A corpus for studying full answer justification

Arnaud Grappy*, Brigitte Grau*, Olivier Ferret†, Cyril Grouin*, Véronique Moriceau*, Isabelle Robba*, Xavier Tannier*, Anne Vilnat*, Vincent Barbier*

*LIMSI-CNRS
BP 133
91 403 Orsay Cedex, France
{firstname.lastname}@limsi.fr
† CEA, LIST, Vision and Content Engineering Laboratory,
Fontenay-aux-Roses, F-92265, France.
olivier.ferret@cea.fr

Abstract

Question answering (QA) systems aim at retrieving precise information from a large collection of documents. To be considered as reliable by users, a QA system must provide elements to evaluate the answer. This notion of answer justification can also be useful when developping a QA system in order to give criteria for selecting correct answers. An answer justification can be found in a sentence, a passage made of several consecutive sentences or several passages of a document or several documents. Thus, we are interesting in pinpointing the set of information that allows to verify the correctness of the answer in a candidate passage and the question elements that are missing in this passage. Moreover, the relevant information is often given in texts in a different form from the question form: anaphora, paraphrases, synonyms. In order to have a better idea of the importance of all the phenomena we underlined, and to provide enough examples at the QA developer's disposal to study them, we decided to build an annotated corpus.

1. Introduction

Question answering (QA) aims at retrieving precise information from a large collection of documents. 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 (Voorhees, 1999). However, to be considered as reliable by users, a QA system must provide 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 trust the system. Moreover, a good justification should be both concise and complete. The aim of a justification is to give the shortest snippet enabling the user to retrieve all the characteristics present in his question without having to read the whole document.

We name justification the set of the linguistic information that allows this verification. This set has to handle all the information given in the question, either the involved entities or their relationships. The justification can then be found in a sentence, a passage made of several consecutive sentences or several passages of a document or several documents. Thus, we are interesting in pinpointing the set of information that allows to verify the correctness of the answer in a candidate passage and the question elements that are missing in this passage.

Besides providing a justification for a user, finding full answer justifications will give QA systems a strategy to evaluate the validity of their answers.

The notion of justification has always been present in QA evaluation campaigns. At TREC[1], an answer was made of a pair (a short string, a document number) and was right if the string was correct and if it was supported (or justified) by the document. At CLEF[2], since 2006, an answer is a triple (a short answer string, a justification passage made of several passages extracted from the document, a document number), and an answer is judged correct by the assessors if the passage and the document allowed to justify the answer. The following example, issued from TREC11 campaign, shows a justification given in a sentence:

> Q: In what country did the game of croquet originate?
>
> S: Croquet (pronounced cro KAY) is a 15th century [ANS French] sport that has largely been...

However, even if the justification is included inside a single sentence, there are some difficulties to exhibit it. There is a variation on the answer type (nationality vs country), between "game" and "sport" and there is a common sense inference to deduce that a sport of some nationality is originated in the related country.

Now, here is an example coming from the same campaign where the justification spreads through two non successive sentences:

> Q: What was the length of the Wright brothers' first flight?
>
> S: Orville and Wilbur Wright made just four brief flights... [...] In a way, the poor stability of the flyer was a tribute to the Wrights' flying skills. " The fact that their first flight was [ANS 120 feet] and their last one was 852 feet shows they were learning, " Watson said.

[1] http://trec.nist.gov/

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

In this example, there exists an anaphoric chain that begins with "Orville and Wilbur Wright", then continues with "the Wrights" and ends at "their" in the sentence that contains the answer. The brotherhood between Orville and Wilbur is not explicit, and should be verified in another document, or an encyclopaedia.
Thus, finding an answer and its justification can be posed as finding candidate passages and to determine:

- if it contains the exact answer;
- which items in the passage match which question items, exactly or with variations;
- the size of the passage that contains the answer and holds the maximum elements of justification;
- the justification items that have to be verified in other documents.

