A question–answer distance measure to investigate QA system progress

Guillaume Bernard[1], **Sophie Rosset**[1], **Martine Adda-Decker**[1] **and Olivier Galibert**[2]

[1]LIMSI. Paris. France, [2]LNE. Paris. France
{rosset,madda,gbernard}@limsi.fr, Olivier.Galibert@lne.fr

Abstract

The performance of question answering system is evaluated through successive evaluations campaigns. A set of questions are given to the participating systems which are to find the correct answer in a collection of documents. The creation process of the questions may change from one evaluation to the next. This may entail an uncontroled question difficulty shift. For the QAst 2009 evaluation campaign, a new procedure was adopted to build the questions. Comparing results of QAst 2008 and QAst 2009 evaluations, a strong performance loss could be measured in 2009 for French and English, while the Spanish systems globally made progress. The measured loss might be related to this new way of elaborating questions. The general purpose of this paper is to propose a measure to calibrate the difficulty of a question set. In particular, a reasonable measure should output higher values for 2009 than for 2008. The proposed measure relies on a distance measure between the critical elements of a question and those of the associated correct answer. An increase of the proposed distance measure for the 2009 evaluation as compared to 2008 could be established. This increase correlates with the previously observed degraded performances. We conclude on the potential of this evaluation criterion: the importance of such a measure for the elaboration of new question corpora for questions answering systems and a tool to control the level of difficulty for successive evaluation campaigns.

1. Introduction

The questions-answering (QA) task consists of providing short, relevant answers to natural language questions. QA research has focused on extracting information from text or spoken sources, providing the shortest relevant text in response to a question. For example, the correct answer to the question *Besides France and Germany, where have we seen cases of mad cow-like disease affecting goats?* is *Belgium*[1] instead of a list of documents. This simple example illustrates the two main advantages of QA over current search engines: First, the input is a natural-language question rather than a keyword query; and second, the answer provides the desired information content and not simply a potentially large set of documents or URLs that the user must plow through.

In the QA domain progress has been observed via evaluation campaigns ((Dang et al., 2007; Mitamura et al., 2008; Forner et al., 2008; Turmo et al., 2008)). The QAst (Questions-Answering on Speech Transcriptions) campaigns focus on evaluating QA systems on speech transcriptions. Oral sentences have different features than the written one (long sentences for instance), and the aim is to evaluate the systems on this type of data. Moreover, the system are evaluated on three different languages: French, English and Spanish.

In the QAst 2009 evaluation (Turmo et al., 2009), a new procedure for building the question corpus has been proposed. In the previous QAst evaluations (Turmo et al., 2008), the questions were created by the evaluators from the documents. In 2009, the objective was to build more spontaneous questions. Native speakers were requested to read excerpts of documents and to ask, using speech, questions about information related to but not included in these excerpts. Because of this new building procedure, the correct answer to a question can be potentially far away from the excerpt use to create the question, specially with the long sentences found in oral transcriptions. Thus, we aim to evaluate whether this new building procedure has an impact on the results obtained on the QAst 2008 campaign.

In this paper, we propose a new measure based on the distance between the answer to a question and its elements, to evaluate whether the difficulty of the task had changed as a result. First, we compare the results obtained on the 2008 and 2009 QAst evaluations. We then motivate and describe our measure, which is applied on the questions corpus of 2008 and 2009 for each language (French, English and Spanish). We analyze the results and finally we conclude on the potential of this measure to assist in the building of new questions corpus in evaluation campaigns.

2. Observations on QAst 2008 and 2009 results

A first observation comes from the general results obtained by all the participants: they all went down (Turmo et al., 2009). There was three similar tasks between the QAst 2008 and 2009 evaluations: question-answering on English EPPS data, Spanish EPPS data and French broadcast news. In 2009 two question sets were proposed: one with written questions and one with manually transcribed spoken questions. Table 1 shows the results obtained by the 2008 version of our systems and the 2009 update of the same systems on the test corpus of QAst 2008. The results on each of the tasks have improved with the 2009 version. The greater gap for the English and Spanish tasks can be explained in part because of the different type of data: English and Spanish tasks use a corpora built from European Parliament plenary sessions and the French task uses a broadcast news corpora.

[1] This question is extracted from the QAST 2008 development set and this is the corresponding answer found in the document collection.

