Evaluating Multilingual Question Answering Systems at CLEF

Pamela Forner[1], Danilo Giampiccolo[2], Bernardo Magnini[3], Anselmo Peñas[4], Álvaro Rodrigo[5], Richard Sutcliffe[6]

1-2 Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy
3 Fondazione Bruno Kessler (FBK), Trento, Italy
4-5 NLP & IR Group, UNED, Madrid, Spain
6 University of Limerick, Limerick, Ireland

forner@celct.it, giampiccolo@celct.it, magnini@fbk.eu, anselmo@lsi.uned.es, alvarory@lsi.uned.es,
richard.sutcliffe@ul.ie

Abstract

The paper offers an overview of the key issues raised during the seven years' activity of the Multilingual Question Answering Track at the Cross Language Evaluation Forum (CLEF). The general aim of the Multilingual Question Answering Track has been to test both monolingual and cross-language Question Answering (QA) systems that process queries and documents in several European languages, also drawing attention to a number of challenging issues for research in multilingual QA. The paper gives a brief description of how the task has evolved over the years and of the way in which the data sets have been created, presenting also a brief summary of the different types of questions developed. The document collections adopted in the competitions are sketched as well, and some data about the participation are provided. Moreover, the main evaluation measures used to evaluate system performances are explained and an overall analysis of the results achieved is presented.

1. Introduction

Under the promotion of the TREC-8 (Voorhees and Tice, 1999) and TREC-9 (Voorhees, 2000) Question Answering tracks, research in Question Answering (QA) received a strong boost. The aim of the TREC QA campaigns was to assess the capability of systems to return exact answers to open-domain English questions. However, despite the great deal of attention that QA received at TREC, multilinguality was outside the mainstream of QA research.

Multilingual QA emerged as a complementary research task, representing a promising direction for at least two reasons. First, it allowed users to interact with machines in their native languages, contributing to easier, faster, and more equal information access. Second, cross-lingual capabilities enabled QA systems to access information stored only in language-specific text collections.

Since 2003, a Multilingual Question Answering Track has been carried out at the Cross Language Evaluation Forum (CLEF) [1]. The introduction of multi-linguality represented not only a great novelty in the QA research field, but also a good chance to stimulate the QA community to develop and evaluate multilingual systems.

During the years, the effort of the organizers was focused on two main issues. One concern was to offer an evaluation exercise characterized by cross-linguality, covering as many languages as possible. From this perspective, major attention was given to European languages, adding at least one new language each year. However, the offer was also kept open to languages from all over the world, as the inclusion of Indonesian shows.

The other important issue was to maintain a balance between the established procedure – inherited from the TREC campaigns – and innovation. This has allowed newcomers to join the competition and, at the same time, offered "veterans" more challenges.

An additional merit of the QA track at CLEF is the creation of reusable multilingual collections of questions and related answers, which represent a useful benchmark resource.

This paper is organized as follows: Section 2 gives a brief description of how the task has evolved over the years, the way in which the data sets have been created, giving also a brief overview of the different types of question developed, the document collections adopted and some data about participation; Section 3 gives a brief explanation of the different measures adopted to evaluate system performance; in Section 4 some results of the participants are discussed highlighting some important features; and finally in Section 5 some conclusions are drawn.

2. The QA Track at CLEF

The QA task consists of taking a short question and a document collection as input and producing an exact answer as output.

In the QA track at CLEF, the systems were fed with a set of questions and were asked to return one or more exact answers per question, – where exact means that neither more nor less than the information required is returned. The answer needed to be supported by the docid

[1] http://www.clef-campaign.org

of the document in which the exact answer was found, and depending on the year, also by portion(s) of text, which provided enough context to support the correctness of the exact answer. Table 1 summarizes all the novelties that have been introduced in the main task over the years of QA campaigns. Each year, a main task was proposed, which constantly evolved, becoming more and more challenging by addressing different types of questions and requiring different types of answer format as output.

