Calvo Lance, M.; Buscaldi, D.; Rosso, P. (2012). Voice-QA: evaluating the impact of misrecognized words on passage retrieval. En Advances in Artificial Intelligence - IBERAMIA 2012. Springer Verlag (Germany). 462-471. doi:10.1007/978-3-642-34654-5_47.

Voice-QA: Evaluating the Impact of Misrecognized Words on Passage Retrieval

Marcos Calvo[1], Davide Buscaldi[2], and Paolo Rosso[1]

[1] Grup d'Enginyeria del Llenguatge Natural i Reconeixement de Formes (ELiRF)
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain
{mcalvo,prosso}@dsic.upv.es
[2] Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier
118 Route de Narbonne, F-31062, Toulouse Cedex 9, France
davide.buscaldi@irit.fr

Abstract. Question Answering is an Information Retrieval task where the query is posed using natural language and the expected result is a concise answer. Voice-activated Question Answering systems represent an interesting application, where the question is formulated by speech. In these systems, an Automatic Speech Recognition module can be used to transcribe the question. Thus, recognition errors may be introduced, producing a significant effect on the answer retrieval process. In this work we study the relationship between some features of misrecognized words and the retrieval results. The features considered are the redundancy of a word in the result set and its inverse document frequency calculated over the collection. The results show that the redundancy of a word may be an important clue on whether an error on it would deteriorate the retrieval results, at least if a closed model is used for speech recognition.

Keywords: Voice-activated Question Answering, Passage Retrieval, Term Informativeness

1 Introduction

Question Answering (QA) is an Information Retrieval (IR) task in which the query is posed in natural language and the expected result is a concise answer. Currently, most QA systems accept written sentences as their input, but in the last years there has been a growing interest in systems where the queries are formulated by voice; as can be seen in, for example, [2] and [5]. In fact, due to this interest some Evaluation Conferences, such as the CLEF (Cross-Language Evaluation Forum) competition have included a voice-activated Question Answering task in different languages [7].

In general, as it is shown in Figure 1, a QA system is composed by an analysis module, which determines the type of the question; a Passage Retrieval (PR)

module, which uses IR techniques to retrieve passages where the answer might be contained; and an answer extraction module, which uses NLP techniques or patterns to extract the answer from the passages. In addition, if the input to the system are utterances, an Automatic Speech Recognition (ASR) module can be used to transcribe the vocal input. One option is to "plug" the ASR before the QA modules, in such a way that the input to the QA system is the sentence (or the n-best sentences) given by the ASR. Figure 1 shows the architecture of such a system, where the output is given back to the user by means of a Text-To-Speech synthesizer (TTS).

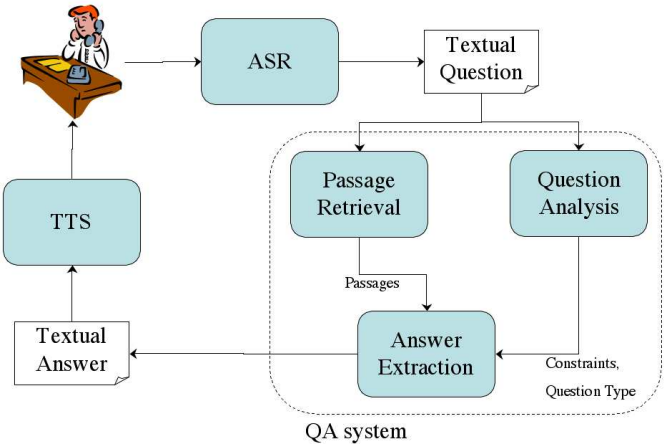

Fig. 1. Modules of a voice-activated Question Answering system

In these systems, recognition errors can strongly modify the meaning of the query. In fact, these errors are crucial in the case of Named Entities (NEs), since they are usually very meaningful words. Unfortunately, NEs are often very difficult to be recognized properly, sometimes because they are in a language different to the user's one, which makes this fact one of the biggest open challenges in voice-activated QA. From an IR perspective, NEs can be characterized by their high IDF (Inverse Document Frequency) and redundancy in the retrieved passages. Thus, our hypothesis is that recognition errors on words with a high IDF and that are redundant in the retrieved passages are key, since the object of the question is lost.

Our aim is to study the correlation between the recognition errors on question words with the above characteristics and the resulting errors in the PR module. We limited our study to this phase and did not take the full QA system because the errors in the other modules are so important that can mask the retrieval errors [4]. We computed the IDF of the words of the original sentence that were misrecognized by the ASR both over the document collection and the passages

retrieved by the PR engine using the correct sentence. This experiment was performed for several language models with a different number of NEs.