In order to have a better idea of the importance of all the phenomena we underlined, and to provide enough examples at the QA developer's disposal to study them, we decided to build a corpus. We will first browse the corpora provided by evaluation campaigns (see section 2.) to show that we cannot depart from them, we will also present some work about the elaboration of QA corpora; then we will present our methodology (see section 3.) to build a corpus that fill in our requirements, either for selecting passages and for annotating them. Finally we will present a quantitative study of this corpus before concluding.

2. Corpora provided by evaluation campaigns and other existing QA corpora

Even if QA evaluation organizers build corpora to favour the test and the evaluation of systems, the corpora provided to the participants are the lists of questions, plus their short answers, and eventually the passages returned by participants. These corpora remain incomplete (all questions are not answered) and do not allow an in-depth analysis of the justification problem because they are free of annotations that could allow to rely question elements to answer elements.
AVE (Answer Validation Exercise[3]) is a task that was introduced at QA@CLEF in 2006. The aim of AVE is to automatically validate the correctness of the answers given by QA systems. In AVE, the corpus made of pairs hypothesis-text was build semi-automatically from responses provided by the QA evaluation campaign. It contained about 3,000 pairs to judge. Snippets are made of maximum 250 characters, which was the size of each justification passage fixed by the QA organizers. Here is an except of the corpus:

> Original question: 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 2006 corpus, human assessors detected 623 *validated* answers and 2441 *incorrect* answers (Herrera et al., 2006). In fact, even correct answers that were not fully justified by the snippet were considered as validated, often because the document supported the answer. Thus, the corpus provided by AVE did not handle all the phenomena we mentioned before and cannot be useful for development.
As already said, passages are provided by systems, and they are filtered by the different criteria the systems are able to handle. As a result, right answers are only those the systems are able to find, and they cannot be representative of the linguistic diversity that exists in documents.
The AVE task is close to the Pascal Recognizing Textual Entailment Challenge[4] (RTE) that defines *"textual entailment"* as the task to decide, given two fragments of text, if the meaning of one can be deduced from the other (Bar-Haim et al., 2006). Participants to this challenge receive a set of pairs constituted of a text (T) plus a hypothesis (H), and must determine if the hypothesis is entailed by the text. The text is also made of a short fragment of text, generally a sentence.
Vanderwende et al. (Vanderwende and Dolan, 2006) made a study on the RTE1 corpus. It was similar in spirit to our present work in that they aim to isolate the class of T-H pairs whose categorization can be accurately predicted based solely on syntactic clues and annotate the syntactic phenomena that occurs between the T-H pairs. They found that only 9% of the true test items can be handled by syntax, and 18% if using a thesaurus. As in our work, such studies permit to find out system requirements to handle the entailment task, and in what proportion it could affect the system.
Because of the lack of adapted resources for developing QA systems, some work were dedicated to build QA corpora ((Cramer et al., 2006) for corpus in German, (Rosset and Petel, 2006) for oral questioning of a QA system for example). However, they were not dedicated to annotate the characteristics shared by questions and answering passages. Other work concerned more precisely the annotation of linguistic phenomena in QA task: (Boldrini et al., 2009) for anaphora resolution, and discourse properties (Varasai et al., 2008) for text summarization and QA. In these two cases, the corpora were dedicated to account for some kind of phenomenon, and not to be representative of a whole task.

3. Corpus elaboration and annotation methodology

The corpus we built is in French, even if the methodology can be applied to all languages. We departed from a subset of questions selected in French EQueR campaign (Ayache et al., 2006) and in CLEF campaigns. We eliminated definition questions, boolean questions and questions about a single named entity requiring the location or the time. This set is made of around 290 questions.
The document collection of EQueR campaign regroups news from "Le Monde" and "Le Monde Diplomatique"

