The Multilingual Question Answering Track at CLEF

Bernardo Magnini[1], Danilo Giampiccolo[2], Lili Aunimo[3], Christelle Ayache[4], Petya Osenova[5], Anselmo Peñas[6], Maarten de Rijke[7], Bogdan Sacaleanu[8], Diana Santos[9], Richard Sutcliffe[10]

[1]ITC-irst, Centro per la Ricerca Scientifica e Tecnologica, Italy, *magnini@itc.it*
[2]CELCT, Italy, *giampiccolo@celct.it.*
[3]University of Helsinki, Finland, *aunimo@cs.helsinki.fi*
[4]ELDA/ELRA, France, *ayache@elda.org*
[5]BTB, Bulgaria, *petya@bultreebank.org*
[6] UNED, Spain, *anselmo@lsi.uned.es*
[7]University of Amsterdam, The Netherlands, *mdr@science.uva.nl*
[8]DFKI, Germany, *Bogdan.Sacaleanu@dfki.de*
[9]Sintef, Norway, *Diana.Santos@sintef.no*
[10]University of Limerick, *Ireland, Richard.Sutcliffe@ul.ie*

Abstract

This paper presents an overview of the Multilingual Question Answering evaluation campaigns which have been organized at CLEF (Cross Language Evaluation Forum) since 2003. Over the years, the competition has registered a steady increment in the number of participants and languages involved. In fact, from the original eight groups which participated in 2003 QA track, the number of competitors in 2005 rose to twenty-four. Also, the performances of the systems have steadily improved, and the average of the best performances in the 2005 saw an increase of 10% with respect to the previous year.
This report describes the task in general and, in more detail, the methodology used for preparing the data-sets as well as the resources available for training systems. The approaches and results achieved by participating groups are also briefly discussed.

1. Introduction

Despite the attention that Question Answering (QA) has received in recent years, multilinguality has been outside the mainstream of QA research, which is still mainly focused on the English language. Multilingual QA has emerged only in the last few years as a complementary research task, and represents a promising direction for at least two reasons. First, it allows users to interact with machines in their native languages, thus contributing to easier, faster, and more reliable information access. Secondly, cross-language capabilities enable QA systems to access information stored only in language-specific text collections.

For these reasons, in 2003 a pilot evaluation campaign was launched under the CLEF umbrella for the evaluation of QA systems for languages other than English. The general aim of the Multilingual Question Answering Track (QA@CLEF, http://clef-qa.itc.it) was to set up a common and replicable evaluation framework to test both monolingual and cross-language Question Answering systems that process queries and documents in several European languages. In addition, the QA@CLEF initiative intends to stimulate attention to a number of challenging issues for research in multilingual QA, including searches in multilingual document collections, collection and combination of answers found in documents from different languages, use of heterogeneous multilingual data collections (such as the Web and XML data) for answer generation, and interpretation of questions in different languages. Over the years, the series of QA competitions at CLEF has registered a steady increment in the number of participants and languages involved. In fact, in the first 2003 campaign, eight groups from Europe and North America participated in nine tasks: three monolingual; Dutch, Italian and Spanish, and five bilingual tasks, where questions were formulated in five source languages -Dutch, French, German, Italian- and answers were searched for in an English corpus collection. In 2004, eighteen groups took part to the competition, submitting forty-eight runs. Nine source languages –Bulgarian, Dutch, English, Finnish, French, German, Italian, Portuguese and Spanish- and seven target languages (there was no corpus available for Bulgarian and Finnish, therefore, they were not included) were considered in the task. In 2005, the number of participants increased to twenty-four, sixty-seven runs were submitted and ten source languages (the same nine languages used the previous year plus Indonesian) and nine source languages (the same that were used previously except for Indonesian which had no corpus) were exploited in eight monolingual and seventy-three cross-language tasks.
Novelties included the type of questions used in the exercise and the metrics used in the evaluation.

