interspeech 2008 incorporating SST 08

Accommodating Explicit User Expressions of Uncertainty in Voice Search or Something Like That

Tim Paek, Yun-Cheng Ju

Microsoft Research, One Microsoft Way, Redmond, WA 98052

{timpaek|yuncj}@microsoft.com

Abstract

Voice search applications encourage users to "just say what you want" in order to obtain useful mobile content such as automated directory assistance (ADA). Unfortunately, when users only remember part of what they are looking for, they are forced to guess, even though what they know may be sufficient to retrieve the desired information. In this paper, we propose expanding the capabilities of voice search to allow users to explicitly express their uncertainties as part of their queries, and as such, to provide partial knowledge. Applied to ADA, we highlight the enhanced user experience uncertain expressions afford and delineate how we performed language modeling and information retrieval. We evaluate our approach by assessing its impact on overall ADA performance and by discussing the results of an experiment in which users generated both uncertain expressions as well as guesses for directory listings. Uncertain expressions reduced relative error rate by 31.8% compared to guessing.

Index Terms: voice search, user uncertainty, something

1. Introduction

Voice search applications (e.g., [7][8][9]) encourage users to "just say what you want" in order to obtain useful mobile content such as business listings, driving directions, movie times, etc. Because certain types of information require recognition of a large database of choices, voice search is often formulated as a both a recognition and information retrieval (IR) task, where a spoken utterance is first converted into text and then used as a search query for IR [14]. Automated directory assistance (ADA) [1][2][14] exemplifies the challenges of voice search. Not only are there millions of possible listings (e.g., 18 million in the US alone), but users also do not frequently know, remember, or say the exact business names as listed in the directory [2][13]. In some cases, users think they know but are mistaken (e.g., "Le Sol Spa" for the listing "Le Soleil Tanning and Spa"). In other cases, they remember only part of the name with certainty (e.g., listing starts with "Le" and contains the word "Spa"). In these cases, what they remember may actually be sufficient to find the listing. Unfortunately, in current voice search applications, users are forced to guess and whatever partial knowledge they could have provided is lost.

In this paper, we propose expanding the capabilities of voice search to enable users to explicitly express their uncertainties as part of their queries, and as such, to allow systems to leverage any partial knowledge contained in those queries. This paper divides into three sections. In Section 2, we present a data analysis of user uncertainty in ADA and explore different types of uncertain expressions found in transcribed call data. In Section 3, we illustrate how uncertain expressions using "something" as wildcards can improve user experience and delineate our approach to handling these expressions. In particular, we describe both language modeling techniques for training n-grams, as well as IR techniques for finding likely exact and approximate matches. In Section 4, we evaluate our approach by assessing its impact on overall ADA performance and by discussing the results of a novel experiment in which users generated both uncertain expressions as well as guesses for directory listings. Finally, we conclude with a discussion of possible extensions and opportunities for future research.

2. User Uncertainty

Before considering how uncertain expressions can improve user experience, it is worth examining how often users are uncertain about their queries. In order to investigate this question with respect to ADA, we performed a data analysis of 1.7 million training sentences used by Tellme, a Microsoft subsidiary, for training various ADA language models (e.g., [5][8]). Unfortunately, it was not possible to separate all human transcribed calls from system generated data, such as recognition results and heuristically generated variations of listings [17]. Given this qualification, we found that only 27.8% of the training sentences matched the exact listings. In other words, in order to obtain the high levels of automation that Tellme achieves for ADA [3], roughly 72% of their training sentences did not match the exact listings. Among the 72%, we searched for sentences containing the word "something", since people tend to express uncertainty with this word. We found 88 utterances, all of which were real transcribed calls. Overall, this represented only 0.5% of the training sentences, given a conservative estimate of the percentage of transcriptions in the data.

What is important about 0.5% is that this number constitutes an anchor from which trade-offs decisions can be made. In particular, if explicit expressions of uncertainty occur at this rate, we should not *lose* much more than 0.5% in overall performance in order to *gain* the handling of these expressions. Of course, 0.5% constitutes just a lower bound, and is likely to increase when users realize that uncertain expressions can be utilized for searching. We return to this trade-off issue again in Section 4.