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

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

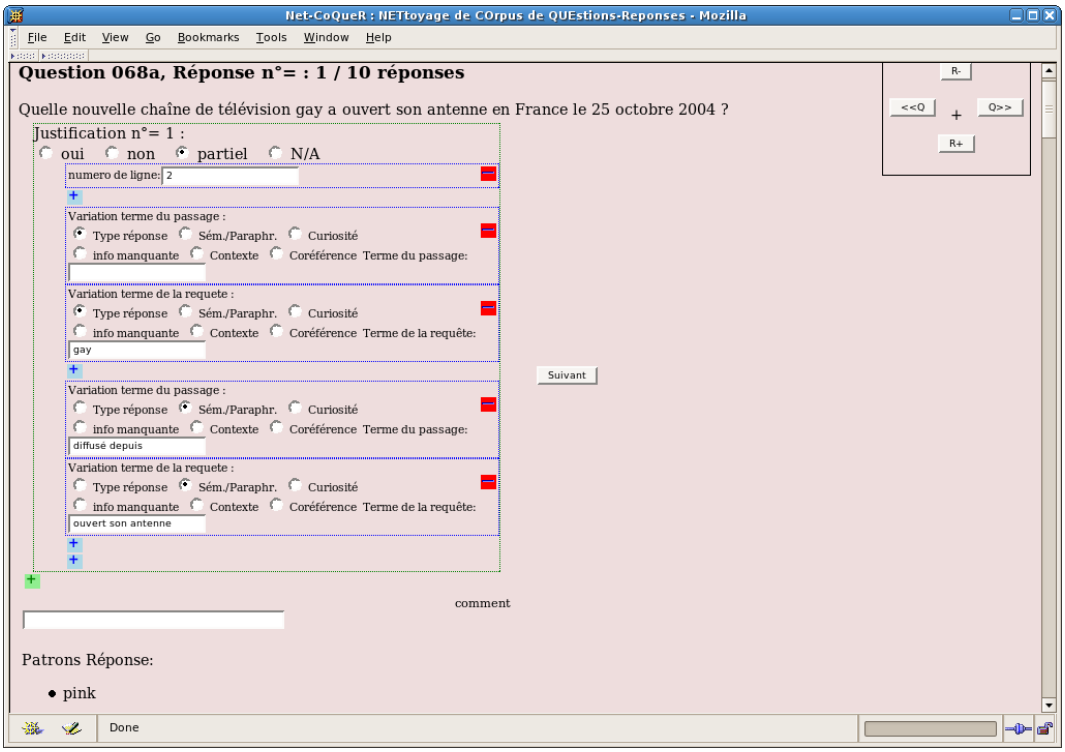

Figure 1: Annotations for partial justifications

newspapers, debates of the French senate and "SDRT" newswires. At EQueR, participants provided long answers (up to 250 characters). A study about the similarity of these passages and the questions (Grau et al., 2008) showed that 40% right passages contains all the words of the question, without variations, and that 35% only hold one difference, either a variation or a missing element: a missing type (30%), a missing verb (22%), a variation of the verb (27%). Thus, to avoid a too great dependency between questions and the document collection, we chose another document collection, the French version of the Wikipedia. The advantage of Wikipedia is that it is free of rights and it would be possible to make public our corpus.

3.1. Selection of documents

In order to avoid the bias introduced by the application of a QA system to select passages, we emphasized on recall when selecting documents associated to the questions. The selecting method follows manual (M) and automatic (A) steps:

1. tagging of the questions (A);
2. ordering of question words according to their significance (A based on their POS tag, then M);
3. construction of different queries by omitting each time a significant word, other than a named entity[5], and by adding the right answers (A);
4. selection of POS tagged documents with Lucene[6] (A).

