Spanish Question Answering Evaluation

Anselmo Peñas, Felisa Verdejo and Jesús Herrera

Dpto. Lenguajes y Sistemas Informáticos, UNED
{anselmo,felisa,jesus.herrera}@lsi.uned.es

Abstract. This paper reports the most significant issues related to the launching of a Monolingual Spanish Question Answering evaluation track at the Cross Language Evaluation Forum (CLEF 2003). It introduces some questions about multilingualism and describes the methodology for test suite production, task, judgment of answers as well as the results obtained by the participant systems.

1 Introduction

Evaluation forums as the Text REtrieval Conference (TREC[1]), NTCIR project[2] or the Cross-Language Evaluation Forum (CLEF[3]) have shown their capability to stimulate research, to establish shared working lines, and to serve as a meeting point for their respective communities. These forums permit the comparison of different systems evaluated under the same conditions. Thus, some evidences about which are better approaches can be extracted. In this kind of evaluation, test suites must be produced to serve as the common evaluation exercises for every system under competition. Test suites generation requires a considerable effort that is justified in such evaluation forums. At the end, these test suites remain as a very valuable resource for future systems evaluation.

Question Answering (QA) research has been promoted and evaluated in such way since TREC-8 in 1999. Now, the Cross-Language Evaluation Forum (CLEF 2003) has brought new challenges: to consider different languages than English and to perform translingual QA [3]. The UNED NLP Group (Spanish Distance Learning University), as Spanish member of the CLEF consortium, is responsible for the Spanish test suite generation in all the QA tasks that involve Spanish, and is also responsible for the results assessments when Spanish take part as target language. We report here the most significant issues related the Monolingual Spanish evaluation task which has been launched in CLEF 2003.

Sections 2 and 3 describe the usual methodology for a QA evaluation based on systems comparison and introduce the challenge of multilingualism. Sections 4 and 5 describe the production of the Spanish test suite. The task, the assessment process and the results for the first monolingual Spanish QA evaluation are described in sections 6, 7 and 8 respectively.

[1] http://trec.nist.gov
[2] http://research.nii.ac.jp/ntcir/index-en.html
[3] http://www.clef-campaign.org

2 Evaluation Methodology

Since a QA system must answer a question in natural language, the evaluation methodology has been the following:

1. *Test suite production*. Mainly, the compilation of the document collection and the formulation of some hundreds of questions over that collection.
2. *Participant systems answering*. All the systems must answer the questions and return their responses in a limited time.
3. *Judgment of answers by human assessors*. Systems answers are judged as correct, incorrect, non-exact, not supported by a document, etc.
4. *Measuring of systems behaviour*. Mainly, the percentage of questions correctly answered, the percentage of questions without answer correctly detected, and some measures such the Mean Reciprocal Rank (MRR) [6] or the Confidence-Weighted Score [8] aimed to give more value to the systems with more precise and confident answers.
5. *Results comparison*.

This methodology permits systems comparison but introduces some restrictions that must be considered in the tasks definition:

1. *Quantitative evaluation constrains the type of questions*. Answers must be valuable in terms of correctness, completeness and exactness in order to measure and compare systems behaviour. Thus, it is not possible to ask any kind of questions.
2. *Human resources* available for the test suite generation and the assessment of answers. Usually, this is an unfunded work that requires some volunteers, which determine not only the possible kind of questions we can evaluate, but also the number of questions and the number of answers per question we can allow.
3. *Collection*. Unfortunately, most of the times the use of a collection is determined by its availability. However, the collection determines the searching domain and so the systems behaviour. Looking for answers in a news collection is a different problem than looking for them in a patents collection. Also processing is different in specific domains or in unrestricted domains. Finally, the comparison between systems working over different languages requires, at least, the availability of comparable multilingual collections.
4. *Roadmap versus state of the art.* There is a good idea of what systems should do in future [1]. However, its necessary to determine when is possible to incorporate new features. Thus, the definition of the evaluation task become a compromise between what is desirable and what is realistic to expect from QA systems. Are systems already able to adjust their confidence in answer, to use encyclopaedic knowledge, to make inferences, to answer temporary questions, to evaluate consistency between different answers, to consider different sources and languages, etc.?
5. *Research direction*. There are some issues related to future system behaviour that are affected by the evaluation tasks definition. For example, systems are tuned according the evaluation measures in order to get better results. In this way, evaluation measures have evolved to give more value to the systems with desirable features (e.g. better answer validation). Another example that shows

how the evaluation task definition affects systems behaviour is the decision of permitting or not the use of external resources as the web, which could serve to improve systems results without improving their own processing skills.

