Towards an automatic validation of answers in Question Answering

Anne-Laure Ligozat, Brigitte Grau, Anne Vilnat, Isabelle Robba, Arnaud Grappy

Towards an automatic validation of answers in Question Answering

Anne-Laure Ligozat, Brigitte Grau, Anne Vilnat, Isabelle Robba, Arnaud Grappy
LIMSI-CNRS
91403 Orsay CEDEX
France
firstName.name@limsi.fr

Abstract

Question answering (QA) aims at retrieving precise information from a large collection of documents. Different techniques can be used to find relevant information, and to compare these techniques, it is important to evaluate QA systems. The objective of an Answer Validation task is thus to judge the correctness of an answer returned by a QA system for a question, according to the text snippet given to support it. We participated in such a task in 2006. In this article, we present our strategy for deciding if the snippets justify the answers: a strategy based on our own QA system, comparing the answers it returned with the answer to judge. We discuss our results, then we point out the difficulties of this task.

1. Introduction

Question answering (QA) aims at retrieving precise information from a large collection of documents, typically the Web. The hypothesis sustained through the development of QA systems is that users generally prefer to receive a precise answer to their questions, instead of a list of documents to explore, as traditional search engines return [7]. However, to be considered as reliable by users, a QA system must be able to give them elements to evaluate the answer. The objective of a QA system should not only be to find answers to questions, but also to express them in such a way that the user may know if he can accept the answer. These elements of justification give the user a mean to control that the suggested answer corresponds to what was looked for.

Moreover, a good justification should be both concise and complete. The aim is to give a short snippet enabling the user to retrieve all the caracteristics present in his question whithout having to read the whole document.

2. Answer validation

The Pascal Recognizing Textual Entailment Challenge [1] (RTE) defines *" textual entailment"* as the task to decide, given two fragments of text, if the meaning of one can be deduced from the other [2, 1]. Participants to this challenge receive a set of pairs constituted of a text plus a hypothesis, and must determine if the hypothesis is entailed by the text.

This task is close to the task named AVE (Answer Validation Exercise [2]) introduced at QA@CLEF [3] in 2006. The aim of AVE is to automatically validate the correctness of the answers given by QA systems. The final objectives are to improve the performance of QA systems by developing methods for automatic evaluation of answers, and to make answer assessment semi-automatic. An example containing the original question, the hypothesis and the text is given below:

> Original qestion: Who was Yasser Arafat?
>
> Hypothesis: Yasser Arafat was **Palestine Liberation Organization Chairman**
>
> Snippet: President Clinton appealed personally to **Palestine Liberation Organization Chairman** Yasser Arafat and angry Palestinians on Wednesday to resume peace talks with Israel

In AVE, the corpus of pairs hypothesis-text was build semi-automatically from responses of the QA evaluation campaign. It contained about 3,000 pairs to judge. AVE participants were evaluated on their capacity to predict the correctness of the answer (assessed by human validators), and thus had to return a value of implication (*YES* or *NO*) for each input pair. In this year corpus, human assessors

[1] http://www.pascal-network.org/Challenges/RTE

[2] `http://nlp.uned.es/QA/AVE/`

[3] Cross Language Evaluation Forum http://clef-qa.itc.it/CLEF-2006.html

detected 623 *YES* answers and 2441 *NO* answers. Thus there was a great unbalance between correct and incorrect hypotheses, which is not the case in RTE campaigns. For this reason, in AVE only the positive answers are taken into account in the evaluation. Thus, the precision, recall and f-measure of participants to AVE were calculated by the following formula:

$$precision = \frac{\#predicted\ as\ YES\ correctly}{\#predicted\ as\ YES}$$

$$recall = \frac{\#predicted\ as\ YES\ correctly}{\#YES\ pairs}$$

$$f\text{-}measure = \frac{2*precision*recall}{precision+recall}$$

3. Validating answers with a QA system

In 2006, after several participations to QA campaigns, we decided to participate to AVE as well, with the objective to use our own QA system for French, FRASQUES. The relevance of a justification with respect to a hypothesis is evaluated according to the information about the answer deduced from the question, and to the answer given by our system.

