Query-by-Example Spoken Document Retrieval – The Star Challenge 2008

Haizhou Li, Khe Chai Sim, Vivek Singh, Kin Mun Lye
Institute for Infocomm Research (I^{2}R)
Agency for Science, Technology and Research (A*STAR), Singapore.
{hli,kcsim,svivek,lyekm}@i2r.a-star.edu.sg

Abstract - **In this paper, we give an update of recent research activities in HLT department of I^{2}R in query-by-example spoken document retrieval (SDR) and report an evaluation campaign, the Star Challenge 2008, which was organized by A*STAR, Singapore. It is suggested that low-level feature-based approach, which does not rely on error-prone speech transcripts, is a promising solution to query-by-example multilingual spoken document retrieval.**

I. INTRODUCTION

With the rapid expansion of audio media sources such as radio, TV and telephony recordings, there is an increasing demand for automatic indexing and retrieval of spoken documents. SDR is essentially the task of retrieving excerpts from a large collection of spoken documents based on a user's request. Real world search need over spoken documents has prompted us, the Human Language Technology (HLT) department of I^{2}R, to look into several research problems of spoken document retrieval in greater detail, in particular, 1) rich transcription of spoken document – the technique to detect speech events, to convert speech signals from digital sound waves to computer readable content; 2) document characterization – the technique to characterize spoken documents for search purposes.

Abacus is a multilingual speech recognition platform which has been developed in HLT department since 1999. It is a phone-based speech recognizer that supports both grammar-based spoken dialogue application and very large vocabulary continuous speech recognition with n-gram language model. Abacus is known for its unique features in handling multilingual/mixed-lingual/code-switch speech, and in supporting Asian languages such as Chinese, Japanese, Korean and English. It has served as the backbone that supports HLT's participation in National Institute of Standards and Technology (NIST) evaluations such as Rich Transcription, Speaker Recognition, and Language Recognition since 2005, and a multi-year industrial project – Audio Keyword Mining System (AKMS) since 2003. AKMS advocates a phone, syllable, and word multi-level indexing scheme for effective spoken term detection.

Moving from spoken term detection towards spoken document retrieval, Abacus is now bracing itself for a greater challenge. In recent years, we are particularly interested in the research problems for query-by-example SDR. 1) Do we need text as the pivot? 2) Do we need lexical words as the indexing items? In this paper, we will first discuss the research problems from I^{2}R's perspective and finally briefly introduce The Star Challenge 2008 – a query-by-example search challenge.

II. DO WE NEED TEXT?

In a query-by-example information retrieval, given a collection of spoken documents, the task is to find documents in the collection which are similar in subject matter to an exemplar, which is a spoken document itself. One obvious way to perform query-by-example retrieval is to run automatic speech recognition (ASR) on the recordings to obtain 1-best transcripts for both queries and documents, and use these transcripts as the pivot for text retrieval in the way that TREC-9 spoken document retrieval (SDR) was carried out. This approach suffers from fact that the 1-best transcripts are far from perfect, especially as far as conversational speech is concerned. The decoding errors in both query exemplar and documents reduce the chance of correct retrieval.

To overcome the decoding errors, one way is to work with not only one transcription hypothesis for each utterance, but also several hypotheses presented in a lattice data structure. A lattice is a connected directed acyclic graph in which each edge is labeled with a term hypothesis and a likelihood value [1]; each path through a lattice gives a hypothesis of the sequence of terms spoken in the utterance. Since the information in a lattice has a statistical interpretation, a retrieval model based on statistical inference, such as the statistical modeling retrieval approach [2], will be a more natural and more principled approach to lattice-based retrieval. We advocate the statistical lattice-based retrieval method for the query-by-example task [3,4]. In this method, we generate a lattice for each speech segment in the collection corpus and the query exemplars, and compute the expected word count – the mean number of occurrences of a word given a lattice – for each word in each lattice. Using these expected counts, a statistical language model is estimated for each spoken document and each query, and a document's relevance to a query can then be computed as a Kullback-Leibler divergence between the document model and query model. To mitigate the problem of noise in the retrieval process caused by non-content words in queries, we can perform stop word removal, or stopping, in the same way as in text-based information retrieval (IR).

Intuitively, text is for human reading. In the query-by-example SDR task, the objective is not to decode the text, but rather to retrieve relevant spoken documents. This allows us to explore retrieval techniques without explicit use of text. Recent efforts have pursued different solutions along this direction [5,6]. We propose the lattice-based retrieval method that extracts the word statistics directly from the lattice. Given two spoken documents that are similar in content, we have good reason to expect that a decoder derives similar statistics from them. We conduct a series of experiments to show that we do not need text as the pivot for query-by-example SDR.

