Query Parsing in Mobile Voice Search

Junlan Feng
AT&T Labs Research
180 Park Avenue
Florham Park, NJ 07920
junlan@research.att.com

ABSTRACT

Mobile voice search is a fast-growing business. It provides users an easier way to search for information using voice from mobile devices. In this paper, we describe a statistical approach to query parsing to assure search effectiveness. The task is to segment speech recognition (ASR) output, including ASR 1-Best and ASR word lattices, into segments and associate each segment with needed concepts in the application. We train the models including concept prior probability, query segment generation probability, and query subject probability from application data such as query log and source database. We apply the learned models on a mobile business search application and demonstrate the robustness of query parsing to ASR errors.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Applications

Keywords

Query parsing, Mobile voice search

1. INTRODUCTION

Voice search is essentially an integration of automatic speech recognition (ASR) and text or database search. In this paper, we describe a query parser between ASR and Search. As expected, the ASR and Search components perform speech recognition and search tasks. The role of query parsing is three fold: (a) segmenting the automatic speech recognition (ASR) output (1-best and word lattices) into meaningful segments, (b)associating each segment with needed concepts in the application, and (c)identifying the main subject of the query such as *night clubs* in *night clubs open on christmas day*.

There are two research literatures closely relevant to our work , namely query segmentation and named entity extraction(NEE). [3] proposed an unsupervised approach to query segmentation based on a generative language model, where the task was to segment the query into segments of text. The argument is that segments carry implicit word proximity and ordering constraints, and hence can help improve retrieval accuracy. It used an expectation-maximization (EM) algorithm to estimate the model's parameters. NEE attempts to identify entities of interest in text. Typical entities include *locations, persons, organizations, dates, times monetary amounts and percentages* (Kubala et al., 1998). Most approaches for NEE tasks rely on machine learning approaches using annotated data.

Our task of voice query parsing confronts a combination of challenges in both NEE and query segmentation. Its complexity is beyond query segmentation. In addition, application concepts in the parser are broader than named entities. Furthermore, we face challenges posed by error-prone ASR and mobile context, in which users expect that search performs highly effective with location and time awareness.

In the rest of the paper, we will describe a scalable statistical approach to query parsing in Section 2. We then present experimental results in Section 3. Finally, we conclude the paper in Section 4.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

2. A STATISTICAL APPROACH

We formulate the query parsing task as follows. The query parser takes ASR 1-best and ASR lattices as input. For ASR lattices, we use the form of Word Confusion Networks (WCNs), represented as Q_{wcn} [2]. Figure 1 shows an example of WCN. There are one or multiple arcs between a pair of consecutive nodes. Symbols on these arcs are alternative words for the given word position. Numbers on the arcs are negative log posterior probabilities of the associated word. ASR 1-best is a special case of WCN, where there is only one word for each word position.

The parsing task is to segment $Q_{wcn} = q_1, q_2, \ldots, q_i, \ldots, q_n$ into a sequence of concepts. Each q_i is a set of possible words on the arcs of the ith word position $q_i = \{w_{a(i)} | 1 \leq a(i) \leq na_i\}$, where na_i is the number of available arcs. Each concept can possibly span multiple words. Let $W_i^j = w_{a(i)}, \ldots, w_{a(j)}$ be one possible word sequence from the ith word to the jth word. $a(i)$ and $a(j)$ are indices of the arcs. Let $S = s_1, s_2, \ldots, s_k, \ldots, s_m$ be one of the possible segmentations comprising of m segments, where $s_k = W_i^j$. The corresponding concept sequence is represented as $C = c_1, c_2, \ldots, c_k, \ldots, c_m$.

$$(S^*, C^* | Q_{wcn}, D) = \underset{\{S,C|Q_{wcn}\}}{argmax} \; P(S, C|D) \cdot P_{cf}(S|D)^{\lambda_{cf}} \quad (1)$$

$$= \underset{\{S,C|Q_{wcn}\}}{argmax} \; P(S|C, D) \cdot P(C|D)^{\lambda_c} \cdot P_{cf}(S|D)^{\lambda_{cf}} \quad (2)$$

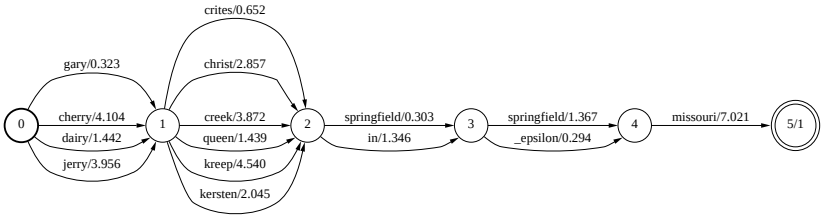

Figure 1: An example confusion network for "Gary crities Sprin gfield Missouri"