	2003	2004	2005	2006	2007	2008	2009
Target lang.	3	7	8	9	10	11	9
Collection	News 1994		+ News 1995		+ Wikipedia Nov. 2006		JRC-Acquis
Nr of question	200						500
Type of question	200 Factoid		+ Temp. restrict +Defn	- Type of question + List	+ Linked question + Closed lists		- Linked + Reason + Purpose +Procedure
Supporting info	Document			Snippet			Paragraph
Size of answer	Snippet		Exact				Paragraph

Table 1: Evolution of the task of QA at CLEF campaigns

In all campaigns, the QA track was structured in both monolingual and bilingual tasks. The success of the track showed an increasing interest in both *monolingual non-English* QA – where questions and answers are in the same language – and in *cross-lingual* QA – where the question is posed in a language and the answer must be found in a collection of a different language.

	2003	2004	2005	2006	2007	2008	2009
QA Tasks	Multiple Language QA Main Task					ResPubliQA	
				Answer Validation Exercise (AVE)			GikiCLEF
				Real Time	QA over Speech Transcriptions (QAST)		
				WiQA		WSD QA	

Table 2: Pilot tasks at QA at CLEF campaigns over the years

QA at CLEF was also an opportunity to experiment with several pilot tasks, as Table 2 shows, i.e. Real Time Question Answering (Noguera et al., 2007), Answer Validation (Peñas et al., 2006), Question Answering over Speech Transcripts (Lamel et al., 2007), Word Sense Disambiguation for Question Answering, Question Answering using Wikipedia (Jijkoun & de Rijke, 2007), and GikiCLEF (Santos & Cabral, 2009).

The common goal of these pilot tasks was to investigate how QA systems and technologies are able to cope with different types of questions from those proposed in the main task, experimenting with different scenarios.

2. 1 Data Collection

The procedure for generating questions did not significantly change over the years. For each target language, a number of questions (ranging from 100 to 200 depending on the campaign) were manually produced, initially using the topics of the Ad-Hoc track at CLEF. The use of topics was originally introduced to reduce the number of duplicates in the multilingual question set. Together with the questions, a gold standard was also produced, by manually searching for at least one answer in a document collection. The questions were then translated into English, which acted as lingua franca, so that they could be understood and reused by all the other groups. Once the questions were collected in a common format, native speakers of each source language, with a good command of English, were recruited to translate the English version of all questions into their own languages, trying to adhere as much as possible to the original.

The introduction of back translation to create cross-lingual question-answer pairs – a paradigm developed in 2003 and used ever since – is one of the most remarkable features of QA at CLEF.

Questions were classified according to different types. In the first campaigns, only two types of questions were considered, i.e.:

– factoid questions, i.e. fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc., for example:

Ex. 1:

- *Who was Lisa Marie Presley's father?*
- *What year did the Second World War finish?*
- *What is the capital of Japan?*
- *What party did Hitler belong to?*

– definition questions, i.e. questions like "What/Who is X?”, for example:

Ex. 2:

- *Who is Lisa Marie Presley?*
- *What is Amnesty International?*
- *What is a router?*
- *What is a tsunami?*

In 2006, list questions were introduced for the first time. They consisted of both "closed lists", i.e. questions that require one answer containing a determined number of items (see Example 3), and “open lists”, where many correct answers could be returned (see Example 4).

Ex. 3:
Q: Name all the airports in London, England.
A: Gatwick, Stansted, Heathrow, Luton and City.

Ex. 4:
Q: Name books by Jules Verne.
A: A Journey to the Centre of the Earth, From the Earth to the Moon, Twenty Thousand Leagues Under the Sea, Around the World in Eighty Days, Eight Hundred Leagues on the Amazon.

In 2007, with the introduction of topic-related questions, the procedure followed to prepare the test set changed considerably. First of all, each organizing group, responsible for a target language, freely chose a number of topics. For each topic, one to four questions were generated. The topic-related questions consisted of clusters of questions which were related to the same topic. The requirement for related questions on a topic necessarily implies that the questions refer to common concepts and entities within the domain in question. Unlike in the previous campaigns, topics could be not only named entities or events, but also other categories such as objects, natural phenomena, etc.

Topics were not given in the test set, but could be inferred from the first question/answer pair. For example, if the topic was *George W. Bush*, the cluster of questions related to it could have been:

Ex. 5:
Q1: Who is George W. Bush?;
Q2: When was he born?;
Q3: Who is his wife?