	French	
	Acc(%)	Δ
2008	45	+5
2009	50	
	English	
	Acc(%)	Δ
2008	33	+19
2009	52	
	Spanish	
	Acc(%)	Δ
2008	33	+23
2009	56	

Table 1: Variation of the results on the test corpus of QAst 2008 between the 2008 and 2009 systems. The Δ measures the difference between the 2008 and 2009 systems results.

Table 2 shows the results obtained with our 2009 system on the QAst 2009 test corpus with written and spoken questions. There are almost no differences between the results on these two question types. However, there is a big loss compared to the results obtained on the QAst 2008 test corpus.

	French	
Modality	Acc(%)	Δ
written	28	0
spoken	28	
	English	
Modality	Acc(%)	Δ
written	27	-4
spoken	23	
	Spanish	
Modality	Acc(%)	Δ
written	36	0
spoken	36	

Table 2: Variation of the results between written and spoken questions on the QAst 2009 test corpus with the 2009 systems. The Δ measures the difference between the two modalities.

This loss is shown more clearly in Table 3 which compares the results obtained by the 2009 version of the systems on QAst 2008 and 2009 test corpus.

Moreover, all the other participants to both evaluation campaigns observed a general performance loss for their English system.

Table 4 shows the results obtained by the others participants on the test corpus of QAst 2008 and 2009 campaigns. For all the English systems, the loss goes from 5% to 10% absolute. One hypothesis could be that the modality of the questions corpus (written or oral) has an impact on the results of the systems. But Table 2 and Table 4 shows that the results obtained on the two modalities are quite similar.

	French	
	Acc(%)	Δ
QAst 2008 test corpus	50	-22
QAst 2009 test corpus	28	
	English	
	Acc(%)	Δ
QAst 2008 test corpus	52	-25
QAst 2009 test corpus	27	
	Spanish	
	Acc(%)	Δ
QAst 2008 test corpus	56	-20
QAst 2009 test corpus	36	

Table 3: Variation of the results on the QAst 2008 and 2009 test corpus with the 2009 version of the systems. The Δ measures the difference between the QAst 2008 and QAst 2009 results.

System	Questions	All	
		MRR	Acc
INAOE 2008	Written	0.38	33%
INAOE 2009	Written	0.36	28%
	Spoken	0.34	26%
UPC 2008	Written	0.37	34%
UPC 2009	Written	0.28	21%
	Spoken	0.12	8%

Table 4: Results for the other systems on English.

The same important differences in results are observed between the 2008 and 2009 results for the written modality.

Observing the two question sets (see (Turmo et al., 2009) for details), we noticed that the written questions were corrected versions of the spoken ones. In consequence we consider that the way the questions has been collected has had a more fundamental influence.

3. Comparison between 2008 and 2009 corpus

To comprehend these differences in performance, we compared the 2008 and 2009 test corpora. We believe that the performance loss between the 2008 and 2009 evaluations can be explained in part by a greater distance between the answers and the questions elements for the 2009 test data. Quantifying the difference required us to design a distance measure between the question elements as found in the documents and the answer. The aim is also to have a measure who can be used again on every questions corpus.

3.1. A distance measure for questions corpus

We aim to evaluate the distance between the elements of a question and its correct answer. In the QAst evaluation campaigns, only the correct answer (there can be several in some cases) is given, along with the document where this answer can be found. As such, we do not know the excerpts of the document used to create the questions. These excerpts contain the elements of the questions, or transformations of these elements, which were used to build the questions. Also, we know the document where the answer can be found, but there is often several occurrences of a same answer in a document. Because we do not know where the elements used to build the questions are, we need an approach who evaluate the global repartition of the occurrences of each elements and each answer to a question in a document.

For each question of the corpora we measured the *global* distance between the elements of the question and occurrence of the correct answer. The global distance is computed as the average of distances between the elements of the question found in the document and the answer. Only question elements considered important by our system are kept. The elements considered pertinent in the question are named entities (standard, extended and nonspecific) and multi-words expressions. In the following question, *Where did Missus Sennett criticize the Ombudsman ?*, three elements are considered important: *Missus Sinnott*, *criticize* and *Ombudsman*. Questions elements are either words or groups of words. Having the global distance for each occurrence of the correct answer to a question, the system choose the occurrence with the lowest distance as the distance of the question. This distance is measured in term of words.