3.1. FRASQUES our French QA system

We will first present briefly our French QA system, whose architecture is divided into four main components:

- Question analysis: it processes a syntactic analysis of the question to detect some of its characteristics such as: 1) its keywords, 2) its main verb, 3) the expected answer type, which can be a named entity (person, country, date...) or a general type (like *conference* or *address*...), 4) the focus of the question, which we defined as the entity about which a characteristic is required and which has to be found in the sentence containing the answer.

- Document selection: the search engine Lucene [4] searches the collection to return relevant documents.

- Document processing: it uses Fastr [5] to recognize linguistics variants of the question terms: for example, "Europe's currency" will be recognized as a variant of "European currency". Then the named entities of the documents are tagged with around 20 named entity types. The sentences containing at least one variant of the question terms are kept.

- Answer extraction: it extracts precise answers from the sentences. The extraction strategy depends on the expected type of the answer. If the answer is a named entity, the named entity of the expected type which is closest to the question words is selected. Otherwise, patterns of extraction are used. This patterns were written in the Cass [6] format, a syntactic analyzer used here for answer extraction instead of syntactic analysis. These rules express the possible position of the answer with respect to the question characteristics such as the focus or the expected answer type. Thus, Cass tags the answers in the candidate sentences.

3.2. Answer validation system

The answer validation system (see figure 1) uses three of the four modules of FRASQUES. The input of the answer validation is a pair hypothesis-snippet, along with the original question *Q* and the answer to judge *A1*. The question is analyzed by the Question analysis module. Then, the Document processing module is used, but on the snippet to judge instead of the output of the search engine. The Answer extraction module extracts the anwer(s) *A2* that is found by our system in the snippet. Finally, the answer *A1* is evaluated, and the system returns YES if the answer is considered as justified or NO otherwise, with a confidence score.

The decision algorithm proceeds in two main steps. During the first one, we try to detect quite evident mistakes, such as the answers which are completely enclosed in the question, or which are not part of the justification. When the question contains a date, its temporal context is compared to the temporal context of the snippet. Until now the temporal context is made of the dates recognised in the document description or in the snippet itself. If temporal contexts are inconsistent the pair is negatively judged.

The second step proceeds to more precise verifications. In an ideal case, a correct justificative snippet should contain a declarative reformulation of the question plus the answer, and all the terms of the question should be present in the snippet, linked by the same relations.

Therefore, the question terms or their variations are searched in the snippet. And a score is calculating according to their importance: the focus or the proper names play an important role while the expected answer type or the main verb are less often present in the snippet. Finally, a last verification, consists in determining if the answer is of the correct expected type (a verification module using external source of knowledge is currently being developed).

All these criteria lead us to calculate 2 scores. The first of them concerns the correlation between the answer extracted from the hypothesis *A1* and the answer found by our system FRASQUES: *A2*. When both answers are completely different the hypothesis is refuted. When they are similar, or when FRASQUES did not obtain any answer, the decision is made by taking into consideration the presence/absence

