NLEL at RespubliQA 2010

Santiago Correa, Davide Buscaldi, and Paolo Rosso

Natural Language Engineering Lab., ELiRF
Universidad Politécnica de Valencia, Camino de Vera s/n, Valencia, España
santcg@gmail.com{scorrea,dbuscaldi,prosso}@dsic.upv.es
{dbuscaldi,prosso}@dsic.upv.es
http://www.dsic.upv.es/grupos/nle

Abstract. This report describes the participation of the NLEL Lab. from the Universidad Politécnica of Valencia to the RespubliQA task at CLEF 2010. The system designed for this participation is based on the one used in our previous participation, with some modifications required in order to adapt it to the new guidelines. The system participated to both the "Paragraph Selection" (PS) and "Answer Selection" (AS) sub-tasks.

Keywords: Question Answering, n-gram based Passage Retrieval

1 Introduction

The participation to the PS sub-task was centered around the JIRS n-gram based passage retrieval system [6]. In order to participate in the AS sub-task, it was necessary to integrate into the system an Answer Extraction module, which was developed originally for the QUASAR QA system [5], which participated in past CLEF-QA editions, from 2005 to 2007. In the following sections we describe the characteristics of the QA system in both PS and AS configurations.

2 JIRS Passage Retrieval System

JIRS[1] is an n-gram based passage retrieval system that has been developed specifically for the Question Answering task. An n-gram is a sequence of n adjacent terms extracted from a sentence or a question. JIRS is based on the premise that in a sufficiently large document collection, question n-grams should appear near the answer at least once. JIRS represents the core of the system, since it was used both in the PS and AS sub-tasks.

The architecture of JIRS is shown in Figure 1. The user question is passed to a search engine that returns relevant snippets of a documents collection in which relevant terms from the question occur. The n-gram extraction module will return all the n-grams of size 1 to n, where n is the number of terms of the question. This process is done both for the question and for each of the snippets

[1] http://sourceforge.net/projects/jirs/

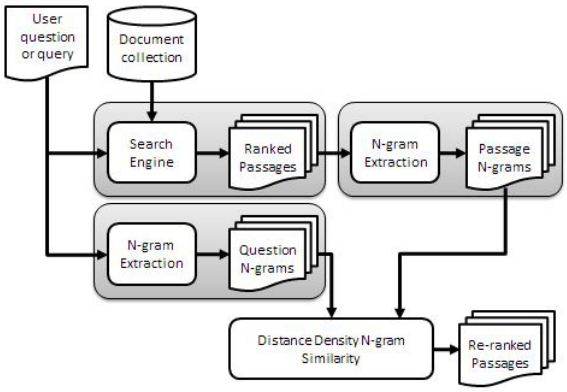

Fig. 1. Architecture of *JIRS* Passage Retrieval system

retrieved by the search engine. Once obtained the n-grams of the question and the snippets, a comparison is made to calculate a similarity value between them. This similarity value is used to sort the list of passages that will eventually be returned to the user. The similarity between the question and the retrieved passages is defined in Equation 1.

$$Sim(p,q) = \frac{\sum\limits_{\forall x \in Q} h(x,P)\frac{1}{d(x,x_{max})}}{\sum_{i=1}^{n} w_i} \tag{1}$$

Where, $Sim(p,q)$ is the function that measures the similarity of n-grams sets of the question q with respect to the n-grams sets of the passage p. P is the n-gram set of the heaviest passage p (i.e., the one with most weight) whose terms are in the question; Q is the set of j-grams that are generated from the question q and n is the total number of terms in the question. There are three special and particular terms functions:

- w_i is the weight of the i-th term of the question which is determined by:

$$w_i = 1 - \frac{log(n_i)}{1 + log(N)} \tag{2}$$

  Where n_i is the number of sentences in which the term t_i occurs and N is the number of sentences in the collection;
- the function $h(x,P)$ measures the weight of each n-gram and is defined as:

$$h(x,P_j) = \begin{cases} \sum_{k=1}^{j} w_k & \text{if } x \in P_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

  Where w_k is the weight of the *k-th* term (see Equation 2) and j is the number of terms that compose the analyzed n-gram;

– and the factor $\frac{1}{d(x,x_{max})}$ that is a distance factor which reduces the weight of the n-grams that are far from the heaviest n-gram. The function $d(x, x_{max})$ determines numerically the value of the separation according to the number of words between a n-gram and the heaviest one. That function is defined as shown in Equation 4 :

$$d(x, x_{max}) = 1 + k \cdot ln(1 + L) \tag{4}$$

Where k is a factor that determines the importance of the distance in the similarity calculation and L is the number of words between a n-gram and the heaviest one (see Equation 3).

