On Interfaces for Mobile Information Retrieval

Edward Schofield[1] and Gernot Kubin[2]
[1] Telecommunications Research Center Vienna (ftw)
Vienna, Austria
schofield@ftw.at

[2] Signal Processing and Speech Communication Labs
Graz University of Technology
Graz, Austria
g.kubin@ieee.org

May 2002

Abstract

We consider the task of retrieving online information in mobile environments. We propose question answering as a more appropriate interface than page browsing for small displays. We assess different modalities for communicating using a mobile device with question-answering systems, focusing on speech. We then survey existing research in spoken information retrieval, present some new findings, and assess the feasibility of the endeavor.

1 Introduction

The telecommunications industry worldwide has scrambled to bring what is available to networked computers to mobile devices. The degree of proliferation of Web content has not, so far, been matched by content for mobile devices; nor is this necessarily a bad thing. This short paper argues that there are potentially better interfaces for finding information with small devices than through a page-browsing or card-flipping metaphor. We review, as an alternative, automatic question-answering systems—which respond to questions, rather than keywords, extracting answers, rather than documents—and discuss speech recognition for this task.

2 Information needs and searching behavior

A vaguely defined need for information is often expressed by "Tell me about Topic X." A more precise need for information is more often expressed as a question. But because current search engines return only documents, not answers, information seekers online tend to express all their information needs, vague or precise, in the first form. Some studies on Internet searching behavior [1] have revealed that the vast majority of searches are for one or two keywords, after which users must sift through several documents, perhaps many, to locate the information they desire. The modest goals for current search engines impose some obvious inconveniences: sifting through documents, particularly for specialized information, takes time and distracts attention from the task at hand.

The difficulty of finding online information with a mobile device is much greater. The screens of mobile phones and PDAs range from small to tiny—too small to comfortably read and browse hyperlinked documents. A promising alternative is to specify information needs more precisely in the form of natural-language questions, and have an automated question answerer scour the online data for you. The interest in question-answering (QA) systems, originally designed for easy access to structured databases, has recently grown with the proliferation of Web content and the inclusion of QA competitions in the annual TREC Text Retrieval Conference [2].

QA systems offer considerable opportunity for mobile information retrieval. They are designed to relieve the user of the burden of searching; they therefore circumvent the difficulties of presenting interim search results on small screens. Systems competing in TREC competitions return answers of either 50 or 250 bytes—morsels that can be elegantly shown on a scrollable 5-line mobile display. In contrast, the data sources from which QA systems extract answers need not be formatted for mobile displays; they can arbitrary web pages or databases.

3 The question of input

Users could pose questions to mobile devices in at least three ways: using a keypad, a stylus, or speech. Text-entry rates for the multi-tap method on older mobile phones are commonly 7-15 wpm; with predictive-text facilities this rate roughly doubles [3]. Key tapping would therefore allow the entry of a typical 10-word question in 20–40 seconds, with continuous visual attention. Handwriting with a stylus can be doodled at comparable speeds [4]. This would suffice to satisfy some information needs. For others, a faster and easier interface would be preferable.

Some early research into spoken interfaces for information retrieval was conducted at Xerox labs in 1993. Kupiec and others [5] observed that the true combinations of words in spoken queries co-occurred in documents in closer proximity than misrecognized combinations. They designed a system exploiting this that simultaneously disambiguated word hypotheses and found relevant text for retrieval. Their prototype recognized isolated words from a vocabulary of 100 000 with encouraging results.

More recently Google Labs deployed a prototype of a voice-operated telephone interface to its popular search engine. This recognized isolated keywords and displayed matching documents as a standard web page. For indistinct keywords

Table 1: Cross entropy (bits/word) for bigram models

	TREC 10 Test	Gutenberg Test
Question corpus	6.9	9.7
General text corpus	8.1	8.2

the speech recognizer fed the search engine several phonetically similar variants. Such a system could find a useful application in the disabled market.

We argue, however, that information needs cannot be adequately expressed with keywords, and that responding only to keywords limits a system's potential to satisfy those needs. We also argue that the additional contextual information in questions would be valuable for disambiguating spoken input and could facilitate more accurate recognition.

3.1 Posing questions with speech

We first outline some fundamental limits on the usefulness of speech as a modality. First, speech is public, potentially disruptive to people nearby and potentially compromising of confidentiality. Second, the cognitive load imposed by speaking can interfere with simultaneous problem-solving tasks like preparing documents [6]. Third, speech recognizers make errors—and will do so for the foreseeable future [7]. Successful spoken interfaces must accommodate these limitations. In particular, they must be robust to errors in recognition by requesting clarification from users in a seamless way.

The advantages of speech as a modality are more obvious. It is rapid: commonly 150–250 wpm [8]. It requires no visual attention. It requires no use of hands. All mobile phones, and many PDAs, are equipped with microphones. Standardized protocols for distributed speech recognition could also allow mobile devices, in a few years, to parameterize spoken input and outsource memory-hungry recognition tasks to remote servers [9].

Our research has focused on language modeling for questions, since accurate language models are important prerequisites for accurate speech recognition when vocabularies are large [10]. We have analysed logs of around 500 000 questions compiled from various sources including Usenet Frequently Asked Questions and the Excite and Ask Jeeves search engines.