2.1. Types of uncertain expressions

In examining the 88 utterances containing "something", we observed the following types of uncertain expressions (with examples in italics):

1. Uncertainty about specific words in the name.
   - *... Air Something Boeing*
   - *U. J. C. United Jewish Something ...*
2. Uncertainty about which of several specific words belong in the name.
   - *Adirondack Steak And Grill Or Grill And Steak ...*
   - *The Skating Rink Hot Wheels Skate or Something*

3. Uncertainty about the specific address.
   - *Arch Diocese On Fifty Something Hundred Street Catholic Services*
   - *Paper House On Seventy Something In Amsterdam*
4. General uncertainty about the entire query.
   - *Jillian's Game Palace Or Something*
   - *N. C. I. Concord Cedar Junction Prison Or Something Like That*
5. Uncertainty about the name with supplementary category information.
   - *It's A Hospital I Think It's Moses Something Or Other*
   - *... Travel Trailer It's A Trailer I Don't Know They Sell Trailers ...*

With the first type of uncertain expression, users often replaced a word they did not remember with "something" or "something like that" as placeholders. With the second type, users suggested possible choices for parts of the name using an alternation such as "steak and grill *or* grill and steak". With the third type, because Tellme allows users to provide address information to constrain their search, users frequently expressed uncertainty about the exact address, mostly streets.

The first three types of uncertain expressions, in which users were uncertain about specific parts of their queries, constituted 48.8% of the utterances. The rest were expressions conveying general uncertainty – in some cases, users were uncertain about the entire query itself, using editing expressions [11] such as "something like that" to convey doubt. In other cases, users provided information about the general category of their search, such as "I don't know they sell trailers". Finally, the rest of the general expressions were too ambiguous to classify (e.g., "Something Fun").

3. Handling Uncertain Expressions

From the perspective of user experience, a voice search application capable of accommodating explicit user expressions of uncertainty, such as those described in the previous section, has tremendous appeal. In this section, we first highlight the user experience of finding likely matches to uncertain expressions. We then delineate our approach to implementing this user experience with respect to language modeling and information retrieval.

3.1. User experience

Because of possible trade-offs in overall performance due to handling all types of uncertain expressions, we decided to focus on supporting just the first and most common type; namely, uncertain expressions about specific words in the name. In particular, we added support for both recognition of "something" in a query as well as retrieval of likely matches using "something" as a wildcard (*) in a regular expression. Noticing how users sometimes knew part of the spelling of words, such as in "U.J.C. United Jewish Something", we decided to handle initial characters of words as well so that users could say "United Jewish C Something".

A telephony-based ADA system for handling uncertain expressions would allow users to engage in more natural dialogues such as the following:

System: City and state please.
User: Seattle, WA.
S: Say the name of the business you want …
U: Le Sol Spa.
S: Esperanza, located on …
U: Le Sol Sun Spa.
S: Sorry, I didn't get that. Please try again. You can say "something" for any part of the name you're not sure of, such as "chateau something winery." You can also spell the first letter along with "something", such as "c something winery."
U: Le S Something Spa.
S: Le Soleil Tanning & Spa, located on …

In the example above, the user initially provided incorrect listings. Upon learning how "something" could be used, he then specified whatever partial knowledge he had; namely, that the listing starts with "Le", followed by a word beginning with "S", and containing the word "Spa". This partial knowledge was indeed enough to obtain the desired listing. Although the system informed the user of its capabilities only after it had failed several times, the question of how best to inform users, whether at the beginning, at random, or at possible error points, requires further usability study.

Voice search applications with a user interface can offer even richer user experiences. In the Search Vox multimodal interface [16], we displayed not only the top matches for uncertain expressions, but also the query itself for users to edit, in case they wanted to refine their queries using text. Figure 1 shows a screenshot of the Search Vox results for the spoken utterance "Le S Something Spa", from the previous example, as well the more general expression "Le Something Spa". Note that the system not only retrieved exact matches for the utterances as a regular expression query, but also approximate matches. We discuss in more detail later.