[5]Named entities are less likely to vary or be missing in the document

[6]http://lucene.apache.org/

3.2. Annotation scheme

The annotation scheme is dedicated to classify the phenomena to handle when retrieving a full justification, and not to annotate these phenomena precisely. Annotators have first to decide if the document contains a full exact justified answer "oui (yes)", a partial justified answer "partiel (partial)" or an incorrect answer "non (no)". Figure 1 shows the annotation scheme and an excerpt of its application to the following example:

> **Question**: Quelle nouvelle chaîne de télévision gay a ouvert son antenne en France le 25 octobre 2004?
> (What new gay television channel opened its antenna in France on October 25, 2004?)
>
> **Query**: chaîne France (channel France)
> **Answer**: Pink TV
>
> **Answering passage**, line 2: Cette catégorie regroupe les émissions de télévision diffusées sur la chaîne française Pink TV depuis le 25 octobre 2004.
> (This category includes television broadcast on French channel Pink TV since October 25, 2004)

A full exact justified answer is a sentence that fully justifies the answer, with the exact terms of the questions. A partial justification holds when the document justifies the answer with variations, sparse or missing terms. An annotation describes a partial justification to an answer whose location is given inside the document with its line number (field "numero de ligne"). Then partial justifications are annotated according to the following cases:

- missing of the answer type, fully or totally ("Type réponse" button);
- semantic variation or paraphrase ("Sem/Paraphr." button);
- missing element in the whole document ("info manquante" button);
- element in the document context ("Contexte" button), often date, title, location;
- presence of an anaphora ("Coréférence" button);
- a justification requiring an elaborated reasoning is stamped as "Curiosité" (curiosity).

For each annotation type, a parallel information is given in order to put in relation terms of the passage and terms of the question. When terms are not in the same sentence as the answer, the line number is given with the terms, preceded by # in the field "terme". A justification can be annotated with several types of variations, and a document can contain several answers justified. In such a case, each one will be annotated.
In the example of figure 1, the answering passage does not contain that Pink TV is a gay channel, it is just said it is a channel. Thus, an annotation relative to a missing type is added. Secondly, the text does not explicit the opening time, but gives a time of program broadcasting. Thus, we can infer that the channel launched its program at this time, as it is the same as required in the question. Thus, a paraphrase annotation is created between these two expressions. Let us see another example, that involves a lot of annotations (see Figure 2).

> **Question**: Dans quelle grande capitale la Tour Eiffel fut érigée en 1889 ?
>
> (In which big capital was erected the Eiffel Tower in 1889?)
>
> **Passage** in the Wikipedia page about the Eiffel Tower:
>
> line 2: La *tour Eiffel* est une tour de fer puddlé *construite* par Gustave Eiffel et ses collaborateurs pour l'Exposition universelle de *1889*.
> line 3: Situé l' extrémité du Champ-de-Mars, en bordure de la Seine, ce *monument parisien*, symbole de la France et de sa *capitale* est l'un des sites les plus visités du pays.
>
> (line 2: The *Eiffel Tower* is an iron tower *built* by Gustave Eiffel and his collaborators for the Universal Exposition of *1889*.
>
> line 3: Sited on the border of Champs-de-Mars, along the Seine river, this *Parisian monument*, symbol of the France and of its *capital*, is one of the most visited sites of the country.)

As we can see, the answer is given line 3, by the term Parisian. So, a semantic variation is annotated on the answer (réponse). The Eiffel Tower is replaced by an anaphora with the term monument in the answering sentence. Justification elements are scarsed along line 2 and line 3. Thus, elements in line 2 ought to be annotated as contextual elements (*1889* and *built*), plus a paraphrase for *built*. However we can see that the annotator omitted to annotate the contextual characteristic of *built*. This kind of complex annotations leads to low our annotator agreement.

4. Corpus characteristics

There are 291 questions, with average 1.6 queries by question. Only the first ten passages are annotated. The corpus was annotated by seven annotators.