These considerations are present in the evaluation task definition. Since 1999, QA at TREC has evolved increasing the collection size, the number, types and difficulty of questions, and restricting the number and exactness of answers. Systems have been able to adapt to these challenges and get better results in each participation. Now, CLEF 2003 has launched a new challenge: multilingualism.

3 The Challenge of Multilingualism

Multilingualism is one of the *research directions* that can be promoted according to the current state of the art in Question Answering. The definition of new evaluation tasks with multilingual features supposes a new challenge that must be accompanied by an appropriate methodology. From the evaluation point of view, multilingualism introduces new questions that affect the evaluation methodology:

- How to ensure that fully multilingual systems receive the best evaluation?
- What multilingual tasks can be proposed with the current state of the art?
- What is the possible roadmap to achieve fully multilingual systems?
- Which resources are needed for the evaluation purposes?

These questions are very interrelated and we must give answers carefully (although they will evolve during the next years). For example, to ensure that fully multilingual systems receive the best evaluation, we could have considered a proposal in the following way:

1. Build a unique multilingual collection, with documents in all the languages under evaluation.
2. Build a unique set of questions in different languages. For example, if we have 250 questions and 5 languages we can formulate 50 questions in each language.
3. Ensure answers in just one of the target languages. Otherwise, if all the questions are formulated in all the languages and they have answers in all the languages, then a monolingual system can achieve the same results as a fully multilingual system. For example, if we have 250 questions and 5 languages, we can ensure that 10 questions in language A would have answer only in documents of language B, for all the 25 combinations of two languages A, B.
4. Run a unique evaluation task for all the systems and compare their results. Those that work with more languages would be able to get more answers and better results.

This methodology introduces some important difficulties:

- How to ensure answers in only one language? If we use comparable multilingual collections then a very hard pre-assessment is needed. If we use collections with different domains for each language then results could be biased. We would have to find some criteria based on things like dates or locality of events.

- How to find and appropriated balance among all languages and the type and difficulty of questions? The evaluation could reward systems centred in one of the languages if it is easier to find more answers in one language than in the others.
- The human assessment of correctness and exactness of answers must be performed by native speakers. Since systems would give an answer in any language, the human assessment process needs additional coordination to send each answer to the appropriate human assessor.

However, with the current state of the art systems it is not realistic to plan an evaluation like this in a very short term: Are systems able to answer a question in any language? Are systems able to find answers in sources of any language? A naive approach consists in the translation of questions by means of an automatic Machine Translation system and then use a monolingual QA system. So, we can expect systems to process questions in several languages and find answers in a different one, but very few systems will deal with more than one target language in order to find answers in more than one different collection.

In this way, to perform a separate evaluation for each target language seems to be more realistic in the very short term, and avoids the mentioned difficulties. This has been the option followed by CLEF 2003, in which the central issue has been to develop a methodology for the production of questions in several languages. However, we must follow closely the systems evolution in order to introduce global measures rewarding systems that consider as many languages as possible.

4 Spanish Test Suite

4.1 Spanish Document Collection

The collection used in the Monolingual Spanish QA task 2003 corresponds to the Spanish collection of CLEF 2002 campaign. This document set contains more than 200.000 international news from EFE Press Agency during the year 1994 (about 500 Mb). News cover a wide range of topics (sports, society, politics, etc.) so it is considered an unrestricted domain collection.

4.2 Spanish Questions Set

The questions set has been produced in coordination with the Italian ITC-IRST, UNED (Spanish Distance Learning University) and the University of Amsterdam. As a result of this coordinated work, the DISEQuA[4] corpus [2] has been created with 450 questions and answers translated into English, Spanish, Italian and Dutch.

Before starting the generation of questions, the TREC 2002 set of questions was studied with the aim to determine their style and the difficulties to find the answers.

