Dynamic Adaptation of Language Models in Speech Driven Information Retrieval *

César González-Ferreras and Valentín Cardeñoso-Payo

Departamento de Informática, Universidad de Valladolid, 47011 Valladolid, Spain
{cesargf,valen}@infor.uva.es

Abstract. This paper reports on the evaluation of a system that allows the use of spoken queries to retrieve information from a textual document collection. First, a large vocabulary continuous speech recognizer transcribes the spoken query into text. Then, an information retrieval engine retrieves the documents relevant to that query. The system works for Spanish language. In order to increase performance, we proposed a two-pass approach based on dynamic adaptation of language models. The system was evaluated using a standard IR test suite from CLEF. Spoken queries were recorded by 10 different speakers. Results showed that the proposed approach outperforms the baseline system: a relative gain in retrieval precision of 5.74%, with a language model of 60,000 words.

Key words: speech recognition, information retrieval, speech driven information retrieval, language model adaptation, out of vocabulary words

1 Introduction

Using the web to access information is becoming mainstream. People are used to access the world wide web using a personal computer. Moreover, when users access the web, searching is usually the starting point. As more information becomes available, better information retrieval technology is required.

There is also a proliferation of mobile devices that allow access to the web anytime and everywhere. However, their user interface is limited by small displays and input devices (keypad or stylus). Speech can be used to overcome those limitations and provide a more usable interaction. Furthermore, using speech as the input to an information retrieval engine is a natural and effective way of searching information in a mobile environment.

This paper reports on the evaluation of a system that allows users to search information using spoken queries. The front end is a large vocabulary continuous speech recognizer (LVCSR) which transcribes the query from speech to text and puts it through an information retrieval (IR) engine to retrieve the set of documents relevant to that query. A two-pass approach is proposed in order to increase performance over the baseline system. In the first pass a set of documents relevant to the query are retrieved and used to dynamically adapt the

* This work has been partially supported by *Consejería de Educación de la Junta de Castilla y León* under project number VA053A05.

language model (LM). In the second pass the adapted language model is used and the list of documents is presented to the user. The system is designed for Spanish language. The performance of the system was evaluated using a test suite from CLEF, which is an evaluation forum similar to TREC. We recorded 10 speakers reading the queries. Results of different experiments showed that the proposed approach outperforms the baseline system. We report a relative reduction in OOV word rate of 15.21%, a relative reduction in WER of 6.52% and a relative gain in retrieval precision of 5.74%, with a LM of 60,000 words.

The structure of the paper is as follows: Section 2 presents some related work; Section 3 explains the system in detail; Section 4 describes the experiments and the analysis of results; in Section 5 we discuss about factors that affect system performance; conclusions and future work are presented in Section 6.

2 Related Work

This is the first work, to our knowledge, on speech driven information retrieval for Spanish language. Experiments for other languages have been reported in the bibliography. All the experiments employed a similar methodology: a standard IR test suite (designed to evaluate IR systems using text queries) was used; some speakers reading the queries were recorded; finally, system performance was evaluated and compared with the results obtained using text queries.

First experiments in speech driven information retrieval were described in [1], for English language. Results showed that increasing WER reduces precision and that long spoken queries are more robust to speech recognition errors than short ones. A system designed for mobile devices was presented in [2], and several experiments in Chinese were reported. Retrieval performance on mobile devices with high-quality microphones (for example PDA) was satisfactory, although the performance over cellular phone was significantly worse. Some experiments in Japanese were presented in [3]. The performance of the system was improved using the target document collection for language modeling and a bigger vocabulary size. More experiments with the same test collection were described in [4]. Techniques for combining outputs of multiple LVCSRs were evaluated and improvement in speech recognition and retrieval accuracies was achieved.

3 System Description

The objective of the system is to retrieve all the documents relevant to a given spoken query. The system is based on a two-pass strategy, as shown in Fig. 1. In the first pass, the spoken query made by the user is transcribed into text by the speech recognizer, using a general LM. Next, a list of documents relevant to that query are retrieved by the information retrieval engine. Then, using those documents, a dynamic adaptation of the LM is carried out. In the second pass, the speech recognizer uses the adapted LM instead of the general LM. Finally, the list of documents relevant to the query is obtained and presented to the user. In the following sections we describe in detail each component of the system.