4.1. Coherence of the annotations

In order to evaluate the annotators coherence, 571 passages were annotated globally by each possible couple of annotators. All the 21 couples shared a sub-corpus issued from 3 questions. As each question can generate several queries and as ten passages maximum by query are annotated, two annotators shared in average 23 passages. The annotator agreement was evaluated by calculating the kappa value for the three classes of annotations (YES, NO, PARTIAL). The kappa value is calculated on the confusion matrices built for each couple of annotators. We obtained a mean value k=0.63, weighted or not by the number of the annotations for each class, which is a good agreement. Differences essentially hold between NO and PARTIAL answers: some annotators considered that when it will require too much knowledge to build a justification, the answer was NO, while others answered PARTIAL for these cases that are annotated as "curiosity".
In a second step, when two annotators share PARTIAL decisions, a kappa value was calculated on their types of annotation. In that case, the mean value was 0.53 at first. Differences essentially came from a different interpretation between "elements in context" and "missing elements" that some annotators have misinterpreted as missing elements in the sentence that contains the answer. Each annotator has verified his annotations to correct them, according to a more detailed annotation guide. The mean value of kappa is now 0.59, the mean is weighted by the different numbers of annotation in each class. This value now corresponds to a rather good agreement (the low limit is given at 0.61 in the litteracy). Even if the kappa value increased, there are still a lot of differences, which come from the complexity of the annotation task.
In order to have a more precise evaluation of the annotator agreement, we wanted to compute the kappa value for the different kinds of annotations. In this calculus, confusion matrices for each couple are computed in the following way (see Table 4.1.):

- the number of times the two annotators have annotated a paraphrase,
- the number of cases each annotator has annotated a paraphrase and not the other one,
- the fourth number might represent the number of cases the two annotators do not find a paraphrase. In order