The requirement that questions are grouped by topics implied that the questions related to the same topic could refer to common concepts and entities within the cluster. In a series of questions this is expressed especially by co-reference – a well known phenomenon within Natural Language Processing which nevertheless had not been considered in previous QA at CLEF exercises.

In 2009 significant changes were introduced in the task. The exercise consisted of extracting not an exact answer but an entire paragraph of text, containing the information needed to satisfy the query, from a set of legal European Union documents. Moreover, new types of questions were introduced, i.e.:

- Purpose questions, asking for the aim, goal or objective of something, for example:

Ex. 6:
Q: Why have the imports of live poultry from Romania been suspended?
P: (2) Commission Decision 2005/710/EC of 13 October 2005 concerning certain protection measures in relation to highly pathogenic avian influenza in Romania [5] provides that Member States are to suspend imports of live poultry, ratites and farmed and wild feathered game and hatching eggs of those species from the whole territory of Romania and of certain products from birds from parts of that territory.

- Procedure questions, asking for a set of actions which is the official or accepted way of doing something, for example:

Ex. 7:
Q: How do you find the maximum speed of a vehicle?
P: The maximum speed of the vehicle is expressed in kilometers per hour by the figure corresponding to the closest whole number to the arithmetical mean of the values for the speeds measured during the two consecutive tests, which must not diverge by more than 3 %. When this arithmetical mean lies exactly between two whole members it is rounded up to the next highest number.

2.2 Document Collections

Before 2009, the target corpora in all languages, released by ELRA/ELDA, consisted of large, unstructured, open-domain text collections which were comparable because they were made up of newspaper and news agency articles referring to the same time span (1994/1995). The texts were SGML tagged and each document had a unique identifier (docid) that systems had to return together with the answer, in order to support it.

The choice of a different collection was a matter of long discussion, copyright issues remaining a major obstacle. A step towards a possible solution was made by the proposal of the WiQA pilot task, which represented a first attempt to set the QA competitions in their natural context, i.e. the Internet. As nowadays such large information sources are available on the web, this was considered a desirable next level in the evolution of QA systems. An important advantage of Wikipedia was that it is freely available in all languages considered.

In 2007, the Wikipedia was also adopted in the main task, beside the data collections composed of news articles. The “snapshots” of Wikipedia were made available for download both in XML and HTML versions. The answers to the questions had to be taken from actual entries or articles of Wikipedia pages. However, the variations in the size of each Wikipedia, depending on the language, were problematic.

In 2009, a subset of the JRC-Acquis Multilingual Parallel Corpus was used. JRC-Acquis [2] is a freely available parallel corpus of European Union (EU) documents, mostly of a legal nature. It comprises the contents, principles and political objectives of the EU treaties; EU legislation; declarations and resolutions; international agreements; acts and common objectives. Texts cover various subject domains, including economy, health, information technology, law, agriculture, food, politics and more. JRC-Acquis is being used again in 2010,

[2] http://wt.jrc.it/lt/Acquis/

along with EUROPARL[3].

2.3 Participation and languages involved

At a first glance one can say that over the years the series of QA evaluation exercises at CLEF has registered a steady increment in the number of participants and languages involved, which is particularly encouraging as multilinguality is one of the main characteristics of these exercises. Table 3 gives an overview of participation, languages and runs, covering all the years of QA campaigns.

	Participants	**Submitted Runs**	**Monolingual Runs**	**Cross-lingual Runs**	**Activated Tasks**	**Tasks chosen by at least 1 participant**	**Target Languages**
2003	8	17	6	11	9	6	3
2004	18	48	20	28	56	19	7
2005	24	67	43	24	81	23	8
2006	30	77	42	35	24	24	9
2007	22	37	20	17	37	18	10
2008	21	51	31	20	43	20	11
2009	12	28	26	2	110	7	10

Table 3: Statistics about QA at CLEF campaign over the years

In the first campaign, eight groups from Europe and North America participated in nine tasks, of which six were enacted. Monolingual tasks were in Dutch, Italian and Spanish, with three bilingual tasks – French, Italian and Spanish into English.