This paper is structured as follows: Section 2 describes the task, in terms of the questions which are posed to systems, the expected answers and the evaluation metrics adopted for assessment; Section 3 reports on the procedure adopted for the construction of the question set; Section 4 states how many monolingual and cross-lingual tasks were activated in 2005, as well as the number of participating systems; Section 5 reports on the main results achieved by systems and section 6 describes the important role of multilingual resources in training question answering systems.

	BG target		DE target		EN target		ES target		FI target		FR target		IT target		NL target		PT target	
	R	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**
BG source	**2**	**2**	-	-	1	1	-		-	-	-	-	-	-	-	-	-	-
DE source	-	-	**3**	**2**	1	1			-	-	-	-	-	-	-	-	-	-
EN source	-	-	3	2			3	2	-	-	1	1	-	-	-	-	1	1
ES source	-	-	-	-	1	1	**13**	**7**	-	-	-	-	-	-	-	-	-	-
FI source	-	-	-	-	2	1	-	-	**2**	**1**	-	-	-	-	-	-	-	-
FR source	-	-	-	-	4	2	-	-	-	-	**10**	**7**	-	-	-	-	-	-
IN source					1	1												
IT source	-	-	-	-	2	1	2	1	-	-	1	1	**6**	**3**	-	-	-	-
NL source	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**3**	**2**	-	-
PT source	-	-	-	-	-	-	-	-	-	-	1	1	-	-	-	-	**4**	**3**

Table 1: Participants (P) and runs (R; monolingual in bold) at QA@CLEF 2005.

2. Task Definition

In all the the campaigns which have taken place so far, the main task has basically remained unchanged: two-hundred questions are provided as an input, and exact answer-strings is required as an output. The target corpora in all the languages are collections of newspapers and news agencies' articles, whose texts had been SGML tagged. Each document has a unique identifier (docid) that systems have to return along with the answer in order to be able to support it. The corpora, released by ELDA, are large, unstructured, open-domain text collections.

Participants are allowed to submit only one response per question and two runs per task, which are judged by human assessors according to correctness and exactness. An answer is correct when it is clear and pertinent, and is exact when it provides nothing more or less than the required amount of information. In the last two campaigns, only exact answers were allowed, and the responses were judged as Right, Wrong, ineXact or Unsupported (when the answer-string contained a correct answer but the returned docid did not support it).

A partial analysis of the inter-tagger agreement shows that exactness still poses a major problem in evaluation in the evaluation of responses. Disagreement between judges is mostly due to this parameter.

In 2004, definition questions were introduced for the first time, even though they were considered particularly difficult because they often raise problems in the assessment of exactness. Surprisingly, they generally scored quite well, thus proving that they are less of a challenge than previously thought. This is probably due to the fact that the answer often contained the extension of an acronym (ex. for organization) or the apposition of a proper name (ex. for people).

The main measure used for the evaluation is accuracy, i.e. the fraction of right answers. The answers are usually returned unranked (i.e. in the same order as in the test set), but a confidence value, which could range from 0 to 1, may be added to each string and be used to calculate an additional Confidence-weighted Score (CWS).

3. Test Set Preparation

As mentioned already, the track has steadily grown during the years, and the 2005 campaign was the biggest: nine target languages and ten source languages were used to perform eight monolingual and seventy-three cross-language tasks.

Over the years, a procedure for the preparation of the test set has been consolidated (Magnini 2004). The questions in the test sets address large open domain corpora, mostly represented by the same comparable document collections: *NRC Handelsblad* (years 1994 and 1995) and *Algemeen Dagblad* (1994 and 1995) for Dutch; *Los Angeles Times* (1994) and *Glasgow Herald* (1995) for English; *Le Monde* (1994) and *SDA French* (1994 and 1995) for French; *Frankfurter Rundschau* (1994), *Der Spiegel* (1994 and 1995) and *SDA German* (1994 and 1995) for German; *La Stampa* (1994) and *SDA Italian* (1994 and 1995) for Italian; *PÚBLICO* (1994 and 1995) and *Folha de Sao Paulo* (1994 and 1995) for Portuguese; and *EFE* (1994 and 1995) for Spanish. In 2005, two new corpora were added, *Aamulehti* (1994-1995) for Finnish, and *Sega* and *Standard* for Bulgarian (2002). Unlike the other collections, which cover the same time span, the Bulgarian corpora dates back to 2002, and therefore, the information that it contains is only partially comparable with the other corpora.