[4] `http://lucene.apache.org/`

[5] `http://www.limsi.fr/Individu/jacquemi/FASTR/`

[6] http://www.sfs.nphil.uni-tuebingen.de/~abney/

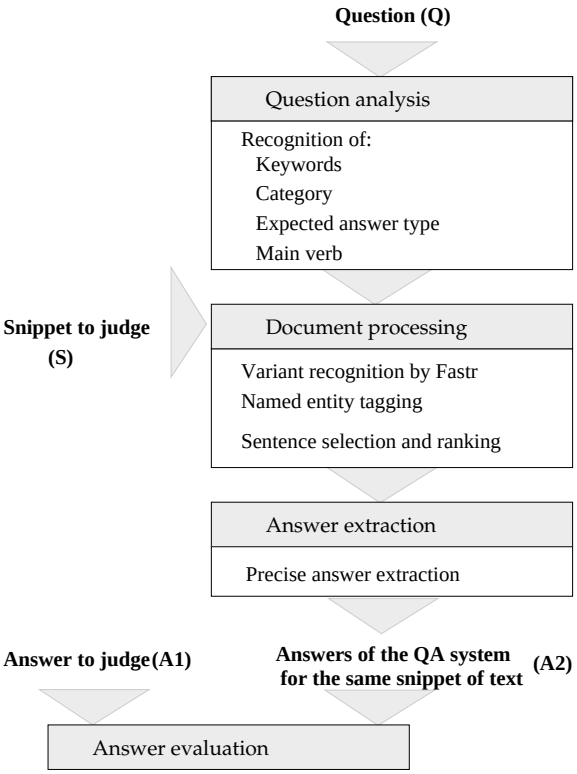

Figure 1. Architecture of the answer validation system

of particular terms. This first score is positive, either when both answers are similar or when the important terms are present in the snippet; it is negative in the other cases.

The second score takes into consideration the number of terms present in the snippet and their value. It is positive when at least one of the term is present, negative otherwise.

Finally a snippet is considered to be an acceptable justification if:

- if *A2* is empty and the score determined by the terms is positive, and in this case, this score is the final score,
- if *A1* and *A2* are similar and the second score is positive, and then the final score is the highest of both scores,
- if both scores are opposite, we choose the highest which it is positive.

3.3. Results

The evaluation corpus contained 3266 pairs, but among them, 202 have not been judged. Since hypotheses were generated automatically, they contain many syntax errors, hence we used only the question plus the answer extracted from the hypothesis.

	# Yes	# No	Total
Evaluated	623	2441	3064
Our results	672	2392	3064
Our correct results	360	2128	2448
Precision	0.54	0.73	0.8
Recall	0.58	0.87	0.8
f-measure	0.56	0.79	0.8

Table 1. New AVE results on CLEF 2006 data

During our participation to AVE, many bugs remained in our programs that have since been corrected. A precise examination of these values enabled us to see that the jugements were sometimes incorrect: some of the positive values are erroneous because the snippet does not contain any justification of the answer. We changed 82 positive values in negative ones.

Table 1 does not present our official results but our new results after diverse improvements of our programs. The first line gives the number of justified pairs (YES), and not justified pairs (NO) evaluated by the human judges on the corpus. The second line contains our results and the third one our correct results. The three last lines are the precision, recall and f-measure of these results following the formula given section 2.

Among our NO answers, we distinguish sure ones from the others. A NO answer is sure if a criteria of the first step is not satisfied (see the preceding section), moreover such NO answers receive the highest confidence score. We found 1637 pairs of "sure NO". Among them, 1415 were correctly judged, so the precision for these answers is 0.87.

Another observation is that for our system it is generally easier to decide that an answer is not justified than justified. Both precision and recall for our NO answers are quite good. On the other hand, our YES answers obtained less good results (but not so far of these of the best system this year in AVE, which obtained an f-measure of 0.6063 on the Spanish task, see section 4, for the related works).

To improve the results of our system we used the evaluated answer corpus to detect the presence of the important terms of the question (obtained in the question analysis step) in the positive and negative answer snippets. Table 2 shows for each term, the number of pairs in which it was searched (first column), the percentage of positive answer snippet in which it was found and also the percentage of negative answer snippet in which it was found (second and third columns). For example, the first line can be interpreted as follows: for the 1131 questions for which we detected a focus, 89% of the correct snippets contained this focus, while only 49% of the incorrect snippets did. This corpus

	# of pairs	found in % of	
		positive answer	negative answer
Focus	1131	89	45
General type	1040	50	35
Main verb	1065	28	13
NP	1187	94	51

Table 2. Presence of the important question terms in the positive and negative answers

study needs still to be refined, but it already confirmed our intuition concerning the importance of a good extraction of the focus, or a good proper name recognition.

