

# A corpus of real-life questions for evaluating robustness of QA systems

Laurianne Sitbon<sup>1,2</sup> Patrice Bellot<sup>1</sup> Philippe Blache<sup>2</sup>

(1) Laboratoire d'Informatique d'Avignon - University of Avignon

(2) Laboratoire Parole et Langage - University of Provence

{laurianne.sitbon, patrice.bellot}@univ-avignon.fr, blache@lpl-aix.fr

Many evaluation campaigns on question answering (QA) systems have been organized for years. The international TREC<sup>1</sup> conference proposes a track about it, the European CLEF<sup>2</sup> campaign proposes cross evaluation in eight languages, NTCIR<sup>3</sup> includes a QA track in three languages and the French Technolanguage EQUER<sup>4</sup> evaluation focused on 500 corpus based questions. But the questions asked in those campaigns are checked for being well formed and enough complete. We aim to test QA systems on their ability to answer questions spontaneously typed by people without thinking deeply to the grammatical and lexical forms they might use. This is designed to test QA system robustness in more real life uses. Related experiments already done in document retrieval field test the robustness of search engine with automatic transcription of spoken queries (Crestani, 2000) or with automatically degraded text entries (Ruch, 2002). Moreover, we aim to test our QA system (Gillard et al., 2006) with questions written by dyslexic adults or children, or non-native speakers. This public is intuitively the most concerned with these problems of robustness.

## 1 Corpus constitution

A web-based approach was used to acquire data for the corpus. The motivation for such an approach is two fold. Firstly, it lets users make the experiment in relaxing conditions, when and where they wish to. Secondly, it permits to collect data from a wide population, especially for dyslexics individuals already solicited for psycholinguistics experiments. It removes geographical constraints often restraining the quantity of data in this area.

The experiment is composed of 20 questions selected from EQUER French evaluation campaign. The selected questions were some right answered by SQuALIA (Gillard et al., 2006). 8 of them contain proper nouns. 2 of them contain foreign low frequency proper nouns. The covered focuses are: person name (5 questions), number (5 questions), date (3 questions), location (2 questions), money, distance, age, journal name and military grade.

The main issue is getting enough spontaneous production of questions while each question must be equivalent to its corresponding one in the evaluation campaign question set. The first obvious tip is that nothing must be written on screen referring to the question subject. Proper nouns

---

<sup>1</sup><http://trec.nist.gov/data/qa.html>

<sup>2</sup><http://clef-qa.itc.it/2005/>

<sup>3</sup><http://www.slt.atr.jp/CLQA/>

<sup>4</sup><http://www.technolanguage.net/article61.html>

must not appear with their correct spelling. That is why we use voice instructions. However, we might not influence the produced syntax of the question by dictating it. Offering the answer like jeopardy game question has been considered, but it supposes knowledge and it can sometimes leads to many different questions. We finally decided to make them hear a description of the answer. For example, "who is the French president" is instructed with "ask for the name of the president of France".

The data have been mainly collected from adults at diverse level of graduate studies. 9 participants are native French speakers, 6 are non-native (Chinese, German and Spanish living in France), and 2 are dyslexics native speakers.

## 2 Observations

The spelling mistakes occurring in typed questions fall various categories. The first is missing "?" at the end of the question. Almost half user made this mistake at least once. This can not have an impact on our system because it doesn't process the punctuation. However, a more wide system able to detect if a users query is a question may fail and return documents instead of short answers. Most users also produce accentuation mistakes. For non-dyslexics native users, this is generally due to the keyboard used instead of the user ability to write well. 6 of the native French users and all non-native and dyslexic users produced at least one syntactical mistake. When the questions are processed as bags of lemmatized words the impact is low. Each user made at least one orthographic mistake on a proper noun. The rate of proper names misspelling on questions typed by non-native users is 39%. Half native users and all dyslexic and non-native users misspell nouns. The latter are the most productive with almost a third of the questions.

## 3 Evaluation of SQuALIA with spontaneous questions

QA systems based on a numeric approach classically answer a question within 4 steps :

- retrieving documents potentially containing the answer,
- detecting the expected answer type (definition, place, person name, ...),
- identifying most relevant passage inside documents (according to question terms and expected answer type),
- scoring candidate answer (all entities corresponding to expected answer type) in these passages according to a distance to the terms of the question.

A question is consider well answered if a correct answer appears in the list of 5 answers proposed by the system. The automatic judgment of correct answers and the multiplicity of correct answers for many questions lead us to differentiate non ambiguous questions. For EQUER campaign, there was no set of patterns of correct and supported answers yet (as for TREC QA campaign <sup>5</sup>) so we constructed our own <sup>6</sup>.

---

<sup>5</sup>[http://trec.nist.gov/data/qa/2004\\_qadata/04.patterns.zip](http://trec.nist.gov/data/qa/2004_qadata/04.patterns.zip)

<sup>6</sup>This reference has been validated and is now available

If the expected answer type is not detected or wrong, the question answering system is likely to fail. For most of our users, the expected answer type computed by the system is incorrect or missing for 20% to 40% of typed questions. The errors are partially due to spelling errors and also to the words chosen by the user to formulate the question. Indeed, we notice that 18% of these questions did not contain any mistake. This means that the words chosen by the user in these questions implies a formulation unexpected by the system for that type of question. A more robust detection process must be studied and implemented.