From these news collections, 100 questions are produced in each target language and at least one answer is searched for in relevant documents. The questions are then translated into English, in order for them to be understood and reused by the other groups. If possible, the difficulty of the test sets is balanced, according to such different answer types as TIME, MEASURE, PERSON, ORGANISATION, LOCATION, and OTHER.

Once the questions have been formulated, translated into English and collected in a common XML format, native speakers of each source language, who have a good command of the English language, translate the English versions of the other questions with adherence to the original versions. The process is extremely challenging as there are always many cultural and liguistic discrepancies.

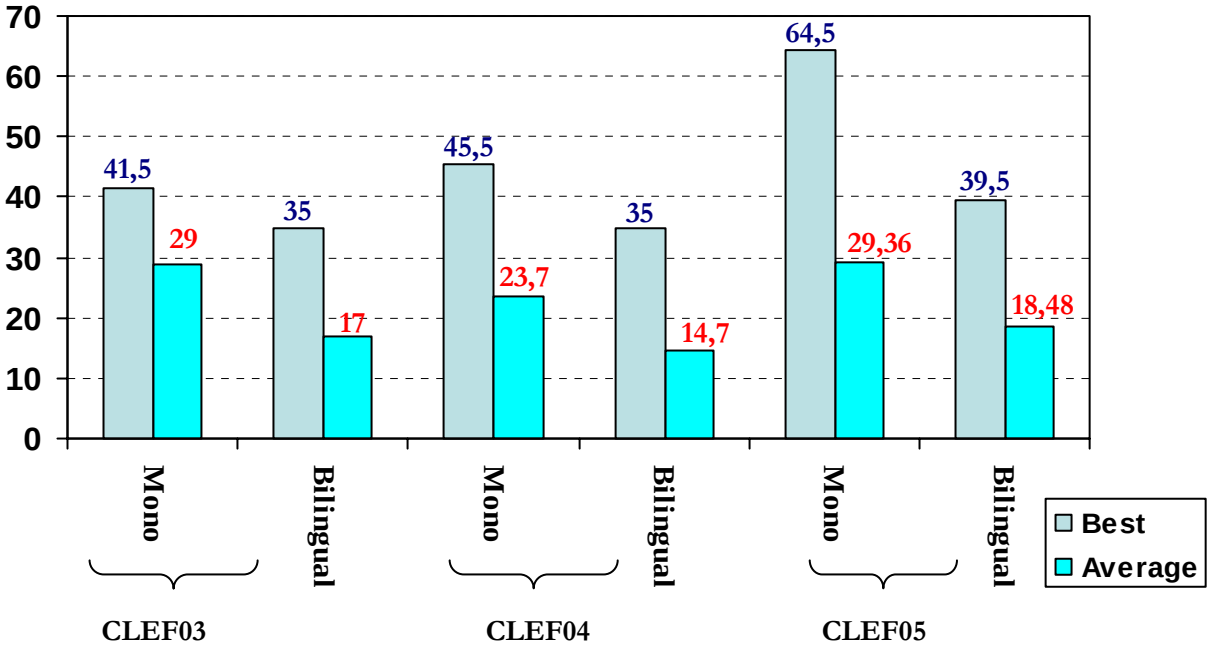

Figure 1: Best and average results in the QA@CLEF campaigns

Anyway, the upside is that manual translation reveals cross-cultural as well as cross-language problems, since QA systems are designed to work in the real world.