In 2004, the CLEF QA community grew and eighteen groups tested their systems, submitting 48 runs. Nine source languages – Bulgarian, Dutch, English, Finnish, French, German, Italian, Portuguese and Spanish – and 7 target languages (all the source languages but Bulgarian and Finnish, which had no corpus available) were exploited to set up more than 50 tasks, both monolingual and bilingual. As can be noticed from the table not all the proposed tasks were then carried out by the participants.

In 2005, the positive trend in terms of participation was confirmed: the number of participants rose to twenty-four and 67 runs were submitted. Ten source languages – the same as those used in the previous year plus Indonesian – and 9 target languages – the same used as sources, except Indonesian which had no corpus available – were exploited in 8 monolingual and 73 cross-language tasks.

In 2006, a total of 30 participants was reached. Eleven source languages were considered – Bulgarian, Dutch , English, French, German, Indonesian, Italian, Polish, Portuguese, Romanian and Spanish. All these languages were also considered as target languages, except for Indonesian, Polish and Romanian. Twenty-four tasks were proposed, divided into 7 monolingual and 17 cross-lingual tasks.

After years of constant growth, the number of participants decreased in 2007 due to the new challenges introduced in the exercise. Also the number of submitted runs decreased appreciably, from a total of 77 – registered in the previous campaign – to 37.

In 2008, the number of participants remained almost the same as in 2007 even though the number of submitted runs increased to 51.

Year	**Monolingual Runs**	**Cross-Lingual Runs**
2003	IT(2), NL(2), SP(2)	ES-EN(2), FR-EN(5), IT-EN(2)
2004	DE(1), ES(8), FR(2), IT(3), NL(2), PT(3)	BG-EN(1),BG-FR(2), EN-FR(2), EN-NL(1), ES-FR(2), DE-EN(3), DE-FR(2), FI-EN(1), FR-EN(6), IT-EN(2), IT-FR(2), NL-FR(2), PT-FR(2)
2005	BG(2), DE(3), ES(13), FI(2), FR(10), IT(6), NL(3), PT(4)	BG-EN(1); DE-EN(1), EN-DE(3), EN-ES(3), EN-FR(1), EN-PT(1), ES-EN(1), FI-EN(2) FR-EN(4), IN-EN(1), IT-EN(2), IT-ES(2), IT-FR(2), PT-FR(1)
2006	BG(3), DE(6), ES(12), FR(8), IT(3), NL(3), PT(7)	EN-DE(2), EN-ES(3), EN-FR(6), EN-IT(2), EN-NL(3), EN-PT(3), ES-EN(3), FR-EN(4), FR-ES(1), DE-EN(1), ES-PT(1), IT-EN(1), PT-ES(1), RO-EN(2), IN-EN(1) PL-EN(1), PT-FR(1)
2007	DE(3), ES(5), FR(1), IT(1), NL(2), PT(7), RO(3)	DE-EN(1), EN-DE(1), EN-FR(1), EN-NL(2), EN-PT(1), ES-EN(1), FR-EN(2), IN-EN(1), NL-EN(2), PT-DE(1), RO-EN(1)
2008	BG(1), DE(6), ES(6), EU(1), FR(1), NL(2), PT(9), RO(4)	DE-EN(3), EN-DE(3), EN-EU(1), EN-ES(2), EN-FR(1), EN-NL(2), ES-DE(2), ES-EU(2), FR-ES(1), NL-EN(1), PT-FR(1), RO-EN(1)
2009	DE(2), EN(10), ES(6), FR(3), IT(1), RO(4)	EU-EN(2)

Table 4: Languages at QA@CLEF. Number of runs for each monolingual language and for each cross-lingual language pair.

2009 was the year of the first experimentations with a new document collection and a new domain. Participation further decreased probably due to the new challenges introduced. Monolingual tasks were chosen by most participants, who probably were not motivated enough to undertake a cross-language task.

3. Evaluation

Each participant is required to return for each question in the test set at least one answer and a text for supporting the correctness of the answer. Until 2005, the supporting information was the id of the document, while starting from 2006 each system had to return a supporting snippet (no more than 500 bytes) containing the answer.

