Evaluating Robustness Of A QA System Through A Corpus Of Real-Life Questions

Laurianne Sitbon[1,2] **Patrice Bellot**[1] **Philippe Blache**[2]

(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

Abstract

This paper presents the sequential evaluation of the question answering system SQuaLIA. This system is based on the same sequential process as most statistical question answering systems, involving 4 main steps from question analysis to answer extraction.The evaluation is based on a corpus made from 20 questions taken in the set of an evaluation campaign and which were well answered by SQuaLIA. Each of the 20 questions have been typed by 17 participants natives, non natives and dyslexics. They were vocally instructed the target of each question. Each of the 4 analysis step of the system involves a loss of accuracy, until an average of 60% of right answers at the end of the process. The main cause of this loss seem to be the orthographic mistakes users make on nouns.

The most natural way for users to query search engines for specific information is to type a question. From this natural interaction the interest for efficient question raises Many evaluation campaigns on question answering (QA) systems have been organized for years. The international TREC[1] (Text REtrieval Conference) conference proposes a track about it , the European CLEF[2](Cross Language Evaluation Forum) campaign proposes cross evaluation in eight languages, NTCIR[3](NII Test Collection for IR systems) includes a QA track in three languages and the French Technolangue EQueR[4](Evaluation en Questions Réponses) evaluation focused on 500 questions made from the data (Ayache et al., 2006). But the questions asked in those campaigns are manually checked for being well formed and complete enough. 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 have been testing the robustness of search engines with automatic transcription of spoken queries (Crestani, 2000) or with automatically degraded text entries (Ruch, 2002). The *Confusion track* (Kantor and Voorhees, 1997) of TREC evaluation campaing focuses on difficult queries for document retrieval. The idea of analysing the performance of a QA system at the different steps of the process has been proposed by (Moldovan et al., 2003) where they analyse the causes of failure with standard questions only. The work we present here focuses on the evaluation of our QA system (Gillard et al., 2007) through a corpus created on purpose. This paper will first present the objectives, the protocol and some observations made on this corpus of semi-spontenaously typed questions. The second section will provide a step by step evaluation of our QA system with these new data.

[1] http://trec.nist.gov/data/qa.html
[2] http://clef-qa.itc.it/2005/
[3] http://www.slt.atr.jp/CLQA/
[4] http://www.technolangue.net/article61.html

1. Corpus constitution

In order to be able to compare our results with results obtained with standard data, we must ensure that the system is potentially able to answer each question. For this purpose we need to guide the user to make him ask a single question on a specific target we know the system can answer. Because many formulations can be used to ask a question and because the words and the syntax chosen by the user can affect the performances of the system, the dictation of a specific question is not suitable.

1.1. Experimental protocol

A web-based approach has been used to acquire data that constitute a 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 amount 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., 2007). 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 set of questions is presented in Table 5 (see appendices).

The main issue is getting enough spontaneous composition 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 topic of the question. Proper nouns must not appear in their correct spelling. That is why we choose to use voice instructions. However, we might not influence the produced syntax of the question by dictating it. Offering the answer as in a jeopardy game has been considered, but it supposes knowledge and it would often lead to many different questions. We finally decided to make users 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 gratuate studies. 9 participants are native French speakers, 6 are non-native (Chinese, German and Spanish living in France) and 2 are dyslexics native speakers who are intuitively the most concerned with robustness issues.

1.2. Observations

A set of questions typed by users for one standard question is given in Table 4 of the appendice section. This is a good example of the tendance of users to employ various formulations for the same target. All typed questions have been manually annotated in order to determine if they contain at least one mistake falling into one of the 5 categories : missing "?", accentuation, syntax, name (proper noun) spelling, noun spelling. Almost half user fail adding the "?" at the end of at least one question. This cannot have an impact on our system because it doesn't process the punctuation. However, a more wide system able to detect whether 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 use of a *querty* keyboard instead of the user's ability to write well. The percentage of users and the percentage of questions concerned by one of the 3 last categories of mistakes are given in Table 1.

	syntax		Name		noun	
	pers	sent	pers	sent	pers	sent
F (9)	67%	12%	100%	18%	44%	4.5%
NF (6)	100%	36%	100%	39%	100%	28,5%
D (2)	100%	25%	100%	30%	100%	17.5%

Table 1: Distribution of mistakes for each category of users (french speakers (F), non native french speakers (NF) and dyslexics (D) : percentages of users and percentages of sentences involved in each of the three categories of mistakes : syntactical mistake, name mispell or noun spelling mistake.

