Language Modelization and Categorization for Voice-Activated QA

Joan Pastor, Lluís-F. Hurtado, Encarna Segarra, and Emilio Sanchis

Grup d'Enginyeria del Llenguatge Natural i Reconeixement de Formes,
Department de Sistemes Informàtics i Computació,
Universitat Politècnica de València, València, Spain
{jpastor,lhurtado,esegarra,esanchis}@dsic.upv.es
http://dsic.upv.es/users/elirf

Abstract. The interest of the incorporation of voice interfaces to the Question Answering systems has increased in recent years. In this work, we present an approach to the Automatic Speech Recognition component of a Voice-Activated Question Answering system, focusing our interest in building a language model able to include as many relevant words from the document repository as possible, but also representing the general syntactic structure of typical questions. We have applied these technique to the recognition of questions of the CLEF QA 2003-2006 contests.

Keywords: Voice-Activated Question Answering, Automatic Speech Recognition, Language Models, Named Entities Recognition.

1 Introduction

In recent years the interest in the development of applications for accessing to large non structured repositories of information of different types, such as text, audio, video, or images has increased. Evaluation conferences, such as TREC[1] and CLEF[2], organize multiple tracks in order to compare the behavior of the different approaches proposed. One of these tracks is Question Answering (QA), which goal is to access to information repositories, accepting questions in natural language.

At this moment, most of the QA systems accept written sentences as input, but in the last years the interest in using voice to ask the questions has increased [10,8], as well as in accessing to large audio repositories. In order to develop real world applications, it would be interesting to design speech-driven systems providing access to information from mobile telephones, tablets or other speech interfaces. The Automatic Speech Recognition (ASR) component of this kind of systems has to present some specific features: to be able to deal with a large vocabulary, to provide a good language model that characterizes the type of questions, and to be prepared to correctly recognize some relevant words that

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 475–482, 2011.

have a great influence in the posterior answer searching process. In particular, the recognition of Named Entities (NE) is one of the main problems to be faced in question answering applications with speech input. NEs are key elements for the search process [5], so misrecognition of a spoken NE can produce serious errors in the search results. Some works related to the NE recognition problem are [1,7,6].

In the literature it has shown that, in general, better results in NE recognition [5] imply better performance of the Voice-activated QA system. In a previous work in our laboratory, some experiments were carried out with the QUASAR QA system [11] with the aim of studying the effect on the QA system accuracy of the word error rate introduced by the ASR system, especially from the perspective of the recognition of NEs. The experiments with simulated speech input (i.e., errors in the input questions -200 questions of the CLEF 2005 Spanish monolingual test set- were introduced) showed that when recognition errors do not affect NEs, the QA system performance is still good, even with a WER of 25%. Error rates greater than 30% made the system behavior deteriorate quickly.

In this work, we present an approach to the ASR component of a QA system, focusing our interest in the language modelization for questions, and in the influence of categorization of relevant words, such as NEs, in the system performance. The language modelization proposed is based on keeping the specific characteristics of the syntax of questions, but adding to the ASR vocabulary different sets of relevant words in order to increase the coverage. In order to determine what are the relevant word candidates to be included in the ASR vocabulary, we have used a Part-of-Speech (POS) tagging tool [9] to find NEs and common nouns in the document repository. Based on their frequency we built different sets of relevant words to study the behavior of the recognition process.

We have applied these technique to the recognition of questions of the CLEF QA 2003-2006 contests. The corpus consist of a set of questions and the target collection (the set of documents to be searched in order to find the answer) composed by documents of the EFE (Spanish news agency) of the years 1994 and 1995. Due to the correlation of the performance of the ASR system and the performance of the whole Voice-Activated QA system, we have decided to give the results in terms of the performance of the ASR system. The experimentation carried out shows that the categorized language models outperform the language model learned only with training questions.

This article is organized as follows, Section 2 presents the adaption in the vocabulary and language model of the ASR system for spoken question recognition. Section 3 presents the experimental set-up. Section 4 presents the results of the evaluation of the performance of the ASR depending on the language model and the amount of the relevant words in the vocabulary. Finally, some conclusions are presented in Section 5.

2 Language Model Estimation

In this section, we are going to describe our proposal for language model estimation for Voice-Activated QA. In this work, we have focused on the data

pre-processing for training a suitable language model in order to achieve better performance in Voice-Activated QA. The main idea is to build a language model able to include as many NEs and other relevant words from the document repository as possible, but also representing the general syntactic structure of typical questions extracted from training data (*eg. Who is...?, When did...? What is...?*). To do that, it is necessary to build a categorized language model where the categories are related to the concepts that could be asked by the user.