3.2. Language modeling

As discussed earlier, recent approaches to voice search involve recognition plus IR. For ADA recognition, n-gram statistical language models are typically used to compress and generalize across listings as well as their observed user variations [18]. In order to support n-gram recognition of uncertain expressions, we decided to modify our training data. Given that not enough occurrences of the word "something" appeared in the training sentences for it to be accurately recognized (i.e., 88), we boosted that number artificially by creating pseudo-listings from the original data. For every listing which was not a single word (e.g. "Starbucks"), we added new listings with "*" and "*i*-*" replacing individual words, where *i* denotes the initial letter of the word being replaced. For listings with more than two words, because people tend to remember either the first or last word of a listing, we focused on replacing interior words. Furthermore, to preserve counts for priors, we always added 4 new listings (and 4 duplicates for single word listings). For example, for

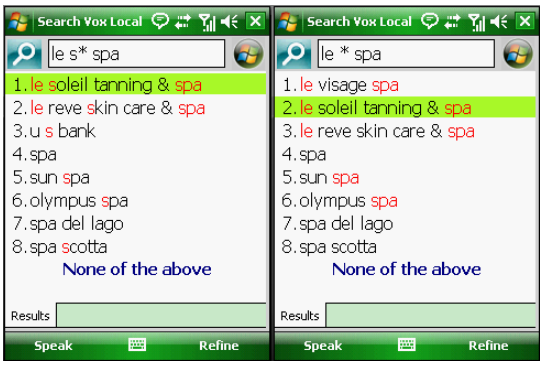

Figure 1. Screenshots of a multimodal voice search application displaying search results for the utterances "Le S Something Spa" and "Le Something Spa".

the listing "Le Soleil Tanning and Spa", we generated "Le *", "Le S*", "* Spa", and "T* Spa". Although this approach of adding new listings with words replaced by "*" and "*i*-*" is certainly a heuristic, we found that it facilitated adequate bigram coverage. Finally, we modified the pronunciation dictionary so that "*" could be recognized as "something".

The advantage of this approach was twofold. First, because we replaced words with "*" and "*i*-*" instead of the word "something", we avoided conflicts with businesses that had "something" as part of their name (only 9 in the Seattle area). Second, by having the recognition produce wildcards (see [10] for more text and inverse text normalization details), we could treat the recognized result in its very condition as a regular expression for search.

3.3. Information retrieval

After obtaining a regular expression from the recognizer (e.g., "Le * Spa"), we needed an index and retrieval algorithm that could quickly find likely matches for the regular expression. We accomplished this by encoding the directory listing as a k-best suffix array [4]. Because a k-best suffix array is sorted by both lexicographic order and any figure of merit, such as the popularity of listings in the call logs, it is a convenient data structure for finding the most *likely*, or in our case, the most *popular* matches for a substring, especially when there could be many matches. For example, for the query "H* D*", the k-best suffix array would quickly bring up "Home Depot" as the top match. Furthermore, because lookup time for finding the *k* most popular matches is close to O(log N) for most practical situations with a worst case guarantee of O(sqrt N), where N is the number of characters in the listings [4], user experience did not suffer from any additional retrieval latencies. Note that before we submitted any regular expression as a search query, we applied a few simple heuristics to clean it up (e.g., consecutive wildcards were collapsed into 1 wildcard).

Besides regular expression queries using a k-best suffix array, which provides popular *exact matches* to the listings, it is also useful to also obtain *approximate matches*. For this purpose, we implemented an improved term frequency – inverse document frequency (TFIDF) algorithm, which is fully described in [18]. Because statistical language models can produce garbled output, voice search typically utilizes approximate search techniques, such as TFIDF, because they treat the output as just a bag of words. This is advantageous when users either incorrectly remember the order of words in a listing, or add spurious words.

In some ways, the two IR methods are flip sides of each other. The strength of finding exact matches is that we can leverage any partial knowledge users may have about their queries (e.g., word order) as well as the popularity of any matches. Its weakness is that it assumes users are correct about their partial knowledge. On the other hand, this is the strength of finding approximate matches; it is indifferent to word order and other mistakes users often make. However, this strength is also its weakness is that it cannot exploit as effectively correct partial knowledge (e.g., initial characters of words), nor popularity for that matter. We return to this issue in the user experiment.

3.4. Related research

Although no other voice search applications that we know of to date accommodate uncertain expressions using "something", some systems do give users the option of saying "I don't know". For example, Tellme's Premium DA [5] engages users in a directed dialogue where users can constrain their business search by providing category or street information, or simply "I don't know", when prompted.

One line of research related to the work presented here focuses on generating alternative user expressions for exact business listings automatically using various methods such as transduction rules [17], statistical translation models [13], and even games [15]. This prior research is not only complementary by benefitting ADA overall, but it may also be possible to leverage similar techniques to generate "*" and "*i*-*" word replacements for training our language model.