[3] http://www.europarl.europa.eu

Each single answer was judged by human assessors, who assigned to each response a unique judgment. The possible judgments were:

- *Right* (R): the answer string consisted of nothing more than an exact answer and it was supported by the supporting text.
- *Wrong* (W): the answer string did not contain a correct answer.
- *Unsupported* (U): the answer was correct, but it was impossible to infer its correctness from the supporting text
- *IneXact* (X): the answer was correct and supported, but the answer string contained either more or less bits than the exact answer.

3.1 Evaluation Measures

Several evaluation measures have been used in the first 7 editions of QA@CLEF. In each competition, one main measure was selected to rank systems' results. Furthermore, various additional measures were used in order to provide more information about systems' performance.

Mean Reciprocal Rank (MRR) was employed in the first campaign as the main evaluation measure, while it remained as a secondary measure in the following editions when more than one answer per question was requested. *MRR* was applied when systems had to return up to three answers per question ranked by confidence, putting the surest answer in the first place. According to *MRR*, the score for each question is the reciprocal of the rank at which the first correct answer is given. Therefore, each question can receive either the value 1, 0.5, 0.333 or 0 (in the case where none of the three answers is correct). The final evaluation score is the mean over all the questions. *MRR* is related to the Average Precision used in Information Retrieval (IR) (Voorhees and Tice, 1999).

The most used evaluation measure in the CLEF QA tracks has been *accuracy*, which is the proportion of questions correctly answered. In the case where more than one answer is given to a question, *accuracy* takes into consideration only the first answer. *Accuracy* was used as the main evaluation measure from 2004 to 2008 (inclusive), while it was exploited as a secondary measure in 2009, where c@1 was introduced.

With *c@1,* all questions must have at least one correct answer in the document collections, and systems can either respond to a question, or leave it unanswered if they are not confident about finding a correct answer. In fact, *c@1* rewards systems' ability to maintain the number of correct answers, while reducing the amount of incorrect ones by leaving some questions unanswered. The main rationale behind *c@1* is that, in some scenarios (for example in medical diagnosis), to leave a question unanswered is preferable to giving an incorrect one. This is effectively a strategy of increasing precision while maintaining recall, an essential provision for any system which is going to be accepted by real users. The formulation of *c@1* is given in Equation (1), where:

n_R: number of questions correctly answered
n_U: number of unanswered questions
n: total number of questions

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n}) \qquad (1)$$

The rest of the measures used at CLEF QA tracks were focused on evaluating systems' self-confidence in the correctness of their responses. The first of these measures was Confidence Weighted Score (CWS) (Voorhees, 2002). In order to adopt this measure, systems had to order their answers from the most confident response, to the least confident one. *CWS* rewards a system for a correct answer early in the ranking, more than for a correct answer later in the ranking.

The formulation of *CWS* is given in Equation (2), where *n* is the number of questions, and *C(i)* (Equation (3)) is the number of correct answers up to the position *i* in the ranking. *I(j)* is a function that returns 1 if answer *j* is correct and 0 if not.

$$CWS = \frac{1}{n}\sum_{i=1}^{n}\frac{C(i)}{i} \qquad (2)$$

$$C(i) = \sum_{j=1}^{i} I(j) \qquad (3)$$

Another measure focused on the evaluation of systems' self confidence was *K1*, which was proposed in a pilot task in CLEF 2004 (Herrera et al., 2005). In order to apply *K1*, QA systems had to return a real number between 0 and 1 indicating their confidence in the given answer. 1 means that the system is totally sure about the correctness of its answer, while 0 means that the system does not have any evidence on the correctness of the answer.

K1 is based on a utility function that returns -1 if the answer is incorrect and 1 if it is correct. This positive or negative value was weighted with the normalized confidence self-score given by the system to each answer. The formulation of *K1* is shown in Formula (4).

$$K1 = \frac{\sum_{i \in \{correct\_answers\}} self\_score(i) - \sum_{i \in \{incorrect\_answers\}} self\_score(i)}{n} \qquad (4)$$

K1 ranks between -1 and 1. However, the final value of *K1* is difficult to interpret: a positive value does not indicate necessarily more correct answers than incorrect ones, but that the sum of scores of correct answers is higher than the sum of scores of incorrect ones.