4. Related works

Penas et al. ([5]) give an overview of the first AVE launched during QA@CLEF 2006 campaign. Different approaches were adopted, and it seems that the approach using logic gave the best results.

Tatu et al. ([6]) use a named entity recognizer, a syntactic parser and a semantic parser to transform both hypothesis and text in a rich logic representation. Then both representations are submitted to COGEX, a natural language logic prover, that decides whether the text entails the hypothesis or not and also gives a justification of this decision. World knowledge coming from eXtended WordNet is also used when the knowledge contained in the logic form is not sufficient to enable the system to answer. Most of errors of this system were due to the fact that in AVE 2006, hypothesis being automatically generated they were very often incorrect from a syntactic point of view, leading to incorrect logic representations. Nevertheless, this system obtained the best results in both languages it participated: English (with an f-measure of 0.4559) and Spanish (f-measure: 0.6063). The adopted approach in this system is not only logic: syntactic, semantic and even world knowlede is largely used. Our approach is quite different, since for French we do not dispose of such knowledge base, logic prover or reliable syntactic parser. Therefore, we adopted an approach based on lexical knowledge and local syntax through the use of pattern

Herrera et al. ([3]) developed an approach based on an SVM (Support Vector Machine) classification. They apply lemmatization and entity recognition on both snippet and hypothesis. Then they determine the entailment between the numeric entities of the hypothesis and those of the snippet, and also the entailment between the named entities. The model of their classifier is then trained on all these features plus the percentage of word, unigrams, bigrams, trigrams of the hypothesis present in the snippet. Both their runs obtained the second and the third place in the Spanish task with an f-measure of 0.5655 and 0.5615). Classification tools are widely used in RTE and AVE campaigns and we plan to use them as well in our system for a specific step like the final choice.

On the same French corpus, Kozareva et al. ([4]) obtained an f-measure of 0.46903, with an approach they used also in RTE: a machine learning textual entailment, which has the ability to function with different languages (and enable them to submit runs in all the different languages of AVE). Our new results are now slightly better of theirs, since we obtained an f-measure of 0.56.

5. Conclusion

In this article, we presented a strategy for answer validation in QA. This strategy is based on our own QA system: the hypothesis and the text snippet are analyzed, and we defined several criteria which enable us to detect whether the snippet justifies the answer or not. In our evaluation of hypothesis-snippet pairs, we distinguish with reasonable precision and recall the cases for which the snippet is most likely to justify the answer. This first experiment in answer validation constitutes a step towards semi-automatic validation of answers in QA. It also helped us improve our system, since some of the criteria we used for answer validation had not been implemented in our QA system.

References

[1] R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor. The second pascal recognising textual entailment challenge. 2006.

[2] I. Dagan and O. Glickman. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France, 2004.

[3] J. Herrera, A. Rodrigo, A. Penas, and F. Verdejo. Uned submission to ave 2006. In *Workshop CLEF 2006*, Alicante, Spain, 2006.

[4] Z. Kozareva, S. Vazquez, and A. Montoyo. Adaptation of a machine-learning textual entailment system to a multilingual answer validation exercise. In *Workshop CLEF 2006*, Alicante, Spain, 2006.

[5] A. Penas, A. Rodrigo, V. Sama, and F. Verdejo. Overview of the answer validation exercise 2006. In *Workshop CLEF 2006*, Alicante, Spain, 2006.

[6] M. Tatu, B. Iles, and D. Moldovan. Automatic answer validation using cogex. In *Workshop CLEF 2006*, Alicante, Spain, 2006.

[7] E. M. Voorhees. TREC-8 Question Answering Track Evaluation. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. Department of Commerce, National Institute of Standards and Technology, 1999.