Figure 1 shows how the language model for the ASR has been built. We use two different corpora to train the language model. One of them is the set including the training questions, from which the syntactic structure is estimated, and the other one is the document repository, from which the additional information to generalize the categorized language model is obtained. It is interesting to note that not all training *queries* are formulated in an interrogative way (*e.g. What is the capital of France?*) but some are in a declarative way (*e.g. Name some tennis players.*), so our model has to be aware of this in order to have a better recognition performance.

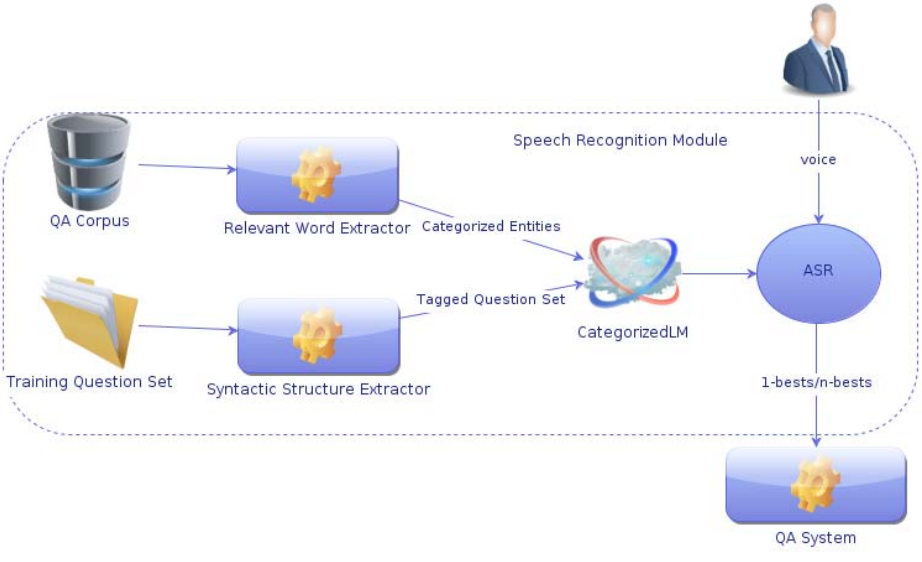

Fig. 1. Speech Recognition Module

2.1 Categorized Language Model

In order to incorporate relevant information to the language model, we have to select which relevant words can be categorized. To do that, after deciding which categories must been selected a POS tagging process is applied in order to obtain the instances of the different categories from the document repository. The key point of our experimentation concerns the amount of relevant words to be included in the language model. The categories used in our language model are:

- Named Entities: usually is the main concept which has been asked by the user.
- Dates and numbers: not all possible combinations of dates (year, month and day) and numbers are included in the training set.
- Common Nouns (CNs): in some cases NEs are not the main concept of the question, or simply there are not in the question. In other cases not only the NE but also some common nouns provides the relevant information to find the answer (e.g Who is the *president* of France?).
- The rest of the words, where, formally, each word belongs to its own class (*i.e.* one class per word).

2.2 Category Members Selection

First, for retrieving the categorized data, we use FreeLing POS Tagging feature [9,2,3]. The words belonging to any of the categories previously described are replaced by their specific tag in the training question set.

Second, we perform the same POS tagging process to the document repository (the target collection where the information has to be retrieved) and we extract the frequency sorted lists of NEs and CNs. The NE list includes more than one million of different elements. Analyzing this set, we have checked that most of these NEs appear just a few times, sometimes due to orthographic mistakes. If we filter out all the NEs that appear less than 10 times, the number of remaining NEs is around $80,000$ and if the threshold is 20 times then the amount of NEs is reduced to around $48,000$. We can assume that the most common NEs in the corpus are the most likely to appear in a question, so they would be added to our NE set, which will be provided as an input to the ASR component. Something similar occurs with the CNs; in this case, the $4,000$ more frequent CNs cover more than the 85% of the CNs present in the training questions.

Each word tagged as NEs or CNs during the tagging process has a confidence score of belonging to that category. Figure 2 shows how the coverage of NEs and CNs increases as more items are included in each category considering only those words with a score of belonging to the category higher than a threshold. It can be seen that the use of different threshold has no influence on the coverage of NEs. Regarding CNs, the more permissive you are the more coverage you get.

2.3 Orthographic Entity Merging

The document repository is a heterogeneous collection of documents that includes all the news published by the EFE agency along two years. For this reason, there are some NEs that appears written in different ways. Usually, these NEs have several orthographic transcriptions with the same phonetic transcription (*e.g, Korea and Corea, Qatar and Catar*).

To avoid this problem, we have merged the entities with the same phonetic transcription keeping the orthographic transcription of the most frequent one. To do that, we have used the Grapheme-to-phoneme tool Ort2Fon [4].