The document retrieval step has not been studied here since documents were furnished with evaluation campaign data. The robustness of this specific task has already been evaluated. (Ruch, 2002) shows that the mean average precision of the system SMART drops by 18,7% on a document retrieval task when at 15% of the words of the queries are automatically corrupted.

The passage extraction and question selection are processed together by computing two scores. The passages are scored according to a density score. The candidate answers are scored according to a compacity score. Both scores are based on words, names entities and proper nouns alignments with the question. A comparison between average scores for initial questions and right or wrong answered typed questions let us conclude that the difference is not significant. We evaluate the impact of users formulation by considering all potential answers issued from this step. This showed that 13% of well tagged questions could no longer be well answered after this step.

Finally, the maximum reachable rate of well answered questions is 72% of questions typed by users. The reached rate is 60% over all users. The correlation between type of mistake made questions and right answers reveals that 59% of questions with a syntactical mistake were well answered while only 31% of questions containing orthographic mistakes were. Surprisingly, the system found the right answer for 56% of sentences with misspelled proper nouns. This rate would be probably lowered by the document retrieval step.

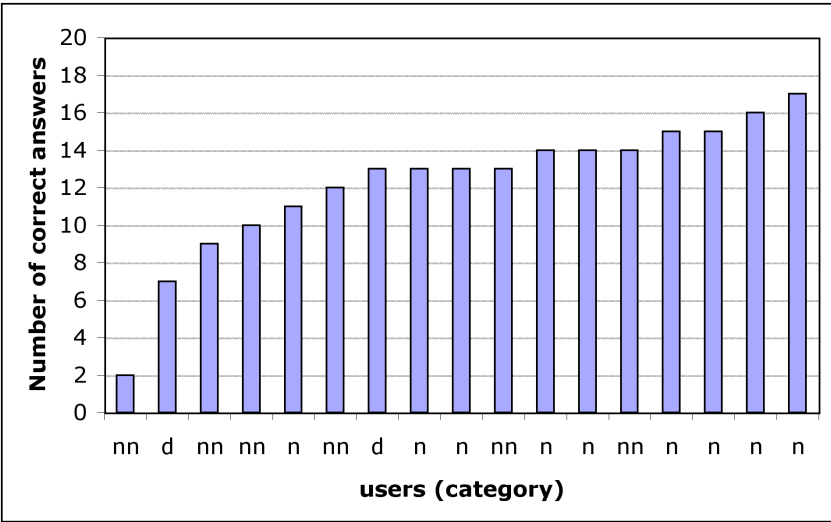


Figure 1: Number of questions correctly answered by user. Users categories are native (n), non native (nn) and dyslexic (d).

The graph on Figure 1 shows the number of questions well answered by the system among the 20 re-typed questions, for each user in the experiment, of one of the categories native, non native or dyslexic. This shows that even for native users the performances of the system decrease (the initial number of correct answers was 20). This means that robust systems must be also designed

for users without special linguistic needs.

## 4 Future work

Not only evaluating the robustness of the QA systems, our corpus of questions might allow a classification of human errors possibly encountered in the input of such systems and so a user profiled QA. This also raises the issue of detecting types of errors made by the user.

We also collected a second corpus of questions typed by dyslexic children. The procedure to collect this corpus is nearly the same as for the previous corpus. There were only five questions. The questions contain only words in MANULEX lexical database (Lété et al., 2004), a lexicon of French word used by children between 8 and 10 years old approximately. 19 dyslexic children under the instructions of a person have typed the questions. This corpus has been so far used in order to evaluate a rewriting method (Sitbon et al., 2007). The evaluation of the whole QA process including rewriting still needs to be done. More generally, the evaluation of the process after a pre-processing with a spell checker or a rewriting system is planned.

A more intensive study of questions typed by non-natives English speakers can also be of interest. 100 French graduate students who learned English for at least 6 years at school will provide data for such a corpus. The experiment is composed of 40 questions from CLEF European evaluation campaign, divided into two sets of 20 questions each. The instructions are also directed with voice. In order to avoid a simple translation of the question, we give the same kind of instructions as for the previous experiment. They are given in French.

## References

(Crestani F., 2000) F. Crestani, 2000. Effects of word recognition in spoken query processing In *Proceedings of the IEEE Advances in Digital Libraries*, Washington, DC, USA.

(Gillard et al., 2006) L. Gillard, L. Sitbon, E. Blaudez, P. Bellot, M. El-Bèze, 2006. The lia at qa@clef2006. Dans les actes de *Cross Language Evaluation Forum (CLEF) 2006*, Alicante, Espagne.

(Lété et al., 2004) B. Lété, L. Sprenger-Charolles, P. Colé, 2004. Manulex : A grade-level lexical database from french elementary-school readers. *Behavior Research Methods, Instruments, and Computers* 36, 156–166.

(Ruch, 2002) P. Ruch, 2002. Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, Taipei, Taiwan, 1–7. Association for Computational Linguistics.

(Sitbon et al., 2007) L. Sitbon, P. Bellot, P. Blache, 2007. Phonetic based sentence level rewriting of questions typed by dyslexic spellers in an information retrieval context . In *Proceedings of Interspeech - Eurospeech 2007*, Antwerp, Belgium.