3 Answer Extraction System

In order to cope with the AS sub-task guidelines, which require that beyond retrieving a paragraph containing the answer to a question in natural language, systems are required to demarcate also the exact answer, we had to fit JIRS with an answer extraction system. This system is based on the QUASAR AE module described in [2], which has been used to participate in previous CLEF-QA tasks. The system has been modified by the addition of two new categories of questions: PERCENTAGE and MODE, and a new question analysis module based on the extraction of constraints by means of *idf* weights.

3.1 Question Analysis Module

This module obtains both the expected answer type (or *class*) and some constraints from the question. The different answer types that can be treated by our system are shown in Table 1.

Each category is defined by one or more patterns written as regular expressions. For instance, the Italian patterns for the category "CITY" are: *.*(che—quale) .*cittá .+ and (qual—quale) .*la capitale .+* . The questions that do not match any defined pattern are labeled with *OTHER*. If a question matches more than one pattern, it is assigned the label of the longest matching pattern (i.e., we consider longest patterns to be less generic than shorter ones).

The Question Analyzer has the purpose of identifying patterns that are used as constraints in the AE phase. In order to carry out this task, the set of different n-grams in which each input question can be segmented are extracted, after the removal of the initial quetsion stop-words. For instance consider the question: "*Where is the Sea World aquatic park?*", then the following n-grams are generated:

```
[Sea] [World] [aquatic] [park]
[Sea World] [aquatic] [park]
[Sea] [World aquatic] [park]
[Sea] [World] [aquatic park]
[Sea World] [aquatic park]
```

Table 1. QC pattern classification categories.

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

```
[Sea] [World aquatic park]
[Sea World aquatic] [park]
[Sea World aquatic park]
```

The weight for each segmentation is calculated in the following way:

$$\prod_{x \in S_q} \frac{\log 1 + N_D - \log f(x)}{\log N_D} \tag{5}$$

where S_q is the set of n-grams extracted from query q, $f(x)$ is the frequency of n-gram x in the collection D, and N_D is the total number of documents in the collection D.

The n-grams that compose the segmentation with the highest weight are the *contextual* constraints, which represent the information that has to be included in the retrieved passage in order to have a chance of success in extracting the correct answer.

3.2 Answer Extraction

The input of this module is constituted by the n passages returned by the PR module and the constraints (including the expected type of the answer) obtained through the *Question Analysis* module described in Section 3.1. The positions of the passages in which the constraints occur are marked before passing them

to the text analyzers (we named them *TextCrawlers* since they move on text like a spider on its web). One of these analyzers is instantiated for each of the n passages with a set of patterns for the expected type of the answer and a pre-processed version of the passage text.

Each TextCrawler begins its work by searching all the passage's substrings matching the expected answer pattern. Let us define C the set of constrains extracted in the Question Analysis phase; then a weight $w(s)$ is assigned to each found substring s, inversely proportional to the text distance of s with respect to the constraints $c_i \in C$. The final weight $w(s)$ is calculated as the product of the distance weights obtained for every constraint in the passage: $w(s) = \prod_{c_i \in C} 1/d(s, c_i)$.

A *Filter* module is based on a set of patterns compiled by hand in order to discard the candidate answers which do not match an allowed pattern or that do match with a forbidden pattern. When the Filter module rejects a candidate, the TextCrawler provide it with the next best-weighted candidate, if there is one. Finally, when all TextCrawlers end their analysis of the text, the *Answer Selection* module selects the answer to be returned by the system. The following strategies apply:

- Simple voting (SV): The returned answer corresponds to the candidate that occurs most frequently as passage candidate.
- Weighted voting (WV): Each vote is multiplied for the weight assigned to the candidate by the TextCrawler and for the passage weight as returned by the PR module.
- Double voting (DV): As simple voting, but taking into account the second best candidates of each passage.
- Top (TOP): The candidate elected by the best weighted passage is returned.