6 of the native French users and all non-native and dyslexic users made 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 spelling mistake on a name. The rate of names misspelled in questions typed by non-native users is 39%. Half native users and all dyslexic and non-native users misspell nouns. The latter mispell nouns in almost a third of the questions. We can notice that the proportion of questions with a noun mispelled is far lower than those with syntax or name mistakes. This is even more surprising from the dyslexics users who were expected to show higher rates.

2. 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 passages inside documents (according to the terms of the question and its expected answer type),
- scoring candidate answer (all entities corresponding to the expected answer type) in these passages according to a distance to the terms of the question.

A question is considered well answered if an answer judged as correct 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[5]) so we constructed our own[6].
Our study analyses step by step the performance loss compared to the process of the standard question and compared to the previous step considering the categories of mistakes and the categories of users.
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 document retrieval system SMART drops by 18.7% on a document retrieval task when at 15% of the words of the queries are automatically corrupted.

2.1. Evaluation of the expected answer type categorisation

Each question is assigned a named entity category in order to make it match with candidate answers. The categories can be at different levels of precision, from very general (*Person Name, Location, ...*) to very specific (*Inventor, River, ...*). If the expected answer type is not detected, wrong or underspecified (assigned a too general category), the question answering system is likely to fail (Sitbon et al., 2006). SQuaLIA bases the categorisation on Sekine's named entity hierarchy (Sekine et al., 2002) that identifies 150 different categories. For most of our users, the expected answer type computed by the system is incorrect or missing for 20% to 40% of typed questions. On average the system was unable to asign an anwer type to 20% of the questions and assigned an under specified answer type to 16% of the questions.
The wrong or missing categorisations are partially due to spelling errors and also to the words chosen by the user to

[5] `http://trec.nist.gov/data/qa/2004_qadata/04.patterns.zip`

[6] This reference has been validated and is now available

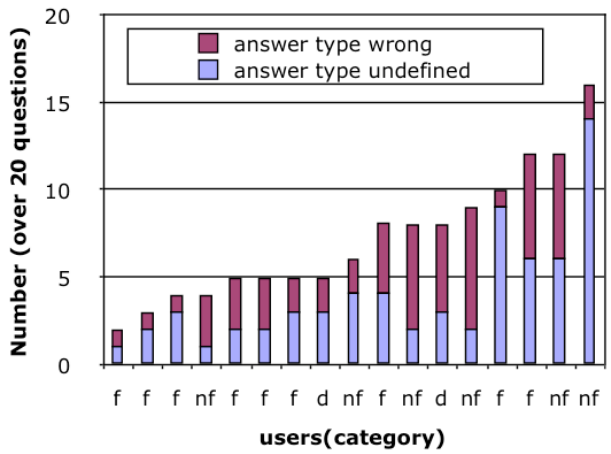

Figure 1: Distribution of errors in expected answer type categorisation (wrong and undefined type) for each user. Users categories are native (n), non native (nn) and dyslexic (d).

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.

syntax	Names	nouns
46%	40.7%	66.7%

Table 2: Correlations between categories of mistakes and answer type failures expressed through the percentage of questions having a mistake of the category being under-specified or not assigned answer type.

Because the system extracts the potential answers only according to the expected answer type, questions that has not been assigned such type cannot be answerd. Thanks to the hierarchy established between names entities categories, underspecified question can still be correctly answered. After the answer type categorisation step, the maximum reachable accuracy of SQuaLIA is 80% of correct answers.