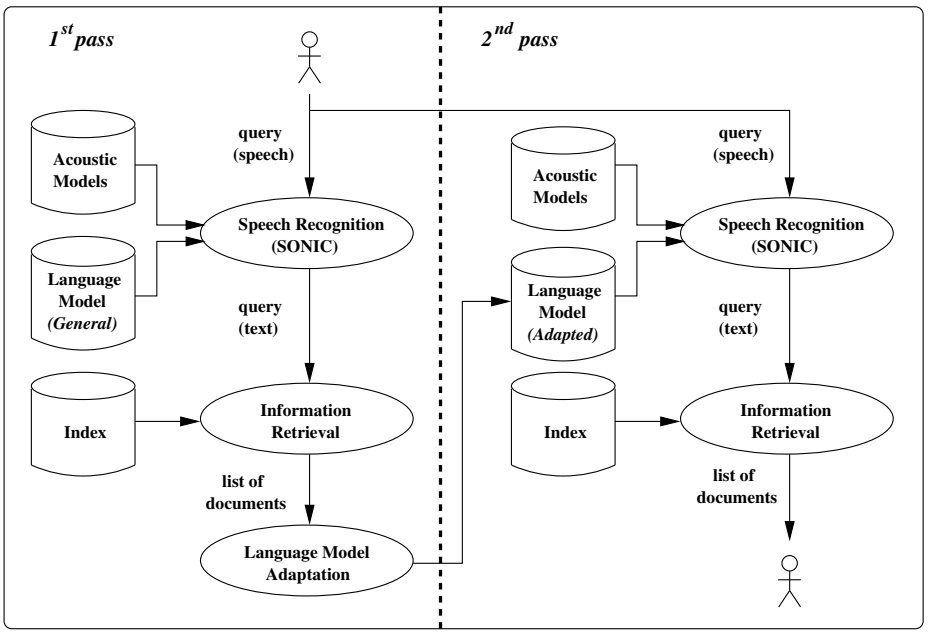

Fig. 1. Architecture of the system, based on a two-pass strategy.

3.1 Speech Recognition

We used SONIC, a large vocabulary continuous speech recognizer from the University of Colorado [5]. It is based on continuous density hidden Markov model (HMM) technology and implements a two-pass search strategy using token-passing based recognition. We trained acoustic and language models for Spanish language.

Acoustic models were triphone HMMs with associated gamma probability density functions to model state durations. Standard feature extraction was used: 12 Mel Frequency Cepstral Coefficients (MFCC) and normalized log energy, along with the first and second order derivatives. We used Albayzin corpus to train gender independent acoustic models [6] (13,600 sentences read by 304 speakers).

Word based trigram language models were created, with three different vocabulary sizes: 20,000, 40,000 and 60,000 words. To train the **general LM** (used in the first pass), the target document collection was used because this can result in an adaptation of the LVCSR to the given task and provides better system performance [3]. EFE94 document collection is composed of one year of newswire news (511 Mb) and has 406,762 different words. The vocabulary was created selecting the most frequent words found in the documents. We used SRILM statistical language modeling toolkit [7], with Witten Bell discounting.

3.2 Information Retrieval

We used a modified version of an information retrieval engine developed for Spanish language [8]. It is based on the vector space model and on term frequency-inverse document frequency (TF-IDF) weighting scheme. We also used a stop word list to remove function words and a stemming algorithm[1] to reduce the dimensionality of the space.

The similarity of a query q with each document d_i in the document collection is calculated as follows:

$$sim(d_i, q) = \sum_{t_r \epsilon q} w_{r,i} \times w_{r,q} \tag{1}$$

$$w_{r,i} = (1 + log(tf_{r,i})) \times log\left(\frac{N}{df_r}\right) \tag{2}$$

$$w_{r,q} = log\left(\frac{N}{df_r}\right) \tag{3}$$

where $w_{r,i}$ is the weight of the term t_r in the document d_i; $w_{r,q}$ denotes the weight of the term t_r in the query q; $tf_{r,i}$ represents the frequency of the term t_r in the document d_i; df_r denotes the number of documents in the collection that contain the term t_r; N is the total number of documents in the collection.

3.3 Language Model Adaptation

The system dynamically adapts the LM to the query made by the user. A two-pass strategy is applied: in the first pass, a general LM is used for speech recognition and the documents retrieved are used to train an adapted LM; in the second pass, the adapted LM is used to obtain the final list of documents. We compared two different approaches to create the adapted language model:

- **Topic LM**: a LM trained using the 1000 documents obtained in the first pass. The vocabulary was built selecting the most frequent words found in that documents.
- **Interpolated LM**: a combination of the *general LM* and the *topic LM*. Because of the limited amount of data available to train the topic LM, we decided to merge it with the general LM. Linear interpolation was employed to combine both models. The interpolation coefficient was computed using the EM algorithm (training data was divided into two portions: one was used to train the topic LM and the other was used to estimate the interpolation coefficient). The vocabulary was built as follows: first, words from the adaptation data were selected based on frequency; second, if the desired vocabulary size was not reached, words from the general corpus were added, based also on frequency.