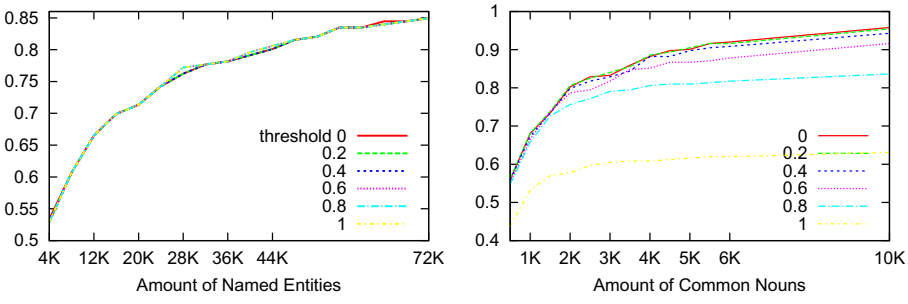

Fig. 2. NE/CN Coverage

There are some especial cases where two different entities has the same phonetic transcription (*e.g Baldi and Valdi*). In this case, due to the fact that the ASR is unable to discriminate between them, the most frequent one is used.

3 Experimentation

3.1 Evaluation Metrics

The most commonly used parameter to measure the performance of ASR is the Word Error Rate (*WER*). Another criterion which indicates the performance of the language model regardless of the ASR used is the percentage of Out-of-Vocabulary words (*OOV*). In addition to WER and OOV, we have defined some other measures related to the recognition of relevant words:

- *NE_ACC*: Named Entities accuracy. A good indicator of the performance of the global Voice-Activated QA system would be determine the set of test NEs properly recognized by the ASR.
- *CN_ACC*: Common Nouns accuracy. In the way that NEs accuracy, CNs accuracy would be a good indicator of the performance of the Voice-Activated QA system.
- *WER_SW*: WER without considering stop words. Usually, stop words are not taken into account by the QA systems.

3.2 Experimental Setup

In our experiments, we have used the questions from the CLEF QA 2003-2006 contests in Spanish. The document repository (the set of documents to be searched in order to find the answer) is composed by documents of the EFE

(Spanish news agency) of the years 1994 and 1995. The set of questions amounts to 1,800 questions divided in two subsets: 1,600 for training and 200 for test. The 200 test questions were acquired by an specific user and are used as input of the ASR.

For the experimentation, three different language models were applied:

- *Single NEs model*: using an incremental number of NE between 4,000 and 48,000 in order to check how an increase of the amount of NE affects to the performance of the recognizer.
- *NEs Modified model*: using the same number of entities as in the previous model, but including the phonetic approach described in section 2.3.
- *NEs/CNs Modified model*: using the same entities as in the *NE Model Modified Model* and including an amount of 4,000 CNs.

4 Results

Table 1 presents the results of the experimentation, It is also included, as a baseline, the language model trained only with the training questions without either categorization nor generalization (*Plane training model*). This Table only shows the best results for each one of the language models. Figure 3 shows, for each model, how the number of relevant words in the model affects the different proposed measures.

Table 1. Experimental results summary

	WER	NE_ACC	CN_ACC	WER_SW	OOV
Plane Training model	0.384	0.420	0.768	0.449	0.222
Single NEs Model (best)	0.326	0.551	0.825	0.402	0.133
NEs Modified Model (best)	0.315	0.546	0.817	0.389	0.127
NEs/CNs Modified Model (best)	0.290	0.537	0.871	0.350	0.089

It can be seen that, for all models, WER and WER without stop words (WER_SW) measures gets worse when a few amount of entities is included. While the WER remains stable for *Single NEs model* and *NEs Modified model*, the performance slightly improves for the *NEs/CNs Modified model*. It is important to see that each improvement has a good impact in all system measures.

Figure 3 shows that in the *Single NE model* the NE accuracy decreases significantly until 20,000 NEs, while in the other models it remains stable. This occurs because some entities, which are well recognized in previous experiments, are confused when the amount of NEs is increased. The *NEs/CNs Modified model* provides a more flexible language model which avoids this problem. Even in the *NEs/CNs Modified model*, the accuracy increases when more NEs are added.

It is interesting to see how the *Single NEs model* has the best Named Entity accuracy with the smallest NEs set (4,000), also the other recognition measures work better with the *NE/CNs Modified Model* while the amount of NE/NCs increases.