[4] This corpus is available at http://clef-qa.itc.it and http://nlp.uned.es/QA

73 questions were translated into Spanish and their answers were searched in the Spanish collection. We found that the Spanish document collection was large enough to find most of the answers. Since the TREC 2002 questions were public for potential participants, these Spanish translations were not used for CLEF 2003 edition.

The questions set production has followed a methodology in five steps:

1. *Production of 200 candidate questions in Spanish*. Candidate questions were formulated taking as starting point the topics produced in past editions of CLEF (Cross-Language Information Retrieval tasks at CLEF 2000, 2001 and 2002). In this way, candidate questions were produced without exploring the document collection, trying to avoid any influence in the questions formulation and wording. The type of questions corresponds to short and fact-based answers. Four people were involved in this work in order to include different styles in questions formulation.
2. *Selection of 147 questions with answer*. The answers for the 200 candidate questions were searched in the collection. A question has an answer in the collection if there is a document that contains and supports the correct answer without any inference implying knowledge outside the document. Finally, 147 questions with an answer in the document collection were selected and translated into English in order to share them with the Italian and Dutch QA coordinators.
3. *Processing of questions produced in Italian and Dutch*. A parallel process was followed in both Italian and Dutch languages, producing near 300 more questions translated into English. These 300 questions were translated into Spanish and, again, an answer for each one was searched in the collection. At this point, almost 450 different questions had been produced and translated into English, Spanish, Italian and Dutch. All of them have an answer in at least one collection of CLEF 2002 (except English). This is the corpus that we have called DISEQuA.
4. *Selection of 180 questions with known answer*. From the DISEQuA corpus, 180 questions with answers in the three collections were selected. The respective translation of these 180 questions were used in each of the three Monolingual QA tasks.
5. *Selection of the final 200 questions*. The final 200 Spanish questions set is composed by the 180 questions (in the Spanish version), and 20 more questions without known answer in the Spanish collection. These 20 questions have been used to evaluate systems capabilities to detect questions without answer. These final 200 questions are referred to facts (dates, quantities, persons, organizations, etc.)

4.3 Preliminary Assessment

Dates and numbers change across different news for the same event. Sometimes, the first information is incomplete or not well known yet. Sometimes, there is a changing process or an increasing count along several days. For example, the number of died people in one accident. In these cases, there is more than one answer supported by different documents. Assessors must evaluate an answer without any inference or use of information not contained in the supporting document. For example, some preliminary questions asked for events in 1994, but being the year of the news collection, this year doesn't appear explicitly in the document text and it had to be removed from the questions for the final version.

5 Dealing with Multilingualism in Questions Set Production

The production and translation of questions have some difficulties that we comment in the following subsections.

5.1 Several Spellings

Sometimes there are several possible spellings for an entity during the question translation process. One case corresponds to old or new writing styles. For example, the term “Malasia” corresponds to old style in Spanish, while the term “Malaisia” corresponds to the modern one. In these cases, when both expressions appear in the collection, the modern style has been chosen for production and translation of questions. Another case, are entities with two different *sounds* and both appear in the collection. For example, Oryx and Órice are both used in Spanish, but the Spanish-like sound corresponds to the second one. In these cases, the Spanish-like sound is chosen. When not further criteria are found, the most frequent translation in the collection is chosen.

5.2 Acronyms Translation

Some acronyms change across different languages. For example NATO and OTAN correspond to the English and Spanish versions respectively. In these cases, the acronym is translated. In some cases, there are frequent acronyms in English that correspond to entities that are not referred with acronyms in Spanish. For example, BSE (*Bovine Spongiform Encephalopathy*) is a frequent acronym in English, while in Spanish is more frequent the entire expression *Encefalopatía Espongiforme Bovina*, being not frequent their acronyms (either BSE or EEB). However, instead of using the most frequent expression, in these cases where the source question has an acronym, the translation into Spanish maintain the acronym but in the Spanish version.