[1] Snowball stemmer: http://snowball.tartarus.org

4 Experiments

We measured the performance of the system using CLEF 2001, a standard IR test suite. CLEF organizes evaluation campaigns in a similar way to TREC. Its aim is to develop an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, under standard and comparable conditions [9].

In the following sections we describe the experimental set-up and present the results of the different experiments.

4.1 Experimental Set-Up

We used CLEF 2001 Spanish monolingual IR test suite, which includes a document collection, a set of topics and relevance judgments. The document collection has 215,738 documents of the year 1994 from EFE newswire agency (511 Mb). Topics simulate user information needs and are used to build the queries. There are 49 topics and each of them has three parts: a brief title statement, a one-sentence description and a more complex narrative. Relevance judgments determine the set of relevant documents for each topic, and were created using pooling techniques.

We expanded CLEF 2001 test suite to include spoken queries. We used the description field of each topic as query (mean length of 16 words, ranging from 5 to 33). We recorded 10 different speakers (5 male and 5 female) reading the queries. Headset microphone was used under office conditions, at 16 bit resolution and 16 kHz sampling frequency.

4.2 Results

Spoken queries were processed by the system and the 1000 most relevant documents (sorted by relevance) were retrieved for each query. Mean average precision (MAP) was calculated using relevance judgments. The same methodology of CLEF was used to evaluate the results [9]. Results for different configurations of the system are shown in Table 1:

- **Text**: results using text queries (for comparison purposes).
- **General LM**: results obtained in the first pass, using only the *general LM* (baseline system).
- **Topic LM**: results obtained in the second pass, using *topic LM* as the adapted LM (see Sect. 3.3).
- **Interpolated LM**: results obtained in the second pass, using *interpolated LM* as the adapted LM (see Sect. 3.3).

Three different vocabulary sizes were used: 20k, 40k and 60k words. For each different configuration we report the out of vocabulary word rate (OOV), the word error rate (WER) and the mean average precision (MAP).

Table 1. System performance for different system configurations, using a vocabulary of 20k, 40k and 60k words (OOV: out of vocabulary word rate; WER: word error rate; MAP: mean average precision).

	OOV	**WER**	**MAP**
Text	–	–	0.4475
20k general LM	6.77%	24.2%	0.3013
20k topic LM	3.27%	20.3%	0.3412
20k interpolated LM	3.27%	19.6%	0.3393
40k general LM	2.81%	19.0%	0.3267
40k topic LM	2.38%	18.9%	0.3557
40k interpolated LM	2.08%	17.8%	0.3534
60k general LM	2.17%	18.4%	0.3412
60k topic LM	2.36%	18.5%	0.3608
60k interpolated LM	1.84%	17.2%	0.3591

4.3 Analysis of Results

The results of the baseline system were improved by the use of dynamically adapted LM. There was a reduction in OOV and WER, and the reduction was larger using interpolated LM: a relative reduction in OOV word rate of 15.21% and a relative reduction in WER of 6.52%, for a vocabulary of 60k words. There was an increase in MAP, but in contrast with OOV and WER, the use of topic LM yielded to slightly better results: a relative gain in MAP of 5.74%, for a vocabulary of 60k words.

As a possible explanation, we argue that both adapted LMs provided better estimates than the general LM, because they were trained using documents with a semantic relatedness with the current query. Both adapted LMs obtained better estimates for content words, which affected retrieval performance. However, interpolated LM also provided better estimates for function words (the reason for the reduction in WER), which had no effect in retrieval performance.

Experiments with different vocabulary sizes showed that adapted LM improved performance in all cases. Better absolute results were obtained with a vocabulary of 60k words and higher relative increase with a vocabulary of 20k words. As an interesting result, the performance using 20k topic LM (two-pass strategy) was equivalent to the performance of 60k general LM (one-pass).

We also analyzed the results of each individual query. Most of the queries had small loss of precision while some queries had high loss of precision. It means that in general queries did well, but there were some that did badly.

In Table 2 we compare our results with other systems. For each system, the loss in MAP of spoken queries compared with text queries is calculated. We claim that our system has a performance comparable with the systems in the state of the art, although the comparison is not conclusive, since there are many factors that affect system performance: language, IR test suite, length of the queries, vocabulary size of the recognizer and retrieval model of the IR engine. Moreover, there are some important differences between our experiments and the

experiments reported by other researchers. Barnett et al. [1] used long queries (50-60 words) for the experiment, which are more robust to speech recognition errors than shorter ones. Chang et al. [2] used a gender dependent LVCSR, with speaker and channel adaptation. They also reported a performance for TREC-6 queries significantly better compared to TREC-5 queries for similar settings of the system, but no explanation for this was reported. Fujii et al. [3] and Matsushita et al. [4] used a larger document collection (100 Gb, about 10 million documents) which makes the task more difficult. Overall, the loss in MAP of our system is well inside the margins of those contributions.