The rest of the paper is structured as follows. In Section 2 we explain how a voice-activated Passage Retrieval system works. Then, in Section 3, we describe the Passage Retrieval system that we have used in the experimentation we report in this paper. In Section 4 a discussion about some interpretations of the IDF is provided and in Section 5 we present how we have measured the performance of the PR module. Next, we detail the experimentation performed and discuss the obtained results. Finally, we draw some conclusions.

2 Voice-Activated Passage Retrieval

A voice-activated QA system consists of several connected modules that, working together, aim to find an answer to a question posed by the user in natural language. One of the possible architectures is the one shown in Figure 1, where an ASR system has been joined to a traditional (textual) QA system and the output is optionally given back to the user by means of a Text-To-Speech synthesizer. This is the architecture we used in our experiments.

As shown in Figure 1, among the modules of a QA system it is found the Passage Retrieval (PR) module. The purpose of this module is to extract from a collection of texts a number of them that are relevant for the input question. Another module is the Question Analysis one, which aims to determine the kind of a question (e.g. if the user is expecting a name or a date) and some additional constraints. The Answer Extraction module analyses the passages previously retrieved and using the information given by the Question Analysis module looks up the answer to the original question.

The most critical part in a QA system is the Question Analysis module. In fact, in [4] it is shown that 41.6% of the errors in a Question Answering system derive from an error in the Question Analysis phase, with more than 33% due to the identification of the question type. Answer extraction is also an important source of errors, with 18.7% of the total number of errors in QA. However, Passage Retrieval is shown to be a limited source of errors, as mistakes derived from this module account only for 1.6% on the performance in QA.

For this reason, we have focused our work on the study of the effects of the ASR errors on Passage Retrieval, where the effects of a badly recognized question are directly reflected on the ranking of passages and can be detected. These effects can not be discovered using the complete QA system, since the errors in Question Analysis and Answer Extraction would mask most of the effects of the question recognition over the retrieval phase.

3 The JIRS Passage Retrieval System

In our study, we have used the JIRS Passage Retrieval system. This PR system uses a weighting scheme based on n-grams density. It was proved in [1] that this approach is more effective in the PR and QA tasks than other commonly used IR

systems based on keywords and the well-known TF.IDF weighting scheme. So, JIRS works under the premise that, in a sufficiently large document collection, question n-grams should appear near the answer at least once. The architecture of JIRS is shown in Figure 2.

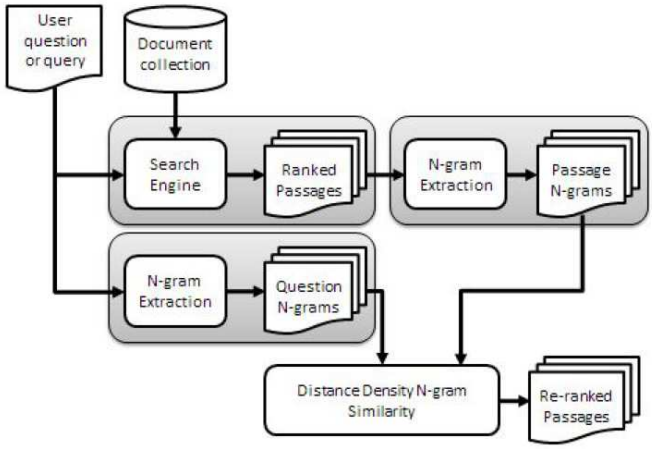

Fig. 2. Structure of the JIRS Passage Retrieval engine

The first step consists in extracting passages which contain question terms from the document collection, which is done using the standard TF.IDF scheme. Subsequently, the system extracts all question k-grams (with $1 \leq k \leq n$, where n is the number of terms of the question) from both the question and each of the retrieved passages. The output of the system is a list of at most M passages (in our experiments we set $M = 30$) re-ranked according to a similarity value calculated between the passages and the question. The similarity between the question q and a passage p is defined in Equation 1.

$$Sim(p,q) = \frac{\sum_{\forall x \in (P \cap Q)} \frac{h(x)}{1 + \alpha \cdot ln(1 + d(x, x_{max}))}}{\sum_{i=1}^{n} w(t_{q_i})} \tag{1}$$

In this equation P is the set of k-grams ($1 \leq k \leq n$) contained in passage p and Q is the set of k-grams in question $q = (t_{q_1}, \ldots, t_{q_n})$; n is the total number of terms in the question. $w(t)$ is the term-weight, determined by:

$$w(t) = 1 - \frac{log(n_t)}{1 + log(N)} \tag{2}$$

Here n_t represents the number of sentences in which the term t occurs and N is the number of sentences in the collection.

