Lexical validation of answers in question answering

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

Lexical 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, typically the Web. Different techniques can be used to find relevant information, and to compare these techniques, it is important to evaluate question answering systems. The objective of an Answer Validation task is to estimate the correctness of an answer returned by a QA system for a question, according to the text snippet given to support it. In this article, we present a lexical strategy for deciding if the snippets justify the answers, based on our own question answering system. We discuss our results, and show the possible extensions of our strategy.

1. Introduction

Question answering (QA) aims at retrieving precise information from a large collection of documents, typically the Web. To be considered as reliable by users, a question answering system must be able to give them elements of justification to evaluate the answer whithout having to read the whole document. Here is an example of such a justification:

> Question: When was the Berlin wall demolished?
>
> Answer: **in 1989**
>
> Justification (supporting snippet): the Berlin wall divided East and West Berlin for 28 years, from the day construction began on August 13, 1961 until it was dismantled **in 1989**.

2. Answer validation

In 2006, an Answer Validation Exercise (AVE) [1] was introduced in the evaluation campaign CLEF, which aimed at automatically validating the correctness of the answers given by QA systems according to their supporting snippets. The goal of this exercise is to improve the performance of QA systems by developing methods for automatic evaluation of answers, and to make answer assessment semi-automatic. The organizers provided answers of QA systems with their supporting snippets, and the participants had to decide if each answer was correct or not according to the snippet. For example, the following couple (hypothesis, snippet) can be given:

> Hypothesis: Yasser Arafat was **Palestine Liberation Organization Chairman** [2]
>
> Snippet: President Clinton appealed personally to **Palestine Liberation Organization Chairman** Yasser Arafat and angry Palestinians on Wednesday to resume peace talks with Israel

The hypothesis is the reformulation of the question *"Who was Yasser Arafat?"* containing the answer given by a system *"Palestine Liberation Organization Chairman"*.

In AVE, the corpus of about 3,000 pairs hypothesis-snippet was built semi-automatically from answers of a question answering evaluation campaign. AVE participants were evaluated on their capacity to predict the correctness of the answer (assessed by human validators), and had to return a value of implication (YES or NO) for each input pair. The precision, recall and f-measure of participants to AVE were calculated by the following formulas:

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

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

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

[1] http://nlp.uned.es/QA/AVE/

[2] In our AVE examples, the answer will be written **in bold**.

3. Validating answers with a question answering system

In 2006, after several participations to question answering campaigns (in monolingual French and English tasks and cross-lingual ones), we decided to participate to AVE. In this campaign, our objective was to use our own question answering 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. Our answer validation system

The answer validation system uses three out of the four modules of the question answering system as figure 1 shows. The input of the answer validation is a pair hypothesis-snippet, along with the original question *Q* and the answer to judge *A1*.

Fisrt, the question is analyzed by the Question analysis module, which processes a syntactic analysis of the question to detect some of its characteristics such as its keywords, the expected answer type (which can be a named entity like person, country, date... or a general type like *conference* or *address*), the focus of the question (which is defined as the entity about which a characteristic is required).

Then, the Document processing module is used, but on the snippet to judge instead of the output of the search engine. This module uses Fastr [3] to recognize linguistic 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 Answer extraction module extract the anwer(s) *A2* that is found by our system in the snippet. 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. These patterns were written in the Cass [4] format, a syntactic parser 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 type of the answer. Cass thus tags the answers in the candidate sentences.

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,

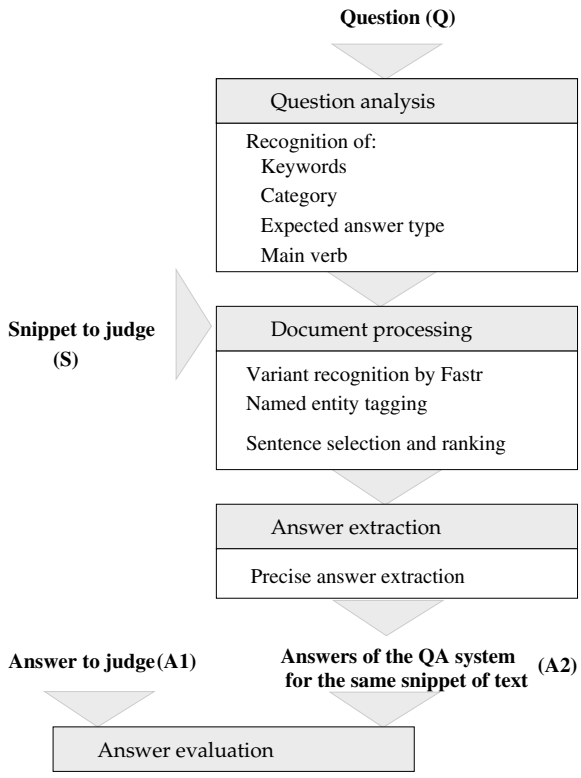

Figure 1. Answer validation system

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, this date is compared to the dates of the snippet, and if they are inconsistent, the pair is judged as a NO.

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, all the terms of the question should be present in the snippet, linked by the same relations.

3.1.1 Question terms

Among question terms, some play a more important role, for instance, the object of the question, that we named *focus*. The focus is the entity about which the question is asked and either a characteristic or a definition of this entity may be searched. In *"Which is the political party of Lionel Jospin?"*, the focus is *"Lionel Jospin"*.

