translators: Google, Freetranslation and Webtranslation[3]. Tables 1-3 compare the results using the different translation machines and applying the different reformulations techniques (refer to section 2.1). The mean reciprocal rank[4] (MRR) and the precision[5] are indicated for each experiment.

Table 1. MRR/Precision using the "bag of words" reformulation

Query	Translator		
	Webtranslation	**Freetranslation**	**Google**
Date (36)	0.086 / 0.389	0.003 / 0.111	0.006 / 0.138
Location (52)	0.330 / 0.461	0.275 / 0.442	0.085 / 0.134
Organization (27)	0.167 / 0.185	0.106 / 0.185	0.012 / 0.037
Person (51)	0.152 / 0.372	0.126 / 0.353	0.030 / 0.196

Table 2. MRR/Precision using the "verb movement" reformulation

Query	Translator		
	Webtranslation	**Freetranslation**	**Google**
Date (36)	0.120 / 0.25	0 / 0	0.037 / 0.055
Location (52)	0.183 / 0.307	0.061 / 0.096	0.027 / 0.096
Organization (27)	0.104 / 0.259	0 / 0	0 / 0
Person (51)	0.149 / .352	0.027 / 0.137	0.076 / 0.137

[2] The Cross-Language Evaluation Forum (CLEF) http://clef-campaign.org/)

[3] www.google.com, www.freetranslation.com and www.imtranslator_webtranslation.paralink.com respectivaly.

[4] An individual question received a score equal to the **reciprocal** of the **rank** at which the first correct response was returned, or 0 if none of the responses contained a correct answer. The score for a sequence of queries is the **mean** of the individual query's **reciprocal** ranks.

[5] precision = number of found answers / number of total questions

Table 3. MRR/Precision using the “components” reformulation

Query	Translator		
	Webtranslation	**Freetranslation**	**Google**
Date (36)	0.071 / 0.111	0.004 / 0.056	0.048 / 0.083
Location (52)	0.138 / 0.154	0.023 / 0.057	0.019 / 0.19
Organization (27)	0.015 / 0.074	0.003 / 0.037	0 / 0
Person (51)	0.016 / 0.137	0.004 / 0.0197	0.009 / 0.039

4.1 The Influence of Data Redundancy

It is well known that English is the most representative language on the Web (68% in accordance with the last report of Global Reach[6]). In order to analyze the effect of data redundancy on our multilingual QA approach, we made an experiment with Spanish questions using English Web documents as data repository. In this experiment we considered the same set of questions that in the previous case, and we employed the Google search engine. We used the Webtranslation machine to translate the questions from Spanish to English. The answer extraction process was leaded by the “bag of words” reformulation. Table 5 shows the results.

Table 5. Spanish-English experiment

Question	MRR	Precision
Date	.091	.444
Location	.264	.596
Organization	.444	.148
Person	.169	.314

5 Results Discussion

The best results were obtained using Webtranslation, which produced the best question translations. This situation is clear enough when we analyzed the results from the tables 2 and 3, where a syntactically well-formed question is required.

As we expected, the best results were obtained using the “bag of words” reformulation. This fact indicates that online translators tend to produce accurate word-by-word translations (using the frequent senses of the words), but they tend to generate syntactically incorrect questions.

An interesting observation from the experiment was that sometimes the translation machines produce better translations that those manually constructed. This is because they use common words. For instance, the question “Which is the name of John Lennon´s wife?” was manually translated to Spanish into “¿Cuál es el nombre de la *mujer* the John Lennon?”, while the automatic translation by Webtranslation was “¿Cómo se llama la *esposa* de John Lennon?”. The noun *mujer* (woman), used in this context, is less frequent than *esposa* (wife).

Another relevant fact is that sometimes the wrong translations facilitate the QA process (i.e., they favor some kind of reformulations). For instance, the question “Who is the President of the Roma football team? was manually translated to “¿Cómo

[6] http://www.glreach.com

se llama el presidente del equipo de fútbol de Roma?, and automatically translated to "¿Cómo se llama el presidente de la Roma de Fútbol?". Although incorrect, the automatic translation preserves the main concepts of the question, and allows to the system easily find the answer.

In addition, our results indicate that the Web redundancy has great influence on the system response. The precision for the Spanish-English experiment was 8% greater than for the English-Spanish experiment (compare table 5 with the second column of the table 1). However, we notice that the MRR was greater on the initial experiment. This indicates that correct answers in Spanish are easily identified. We believe this is because there is less noisy information on Spanish than in English.

References

1. E. Brill, J. Lin, M. Banko, S. Dumais and A. Ng (2001). *Data-intensive question answering*. In Proceedings of the Tenth Text REtrieval Conference (TREC 2001).
2. A. Del Castillo, M. Montes-y-Gómez and L. Villaseñor-Pineda (2004), *QA on the Web: A preliminary study for Spanish Language*. Fifth Mexican International Conference on Computer Science. (ENC´04). pp. 322-328. IEEE Computer Society, ISBN 0-7695-2160-6.
3. L. Perret (2004). Question answering system for the French language. Working Notes for the CLEF 2004 Workshop, 2004.
4. J. L. Vicedo (2002). La Búsqueda de Respuestas: Estado Actual y Perspectivas de Futuro. Revista Iberoamericana de Inteligencia Artificial. Número 22, Volumen 8, Primavera 2004.