The weight of each k-gram $x = (t_{x_1}, \ldots, t_{x_k})$ is calculated by means of the function $h(x) = \sum_{j=1}^{k} w(t_{x_j})$.

Finally, the distance $d(x, x_{max})$ is calculated as the number of words between any k-gram x and the one having the maximum weight (x_{max}). α is a factor, empirically set to 0.1, that determines the importance of the distance in the similarity calculation.

4 Estimating the informativeness of a term

We can intuitively see that, given a set of documents D, a word w that appears in all of them will not be very informative, since it makes no distinction between the documents. However, if w is found in just one document, then it will probably be one of the most informative for that document. This idea, extended to all the range between these two cases, is the one that underlies the IDF formula [3], which can be written as

$$IDF = -log(\frac{|D(w)|}{|D|}) \tag{3}$$

where $|D(w)|$ is the number of documents where the word w is found and $|D|$ is the cardinality of the collection. According to this formula, the higher the IDF for a word w, the more relevant it is in the collection.

From a PR point of view, two interpretations can be given to the IDF depending on the set of documents considered. On one hand, if we calculate the IDF of a word over the whole document collection, this value represents how important is the word in it, which is called *term informativeness*. On the other, if the set is constituted by the passages retrieved by the PR engine given a query in which the word appears, the IDF can be used to calculate the *redundancy* of that word on this result set. In this case, a low IDF indicates a high redundancy. Due to the filter constituted by the PR phase, a redundant word in this set may not be a stop-word, but a term highly related to the query submitted to the system.

It might happen that, given a query containing a word w, it does not appear in any of the documents returned by the PR engine using this query. Thus, in order to avoid zeroes as a result of the division, in the case of the redundancy we have slightly modified the IDF formula by adding one to both elements of the fraction. This is shown in Equation 4.

$$redundancy = -log(\frac{|D(w)| + 1}{|D| + 1}) \tag{4}$$

5 Measuring the performance of the Passage Retrieval module

The output of the PR module is a ranked list of passages. So, it is interesting to know if this ranking would match what a user would expect from the PR system.

Among the IR measures that are commonly used to take into account the position of the passages, we chose the Normalized Discounted Cumulative Gain (nDCG), since it is the one that best models the user's preferences, according to [6].

In order to calculate IR measures such as the nDCG, it is necessary to have a set of *relevance judgments*, which is a set of documents considered to be relevant for the query. In our case, this set was built using hand-made answer patterns and regular expressions to test if a passage contains the answer.

Normalized DCG at position π is defined as:

$$nDCG_\pi = \frac{DCG_\pi}{IDCG_\pi} \tag{5}$$

where $IDCG_\pi$ is the "ideal" DCG obtained by ranking all the relevant documents at the top of the ranking list, in order of relevance, and $DCG_\pi = rel_1 + \sum_{i=2}^{\pi} \frac{rel_i}{\log_2 i}$, where rel_i is the degree of relevance of the result at position i.

6 Experiments and results

For our experiments we have used the questions in Spanish from the CLEF[3] QA 2003-2006 contests. The target collection (the set of documents where the answer should be found) is composed by documents of the EFE (Spanish news agency) of the years 1994 and 1995. The 1800 questions available were split into a set of 1600 for training and the remaining 200 for test. The latter were uttered by a specific user (because this corpus does not include utterances of the written questions) and constitute the input to the ASR.

We have trained a generic Language Model (LM) for the ASR with just the training questions, separating the NEs in a category. Then, we have added more elements to this set according to its frequency in the document collection. So, we can distinguish two types of LMs: the Open Named Entity models, which include only the N most frequent NEs taken from the target collection, and the Closed NE models, which include all the test NEs plus a number of the most frequent NEs taken from the same collection, in order to amount up to N Named Entities. In both cases the minimum number of NEs considered in the category was $4,000$ and the maximum $48,000$. As the original corpus does not have the NEs tagged in any way, previously to this process we automatically tagged the corpus using a POS-tagger.

For this experimentation we used both Open and Closed Named Entity models because they simulate different kinds of real applications of QA systems. On one hand, an Open NE model simulates a situation where the number of NE that the system has in its vocabulary is limited but the users are allowed to ask about whatever they want. On the other, the Closed NE model simulates a more restricted domain where the NEs the user can ask about are limited and known when building the system.

[3] http://www.clef-campaign.org

Once all the test questions were recognized using each one of these models, we considered two outputs from the ASR: the recognized sentences themselves and the Word Error Rate (WER). Then, we performed the Passage Retrieval process, taking the recognized sentences as its input.