5.3 Second Translations

Some final questions have been the result of two translations: one from the source language into English, and a second from English into the target language. English has been chosen as intermediate language for two reasons: First, to build a richer resource for ulterior multilingual QA evaluation, and second, to simplify the translation process between pairs of languages. However, each translation may modify slightly the original question and, finally, introduce some variation in meaning. This problem doesn't affect the monolingual task evaluation, but affects the quality of the question set as a resource for further multilingual QA evaluation. To avoid this problem, translators have considered both, the original question and the English version.

6 Monolingual Spanish Task

The guidelines[5] for the participants in CLEF 2003 were the same in all Monolingual Question Answering tasks (Spanish, Italian and Dutch). Systems could participate in one or both of the following subtasks: exact answer or 50 bytes long string answer. In the exact answer subtask the answer-string must contain exclusively a correct answer for the question. In the 50 bytes long string subtask, the correct answer must be a part of a 50 bytes-sized string, possibly containing irrelevant information. For example, for the question 'What is the capital of France?', either *paris* or *Paris* are always considered as correct answers (whenever the document supports the answer), while answers like:

- Paris is a very large city where
- 100 years ago, Paris was an

are consider correct only in the 50-bytes string subtask (whenever the document supports the answer).

Participants are provided with 200 questions intending to return short and fact-based answers. Then, they have to produce up to two runs without any kind of human intervention to obtain up to three answers per question and run. That means that each run contains one, two or three answers per question.

All the answers produced by participant systems are submitted into one file for each run, that responds to the following structure:

1. Each line of the file contains one single answer, then for each question will be one, two or three associate lines.
2. Each line is conformed by the following fields (in the same order that we quote them and separated by any amount of white space):

Field	Description
quid	Question number, provided by the organizers.
system run-tag	Unique identifier for a system and a run.
answer rank	Shows that the answers are ordered by confidence, and that the system places the surest response in the first position.
score	Integer or real number showing the system confidence in the answer. This field is not compulsory.
docid	Identifier of the supporting document, or the string 'NIL' to affirm that there is not answer in the colletion.
answer-string	Exact answer, or a string containing the answer (in 50-byte answer string task). If the field docid is 'NIL', this column is empty.

[5] Available at http://clef-qa.itc.it

For example, this is part of one response file for the Monolingual Spanish QA task:

```
0013 alicex031ms 1 3003 EFE19940525-14752 1990
0013 alicex031ms 2 2003 EFE19941003-00830 lunes
0013 alicex031ms 3 2003 EFE19940520-11914 1993
0014 alicex031ms 1 2008 EFE19940901-00341 23 millones
0014 alicex031ms 2 2008 EFE19940330-18839 24.854
0014 alicex031ms 3 2007 EFE19941228-14902 8.815.000
0015 alicex031ms 1 2019 EFE19940103-00540 Ejército Republicano Irlandés
0015 alicex031ms 2 2002 EFE19940428-16985 Sociedad Romana Construcciones Mecánicas
0016 alicex031ms 1 0 NIL
```

7 Assessment Process

Human assessors have evaluated the runs produced by the systems, in order to qualify each given answer by assigning them one of the following judgements:

Judgement	Description
Incorrect	The answer-string does not contain a correct answer or the answer is not responsive.
Unsupported	The answer-string contains a correct answer but the document returned does not support that answer.
Non-exact	The answer-string contains a correct answer and the document supports that answer, but the answer contains more than just the answer. (Just for the exact answer runs).
Correct	The answer string consists of exactly a correct answer (or contains the correct answer within the 50 bytes long string) and that answer is supported by the document returned.

A sample of judgements is shown in the following figure:

Question / Answer	Judgement	Comment
M SPA 0002 **¿Qué país invadió Kuwait en 1990?** 0002 alicex032ms 2 4010 EFE19940825-12206 **ONU**	Incorrect	
M SPA 0049 **¿Dónde explotó la primera bomba atómica?** 0049 alicex032ms 2 3012 EFE19941203-01729 **Hiroshima**	Unsupported	Is not possible to infer from the doc. if the given answer is correct
M SPA 0036 **¿En qué año cayó el muro de Berlín?** 0036 alicex032ms 1 2010 EFE19940107-02719 **noviembre de 1989**	Non-exact	The sub string 'noviembre de' exceeds but '1989' is correct
M SPA 0001 **¿Cuál es la capital de Croacia?** 0001 alicex032ms 1 2050 EFE19940127-14481 **Zagreb**	Correct	

The following subsections discuss some criteria, problems and findings during the assessment process.