Our first observation is that the lexical structure of fact-seeking questions tends to be highly constrained. To formalize and test this observation we trained bigram and trigram language models on our question corpus and evaluated the fits of the models on two unseen texts: the test questions from TREC 10 and a sample from the Project Gutenberg archive. For comparison we trained bigram and trigram language models on a larger corpus of sentences drawn from Gutenberg. The results, in Table 1, verify the observation.

A second, partly overlapping, observation is that the syntax of fact-seeking questions tends to be highly constrained. This is described in detail in [11].

Table 2: Inter-word transition frequencies for a small sample of questions

	article	noun	verb	*wh*−	prep.	adj.	other	END
START	0	.03	.23	.70	0	0	.04	0
article	0	.73	0	0	0	.17	.10	0
noun	.01	.25	.25	0	.16	.03	.06	.25
verb	.31	.39	.06	.01	.06	.01	.09	.07
wh−	0	.06	.83	0	.02	.04	.06	0
preposition	.27	.43	.09	.02	0	.05	.14	0
adjective	0	.52	.04	.04	.09	.04	.04	.22
other	.08	.41	.13	.03	.05	.10	.15	.05

Table 2 presents some initial statistics of inter-word transition frequencies, suitable for modeling syntax as a Markov chain between parts of speech.

3.2 The magnitude of the problem

The word-error rate of recognized speech is positively correlated with vocabulary size. Open-domain information retrieval would require the largest vocabulary of proper nouns imaginable—perhaps 200,000. Our initial impression is that, in contrast to the proper nouns, the sets of adjectives and verbs in fact-seeking questions is small, even for a large domain, and that verbs commonly appear in few orthographic forms.

A factor that reduces the difficulty of the speech-recognition task is that the words that are most acoustically ambiguous, and most often misrecognized, are 'function words' like articles and prepositions that contribute relatively little to the semantic content of a sentence. Conversely, longer 'content words' present less difficulty for speech recognizers [12]. It is possible that misrecognitions in function words could be ironed out automatically during the deep semantic and syntactic parsing performed by QA systems. A similar principle was verified in [13] for search queries (in formal language), but for questions (in natural language) this has not yet been investigated.

Spoken question answering on an open-domain is relatively difficult; it can benefit from a restriction on the domain to a specific need, for at least three reasons. First, smaller domains imply smaller vocabularies, with accordingly more accurate spoken interfaces. Second, smaller domains allow tighter linguistic analysis, and accordingly better accuracy in interpreting and finding answers to questions. Third, smaller domains allow more control over the document collections to be searched, permitting semi-automatic indexing and tagging of their content.

4 Conclusions and future directions

This paper has described question answering as an alternative to the metaphor of page browsing for finding information with mobile devices. It outlined the pros and cons of spoken interfaces for retrieving information, surveyed existing research in the field, and presented some initial findings on the structure of fact-seeking questions. In the light of these findings it then assessed the difficulty of automatically recognizing spoken questions.

The largest hurdle for spoken interfaces for information retrieval is the inevitability of misrecognitions, and effective systems require some facility to correct these interactively. A few research fields could possibly rise to this challenge. Multimodal interfaces offer some promise in disambiguating input by combining partial hypotheses from different modalities [14]—for example, spoken input could guide the hypotheses made about word-completion during key-tapping, or vice versa. Another possibility is that paraphrases could clarify ambiguities for speech recognizers, as they can for humans. We will address some of these topics in future work.

References

[1] Bernard J. Jansen and Udo W. Pooch. A review of web searching studies and a framework for future research. *Journal of the American Society of Information Science*, 52(3):235–246, 2001.

[2] Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of the Tenth Text REtrieval Conference (TREC-10)*. Department of Commerce, National Institute of Standards and Technology, 2001.

[3] M. Silfverberg, S. MacKenzie, and P. Korhonen. Predicting text entry speed on mobile phones. In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*, pages 9–16, The Hague, 2000.

[4] W. Soukoreff and I.S. MacKenzie. Theoretical upper and lower bounds on typing speeds using a stylus and keyboard. *Behaviour and Information Technology*, 14:379–379, 1995.

[5] J. Kupiec, D. Kimber, and V. Balasubramanian. Speech-based retrieval using semantic co-occurrence filtering. In *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, NJ, March 1994.

[6] L. Karl, M. Pettey, and B. Shneiderman. Speech-activated versus mouse-activated commands for word processing applications: An empirical evaluation, 1993.

[7] Mark Huckvale. 10 things engineers have discovered about speech recognition. In *NATO ASI Speech Pattern Processing*, 1997.

[8] D. R. Aaronson and E. Colet. Reading paradigms: From lab to cyberspace? *Behavior Research Methods, Instruments and Computers*, 29(2):250–255, 1997.

[9] D. Pearce. An overview of ETSI standards activities for distributed speech recognition front-ends. In *Proceedings of AVIOS 2000: The Speech Applications Conference*, May 2000.

[10] Frederick Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts, 1998.

[11] Edward J. Schofield. Language models for questions. In *Proceedings of the European Association for Computational Linguistics (EACL)*, Budapest, Hungary, April 2003.

[12] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[13] Fabio Crestani. An experimental study of the effects of word recognition errors in spoken queries on the effectiveness of an information retrieval system. Technical Report TR-99-016, International Computer Science Institute, Berkeley, Berkeley, CA, 1999.

[14] Sharon L. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *CHI*, pages 576–583, 1999.