4. Result Discussion

The QA campaigns can be divided into three eras:

- Era I: 2003-2006. Ungrouped mainly factoid questions asked against monolingual newspapers; Exact answers returned.
- Era II: 2007-2008. Grouped questions asked against newspapers and Wikipedias; Exact answers returned.
- Era III: 2009. Ungrouped questions against multilingual parallel-aligned EU legislative documents; Passages returned.

Since the task was quite different in each era, we need to consider the evaluation results separately.

In the first era, monolingual factoid QA showed a steady improvement, starting at 49% in the first year and increasing to 68% in the fourth (2006). Interestingly, the best system was for a different language in each of those years. The improvement can be accounted for by the adoption of increasingly sophisticated techniques gleaned from other monolingual tasks at TREC and NCTIR, as well as at CLEF. However, during the same time cross-lingual QA showed no improvement at all, remaining in the range 35-45%. The bottleneck for cross-lingual QA is Machine Translation and clearly the required improvement in MT systems has not been realized by participants in the task.

In the second era, the task became considerably more difficult because questions were grouped around topics and in particular because they were allowed to use co-reference. Monolingual performance increased from 54-64% during this time while cross-lingual performance decreased from 42% to 19%. These figures can be explained by the fact that the monolingual systems in each case were the same while the first cross-lingual system was from a particularly important group which has consistently achieved very good results at TREC. Unfortunately this group chose not to participate in 2008.

In the third era, the task reverted to one of paragraph retrieval while at the same time the questions and document collection became more difficult. Monolingual performance stayed at a similar level of 61% as did the cross-lingual figure at 18%.

Table 5 summarizes the results in terms of accuracy; these are given as the percent of questions which were answered correctly, to the nearest 1%. In 2003, three attempts were allowed at each question and if one of these was correct, the answer was "exactly right".

As regards language trends in the task (Table 4), the main interest has always been in the monolingual systems, with the majority of teams building a monolingual system in just their own language. Naturally, most groups are also capable of building a good English monolingual system, but these have not been allowed at CLEF except in 2009. However, cross-lingual runs from or to English are allowed, and as the table shows, most of the runs between languages are indeed either from English to the language of the team or the other way around. What follows from this is that a relatively high number of cross-language tasks are activated each year with a very small number of runs (often just one or two) being submitted for each.

Year	**Monolingual**			**Cross-Lingual**		
	Worst	Best	Ans	Worst	Best	Ans
2003	35% ES	49% IT	Exact *	11% FR-EN	45% IT-EN	Exact *
2004	9% ES	46% NL	Exact	7% BG-FR	35% EN-NL	Exact
2005	14% FR	65% PT	Exact	1% IN-EN	40% EN-FR	Exact
2006	0% PT	68% FR	Exact	4% FR-EN	49% PT-FR	Exact
2007	6% PT	54% FR	Exact	3% ES-EN	42% EN-FR	Exact
2008	5% RO	64% PT	Exact	1% DE-EN	19% RO-EN	Exact
2009	11% EN	61% EN	Para	16% EU-EN	18% EU-EN	Para

Table 5: Results at QA@CLEF. These are given as the percent of questions answered exactly right, to the nearest 1%. In 2003, three attempts were allowed at each question and if one of these was correct, the answer was "exactly right". For results in terms of the other measures C@1 (2009), CWS (2004-8), K1 (2005-7) and MMR (2003, 2006) see the individual overview papers.

If several systems perform the same task on the same language pair, direct comparison is of course possible. However, as discussed above, the nature of CLEF means that this is rarely possible. So, can performance on different tasks be compared? Up until 2009, each target language had its own document collection and corresponding set of questions which were then back-translated into the source languages. Thus all tasks of the form XX-YY (with a fixed YY) were answering the same questions (albeit in different source languages) against the same target collection in language YY. This made a measure of comparison possible, mainly in the case where YY was EN since this was a task which was within the means of most groups through their familiarity with English.

In order to take this comparison further, a new strategy was adopted in 2009 whereby a parallel aligned collection was used (Acquis) meaning that the questions and document collection were exactly the same for all monolingual tasks as well as all cross-lingual tasks.