7.1 Correct and Exact Answers

Correctness and exactness of answers are in the opinion of human assessors. A numerical answer is considered more responsive if it includes the unit of measure, but depending on the measure it can be considered as correct or not; for example, '13 euros' and '13' would be positively considered. In case of dates of specific events that ended in the past, both day and year are normally required except if the question refers only to the year; or assessors consider that the year is sufficient. When the system answer contains some misspelling, the supporting documents are explored and if they are the source of that misspelling, the answer is considered as correct.

7.2 NIL Answers

According to the response format, there is no way for systems to explicitly indicate that they don't know or can't find the answer for a question. A NIL answer means that the system *decides* there isn't an answer for that question in the collection. For this reason, a NIL answer is correct if neither human assessors nor systems have found any answer before or after the assessment process. If there is an answer in the collection, NIL is evaluated as incorrect.

7.3 Not Exact, Not Responsive and Not Supported Answers

The assessment process doesn't contemplate to give two qualifications for one answer. When the answer for a question presents simultaneously non-exact and not supported characteristics, it is necessary to choose a unique label. For example, to the question 'Where did the first atomic bomb explode?' one system gave the answer pair '*EFE19941020-11470 Hiroshima Nagasaki'*. This is not exact because the string exceeds the exact answer, and simultaneously, the answer is not supported by the indicated document, since it does not specify nor it is possible to be inferred that Hiroshima is the city where exploded the first atomic bomb. In these cases, we have to distinguish whether the system participates in the exact answer-string or in the 50-byte answer-string subtask. In the exact answer-string subtask, not exact answers must be evaluated as incorrect. In the 50-byte answer-string subtask the unsupported label is chosen for this case.

Analogously, if the answer is not supported and non-responsive it is qualified as incorrect. For example, to the same question above, one system gave the answer pair 'EFE19940613-07857 Japón' which is not completely responsive. Simultaneously, the indicated document does not specify nor it is possible to be inferred that Japan is the country where exploded the first atomic bomb. If the assessor decides that the answer is not responsive, it must be tagged as incorrect. Otherwise, if the assessor considers that the answer is responsive, it must be tagged as unsupported.

7.4 Assessors Discrepancies

Two different human assessors have searched the answers for each question. The first assessment was performed before systems response, in order to select a set of evaluation questions with a known answer. The second one was performed after systems response, during the assessment process of their answers. Assessors for both tasks may change and may have different criteria to decide whether a question has a supported answer or not. For example, the first assessor may consider that is not possible to answer one question without considering inferences and knowledge outside the supporting documents and then the question is tagged with NIL answer. The second assessor may find that a system gives a correct answer instead of NIL for that question. In this case, the initial NIL tag must change in order to consider correctly the system answer. Another more problematic example of discrepancy is the opposite, when the first assessor found a possible answer, a system gave NIL and the second assessor agrees that is not possible to answer the question without considering further inferences and knowledge outside the supporting documents. In this case, the two assessors must discuss it in common and agree (even with a third opinion) whether initial tag must be changed or not.

7.5 Errors in Questions

During the assessment process we detected that one of the questions had a translation mistake. The word 'minister' in question 84 was translated as 'president'. The question was originally 'Who is the Italian minister of Foreign Affairs?', but the Spanish version was '¿Quién es el presidente italiano de Asuntos Exteriores?'. This failure may confuse the participant systems and we had to decide whether to discard the question or not. About errors in questions we can find some experience in past editions of QA at TREC. In TREC QA 2001 [7] eight questions were removed from the evaluation, mostly due to spelling mistakes in the question. However, in TREC 2002 [8] another criterion was taken and they decided to evaluate all the answers despite the remaining errors, arguing that it is difficult to know when to call something an error and it is assumed that systems have to cope with certain user errors. Despite the translation mistake, the meaning of the question remains clear enough, so we decided to follow TREC QA 2002 criterion in this case, understanding that systems must be robust in certain degree. We also studied to count NIL as a correct answer since, strictly, there isn't an answer for this question. However, no systems gave NIL answer for this question and assessment remained as usual.