Finally, one hundred additional questions are selected from the other source languages and are manually verified and searched for answers in the corpus of the respective language, so that, at the end, each language has two-hundred questions.

4. Participants

The positive trend in terms of participation registered in 2004 was confirmed in the 2005 campaign. From the original eight groups which participated in the 2003 QA task, submitting a total of nineteen runs in nine tasks, in 2005, the number of competitors rose to twenty-four representing an increase of 33% with respect to the previous year, when eighteen groups took part in the exercise. The total of submitted runs was sixty-seven.

Most participants in the 2005 competition were from Europe, but groups from both America (Mexico and Canada), and from Asia (Indonesia) were also present.

As shown in Table 1, the systems were tested on only twenty-two of the eighty-one activated tasks (the blank cells represent non-activated tasks).

As in the 2004 campaign, monolingual English was discarded because the task has been examined enough in TREC campaigns. As far as Indonesian is concerned, one task using English as a target was set up.

In the last campaign, the nine monolingual tasks (in bold in the table) were tested by at least one system, with French (FR) and Spanish (ES) as the most preferable languages. As far as bilingual tasks are concerned, fifteen participants chose to test their systems in a cross-language task. English was, as usual, the most frequent target language, having been used in eight cross-lingual tasks by nine participants. Spanish was chosen as a target in a cross-language task by three groups, and so was French. Only one system attempted a cross-language task with Portuguese (PT) as a target, i.e. EN-PT. The other languages was not used for the bilingual bilingual tasks.

5. Results

In comparison to the previous editions, the performances of the systems in the 2005 campaign showed a general improvement (see Fig. 1), although a significant variation remained among target languages. In fact, in 2004, the best performing monolingual system (irrespective of target language) answered 45.5 % of the questions correctly, while the average of the best performances for each target language was 32.1%. In 2005, the best performing monolingual system, irrespective of the target language, answered 64.5 % of the questions correctly (in the monolingual Portuguese task), while the average of the best performances for the target languages was 42%. Comparatively, the cross-language subtasks recorded considerably poorer performances.

A comparison of the best performances for each target language in 2004 and 2005 is shown in Fig. 2, along with 'combination', which represents the score of a virtual system that would be able to return the best answer for each question, choosing among those given by all participating systems.

In addition to accuracy, the organizers also measured the relation between the correctness of an answer and the confidence stated by it, showing that the best systems did not always provide the most reliable confidence score.