As explained before, the ASR that works before the PR may modify the original sentence by introducing recognition errors. Thus, it would be interesting to relate the nDCG values obtained for each of the LMs to the ones achieved if the input to the PR process was composed by the correct test questions. For this reason, we have used as the measure of the Passage Retrieval performance for each LM the value *nDCG_diff* defined as in Equation 6, where nDCG(test_sents) stands for the nDCG obtained using the correct original sentence as the input for the Passage Retrieval module, while nDCG(recognized_sents) means the same but taking the output of the ASR module. The average nDCG obtained for the original test set is 0.584.

$$nDCG\_diff = nDCG(test\_sents) - nDCG(recognized\_sents) \tag{6}$$

Finally, we have calculated the term informativeness and the redundancy of the words of the test queries that were misrecognized by the ASR. These calculations were done over the complete target collection and the passages retrieved by the PR engine using the full correct sentence. In the case of the redundancy, we have calculated a composition within each sentence, both using the mean and max operators. The use of these operators is motivated because in the recognized sentences there may be more than one error, so it is reasonable to consider both the word that would give the largest redundancy (max), and the average redundancy considering all the misrecognized words. Also, for both the redundancy and the term informativeness, and for each LM, we have averaged the results obtained for each sentence. The obtained figures are presented in Table 1.

Table 1. Results for the Closed and Open Named Entities Models

	Closed NE Models					Open NE Models				
		avg redund.					avg redund.			
# NE	WER	mean	max	Term inf	nDCG_diff	WER	mean	max	Term inf	nDCG_diff
4000	0.265	**0.348**	**0.529**	2.329	**0.151**	0.333	0.522	0.755	2.379	0.265
8000	0.298	**0.419**	**0.606**	2.020	**0.183**	0.347	0.530	0.762	2.526	0.262
12000	0.305	**0.432**	**0.614**	2.011	**0.197**	0.351	0.531	0.750	2.455	0.273
16000	0.310	**0.448**	**0.636**	2.102	**0.192**	0.350	0.533	0.759	2.364	0.252
20000	0.310	**0.454**	**0.644**	2.143	**0.192**	0.348	0.534	0.760	2.389	0.248
24000	0.306	**0.456**	**0.648**	2.091	**0.195**	0.342	0.531	0.760	2.292	0.246
28000	0.312	**0.461**	**0.660**	2.159	**0.201**	0.344	0.526	0.755	2.297	0.242
32000	0.319	**0.487**	**0.689**	2.241	**0.208**	0.342	0.533	0.764	2.336	0.232
36000	0.319	**0.489**	**0.691**	2.293	**0.205**	0.344	0.534	0.766	2.388	0.229
40000	0.319	**0.496**	**0.698**	2.330	**0.199**	0.342	0.539	0.768	2.409	0.222
44000	0.321	**0.493**	**0.698**	2.298	**0.203**	0.345	0.536	0.768	2.390	0.226
48000	0.321	**0.493**	**0.698**	2.298	**0.204**	0.345	0.535	0.768	2.375	0.226

The different behaviour with respect to the number of NEs is due to the own nature of the models: in the closed NE models, the smaller the NE set, the lesser the probability of error is. This is opposed to what happens in the open models, where introducing new NEs increases the chances of finding the right NE among the elements of the set and so recognizing it properly.

With regard to the relationship between redundancy and nDCG in the retrieved passages, Table 1 shows that in the closed NE models the lower the redundancy of the misrecognized words, the lower the nDCG difference is (see also Figure 3). Indeed, their Pearson correlation coefficient amounts to 0.9408. In the open NE models (Figure 4) this correlation is not observed. None of the models show a correlation between the nDCG and the term informativeness of the misrecognized words, somehow surprising as we expected that errors on words with high IDF should be more important.

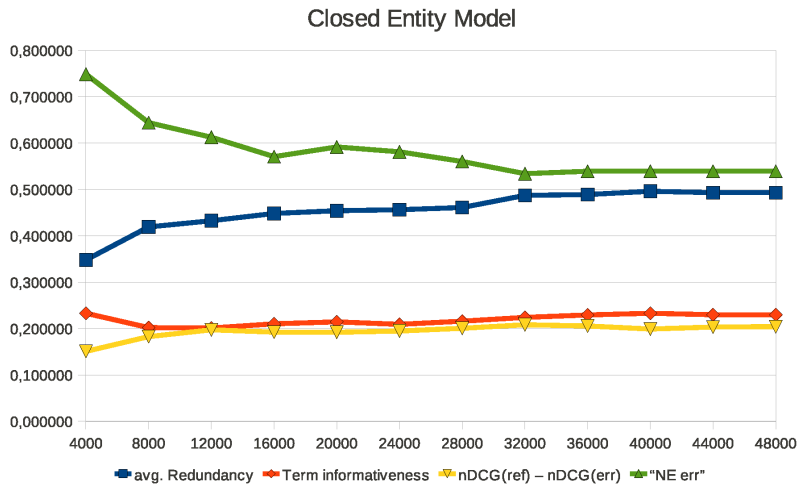

Fig. 3. Closed Entity Model Results. Term informativeness values have been divided by 10. "NE err": error on NEs.