8 Results of the Monolingual Spanish Task

Although up to four groups expressed their intention of participation, the University of Alicante (UA, Spain) [4] is the unique team that has taken part in the monolingual Spanish QA task of this year. UA has submitted two runs for the exact answer subtask. The first run contains 547 answers and the second 546. The following table summarizes the assessment statistics for both runs:

UA System	First Run				Second Run			
Ranking	*1st*	*2nd*	*3rd*	*Total*	*1st*	*2nd*	*3rd*	*Total*
Correct	49	16	26	91	51	15	13	79
Unsupported	0	2	7	9	2	3	4	9
Non-exact	6	1	3	10	6	1	3	10
Incorrect	145	157	135	437	141	156	151	448
Total	*200*	*176*	*171*	*547*	*200*	*175*	*171*	*546*

Responses per question are shown in the following table. UA systems were able to give correct answers for the 40% and 35% of questions respectively.

UA System	First Run	Second Run
Queries with no correct answer	120 (60%)	130 (65%)
Queries with correct answer	80 (40%)	70 (35%)
Mean Reciprocal Rank	0.31	0.30

With regard to the NIL answers, the following table shows that NIL was returned 21 times, being correct 5 of them and incorrect the other 16. From the 20 questions without known answer (before and after the assessment process) 15 of them didn't receive the NIL answer, i.e. NIL was not detected and systems gave wrong answers instead of NIL.

UA System	First Run	Second Run
NIL returned as a response	21	21
NIL correctly detected	5	5
NIL incorrectly responded	16	16
NIL not detected	15	15

9 Conclusions

Question Answering in Spanish is right now an emerging area of interest [5]. The first evaluation of Spanish QA systems has been reported in this paper. This first experience shows the effort put in the generation of useful resources for future multilingual QA evaluation. It also permitted to establish a methodology and some criteria for both, the test suite production and the assessment process. Unfortunately, only one group could adjust their systems on time to take part in the competition. We hope that the results and the resources developed in this first experience will encourage groups to continue their work in order to participate in future editions. Useful resources for Spanish Question Answering are publicly available at http://nlp.uned.es/QA as, for example, the questions and answers in CLEF 2003.

Next edition of the Multilingual Question Answering at CLEF will extend the multilingual tasks to any combination pairs between Spanish, Italian, Dutch, and possibly it will include also English, French and German. Exact answers will be required and the type of questions will include definitions. The three monolingual tasks will continue, and so, the Spanish Question Answering evaluation task.

Acknowledgements

We would like to thank the people that made possible this work: Bernardo Magnini, Alessandro Vallin, Maarten de Rijke, Carol Peters, Donna Harman, Julio Gonzalo and José Luis Vicedo.

This work has been supported by the Spanish government (MCyT, TIC-2002-10597-E).

References

1. Burger, J. et al. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering. NIST. 2002.
2. Magnini, B. et al. Creating the DISEQuA corpus: a test set for Multilingual Question Answering. Evaluation of Cross-Language Information Systems, CLEF 2003. Lecture Notes in Computer Science. Springer-Verlag; 2004a.
3. Magnini, B. et al. The multiple language Question Answering Track at CLEF 2003. Evaluation of Cross-Language Information Systems, CLEF 2003. Lecture Notes in Computer Science. Springer-Verlag; 2004b.
4. Vicedo, J. L. SEMQA: A semantic model applied to Question Answering: Thesis, University of Alicante, Spain; 2002.
5. Vicedo, J. L. Rodríguez H. Peñas A. and Massot M. Los Sistemas de Búsqueda de Respuestas desde una perspectiva actual. Revista De La Sociedad Espańola Para El Procesamiento Del Lenguaje Natural. 2003; 31:351-367.
6. Voorhees, E. M. Overview of the TREC-8 Question Answering Track. Proceedings of Text Retrieval Conference 8. 2000.
7. Voorhees, E. M. Overview of the TREC 2001 Question Answering Track. Proceedings of Text Retrieval Conference 10. 2002.
8. Voorhees, E. M. Overview of the TREC 2002 Question Answering Track. Proceedings of Text Retrieval Conference 11. 2003.