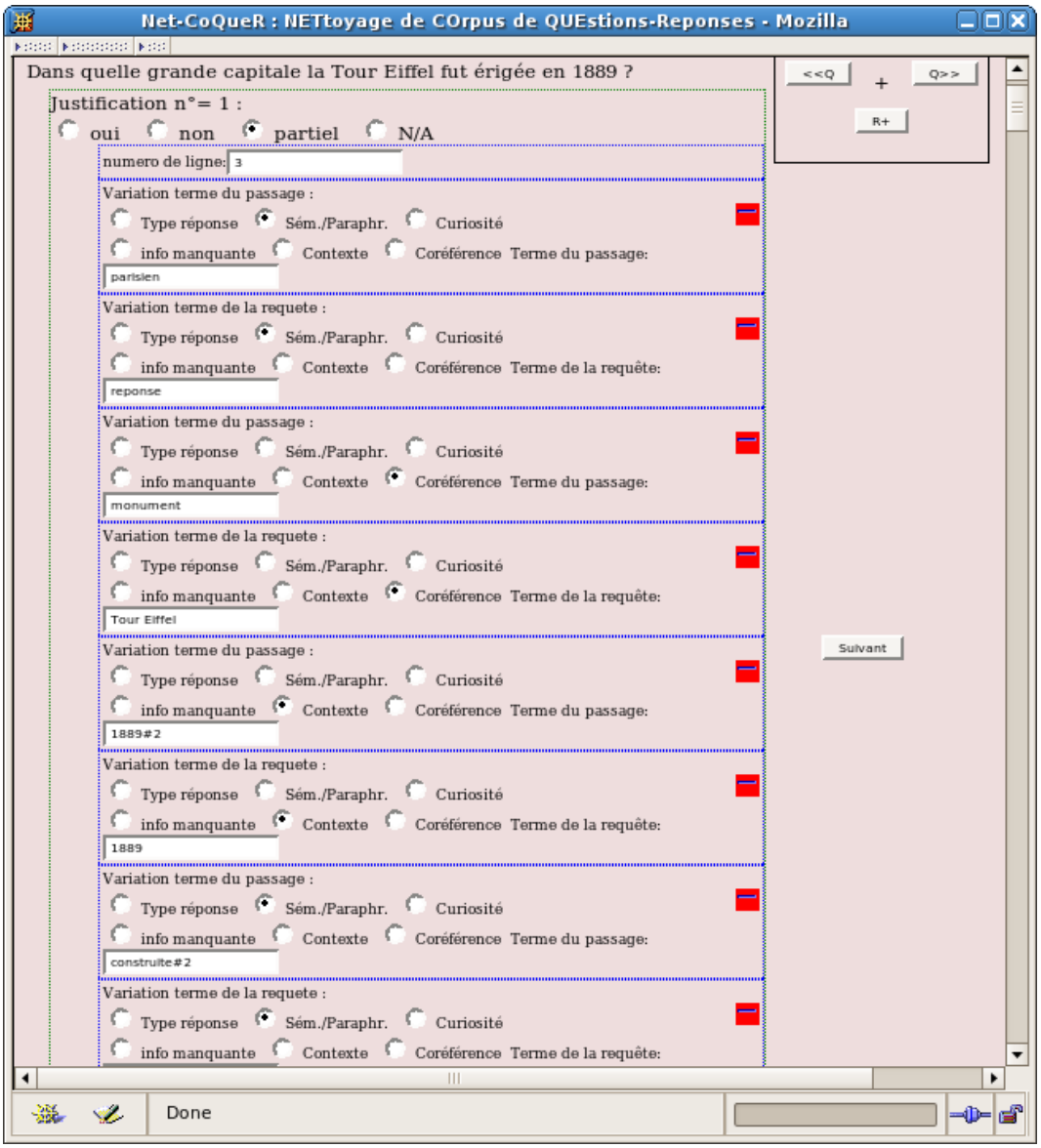

Figure 2: Annotations for partial justifications

to estimate this measure, we choose to consider the total number of annotations done for it. We considered first that it was the total number of annotated documents minus the number of annotated documents, or the total number of all annotations, but these interpretations lead to high values because, in general, annotators agreed with the absence of a phenomenon inside documents. Very often, when they disagree, it is on the kind of annotation in presence of some phenomenon.

	YES	NO
YES	5	1
NO	3	9

Table 1: Confusion matrix for paraphrase annotations

We only calculated a kappa value for "paraphrase" as other types of annotations have a too low frequency by couple of annotators to allow the calculus of a representative kappa. This value is thus 0.79 and corresponds to an almost ideal agreement. When examining the matrices for the different annotation types, we can see that generally, annotators agree for "anaphora" while there are more variations when annotating missing types, missing information, context and curiosity. The three first annotations are sometimes confused, and we will have to clarify another time the annotation guide. Curiosity is a notion more subjective, and this annotation was proposed in order to add an information about the global difficulty for answering some questions.

4.2. Properties of the annotated corpus

Among the 2978 passages, 201 contain full exact justified answers (i.e. YES passages), 2098 are wrong (i.e. NO passages) and 679 hold partial justifications (i.e. PARTIAL passages). Thus 880 passages answer the 291 questions. Our corpus contains more right answers than in the AVE corpus.

When studying the corpus from the question point of view, it exists at least a YES answer for 80 questions (27.5%). All first ten passages to 80 questions (28%) are NO passages, and 125 questions (43%) are only answered by partial justifications, without any YES passages. The unanswered questions will have to be better processed and queries reformulated, if we want to augment our corpus.

We can see that our methodology to build the corpus leads to a corpus that is representative of the phenomena we want to study, in comparison with corpus issued from evaluation campaigns where 40% of the questions are answered by YES passages vs 28% here. It is also representative of the difficulty of the task, as we preserve the independence between the question formulation and the searched collection of documents. Thus, according to this classification, QA systems have to handle linguistic phenomena in order to be able to answer 44% of questions.

Table 2 shows the kinds of variations we found in the corpus.

Annotation type	**#**	**% / 679**
Context	121	18%
Missing element	173	25%
Missing or incomplete answer type	78	11.5%
Paraphrase	400	59%
Anaphora	86	12.5%
Curiosity	142	21%
Total	1000	

Table 2: Repartition of the linguistic phenomena in justifications

Paraphrases represent the most important linguistic phenomena to handle. The other important point showed by this corpus is that systems cannot avoid to handle all these phenomena, and to search for them in passages of more than one sentence length, as showed by the number of context and missing elements.

We also count the kinds of answer (YES, NO, PARTIAL) for questions according to their type of answer (see Figure 3).