Our interpretation of these results is that in the closed NE models the errors on non-NE words, which may have a high redundancy in the result set, are very important. As shown on Table 1, the error on NEs is inversely proportional to redundancy, indicating that NEs are less redundant than some other words.

7 Conclusions and future work

In this paper we attempted to find a relationship between the redundancy and the term informativeness of misrecognized terms on the output of a PR module of a voice-activated QA system and its performance. We used both closed and

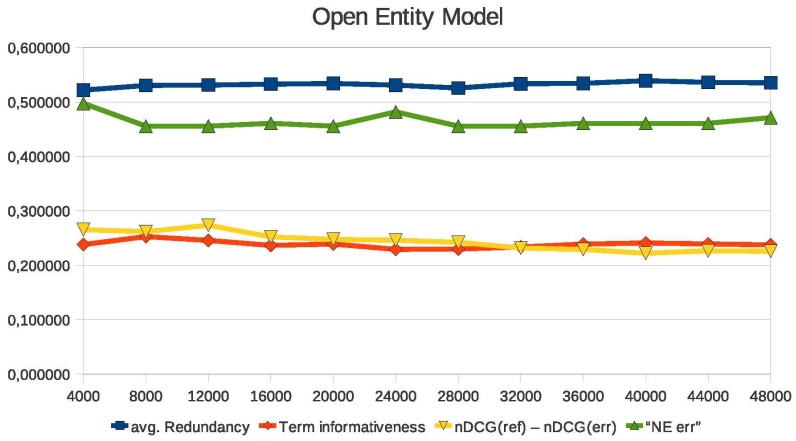

Fig. 4. Open Entity Model results. Term informativeness values have been divided by 10. "NE err": error on NEs.

open NE models as the input to the ASR module. Our results show that the term informativeness, measured as the IDF, is not an indicator of the relevance of the error on that term for the PR process. However, the redundancy of a term in the retrieved passages seems to be an important clue on whether an error on that term will produce a worse result, at least if a closed NE model is used.

As future work, it would be interesting to investigate the relationship between other informativeness measures on the misrecognized words and the nDCG differences, as well as to use other PR engines and compare the results obtained.

Acknowledgments. This work was carried out in the framework of Text-Enterprise (TIN2009-13391-C04-03), Timpano (TIN2011-28169-C05-01), WIQ-EI IRSES (grant no. 269180) within the FP 7 Marie Curie People, FPU Grant AP2010-4193 from the Spanish Ministerio de Educación (first author), and the Microcluster VLC/Campus on Multimodal Intelligent Systems (third author).

References

1. Buscaldi, D., Gómez, J.M., Rosso, P., Sanchis, E.: N-Gram vs. Keyword-Based Passage Retrieval for Question Answering. In: Proceedings of CLEF 2006. pp. 377–384 (2006)
2. Harabagiu, S., Moldovan, D., Picone, J.: Open-domain voice-activated question answering. In: Proceedings of the 19th international conference on Computational linguistics. pp. 1–7. COLING '02 (2002)
3. Jones, K.: Index term weighting. Information Storage and Retrieval 9(11), 619–633 (1973)
4. Moldovan, D., Pasca, M., Harabagiu, S., Surdeanu, M.: Performance Issues and Error Analysis in an Open-Domain Question Answering System. In: Proceedings

of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 133–154. New York, USA (2003)
5. Rosso, P., Hurtado, L.F., Segarra, E., Sanchis, E.: On the voice-activated question answering. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42(1), 75 –85 (2012)
6. Sanderson, M., Paramita, M.L., Clough, P., Kanoulas, E.: Do user preferences and evaluation measures line up? In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 555–562. SIGIR '10, ACM, New York, NY, USA (2010)
7. Turmo, J., Comas, P., Rosset, S., Galibert, O., Moreau, N., Mostefa, D., Rosso, P., Buscaldi, D.: Overview of QAST 2009. In: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, Lecture Notes in Computer Science, vol. 6241, pp. 197–211. Springer (2009)