2.2. Evaluation of passage selection and answer extraction

The passage extraction and answer selection are processed together by computing two scores. The passages are scored according to a density score. This score takes into account for each set of 3 sentences the distance in the whole document to each characteristic object of the question (keywords, named entities and proper nouns). The candidate answers are all named entities in the document that correspond to the expected answer type, having the same category or a category higher or lower in the hierarchy. These candidates are scored according to a compacity score. The compacity score takes into account the distance between the candidate and the closest occurrence of each characteristic object of the question.
We aim to evaluate the impact of spontaneous questions on the answer selection step. A first approach is to look at the evolution of the scores between the processing of standard questions and the processing of spontaneous questions. A comparison of average differences of density and compacity scores of processes leading to both wrong and correct answers between standard questions and users questions is reported in Table 3. Both scores deviations are very close to each other for wrongly and correctly answered questions which cannot let us conclude they are representative of a correlation with the performance loss of the system.

	density	compacity
wrong answer	0.075	0.512
correct answer	0.058	0.306

Table 3: Average differences of density and compacity scores of processes leading to both wrong and correct answers between standard questions and users questions.

We can evaluate the passage selection step another way by looking for the correct answer for each question in all proposed answers instead of the 5 first only. It shows whether at least one of the selected passages contain the correct answer. This evaluation shows that 13% of the questions that was assigned an expected answer type had no chance to have a correct answer no matter how the proposed answers are organised.

2.3. Global evaluation

Lastly, when considering the limits imposed by missing answer type recognition and passage selection, the maximum reachable rate of well answered questions is 72% of questions typed by users. The reached rate is 60% over all users. This means that when a correct answer is retrieved, it appears in the 5 first proposed answers in most cases.
More precisely, the graph on Figure 2 provides the number of correct answers given by the system for each user. The system answers correctly in average to 10 question over 20 for dyslexic and non native users, and to 14 questions over 20 for native 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 when the document retrieval step will be involved in the process.
An important result of this evaluation is that even for native users the performances of the system decrease. This means that robust systems must be also designed for users without special linguistic needs.

3. Another corpus from dyslexic children

In order to study how the system can manage queries from users with strong issues in writing, we also collected another 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

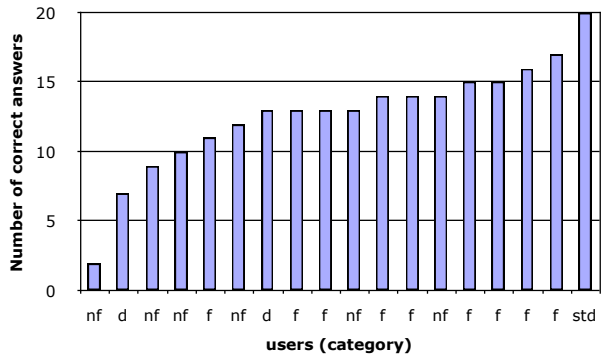

Figure 2: Number of questions correctly answered for each user and standards questions (std). Users categories are native (n), non native (nn) and dyslexic (d).

questions contain only words falling in MANULEX lexical database (Lété et al., 2004), a lexicon of French words used by children between 8 and 10 years old. 19 dyslexic children have typed the questions under the instructions of a person. This corpus has been so far used in order to evaluate a rewriting method (Sitbon et al., 2007). However, this approach provides several hypothesis of the correct form of the question or a weighted vector of the words possibly appearing in the question instead of a correct sentence. The system proposed by (Wolf and Raj, 2002) to deal with similar entries of a document retrieval system with audio queries can provide some solutions.

4. Conclusion and future work

The work presented in this paper leads to the constitution of a reusable corpus[7] of questions typed by real users in almost spontaneous conditions. The corpus is reusable because it is based on standard questions extracted from evaluation campaign dataset. The evaluation of a modular question answering system with these data reveals that human formulations of questions implies decreasing performances of all of its steps. This decrease is related to language mistakes in most cases.

Not only evaluating the robustness of question answering systems, our corpus of questions might allow a classification of human errors possibly encountered in the input of such systems and a user profiled question answering approach. This also raises the issue of detecting types of errors made by the user. A major interest of such classifier would be to allow personalised spell checking and make the system more robust by providing better entries. If such systems can be used in order to correct typos, this may not work as well for users with special needs. (James and Draffan, 2004) highlights the low accuracy of standard spellcheckers for dyslexics.

The evaluation of the whole QA process including rewriting with classical spell checkers or with advanced rewriting scheme still needs to be achieved. More generally, the evaluation of the process after a pre-processing with a spell checker or a rewriting system is planned. We also plan an evaluation reusing standards modules of a standard system[8] and more specifically the information retrieval step.