SV is used for NAME type questions, with DV as a backoff strategy in case of two candidates obtaining the same weight. WV is used for every other type of questions, with TOP as a backoff strategy.

4 Approaches

For the RespubliQA 2010 competition, the NLE Lab.has decided to participate in five monolingual tasks for passages extraction, the distribution of these tasks with the respective approaches used is:

- *English task*: Monolingual and monolingual - Stem participation; introducing these two units is expected to determine whether use of the Stem technique improves the performance of JIRS or not.
- *Spanish task*: Monolingual and monolingual - BM25 participation; introducing these two units is expected to determine whether the use of the BM25 technique improves the performance of JIRS or not.
- *French, Italian and German Tasks*: We present monolingual and multilingual approaches.

The following sections explain each one of the approaches implemented.

4.1 Monolingual approach

The data had to be preprocessed, due to the format of the collection employed in *ResPubliQA* competition, a subset of the *JRC-ACQUIS* and *Europarl* Multilingual Parallel corpus. The documents cover various subject domains: law, politics, economy, health, information technology, agriculture, food and more.

To be able to use the *JIRS* system in this task, the documents were analyzed and transformed for proper indexing. Since *JIRS* uses passages as basic indexing unit, it was necessary to extract passages from the documents. We consider any paragraph included between <p> tags as a passage. Therefore, each paragraph was labeled with the name of the containing document and its paragraph number.

Once the collection was indexed by *JIRS*, the system was ready to proceed with the search for the answers to the test questions. For each question, the system returned a list with the passages that most likely contained the answer to the question, according to the *JIRS* weighting scheme. The architecture of the monolingual *JIRS*-based system is illustrated in Fig. 2.

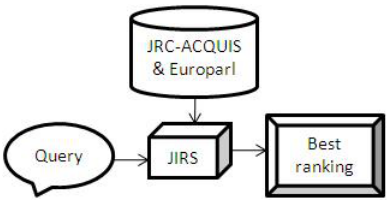

Fig. 2. Architecture of NLEL monolingual system

4.2 Multilingual approach

According to the excellent results obtained in the competition RespubliQA 2009 we decided to implement the multilingual approach also in RespubliQA 2010. This approach used the parallel collection to obtain a list of answers in different languages (Spanish, English, Italian, French and German). The idea of this approach is based on the implementation of 5 monolingual *JIRS*-based systems, one for each language, which process the set of questions in the respective language. For this purpose, we used a parallel sets of questions provided by the competition organisers. The final answer is selected as the one obtaining the best score; if the answer is not in the target language, the identifier of each paragraph (answer) is used to retrieve the aligned paragraph in the target language. The architecture of the multilingual *JIRS*-based system is illustrated in Fig. 3.

4.3 Monolingual - Stem approach

The Monolingual - Stem approach was inspired by the competition of the year 2009 [7], where the best baseline was established using, among others, a corpus

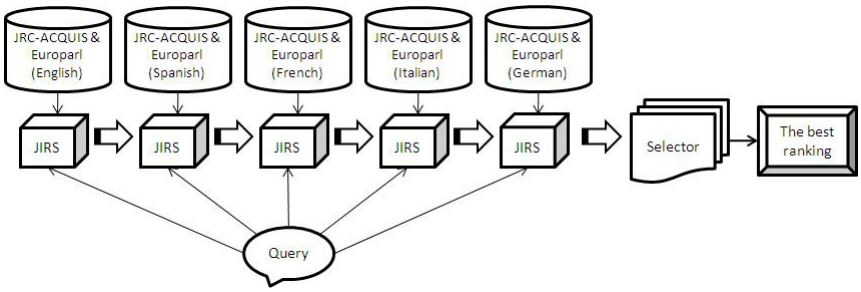

Fig. 3. Architecture of NLEL multilingual system

pre-processed with the *Stem* technique, the outline of that approach can be seen in Fig. 4