4. Evaluation

We now evaluate our approach by first assessing its impact on overall ADA performance and then by discussing the results of a novel user experiment we conducted.

4.1. Impact on ADA performance

No matter how compelling the user experience of using uncertain expressions to search for listings may be, if the trade-off in overall ADA performance is unfavorable it may not be worthwhile. Given that uncertain expressions occurred in at least 0.5% of the transcriptions, assuming that at least twice as many users are likely to start producing these expressions once they are supported, it would not be worthwhile to make changes to ADA *if* the overall performance drops by more than 1%. To assess our impact on performance, we obtained 2317 transcribed ADA utterances from Microsoft Live Search Mobile [7], a multimodal voice search application for Windows Mobile phones. All the utterances were exact business listings so that we could more closely examine any degradation in speech recognition accuracy due to our approach on utterances that should be recognized correctly.

Keeping all other factors besides the language model constant, we evaluated both the Top 1 recognition accuracy and the Top N (where N=10, including the top 1), since some voice search applications such as Live Search Mobile display n-best lists to users for disambiguation. Overall, the Top 1 accuracy dropped from 71.8% to 71.1%, a 1% relative reduction. Furthermore, the Top N accuracy dropped from 80.1% to 79.8%, a 0.4% relative reduction. Because both reductions are almost negligible, we concluded that the heuristic approach we took in Section 3.2 to support recognition of uncertain expressions using "*" and "*i*-*" results in minimal ADA performance loss.

4.2. User experiment

In order to assess whether handling uncertain expressions is better than allowing users to simply guess the listing whenever they are uncertain, we developed a novel experiment similar to the game paradigm described in [15]. In our protocol, subjects were presented with a list of local businesses and asked to cross out any business they were either familiar with or had heard of (e.g., through advertisements). Among the remaining businesses, subjects were asked to circle 10 from which they would like to receive free products or services. They also had to rank these businesses in the order of their preference. When subjects were finished, we took away their list, and asked them to write down all 10 businesses they had just ranked. For recalling the businesses, we instructed subjects to use an underscore wherever they thought they might be missing words, and to also write down their best guess for what those missing words might be. Subjects were given as much time as

	Total (54 Utterances)			Low Frequency Listings (33)			High Frequency Listings (21)		
	Top 1	Top N	Error	Top 1	Top N	Error	Top 1	Top N	Error
Guess	16.7%	18.5%	81.5%	15.2%	18.2%	81.8%	19.0%	19.0%	81.0%
Something	33.3%	44.4%	55.6%	9.1%	24.2%	75.8%	71.4%	76.2%	23.8%

Table 1. Directory Assistance retrieval accuracies for guesses and uncertain expressions using "something" as wildcards, split by whether the target was a high or low frequency listing.

they desired. Afterwards, we recorded subjects reading off their list, asking them to replace the underscores with either "something" or their best guess. Hence, for every "something" expression we had a corresponding best guess. In all, we recruited 15 subjects who recorded 178 utterances, with 54 guesses and 54 "something" expressions. The subjects were all Microsoft employees of various nationalities and accents.

Because the ultimate goal of ADA is to find a target listing, we measured the retrieval accuracy of using either guesses or "something" expressions. As before, we assessed not only the Top 1 accuracy but also the Top N (N=1-10). For guesses, we retrieved approximate matches for the top recognized result, as described in the Section 3.3. For "something" expressions, we first tried to retrieve exact matches for the recognized result and then obtained as many approximate matches as needed to fill N (using the recognized result without wildcards). We found that our language model recognized "something" at 93% accuracy. Table 1 displays the results of the user experiment.

Overall, "something" expressions were roughly 2 times more accurate than guesses for the Top 1 case, and 2.4 times more accurate for the Top N case. Note that whenever the target listing did not occur among the Top N matches, we counted that as an error. With respect to errors, leveraging uncertain expressions dropped the rate from 81.5% to 55.6%, a 31.8% relative reduction. The difference in Top N accuracy between guesses and "something" expressions was statistically significant using McNemar's exact binomial test [6] ($p<.01$). However, the difference in Top 1 accuracy was only significant using a 1-tailed test ($p<.05$).

Because finding the most likely exact matches depends the popularity of the matched listings, we separated the results depending on whether the listing occurred frequently in the call logs (i.e., > 100/month) or not. As expected, the accuracy of "something" expressions was higher for high frequency listings, increasing Top 1 and Top N accuracies by more than a factor of 3. On the other hand, the accuracy of "something" expressions for low frequency listings was only higher in the Top N case. This suggests that uncertain expressions may be best utilized by ADA systems with a GUI for disambiguation.