The two following examples show how the global distance is computed for two questions. In the first example, the correct answer to the question *Which Belgian organization has been declared criminal?* is *Vlaams Blok*. We computed the distances between this answer and each important element of the question which are *Belgian*, *organization* and *criminal*. The corresponding distance values in words are 10, 1 and 2. The global distance for this question is 4.

Which **Belgian organization** *has been declared* **criminal***?*

The **Belgian** *Supreme Court has upheld a previous ruling that declares the Vlaams Blok a* **criminal organization** *and effectively bans it.*

The next example features a longer text segment. The correct answer to the question *Which political leader of Palestine died recently?* is *Arafat*. The important elements of the question are *died*, *Palestine*, *political* and *leader*. The corresponding distance values are 1, 10, 37 and 38, and the global distance is 21.

Which **political leader** *of* **Palestine died** *recently?*

The **death** *of Arafat means that we will now have a new election in* **Palestine**. *The European Union has told Israel that the dialog between the two countries is important to sign a truce. It is necessary to get a new* **political leader** *as soon as possible.*

3.2. Evaluation of the measure

The proposed distance measure was used to investigate the differences between the test corpus of the French, English and Spanish tasks of QAst 2008 and 2009. Table 5 shows the results of that analysis. AD is the average distance obtained for a questions corpus, and SD the standard deviation. A big gap can be noticed between the 2008 and 2009 data on the French and English sets. We see that the mean distance has a strong increase in the QAst 2009 test corpus compared to the previous year, especially on the French corpus. However, we a see a really strong decrease on the results for the Spanish task. As shown in Table 3 there was almost no differences between spoken and written modalities on the 2009 data, the measures do not appear in Table 5.

	French		
	AD	SD	Δ
2008	45	100	+98
2009	143	431	
	English		
	AD	SD	Δ
2008	97	284	+39
2009	136	310	
	Spanish		
	AD	SD	Δ
2008	381	851	-359
2009	22	73	

Table 5: Evolution of the Average Distance on each questions corpus between the 2008 and 2009 evaluations.

Figure 1 shows the distribution of the distances values obtained for each test corpus for the 2008 and 2009 evaluations. In order to have a better representation of the distribution of the distances, we split the values into nine categories, ranging from questions with a distance of zero to questions with a distance superior ton 500. The X axis represents the nine categories and the Y axis the number of questions with a certain distance value. As such, this figure shows for each corpus the number of questions in each categories. It allows us to see the evolution of a corpus from 2008 to 2009.

4. Discussion

As stated before, we believe that the way the questions were created for the QAst 2009 evaluation can partially explain the performance loss observed between the 2008 and 2009 evaluations. Because the speaker had to ask questions about information not contained in the text excerpts, we hypothesized that the distance between the correct answer and the elements of the question was different than in the 2008 evaluation. We built a distance measure to quantify the difference. The proposed distance measure allows to assess the evolution between the test sets of evaluations.

4.1. Correlation between the distance results and the evaluation campaigns results

Using this distance, we compared the test sets for the French, English and Spanish tasks of QAst 2008 and 2009.

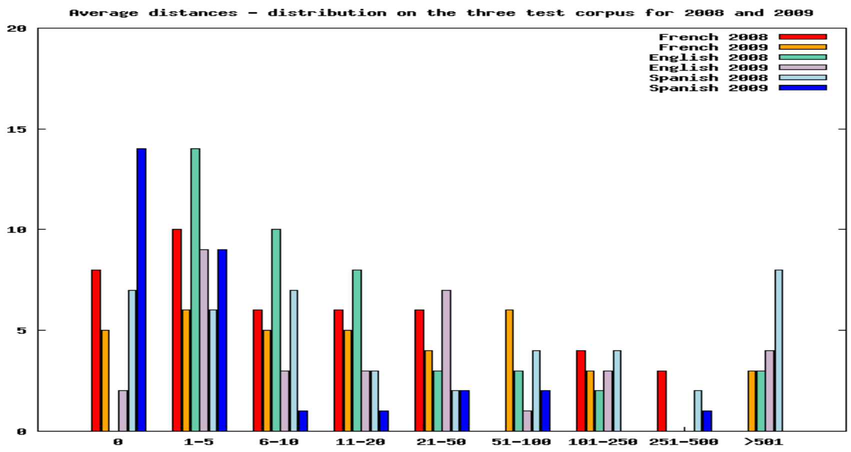

Figure 1: Average distance values - 2008 and 2009 test corpus

As shown in Table 5, the average distance has an increase on the French and English task. However, the Spanish task shows a strong decrease. For each of these three tasks, the standard deviation is very high, indicating that there are strong variations between the distances of a corpus. As such, the mean distance value is not a good indication of the distances of a corpus.