Moreover, some interesting additional experiments were performed at UNED. Firstly, the document collections in all the various target languages were indexed by paragraph, using the same IR engine in each case. The queries in each language were then input to the corresponding IR system, and the top ranking paragraphs returned were used as 'baseline' answers – this was possible because the task that year was paragraph selection, not exact answer selection. Interestingly, many systems returned results which were worse than the baseline, a situation which probably arose because UNED tuned the parameters in their system very carefully.

In the second experiment, UNED compared the performance of the baseline systems across languages. Because all languages were answering the same questions on the same collection, this enabled them to estimate the intrinsic difficulty of the language itself. By applying the resulting difficulty coefficients to the various submitted runs, they were able to make more accurate comparisons between them.

5 Conclusions

Prior to QA at CLEF, almost all QA was in English. Since the task was started in 2003, numerous groups have participated and experiments have been conducted in many different language pairs. The result is that there are now several QA research groups in almost all the European countries and they have sufficient expertise to create systems which can perform complex tasks. In addition to numerous research innovations within systems themselves, there have also been steps forward in the evaluation process itself. These have included the use of several new evaluation measures, the progress towards comparison of systems in different languages, and the development of sophisticated tools for the organization of the tasks.

Another important output has been the multilingual test sets and their associated gold standard answers and document collections. These are made possible by the ingenious paradigm of back-translation which was introduced in 2003 and has been very successfully used at CLEF ever since. Moreover, all this material is available online allowing groups in future to re-use the data produced in order to develop and tune their systems.

Finally, what can be concluded from the results about the QA task itself? Generally, English factoid QA as investigated at TREC over the years is considered to be a solved problem which is no longer worth investigating. Following the activity at CLEF, performance of monolingual non-English systems has improved substantially, to the extent that they are approaching that of the best English systems. Now is the time, therefore, to look at different types of question and different task scenarios, a process which has already started in 2009 with ResPubliQA[4].

Concerning cross-lingual systems, their performance has not shown a comparable improvement over the years to that of monolingual ones because high-performance machine translation remains an unsolved problem, especially where named entities are concerned (e.g. ‘Sur les quais’ translates as ‘On the Waterfront’). Thus translation in the QA domain warrants further investigation if multilingual barriers to text processing are to be overcome.

6 Acknowledgements

The QA evaluation campaigns at CLEF is a joint effort of several institutions and people which collaborated at the creation of the data set in the various languages involved each year.

Also TrebleCLEF Coordination, within FP7 of the European Commission, Theme ICT-1-4-1 Digital Libraries and Technology Enhanced Learning (Contract 215231) support must be acknowledged for its fund for ground truth creation.

Álvaro Rodrigo has been supported by the Education Council of the Regional Government of Madrid and the European Social Fund.

7 References

Forner, P., Peñas, A., Alegria, I., Forascu, C., Moreau, N., Osenova, P., Prokopidis, P., Rocha, P., Sacaleanu, B., Sutcliffe, R., Tjong Kim Sang, E. (2008). Overview of the CLEF 2008 Multilingual Question Answering Track. In C. Peters, T. Mandl, V. Petras, A. Peñas, H. Müller, D. Oard, V. Jijkoun, D. Santos (Eds.), *Evaluating Systems for Multilingual and Multimodal Information Access*, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers.

Giampiccolo, D., Forner, P., Herrera, J., Peñas, A., Ayache, C., Forascu, C., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R. (2007). Overview of the CLEF 2007 Multilingual Question Answering Track. In: C. Peters, V. Jijkoun, T. Mandl, H. Müller, D.W. Oard, A. Peñas, D. Santos (Eds.), *Advances in Multilingual and Multimodal Information Retrieval*, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers.

Herrera, J., Peñas, A., Verdejo, F. (2005). Question Answering Pilot Task at CLEF 2004. *Multilingual Information Access for Text, Speech and Images. CLEF 2004*. Volume 3491 of Lecture Notes in Computer Science, 581--590.

Jijkoun, V., and de Rijke, M. (2007). Overview of the WiQA Task at CLEF 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, M. Stempfhuber (Eds.), *Evaluation of Multilingual and Multi-modal Information Retrieval,* 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers.