III. DO WE NEED LEXICAL WORDS?

Automatic Spoken Document Classification (SDC) is a query-by-example SDR task if we see the test document as the query exemplar and the class of documents as the content for retrieval. Most SDC efforts so far have been devoted to the lexical approach, where text categorization (TC) techniques are applied to the automatic transcripts of spoken documents to derive semantic classes. The transcripts are typically generated from a large vocabulary continuous speech recognizer (LVCSR). In a nutshell, this method simply cascades a LVCSR and a TC module.

However, the task of SDC is more complex than the TC task. By comparing them, we can gain some insights into the SDC task and the inadequacy of the TC methods that have been applied to SDC. In TC, we usually derive the lexical vocabulary from the running text. However, for spoken documents, an additional tokenization step is needed to convert sound wave into a sequence of phonetic units, such as phonemes. This gives rise to two issues: the definition of tokenization unit, and the choice of vocabulary. These two issues have direct impacts on the resulting tokenization and the subsequent SDC performance. To address them, let us study two intrinsic properties of spoken language.

First, to properly select a vocabulary, we need to take into account Zipf's Law [7]. In human languages, some words invariably occur more frequently than others. One of the most common ways of expressing this idea is known as Zipf's Law. This law states that there is always a set of words which dominates most of the other words of the language in terms of their frequency of use. This is true both of words in the general domain and of words that are specific to a particular subject or semantic domain. This is also true both of written words and spoken words. In SDC, we are particularly interested in extracting a vocabulary that is semantically discriminative.

Second, in SDC, the tokenization unit has traditionally been the lexical word. However, since the lexical word is just the written convention of the language, there is no strong reason why we should choose it as the tokenization unit. Therefore, we advocate the use of language independent *acoustic word* (AW) as an alternative. A lexical word is usually defined by its semantic and/or syntactic function while an AW is associated with a sequence of sounds, for instance, phoneme *n*-gram. With this bold proposal, we look forward to deriving semantic classes of spoken documents based on AW statistics.

Now let us try to answer the question "do we need lexical words?" using spoken language recognition (SLR) as a case study. SLR is the process of determining the identity of the language in a spoken document. If we see the spoken documents in the same language as one class, then SLR is a typical spoken document classification (SDC) problem.

The National Institute of Standards and Technology (NIST) has conducted a series of evaluations of SLR technology since 1996. They are focused on language and dialect detection in the context of conversational telephony speech. The I^{2}R team participated in the 2005 and 2007 NIST language recognition evaluation with its state-of-the-art phonotactic solution [8-11], which does away with the lexical words but uses acoustic words instead.

One of the challenges in SDR is to characterize the spoken documents for ease of search, which is known as document characterization. Lexical words are natural choice of indexing items. However, lexical words are language dependent. In the case where spoken language is unknown or multiple languages are present, the choice of lexical words becomes less obvious. Significant improvements in automatic speech recognition (ASR) have provided the necessary instruments that allow for automatically extraction of acoustic words from raw speech corpus [12,13]. We believe that, although common sounds are shared considerably across vocabularies and languages, the phonotactic statistics of such sounds, manifested by the acoustic words, can differ considerably from one document to another due to different usage of vocabularies and languages. Inspired by the promising results in SLR, we believe that acoustic word approach opens up new opportunities in SDR in general, especially as far as multilingual SDR is concerned.

IV. THE STAR CHALLENGE 2008

TheStarChallenge2008
@ Fusionopolis Singapore
Forums | Blogs | Team Area
Fusionopolis | The Star Challenge | A*STAR | Publicity
Competition Tasks & White Paper v.3 NEW
Competition Rounds
View Flash Presentation
Frequently Asked Questions NEW
Registration Closed
as of March 31st
Resource Partner
mda
Agency for Science, Technology and Research
SINGAPORE
fusionopolis

The Multimedia Search Battle Has Begun!

The Star Challenge, on from January to October 2008, is a race to build the next-generation of multimedia search engines. A US$100,000 cash prize awaits the winners of this world-class competition.