Finally, in conducting error analyses, we confirmed our suspicion. As discussed previously, exact matches using "something" expressions work best when users provide correct partial knowledge. Unfortunately, this was not always the case. As such, we are continuing to explore how best to combine exact matches with approximate matches.

5. Conclusion & Future Directions

In this paper, we proposed expanding the capabilities of voice search to enable users to explicitly express their uncertainty as part of their queries, and as such, to allow systems to leverage any partial knowledge contained in those queries. In particular, for ADA, we described our approach to language modeling and information retrieval, and highlighted the natural, enhanced user experience that explicit uncertain expressions afford. Furthermore, in evaluating our approach, we demonstrated that it is possible to handle uncertain expressions with minimal loss to overall ADA performance and that using these expressions can significantly reduce error rate compared to guessing, especially for ADA systems with a GUI for disambiguation.

As future research, we plan to explore more data-driven approaches to generating n-gram training sentences with "*" and "*i*-*" word replacements. For example, in conducting our user experiment, we observed that people tended to forget proper names. By collecting more utterances, it may be possible to learn models of what words people are likely to forget. This could be useful for not only generating "something" data for language model training, but also alternative expressions for listings to improve IR in general. Finally, another research direction is to explore enabling other expressions of uncertainty, such as alternations.

6. References

[1] Acero, A. Bernstein, N., Chambers, R., Ju, Y.C., Li, X., Odell, J., Nguyen, P., Scholz, O. & Zweig G. (2008). "Live Search for Mobile: Web Services by Voice on the Cellphone", in *Proc. ICASSP*, 5256-5259.

[2] Boves, L. Jouvert, D., Sienel, J., de Mori, R., Bechet, F., Fissore, L. & Laface, P. (2000). "ASR for Automatic Directory Assistance: The SMADA Project", in *ASR-2000*, 249-254.

[3] Chang, S., Boyce, S., Hayati, K., Alphonso, I., & Buntschuh, B. (2008). "Modalities and Demographics in Voice Search: Learning from Three Case Studies", in *Proc. ICASSP*.

[4] Church, K., Thiesson, B., and Ragno, R. (2007). "K-Best Suffix Arrays". In *Proc. NAACL HLT*, companion volume, 17-20.

[5] Fully Automated Business Search: 1-800-555-TELL

[6] Gillick, L. & Cox, S. (1989). "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", in *Proc. ICASSP*, 532-535.

[7] http://livesearchmobile.com

[8] http://www.tellme.com/products/tellmebymobile

[9] http://www.vlingmobile.com/downloads.html

[10] Ju, Y.C. & Odell, J. (2008). "A Language Model Approach to Inverse Text Normalization and Data Cleanup for Multimodal Voice Search Applications", submitted for publication.

[11] Levelt, W.J.M. (1989). *Speaking: From Intention to Articulation.* Cambridge, MA: MIT Press.

[12] Levin, E. & Mane, A.M. (2005). "Voice User Interface Design for Automated Directory Assistance", in *Proc. Interspeech*, 2509-2512.

[13] Li, X. Ju, Y.C., Zweig, G., & Acero, A. (2008). "Language Modeling for Voice Search: A Machine Translation Approach", in *Proc. ICASSP*.

[14] Natarajan, P., Prasad, R., Schwartz, R., & Makhoul, J., (2002). "A Scalable Architecture for Directory Assistance Automation", in *Proc. ICASSP*, 21-24.

[15] Paek, T., Ju, Y.C., & Meek, C. (2007). "People Watcher: A Game for Eliciting Human-Transcribed Data for Automated Directory Assistance", in *Proc. Interspeech*.

[16] Paek,T., Thiesson, B., Ju, Y.C., & Lee, B. (2008). "Search Vox: Leveraging Multimodal Refinement and Partial Knowledge for Mobile Voice Search", submitted for publication.

[17] Scharenborg, O., Sturm, J., & Boves, L. (2001). "Business Listings in Automatic Directory Assistance", in *Proc. Eurospeech*, 2381-2384.

[18] Yu, D., Ju, Y., Wang, Y., Zweig G. & Acero, A. (2007). "Automated Directory Assistance System – from Theory to Practice", in *Proc. Interspeech*.