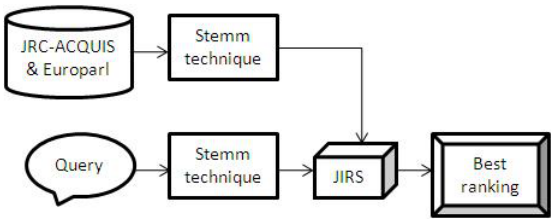

Fig. 4. Architecture of NLEL monolingual - stem system

4.4 Monolingual - BM25 approach

The Monolingual - BM25 approach, was inspired by the competition of 2009 [7], where the best baseline was established through the implementation of, among others, the BM25 technique to find the passages which are expected to be the answer to each question; the scheme that approach can be seen in Fig. 5

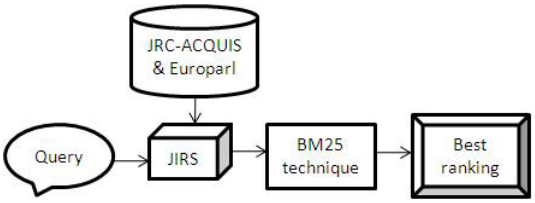

Fig. 5. Architecture of NLEL monolingual - BM25 system

5 Results

In Table 2 shows the results of the paragraph selection task.

Table 2. PS Task Results, Mono: participation uses monolingual approach, Multi: participation uses multilingual approach, Stem: participation uses stem pre-processing, BM25: participation uses BM25 post-processing, ANSWERED: number of questions answered, UNANSWERED: number of questions unanswered, ANSWERED R.C.: number of questions answered with right candidate answer, ANSWERED W.C.: number of questions answered with wrong candidate answer, UNANSWERED R.C.: number of questions unanswered with right candidate answer, UNANSWERED W.C.: number of questions unanswered with wrong candidate answer, UNANSWERED E.: number of questions unanswered with empty candidate, P.A.C.D.: Portion of answers correctly discarded

	EN		ES		FR		IT		DE	
	Mono	Stem	Mono	BM25	Mono	Multi	Mono	Multi	Mono	Multi
ANSWERED	196	198	194	200	191	197	196	199	183	200
UNANSWERED	4	2	6	0	9	3	4	1	17	0
ANSWERED R.C.	128	122	108	39	105	109	124	105	90	88
ANSWERED W.C.	68	76	86	161	86	88	72	94	93	112
UNANSWERED R.C.	2	0	1	0	2	0	2	0	2	0
UNANSWERED W.C.	2	2	5	0	7	3	2	1	15	0
UNANSWERED E.	0	0	0	0	0	0	0	0	0	0
ACCURACY	0.65	0.61	0.55	0.20	0.54	0.55	0.63	0.53	0.46	0.44
P.A.C.D.	0.50	1.00	0.83	0.00	0.78	1.00	0.50	1.00	0.88	0.00
C@1 MEASURE	0.65	0.62	0.56	0.20	0.55	0.55	0.63	0.53	0.49	0.44

In Table 3 shows the results of the answer selection task.

As shown in Table 2, the monolingual approach applied to each of the five languages returns acceptable results, especially in English and Italian. The implementation of the systems: multilingual, monolingual - stem and monolingual - BM25, decreased the overall performance of the system with respect to the monolingual approach that used the JIRS n-grams density weighting scheme. This result confirms that the n-grams density weighting scheme of JIRS fits particularly well the QA task, with respect to term-based weighting scheme, as observed in [1].

It is important to note that the multilingual approach was not able to repeat the results obtained in the RespubliQA-2009 competition. An analysis of the results in RespubliQA-2010 showed that the provided corpus is not perfectly aligned, as it can be observed from Tables 4 and 5: a passage with the same ID in the same document can be different for each of the studied languages. This problem is present in both the JRC-AQUIS and Europarl corpora.