Fig. 1. The Star Challenge 2008 homepage – http://hlt.i2r.a-star.edu.sg/starchallenge

With multimedia search, a person who is interested in a particular topic but does not have the time to watch a lot of video or listen to audio could simply search for the parts that interest him or her. Ideally, we could quickly search video and/or audio by speaking a phrase. But, for now, it is an unsolved problem, especially in a multilingual cyber world. The Star Challenge 2008 identifies several such query-by-example search problems and seeks participation from the community. The evaluation campaign runs two parallel tracks, a voice search track and a video search track. In this paper, we only report the findings in the voice search track.

The Star Challenge 2008 is organized by A*STAR, Singapore as part of a series of events in celebration of the official opening of Fusionopolis, Singapore's science & technology powerhouse to shape the lifestyles and economy of the future. It is organized in two tracks, audio and video tracks. The audio track is focused on two search problems:

1. Search by IPA (Task AT1)

The query is given in International Phonetic Alphabet (IPA), the task is to retrieve from a collection of spoken documents all segments that contain the query IPA sequence regardless of its spoken languages – phonetic search in multilingual speech database.

2. Search by Example (Task AT2)

The query is an utterance spoken by different speakers; the task is to retrieve all segments that contain the query word/phrase/ sentence regardless of its spoken languages – query-by-example in multilingual speech database.

Query-by-example for multilingual spoken document retrieval remains a challenging research problem. The Star Challenge competition serves as a platform for researchers who are interested in query-by-example search techniques to exchange views and to showcase their solutions. It is believed that the competition will bring the state-of-the-art a step forward and spark new ideas.

A. *Evaluation Metric*

Both AT1 and AT2 tasks are evaluated using the Mean Average Precision (MAP) metric as given by:

$$MAP = \frac{1}{L}\sum_{i=1}^{L}\left(\frac{1}{R_i}\sum_{j=1}^{R_i}\text{Precision}(Dij)\right)$$

where L denotes the number of queries and R_i denotes the number of relevant documents corresponding to the ith query. D_{ij} is the set of top-*N* ranked retrieval results containing *j* relevant documents.. Hence,

$$\Pr ecision(D_{ij}) = {}^{j}\!\Big/_{|D_{ij}|}$$

The MAP evaluation metric returns a score between 0.0 and 1.0. A 0.0 score indicates that the system has returned none of the relevant documents while 1.0 score indicates that the system has retrieved all the relevant documents.

B. Competition Rounds

TABLE 1: SUMMARY OF VARIOUS ROUNDS OF THE STAR CHALLENGE COMPETITION.

Rounds	**No of Queries (AT1/AT2)**	**No. of Documents**	**No. of Languages (query/database)**
Round 1	10/10	4300	1/1
Qualifying Round	2/3	2581	2/4
Grand Final	4/4	3234	4/4

The Star Challenge competition consisted of three knock-out rounds. The summary of each round is given in Table 1. The first round of the competition (Round 1) involves only the English speech data. Each task comprises 10 queries and the search database consists of 4,300 documents. The next round of the competition is the Qualifying round which involves 2 languages for the queries (English and Mandarin). The AT1 and AT2 tasks consist of 2 and 3 queries respectively. The search database contains 2,581 documents of 4 different languages: English, Mandarin, Malay and Tamil. Finally, the grand final involves speech data of all four languages for both the queries and the search databases. Both AT1 and AT2 tasks have 4 queries each. The search database contains 3,234 documents.

C. Results

This section presents the analysis of the results of the five teams that made it to the Grand Final. These teams are anonymously identified as Team A to E.

TABLE 2: MAP SCORES OF THE 5 FINALISTS FOR THE AT1 TASKS IN VARIOUS COMPETITION ROUNDS

Team	**MAP Scores**		
	Round 1	**Qualifying**	**Grand Final**
A	0.0417	0.0000	0.0938
B	0.6193	0.2500	0.0833
C	0.5924	0.2500	0.0625
D	0.5803	0.4028	0.0417
E	0.6342	0.0000	0.0000

Table 2 summarizes the MAP scores of the 5 finalists for the AT1 task in various rounds of the competition. Apart from Team A, all team performed reasonably well in Round 1, with MAP scores between 0.5803 – 0.6342. The performance of the qualifying round has a relatively lower MAP score compared to those of Round 1. This suggests that the task is relatively harder due to the multilingual setup of the task. Team A and E did not return any of the relevant documents while the MAP scores for Team B. C and D are 0.2500, 0.2500 and 0.4028 respectively. Finally, the MAP scores of the Grand Final is the lowest of all three rounds. This is expected because the queries now consist of IPA sequences representing speech of four different languages: English, Mandarin, Malay and Tamil. In particular, the last two languages are not commonly studied by the speech community, making the task even more challenging. Team E scored 0.0 for the grand final while the remaining teams scored between 0.0417 – 0.0938.