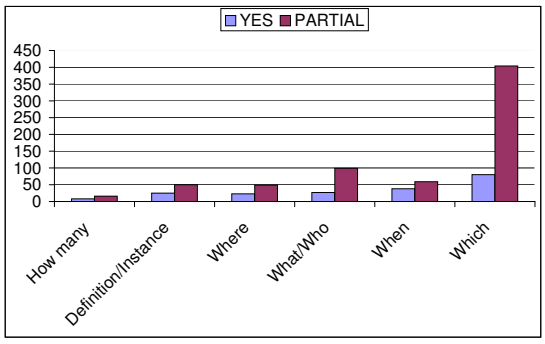

Figure 3: Kinds of answers along with question categories

This allows us to show the proportion of each kind of question we have chosen. The different types of question are not balanced, because we preferred long questions, in order to favour the presence of multiple phenomena, and questions where the type is explicitly given.

At the moment, we plan to continue to homogenize annotations between annotators in order to make the corpus available to the community.

5. Conclusion

We built a corpus of question-passage pairs that are annotated in order to account for answer justifications. Such a corpus aims at being a contribution to the ability to make progress in this specific area. Thus, the principles we chose for collecting answering texts to questions allow us to approach a more realistic distribution of Yes/No/Partial validation of answers. Instead of considering that justifications hold in majority in restricted passages, we collected full texts by applying relaxed queries so that we avoid to create a bias at the retrieval step. When studying the type of justifications found in these texts, we highlighted that the majority of cases are justifications that are spread on several passages of texts, and even some information is sometimes missing in these passages.

We also underlined that a lot of linguistics phenomena have to be handled to evaluate or provide full justifications to the answers given by a system. This corpus, we hope, will be helpful for system development and applications.

6. Aknowledgment

This work has been partly granted by the project ANR-05-BLAN-008501, CONIQUE.

7. References

Christelle Ayache, Brigitte Grau, and Anne Vilnat. 2006. Equer: the french evaluation campaign of questions answering systems. In *5th Conference on Language Resources and Evaluation (LREC 2006)*.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *The Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Ester Boldrini, Marcel Puchol-Blasco, Borja Navarro, Patricio Manuel Martínez-Barco, and Chelo Vargas-Sierra. 2009. Aqa: a multilingual anaphora annotation scheme for question answering. *Procesamiento del lenguaje natural*, (42):97–104.

Irene Cramer, Jochen L. Leidner, and Dietrich Klakow. 2006. Building an evaluation corpus for german question answering by harvesting wikipedia. In *5th Conference on Language Resources and Evaluation (LREC 2006)*.

Brigitte Grau, Anne Vilnat, and Christelle Ayache, 2008. *L'évaluation des technologies de traitement de la langue: les campagnes Technolangue*, chapter 6, évaluation de systèmes de question-réponse. Traité IC2, série Cognition et traitement de l'information. Stéphane Chaudiron and Khalid Choukri, lavoisier edition.

Jesus Herrera, Alvaro Rodrigo, Anselmo Penas, and Felisa Verdejo. 2006. Uned submission to ave 2006. In *Workshop CLEF 2006*, Alicante, Spain.

Sophie Rosset and Sandra Petel. 2006. The ritel corpus - an annotated human-machine open-domain question answering spoken dialog corpus. In *5^{th} Conference on Language Resources and Evaluation (LREC 2006)*.

Lucy Vanderwende and William B. Dolan, 2006. *Machine Learning Challenges*, volume 3944/2006, chapter What Syntax Can Contribute in the Entailment Task, pages 205–216. Springer Berlin - Heidelberg.

Patcharee Varasai, Chaveevan Pechsiri, Thana Sukvari, Vee Satayamas, and Asanee Kawtrakul. 2008. Building an annotated corpus for text summarization and question answering. In *6^{th} Conference on Language Resources and Evaluation (LREC 2008)*.

Ellen M. Voorhees. 1999. 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.