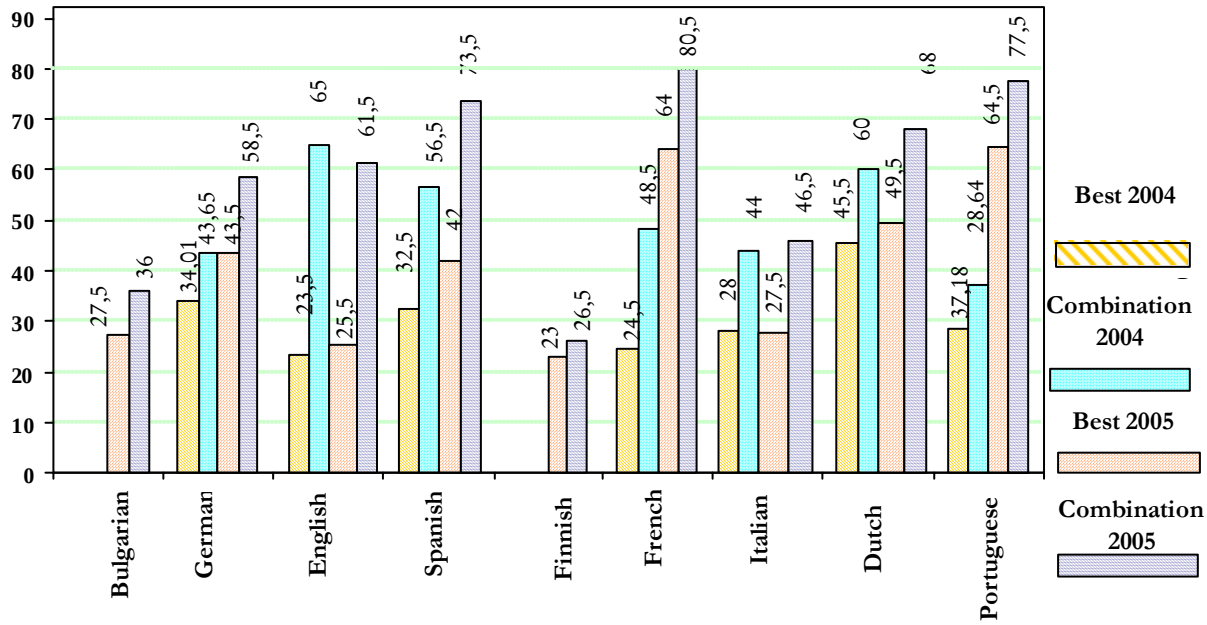

Figure 2: Best Results and Combinations in QA@CLEF 2004 and 2005

6. Resources for Multilingual QA

The collection of questions and answers used in the various editions of QA@CLEF have been collected in a resource, which is available on the web site. The most recent dataset, prepared for the 2005 campaign and called *Multi9-05*, is presented in XML format and is made up of 205 definition questions and 695 factoid, which are quite well balanced according to the type and are divided as follows: 110 MEASURE; 154 PERSON; 136 LOCATION; 103 ORGANISATION, 107 OTHER, 85 TIME.

7. Conclusions

This paper presented the Multilingual Question Answering evaluation campaign organized at CLEF. In the last three years, QA@CLEF has seen an increase in the number of participants and also in the number of languages involved. The rise in the number of participants revealed interesting comparisons among different systems that participated in the same task and compensated one of the drawbacks of the previous campaign. It is also relevant that in 2005, a task with Bulgarian as a target -the language of a new EU member country- was activated, together with a pilot cross-language task with Indonesian as source and English as target.

Since the implementation of the task is now well into its third year, it has been tested thoroughly. Even though it involves nine different institutions from nine different countries, which guarantee their support on a voluntary basis, it has shown that it is able to support the high number of exchanges required for the organization of the task, especially considering the fact that all the entities involved in QA@CLEF are not obligated to participate.

On a more critical note, it clearly appears from a general analysis of the results that, at this stage, Question Answering techniques for European languages demand better NLP tools and resources for the respective languages, as the QA task itself is mainly based on such tools and resources. Furthermore, in a cross-language perspective, the integration of such resources among the different languages is also crucial.

Finally, it must be acknowledged that QA@CLEF, having (at least partially) achieved its goal to promote Question Answering for European languages, now represents quite a large scientific community in Europe and is ready to propose its own ideas for QA, thus paving the way for successive multilingual QA systems.

8. References

The CLEF QA Track coordinators. QA@CLEF 2005 Guidelines, 2005. http://clef-qa.itc.it/2005/guidelines.html.

Herrera, J., Peñas, A., Verdejo, F. (2005). Question Answering Pilot Task at CLEF 1004. In *Proceedings of CLEF 2004. Lectures Notes in Computer Science*. Springer Verlag, (3491), pp. 581-590.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R. (2004). Overview of the CLEF 2004 Multilingual Question Answering Track. In C. Peters, editor. *Results of the CLEF 2004 Cross-Language System Evaluation Campaign.* Bath, U.K: Working Notes for the CLEF 2004 Workshop.

Spark Jones, K. (2003). Is Question Answering a Rational Task? In R. Bernardi and M. Moortgat, editors. *Questions and Answers: Theoretical and Applied Perspectives. Second CoLog NET-ElsNET Symposium*, pp. 24-35.