TABLE 3: MAP SCORES OF THE 5 FINALISTS FOR THE AT2 TASKS IN VARIOUS COMPETITION ROUNDS

Team	**MAP Scores**		
	Round 1	**Qualifying**	**Grand Final**
A	0.2434	0.3600	0.1250
B	0.4063	0.7778	0.3625
C	0.3823	0.5577	0.2083
D	0.2417	0.5096	0.0208
E	0.3241	0.1378	0.0000

Table 3 shows the MAP scores of the 5 finalists for the AT2 task in various rounds of the competition. For Round 1, all the teams showed MAP score performance between 0.2417 – 0.4063. In general, these performances are inferior to those of AT1 task from the same round. This shows that with the knowledge of the query in the form of IPA sequences, the retrieval performance can be improved. Surprisingly, in the qualifying round, all the teams (except Team E) did better compared to Round 1, despite the fact that the queries and databases are in multiple languages. This may be explained by the fact that the two languages found in the queries (English and Mandarin) are commonly studied by the speech community and well-trained phone recognizers in these recognizers are available to perform the retrieval task. Unlike AT1, AT2 does not require explicit knowledge about the IPA representation of the sound in different languages. However, when the task is extended to include languages which are less commonly studied, the retrieval performance deteriorates consistently for all the teams. Apart from Team E, which scored 0.0, the remaining teams achieved MAP scores between 0.0208 – 0.3625. Compared to the results of AT1, the teams generally performed better on the AT2 task under multilingual condition. Most of the teams reportedly used feature-based segmental dynamic time warping techniques [14] to find matching acoustic patterns.

V. Conclusion

The research activities in HLT department of I^2R and the participating teams of The Star Challenge 2008 suggest that low-level feature-based approach, which does not rely on error-prone speech transcripts, is a promising solution to query-by-example multilingual spoken document retrieval.

REFERENCES

[1] D. A. James. The Application of Classical Information Retrieval Techniques to Spoken Documents. PhD thesis, University of Cambridge, 1995.

[2] F. Song and W. B. Croft. A general language model for information retrieval. In Proceedings of CIKM 1999, pages 316– 321, New York, NY, USA, 1999. ACM Press.

[3] T. K. Chia, H. Li, and H. T. Ng. A statistical language modeling approach to lattice-based spoken document retrieval. In Proceedings of EMNLP-CoNLL 2007, pages 810–818, 2007.

[4] Tee Kiah Chia, Khe Chai Sim, Haizhou Li and Hwee Tou Ng, "A Lattice-Based Approach to Query-by-Example Spoken Document Retrieval", *The 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, (SIGIR2008)*, July 20-24, 2008, Singapore

[5] M. A. Siegler. Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance. PhD thesis, Carnegie Mellon University, 1999.

[6] M. Saraclar and R. Sproat. Lattice-based search for spoken utterance retrieval. In Proceedings of HLT-NAACL 2004, pages 129–136, Boston, Massachusetts, USA, May 2004. Association for Computational Linguistics.

[7] Zipf, G.K. Human Behavior and the Principal of Least effort, an introduction to human ecology. Addison-Wesley, Reading, Mass, 1949.

[8] H. Li and B. Ma, “A phonotactic language model for spoken language identification,” in *Proc. ACL*, 2005.

[9] B. Ma, H. Li, and R. Tong, "Spoken Language Recognition Using Ensemble Classifiers", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2053-2062, Sep. 2007.

[10] H. Li, B. Ma and C.-H. Lee, “A vector space modeling approach to spoken language identification,” *IEEE Trans. Audio, Speech and Language Processing*, Vol. 15, No. 1, pp. 271-284, 2007.

[11] Zissman, M.A. Comparison of four approaches to automatic language identification of telephone speech, IEEE Trans. on Speech and Audio Processing, 4, 1 (Jan. 1996), 31-44.

[12] B. Varadarajan, S. Khudanpur, E. Dupoux, Unsupervised Learning of Acoustic Sub-word Units, In Proceedings of ACL 08:HLT, June 2008, pp 165-169

[13] J. G. Wilpon, B. H. Juang, and L. R. Rabiner. 1987. An investigation on the use of acoustic sub-word units for automatic speech recognition. In ICASSP, pages 821–824.

[14] A. Park and J. Glass. 2006. Unsupervised word acquisition from speech using pattern discovery. *Proc.* ICASSP.