Table 2. Comparison between systems in the state of the art and our system (MAP-T: mean average precision using text queries; MAP-S: mean average precision using spoken queries; Ploss: loss in MAP of spoken queries compared with text queries).

	Test Suite	**MAP-T**	**MAP-S**	**Ploss**
Barnett	TIPSTER	0.3465	0.3020	12.84%
Chang	TREC-5	0.3580	0.2570	28.21%
	TREC-6	0.4890	0.4630	5.32%
Fujii	NTCIR-3	0.1257	0.0766	39.06%
Matsushita	NTCIR-3	0.1181	0.0820	30.57%
Our system	CLEF01	0.4475	0.3608	19.37%

5 Discussion

Speech driven information retrieval is a task with an open vocabulary, and better results are obtained with larger vocabulary language models, because of better vocabulary coverage. In our experiments, when 20k LM was used, the majority of the errors were due to OOV words. As we increased the size of the vocabulary, OOV word errors decreased and most of the errors were regular speech recognition errors. However, increasing vocabulary size indefinitely is impractical. We have presented an approach based on dynamic adaptation of language models that obtains a reduction of OOV word rate, and allows the system to improve performance without increasing the vocabulary size.

Each speech recognition error has big impact on retrieval accuracy: keywords with important semantic content may be missing, and some relevant documents are not retrieved. Moreover, words not related with the query may be introduced, making the system retrieve documents not related with the query. Surprisingly, there are some queries that improve when speech recognition errors occur.

There is a mismatch between speech recognition and information retrieval: speech recognition favors frequent words, because they are more likely to be said (higher probability in the LM); whereas information retrieval favors infrequent words, because they usually carry more semantic content (using TF-IDF weighting scheme). The worst case happens with proper nouns: they are effective query

terms, but because of their low frequency, they have low probability in the LM or even are not included in the vocabulary of the speech recognizer.

6 Conclusions

In this paper we describe a system which allows the use of spoken queries to retrieve information from a textual document collection. We used a standard IR test suite to evaluate the performance of the system. The results showed that dynamic adaptation of language models provided better results than the baseline system. We reported a relative reduction in OOV word rate of 15.21%, a relative reduction in WER of 6.52% and a relative gain in retrieval precision of 5.74% for a LM of 60,000 words.

Overall, these results are encouraging and show the feasibility of building speech driven information retrieval systems. Although the performance was not as good as using text input, the system can help in overcome the limitations of mobile devices, and can also be useful in situations where speech is the only possible modality (for example, while driving a car).

As future work, we plan to extend the system with spoken dialog capabilities, because user interaction can provide valuable feedback to improve the retrieval process.

References

1. Barnett, J., Anderson, S., Broglio, J., Singh, M., Hudson, R., Kuo, S.W.: Experiments in Spoken Queries for Document Retrieval. In: Eurospeech. (1997)
2. Chang, E., Seide, F., Meng, H.M., Chen, Z., Shi, Y., Li, Y.: A System for Spoken Query Information Retrieval on Mobile Devices. IEEE Transactions on Speech and Audio Processing **10**(8) (November 2002) 531–541
3. Fujii, A., Itou, K.: Building a Test Collection for Speech-Driven Web Retrieval. In: Eurospeech. (2003)
4. Matsushita, M., Nishizaki, H., Nakagawa, S., Utsuro, T.: Keyword Recognition and Extraction by Multiple-LVCSRs with 60,000 Words in Speech-driven WEB Retrieval Task. In: ICSLP. (2004)
5. Pellom, B., Hacioglu, K.: Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task. In: ICASSP. (2003)
6. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., Nadeu, C.: ALBAYZIN Speech Database: Design of the Phonetic Corpus. In: Eurospeech. (1993)
7. Stolcke, A.: SRILM – an Extensible Language Modeling Toolkit. In: ICSLP. (2002)
8. Adiego, J., Fuente, P., Vegas, J., Villarroel, M.A.: System for Compressing and Retrieving Structured Documents. UPGRADE **3**(3) (June 2002) 62–69
9. Braschler, M., Peters, C.: CLEF Methodology and Metrics. In: Cross-Language Evaluation Forum (CLEF). (2001)