Due to the scheme adopted for the multilingual approach it is necessary to work on a corpus with 100% accuracy in alignment; otherwise, the system is not able to obtain good results, as it can be seen in Table 2. Due to the fact that the

Table 3. AS Task Results, ANSWERED: number of questions answered, UNANSWERED: number of questions unanswered, ANSWERED R.C.: number of questions answered with right candidate answer, ANSWERED W.C.: number of questions answered with wrong candidate answer, ANSWERED M.: number of questions answered with missed candidate answer, ANSWERED I.: number of questions answered with inexact candidate answer, A.E.P.: Answer extraction performance

	EN	ES	FR	IT
ANSWERED	107	150	136	145
UNANSWERED	67	28	40	30
ANSWERED R.C.	10	12	4	6
ANSWERED W.C.	97	138	132	139
ANSWERED M.	20	21	13	18
ANSWERED I.	6	1	11	7
ACCURACY	0.05	0.06	0.02	0.03
C@1 MEASURE	0.07	0.07	0.02	0.03
A.E.P.	0.28	0.35	0.14	0.19

Table 4. Non parallel JRC-Aquis corpus example

File	Passage	Text
jrc31972L0199-de.xml	139	4. BESTIMMUNG DER PEPSINAKTIVITT
jrc31972L0199-it.xml	139	7.3 SE IL PALLONE DELL'APPARECCHIO DI...
jrc31972L0199-en.xml	139	7 . OBSERVATIONS
jrc31972L0199-es.xml	139	3.2 . Ácido clorhídrico 0,075 N .
jrc31972L0199-fr.xml	139	Dfinition : L'unité de pepsine est définie comme...

Table 5. Non parallel Europarl corpus example

File	Passage	Text
EP_TA-20081218-FR_cl.xml	127	4. souhaite vivement entamer des...
EP_TA-20081218-EN_cl.xml	127	4. Expresses its strong willingness to enter...
EP_TA-20081218-ES_cl.xml	127	3. Toma nota de la Comunicación de la...
EP_TA-20081218-DE_cl.xml	127	5. fordert, dass die gegenwrtige Krise nicht...
EP_TA-20081218-IT.xml	127	4. esprime la sua forte volontá di avviare...

answers for the Answer Selection task were extracted from the same passages retrieved in the basic multilingual approach, the results obtained for this task were also poor as shown in Table 3.

6 Conclusions

According to the experiments, the use of techniques such as BM25 and Stem, decrement the performance of JIRS tool for purposes of question answering tasks. It is verified through analysis, that problems with the alignment of the corpus provided poor performance resulting in the multilingual approach used. Additionally, Due to the poor result obtained with the multi-lingual approach, the extraction experiment response has similarly low results. In future work, we plan to implement a filter able to determine the paragraphs alignment of the corpus to improve the performance of multilingual approach.

Acknowledgements

We thank the TEXT-ENTERPRISE 2.0, MICINN (Plan I+D+i) research project (TIN2009-13391-C04-03).

References

1. Buscaldi D., Gmez J. M., Rosso P., Sanchis E. N-gram vs. Keyword-based Passage Retrieval for Question Answering. In: Evaluation of Multilingual and Multimodal Information Retrieval, Revised Selected Papers CLEF-2006, Springer-Verlag, LNCS(4730), pp. 377-384 (2007)
2. Buscaldi D., Rosso P., Gmez J.M., Sanchis E.: Answering Questions with an n-gram based Passage Retrieval Engine. In: Journal of Intelligent Information Systems, 34(2) pp. 113–134 (2009)
3. Correa, S., Buscaldi, D., Rosso, P.: Passage Retrieval and Intellectual Property in Legal Texts. In: FLACOS-2009, Toledo, Spain (2009)
4. Correa, S., Buscaldi, D., Rosso, P.: NLEL-MAAT at CLEF-ResPubliQA. In: Working Notes for the CLEF 2009 Workshop, Corfu, Greece (2009)
5. Gómez J.M., Buscaldi D., Bisbal E., Rosso P., Sanchis E. QUASAR: The Question Answering System of the Universidad Politecnica de Valencia. In: CLEF 2005 Proceedings. Springer Verlag, LNCS(4022), Vienna, Austria.
6. Gómez, J.M., Montes, M., Sanchis E., Rosso, P.: A Passage Retrieval System for Multilingual Question Answering. In: Proc. 8th Int. Conf. on Text, Speech and Dialogue, TSD-2005, Springer-Verlag, LNAI (3658), pp. 343-350 (2005)
7. Pérez, J., Garrido, G., Rodrigo, A., Araujo, L., Peñas, A.:Information Retrieval Baselines for the ResPubliQA Task. In: Working Notes for the CLEF 2009 Workshop,Corfu, Greece, (2009).