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.

5. References

Christelle Ayache, Brigitte Grau, and Anne Vilnat. 2006. Equer : the french evaluation campaign of question answering system equer/evalda. In *5th international Conference on Language Resources and Evaluation (LREC 2006)*, pages 1157–1160, Genoa, Italy, May.

Fabio Crestani. 2000. Effects of word recognition in spoken query processing. In *Proceedings of the IEEE Advances in Digital Libraries*, Washington, D.C, USA.

Laurent Gillard, Laurianne Sitbon, Eric Blaudez, Patrice Bellot, and Marc El-Bèze. 2007. Relevance measures for question answering, the lia at qa@clef2006. *LNCS special issue*, pages 440–449.

Abi James and EA Draffan. 2004. The accuracy of electronic spell checkers for dyslexic learners. *PATOSS bulletin*, August.

Paul B. Kantor and Ellen Voorhees. 1997. Report on the trec-5 confusion track. In *TREC-5 Text REtrieval Conference No5,*, pages 65–74, Gaithersburg, MD , ETATS-UNIS.

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

Dan Moldovan, Marius Pasca, Sanda Harabagiu, and Mihai Surdeanu. 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions On Information Systems*, 21(2):133–154, April.

Patrick 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, pages 1–7, Taipei, Taiwan. Association for Computational Linguistics.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Actes de LREC 2002*, Las Palmas, Canary Islands, Spain.

Laurianne Sitbon, Laurent Gillard, Jens Grivolla, Patrice Bellot, and Philippe Blache. 2006. Vers une prédiction automatique de la difficulté d'une question en langue naturelle. In *13ième conférence Traitement Automatique des Langues Naturelles (TALN)*, pages 337–346, Louvain, Belgique, 10-13 April.

[7] the corpus is available for research purposes on request to the authors

[8] OpenEphyra QA system is now available into open source (`http://www.ephyra.info/`)

Laurianne Sitbon, Patrice Bellot, and Philippe 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.

Peter Wolf and Bhiksha Raj. 2002. The merl spoken-query information retrieval system. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 2, pages 317–320, August.

Appendices

Comment s'appelle le président Tchèchène
Quel est le nom du Président tchetchène ?
Quel est le nom du president de la tchetchenie?
Comment s 'appelle le president de la Tchetchenie?
Quel est le nom du président de la Tchétchénie ?
Comment s'appelle le président tchétchène?
quel est le nom du président de la tchétchénie
Comment s'appelle le président de le Tchetchenie?
qui est le président de la tchetchenie
quel est le nom du président
Qui est le Président de la Tchétchénie ?
Quelle est le nom du président de la Cherchenie?
Quelle est le nom du président Tchétchenne ?
Qui est le président de la tchétchénie?
qui est le presiden de la tchétchénie
Qui est le président de la Tchetchenia?
Quel est le nom du président de la Tchechenie ?

Table 4: Questions typed by the users for the standard question Comment s'appelle le président Tchèchène?

Comment s'appelle le maire de Bastia ?
Combien de personnes souffrent d'acné en Suisse ?
Quelle est la monnaie nationale en Hongrie ?
A combien de kilomètres de Paris se trouve la gare de Tours ?
Comment s'appelle le président Tchétchène ?
Quel âge a l'abbé Pierre ?
Combien y a t il de chômeurs en Europe ?
Qui est le frère de la princesse Leia ?
Combien y a t il d'habitants en Lettonie ?
Quel grade occupe Juan Carlos Rolon dans la marine ?
Quand est mort Kurt Cobain ?
Quel est le nom du roi du Maroc ?
Quelle est la capitale de Terre Neuve ?
En quelle année Hitler est arrivé au pouvoir ?
Qui est le président d'Aérospatiale ?
Combien de personnes sont mortes dans des accidents de la route en 1997 ?
Où se situe San Cristobal de Las Casas ?
Quand a été votée la loi Evin ?
En combien de langues a été publié le "Petit Prince" ?
Quel journal publie chaque année le top 50 des personnalités ?

Table 5: Standard questions used as a base of the corpus of semi spontaneously typed questions.