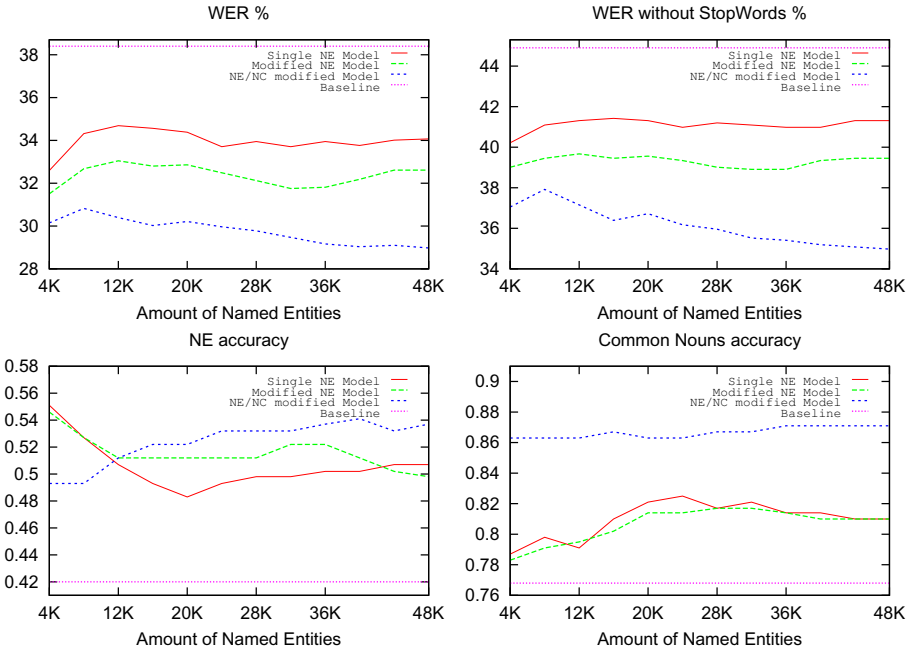

Fig. 3. Language Models results

5 Conclusion and Future Work

In this paper, we have presented an approach to the Automatic Speech Recognition component of a Voice-Activated Question Answering system. We have focused out interest in building a language model able to include as many relevant words from the document repository as possible, but also representing the general syntactic structure of typical questions.

The proposed language models, in which relevant words from the document repository are included, present better results in all the evaluation measures than the language model learned only with training questions (*Plane Training model*).

As future work, we propose first, to take into account non-Spanish Named Entities and their phonetics. Second, we propose to create an interaction mechanism which provides the user with a list of possible NEs to be chosen.

Acknowledgements. Work partially supported by the Spanish MICINN under contract TIN2008-06856-C05-02, and by the Vicerrectorat d'Investigació, Desenvolupament i Innovació of the Universitat Politècnica de València under contract 20100982.

References

1. Akiba, T., Itou, K., Fujii, A.: Language model adaptation for fixed phrases by amplifying partial n-gram sequences. Systems and Computers in Japan 38(4), 63–73 (2007)
2. Atserias, J., Casas, B., Comelles, E., Gónzalez, M., Padró, L., Padró, M.: Freeling 1.3: Five years of open-source language processing tools. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (2006)
3. Carreras, X., Chao, I., Padró, L., Padró, M.: Freeling: An open-source suite of language analyzers. In: Proceedings of the 4th Language Resources and Evaluation Conference (2004)
4. Castro-Bleda, M.J., España-Boquera, S., Marzal, A., Salvador, I.: Grapheme-to-phoneme conversion for the spanish language. In: Pattern Recognition and Image Analysis. Proceedings of the IX Spanish Symposium on Pattern Recognition and Image Analysis, pp. 397–402. Asociación Española de Reconocimiento de Formas y Análisis de Imágenes, Benicàssim (2001)
5. Chu-Carroll, J., Prager, J.: An experimental study of the impact of information extraction accuracy on semantic search performance. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, pp. 505–514. ACM (2007)
6. Harabagiu, S., Moldovan, D., Picone, J.: Open-domain voice-activated question answering. In: Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002, vol. 1, pp. 1–7. Association for Computational Linguistics (2002)
7. Kim, D., Furui, S., Isozaki, H.: Language models and dialogue strategy for a voice QA system. In: 18th International Congress on Acoustics, Kyoto, Japan, pp. 3705–3708 (2004)
8. Mishra, T., Bangalore, S.: Speech-driven query retrieval for question-answering. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5318–5321. IEEE (2010)
9. Padró, L., Collado, M., Reese, S., Lloberes, M., Castellón, I.: Freeling 2.1: Five years of open-source language processing tools. In: Proceedings of 7th Language Resources and Evaluation Conference (2010)
10. Rosso, P., Hurtado, L.F., Segarra, E., Sanchis, E.: On the voice-activated question answering. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews PP(99), 1–11 (2010)
11. Sanchis, E., Buscaldi, D., Grau, S., Hurtado, L., Griol, D.: Spoken QA based on a Passage Retrieval engine. In: IEEE-ACL Workshop on Spoken Language Technology, Aruba, pp. 62–65 (2006)