Four particular roles are distinguished in the questions: focus, expected type of the answer, main verb, and proper names.

In order to confirm our intuitions about these roles, we conducted a corpus study after the evaluation: we measured the presence of each term in correct and incorrect snippets of the AVE corpus. Table 1 shows for each term, the per-

[3] http://www.limsi.fr/Individu/jacquemi/FASTR/
[4] http://www.sfs.nphil.uni-tuebingen.de/~abney/

	found in % of	
	positive answer	negative answer
Focus	89	45
General type	50	35
Main verb	28	13
Proper names	94	51

Table 1. Presence of the important terms extracted from the question in the positive and negative answers

centage of correct and incorrect snippets in which it was found (second and third columns). For example, the first line can be interpreted as follows: for the 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 study confirmed our intuition concerning the importance of a good extraction of the focus, or a good proper names recognition.

3.1.2 Relations between terms

Concerning relations between terms, to control them it is necessary to parse the whole sentence. Since in many cases, such parses are difficult to obtain whith sufficient confidence, we chose to control only the relations which are likely to associate the answer and some particular terms like the focus or the general type. This control is done using extraction patterns, which are written as Cass parser rules.

3.1.3 Answer type

Another verification concerns the expected type of the answer. When this type is a named entity, the system can determine if the answer is of the correct type. In the other cases, either the type is present in the snippet near the answer or it is implicit and does not appear in the snippet. This last possibility is illustrated in the question *"Which animal lays blue eggs?"* which answer is *"Ameraucana chickens lay blue eggs"*. The use of external resources, such as Wikipedia in which many categories and definitions can be found, would be necessary to infer the link between the specific answer and the type to which it belongs.

3.1.4 Score computing

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 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.

3.2 Results

The evaluation corpus contained 3,266 pairs, but among them, 202 were not judged. Since hypotheses were generated automatically, they contain many syntax errors, hence we used only the question plus the answer extracted from the hypothesis.

During our participation to AVE, many errors remained in our programs that have since been corrected. AVE organizers gave the values expected for each pair of the corpus so we could reevaluate our system. 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 into negative ones.

Table 2 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); such NO answers receive the highest confidence score. We found 1035 pairs of "sure NO". Among them, 995 were correctly judged, so the precision for these answers is 0.96.

3.3. Improvement of the lexical criterion

For the above evaluation, the parameters given for each criterion were set manually. In order to improve the use of

Precision	Recall	f-measure
0.55	0.55	0.55

Table 2. FRASQUES **results at AVE 2006**

lexical information, we decided to train our system on part of the AVE corpus. In this corpus, we put aside the "sure NO" pairs, and divided the remaining pairs to have a learning corpus and a training one. For the training, we used the machine learning tool Weka [5]. We considered as criteria the presence of the following terms of the question: focus, general type of the answer, main verb, and multiword expressions. We obtained the following results: precision 0.43, recall 0.84 and f-measure 0.57. These results show that the criteria can be improved by adjusting the scores of each parameter. We are presently implementing theses changes.

4. Related works

Penas et al. ([4]) give an overview of the first Answer Validation Exercice in 2006 [6]. The approaches can be divided into two main categories: logical proofs of the entailment and lexical comparison of hypotheses and snippets.

In the first category, Tatu et al. ([5]) use a named entity recognizer, a syntactic parser and a semantic parser to transform hypotheses and snippets into 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. This system obtained the best results for English (with an f-measure of 0.46) and Spanish (f-measure: 0.60).

In the second category of systems, 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 0.57 and 0.56).

AVE exercise is strongly connected to Textual Entailment ([1, 2]). In Textual Entailment, the approaches are more diverse than in AVE; yet, the best systems also use extensive linguistic and background knowledge, or a very large training corpus.

[5] http://www.cs.waikato.ac.nz/ ml/index.html

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

5. Conclusion

In this article, we have presented a strategy for answer validation in question answering. This strategy is based on our own question answering system: the hypothesis and the text snippet are analyzed by the question answering system, and we have defined several criteria which enable us to detect whether the snippet justifies the answer. 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. We have also presented the possible extensions of our strategy, by using external resources to acquire additional knowledge. Since the answer has to be entirely justified by the snippet, it is important to respect the notion of justification.

This first experiment in answer validation constitutes a step towards semi-automatic validation of answers in question answering. It also helped us improve our system, since some of the criteria we used for answer validation had not been implemented in our question answering 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, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *In Quionero-Candela et al., editors, MLCW 2005,*, volume LNAI Volume 3944, pages 177-190. Springer-Verlag, 2005.

[3] J. Herrera, A. Rodrigo, A. Penas, and F. Verdejo. UNED Submission to AVE 2006. In *Workshop CLEF 2006*, Alicante, Spain, 2006.

[4] A. Penas, A. Rodrigo, V. Sama, and F. Verdejo. Overview of the Answer Validation Exercise 2006 . In *Workshop CLEF 2006*, Alicante, Spain, 2006.

[5] M. Tatu, B. Iles, and D. Moldovan. Automatic Answer Validation using COGEX. In *Workshop CLEF 2006*, Alicante, Spain, 2006.

[6] 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.