Figure 1 shows the distribution of the distance values for each test corpus. We can observe that while the mean distances for the English tests corpus are relatively similar between 2008 and 2009 compared to the French and Spanish corpus, the distribution indicates a strong dispersion. There is also a strong dispersion of the values for French and Spanish. For instance, the Spanish test corpus of 2008 has a lot of values with a great distance: 8 questions with a distance superior to five hundred, while there are 7 questions with a distance value of zero. On the other hand, the test corpus of 2009 has more values with a small distance: 14 questions have a distance of zero while there are no question with a distance superior to five hundred. These distributions of values clearly illustrate the evolution of the three test corpus between 2008 and 2009.

The average distance obtained on the French and English corpus may potentially explain the huge loss between the QAst 2008 and 2009 evaluations. The distance between the elements of a question and its answer have an important effect on the segmentation in snippets of the documents processed by the QA systems. This segmentation is a fundamental aspect of the way the QA systems work. The aim is to simplify the extraction of the answer. Depending on the system, a snippet can be a sentence or a group of lines. When working on oral transcriptions, the snippets are generally build using blocks similar to normal sentences. (Reyes-Barragan et al., 2009) segments the documents into passages of 24 words. Twelve of the words of adjacent passages are included. (Comas and Turmo, 2009) defines the passages as being segments where two consecutive keywords are separate by no more than *w* words. In (Bernard et al., 2009), the documents are selected using a search descriptor which contains the elements of the questions critical in finding the correct answer. The snippets are then extracted using a window's size fixed for each question type. The windows parameter is fixed by tuning on the corpus of the previous years. In (Reyes-Barragan et al., 2009) and (Comas and Turmo, 2009) approaches, the segmentation of the documents needs the question elements to be relatively close between them, or the sentences to have a fixed value. In (Bernard et al., 2009) the segmentation needs the data development corpus to be similar to the data of the test corpus. For the 2009 campaign, the development data used the corpus of the 2008 campaign. Alas, the questions of 2008 and 2009 were created differently.

As such, if the average distance of development data is different from the average distance of the test data, the window's size parameters will not be adapted to the test data. If the parameters are too low, the silence will increase: the window is too small so there are less snippets with an answer close to the elements of the question. On the other hand, if the parameters are too high, the noise will increase: there are a lot more of candidate answers and will be more difficult to evaluate each answer.

The window's size of the 2009 system was fixed using the corpus of the 2008 campaign. With this value, the balance between the noise and the silence is good. Figure 1 shows on the 2009 evaluation for Spanish that the distances are really low. Because the size of the window is too high, there are a lot of candidate answers to treat for the system. As such, it is more difficult to evaluate which one is the correct answer. It could explain why the results were not good on the 2009 test corpus. As such, the window's size parameters need to be fixed to a relatively low value in order to decrease the noise. In a similar way the distance values on the French and English 2009 corpus are much higher. This time the window's size is too small, and so there are less snippets to evaluate. This phenomena might also explain the loss observed on the 2009 campaign.

Finally, it seems that while the new way to build the questions corpus can explained the loss on the results obtained by each system on the 2009 evaluation, it is not the only criteria to explain these results. For instance, the type of data processed for each language could be another criteria: the French task is based on journalistic speeches (Broadcast News), while the English and Spanish are based on Parliamentary talks (EPPS). Features of a language can also be strong criteria to explain these differences in term of results (Bernard et al., 2010).

4.2. Usability for futures evaluations

This measure was used to evaluate the impact of the new way to build questions corpus of the QAst 2009 evaluation campaign. A strong loss between 2008 and 2009 evaluations was observed. The main hypothesis was that the new approach was at least one of the criteria explaining this loss. Because of the building procedure for the questions corpus, it was supposed that the distance between the elements of the question and its answer would increase. Higher distance values could explained the loss in the results between 2008 and 2009. As such, the average distance of a questions corpus was evaluate by our measure distance.