Lamel, L., Rosset, S., Ayache, C., Mostefa, D., Turmo, J., Comas P. (2007). Question Answering on Speech Transcriptions: the QAST Evaluation in CLEF. In: C. Peters, V. Jijkoun, T. Mandl, H. Müller, D.W. Oard, A. Peñas, D. Santos (Eds.), *Advances in Multilingual and Multimodal Information Retrieval*, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers.

[4] http://celct.isti.cnr.it/ResPubliQA

Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu, B., Sutcliffe, R. (2006). Overview of the CLEF 2006 Multilingual Question Answering Track. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, M. Stempfhuber (Eds.), *Evaluation of Multilingual and Multi-modal Information Retrieval*, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers.

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, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, B. Magnini (Eds.), *Multilingual Information Access for Text, Speech and Images*, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers.

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., de Rijke, M. (2003). The Multiple Language Question Answering Track at CLEF 2003. In C. Peters, J. Gonzalo, M. Braschler, M. Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems*, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers.

Noguera, E. Llopis, F. Ferrandez, A. Escapa, A. (2007). Evaluation of Open-Domain Question Answering systems within a time constraint. In *Advanced Information Networking and Applications Workshops*, AINAW '07. 21st International Conference.

Peñas, A., Forner,P., Sutcliffe, R., Rodrigo, Á., Forascu, C., Alegria, I., Giampiccolo, D., Moreau, N., Osenova, P. (2009). Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In Borri, F., Nardi, A. and Peters, C. (Eds.), *Cross Language Evaluation Forum: Working Notes of CLEF 2009,* Corfu, Greece, 30 September - 2 October.

Peñas, A., Rodrigo, Á., Verdejo, F. (2007). Overview of the Answer Validation Exercise 2007. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras, and D. Santos, (Eds.), *Advances in Multilingual and Multimodal Information Retrieval,* LNCS 5152, September 2008.

Peñas, A., Rodrigo, A., Sama, V., Verdejo, F. (2006). Overview of the Answer Validation Exercise 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, M. Stempfhuber (Eds.), *Evaluation of Multilingual and Multi-modal Information Retrieval*, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers.

Rodrigo, Á., Peñas, A., Verdejo, F. (2008). Overview of the Answer Validation Exercise 2008. In C. Peters, T. Mandl, V. Petras, A. Peñas, H. Müller, D. Oard, V. Jijkoun, D. Santos (Eds.), *Evaluating Systems for Multilingual and Multimodal Information Access*, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers.

Santos, D. and Cabral, L.M. (2009). GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia. In Borri, F., Nardi, A. and Peters, C. (Eds.), *Cross Language Evaluation Forum: Working Notes of CLEF 2009*, Corfu, Greece, 30 September - 2 October.

Turmo, J., Comas, P., Rosset, S., Lamel, L., Moreau, N., Mostefa, D. (2008). Overview of QAST 2008. In In C. Peters, T. Mandl, V. Petras, A. Peñas, H. Müller, D. Oard, V. Jijkoun, D. Santos (Eds.), *Evaluating Systems for Multilingual and Multimodal Information Access*, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers.

Turmo, J., Comas, P.R., Rosset, S., Galibert, O., Moreau, N., Mostefa, D., Rosso, P. Buscaldi, D. (2009). Overview of QAST 2009. In Borri, F., Nardi, A. and Peters, C. (Eds.), *Cross Language Evaluation Forum: Working Notes of CLEF 2009*, Corfu, Greece, 30 September - 2 October.

Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, D., Santos, D., Sutcliffe, R. (2005). Overview of the CLEF 2005 Multilingual Question Answering Track. In C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G.J.F. Jones, M. Kluck, B. Magnini, M. de Rijke (Eds.), *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum,* CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers.

Voorhees, E.M., Tice, D.M. (1999). The TREC-8 Question Answering Track Evaluation. In *Proceedings of the Eight Text Retrieval Conference (TREC-8).*

Voorhees, E. M. (2000). Overview of the TREC9 Question Answering Track. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)* .

Voorhees, E. M. (2002). Overview of the TREC 2002 Question Answering Track. In *Proceedings of the Eleventh Text Retrieval Conference (TREC-11).*