As discussed in 4.1., the results of this measure show that this new way of building questions does not always imply a greater distance between the elements of the question and its answer. While the average distance does increase on the French and English 2009 corpus, we observe a surprisingly strong decrease on the Spanish 2009 corpus. Moreover, it also shows that for each language, there is a difference between the 2008 and 2009 average distance. This difference is very strong in the case of the French and Spanish tasks. As such, it implies that the questions corpus of 2008 do not evaluate the systems on the same criteria than those of 2009.

This measure can be used as a criteria to evaluate the evolution of an evaluation campaign when building a new questions corpus. If the aim of a campaign is to evaluate the systems on the same features than the last iteration in order to analyze the progress made, this measure can provide interesting data on the average distance between elements of a question and its answer.

This approach was developed using the critical elements representation of the LIMSI system, but it can clearly be generalized on other system outputs. The measure only need to be adapted to another representation of the critical elements of a question to be used.

5. Conclusion and perspectives

There has been a huge loss in systems results between the QAst 2008 and QAst 2009 test corpus. One reason for this difference could be rooted in the new methodology used to build the questions corpus. To evaluate this hypothesis, a new measure was built to measure the average distance of each question of an evaluation corpus between the elements of the question and its answer. This measure was applied on the three common tasks of the 2008 and 2009 QAst evaluation, which featured three languages: French, English and Spanish. As stated in section 3., the methodology difference ended up with questions where the distances between the elements of the question found in the documents and the answer are much greater than for the 2008 evaluation only on the French and English task. On the contrary, the measure on the Spanish task shows a strong decrease of the average distance.

As such, while it can be supposed that this new way of building questions can imply an increase of the distance between the elements of the question and its answer, it is not always the case. As such, the decrease of the systems performances on the QAst 2009 evaluation can not be explained only because of a greater distance. Therefore, other measures are needed to identify the problems encountered into this evaluation. For instance, it could be interesting to evaluate the presence of referential expressions. Evaluating the features of the different languages could also explained the differences between Spanish and French and English.

Finally, this measure shows great potential into evaluating the differences between several iterations of an evaluation campaign. For instance, it can be used to evaluate the evolution of a campaign from one edition to another. This point is particularly important if the aim of the evaluation is only to evaluate the progression of the candidate systems, and not adding new features. As such, it could be interesting to developed other measures to evaluate the evolution of a campaign.

6. Acknowledgments

This work has been partially financed by OSEO under the Quaero program.

7. References

Guillaume Bernard, Sophie Rosset, Olivier Galibert, Eric Bilinski, and Gilles Adda. 2009. The limsi participation to the qast 2009 track. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, October.

Guillaume Bernard, Sophie Rosset, and Martine Adda-Decker. 2010. Etude des caractristiques des collections de documents pour les valuations de systmes de questions-rponses. In *Journes d'tudes sur la parole (JEP'10)*, Mons, Belgique, May.

Pere Comas and Jordi Turmo. 2009. Robust question-answering for speech transcripts: Upc experience in qast 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, October.

Hoa Trang Dang, Jimmy Lin, and Diane Kelly. 2007. Overview of the trec 2007 question answering track. In *Text Retrieval Conference TREC-15*, Gaithersburg, MD, USA, November.

Pamela Forner, Anselmo Peas, Iaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe, and Erik Tjong Kim Sang. 2008. Overview of the clef 2008 multilingual question answering track. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September.

Teruko Mitamura, Eric Nyberg, Hideki Shima, Tsuneaki Kato, Tatsunori Mori, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, Tetsuya Sakai, Donghong Ji, and Noriko Kando. 2008. Overview of the ntcir-7 aclia tasks: Advanced cross-lingual information access. In *Proceedings of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, Tokyo, Japan.

Alejandro Reyes-Barragan, Luis Villasenor-Pineda, and Manuel Montes y Gomez. 2009. Inaoe at qast 2009: Evaluating the usefulness of a phonetic codification of transcriptions. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, October.

Jordi Turmo, Pere Comas, Sophie Rosset, Lori Lamel, Nicolas Moreau, and Djamel Mostefa. 2008. Overview of qast 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September.

Jordi Turmo, Pere Comas, Sophie Rosset, Olivier Galibert, Nicolas Moreau, Djamel Mostefa, Paolo Rosso, and Davide Buscaldi. 2009. Overview of qast 2009. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, October.