For a given Q_{wcn} ,our goal is to search for the best segmentation and concept sequence (S^*, C^*) as defined by Equation 1, which is rewritten using Bayes rule as Equation 2 with extra parameters λ_c and λ_{cf}. D represents mobile context information such as location, speed, history usage , and time of the mobile device. It impacts the meaning of the query. For instance, *Glendale Restaurant* means different in neighborhoods close to the business named as *Glendal Restaurant* than anywhere else. Features of a mobile devices also impact ASR modeling and performance.

There are three components in Equation 2. $P(C|D)$ is the prior probability of the concept sequence. We use λ_c to scale the prior $P(C|D)$. $P(S|C, D)$ is the segment sequence generation probability. $P_{cf}(S|D)$ is the posterior probability of the word sequence of S on Q_{wcn}. λ_{cf} is used to adjust the influence of ASR posterior probabilities. The values of both λ_c and λ_{cf} are determined empirically. In this paper, we focus on a simplified model without including D as a modeling factor. Hence $P(C|D)$, $P(S|C, D)$ and $P_{cf}(S|D)$ become $P(C)$, $P(S|C)$ and $P_{cf}(S)$. We will describe how these probabilities are learned from data later in this section.

In [1], we proposed to re-rank ASR WCNs to prefer paths containing a query subject. We defined a query subject as the core concept of the query, which is the must match part. Each valid query has a query subject. For examples, *night club* is the query subject in *night clubs open christmas day*. Query *open* doesn't have the query subject. We represent the query subject probability as $P_{sb}(S)$ and introduce it as the forth component to the parsing optimization. Equation 2 hence is extended to Equation 3.

$$(S^*, C^*|Q_{wcn}) =$$

$$\underset{\{S,C|Q_{wcn}\}}{argmax} \; P(S|C) * P(C)^{\lambda_c} * P_{cf}(S)^{\lambda_{cf}} * P_{sb}(S)^{\lambda_{sb}} \qquad (3)$$

We approximate the prior probability P(C) using an ngram model on the concept sequence. Training examples of concept sequences can be created from annotated queries.

We model the segment sequence generation probability $P(S|C)$ using independence assumptions, assume each segment in S is generated independently by C. Contextual reliance on concept level is captured in $P(C)$. A corpus of instantiations of the concept c_k are needed to infer conditional probabilities $P(W_i^j|c_k)$. This corpus can be a union of query logs, database field values and human generated examples. There are many ways to model $P(S|C)$. In this paper, we take a simple approach, approximating $P(W_i^j|c_k)$ as relative frequency.

We estimated the query subject probability $P_{sb}(S)$ through mining query logs, which latently encapsulate the most likely subject phrases. Subject phrases are phrases appearing often as a complete query. More details were reported in [1].

3. EXPERIMENTS

We applied the proposed approach on a mobile voice search application, Speak4It. It is a system developed by yellowpages.com and AT&T Labs-Research, which allows users to speak local search queries in a single utterance and returns information of relevant businesses.

Our training data consists of 18 million web queries to http://www.yellowpages.com/, where a query comprises two fields, SearchTerm and LocationTerm, 11 million unique business entries, and 15 thousand annotated voice queries. The parsing task is to parse a voice query into two-layered concepts. The taxnomy includes 2 coarse grained concepts (*SearchTerm and LocationTerm*) and 8 fine grained concepts (*e.g. Landmark*). We tested our approaches on 1000 randomly selected voice queries from a newer time period than the training data. We measure the parsing performance using concept extraction accuracy via exact string match.

We report the first level parsing performance in Table 1. The *Transcription* column presents the parser's performances on human transcriptions (i.e. word accuracy=100%) of the speech. The *1-best* and *WCN* respectively corresponds to ASR 1-best and WCN output. ASR word accuracy is 67.2%. The promising aspect is that we improved *SearchTerm* extraction accuracy by 2.0% when using WCN as input. Performance on the second level concepts will be published in near future. Though 63.0% SearchTerm extraction accuracy on ASR output is low, search performance is much higher for its robustness to certain ASR errors such as *restaurant* being misrecognized as *restaurants*.

Slots	1-best	WCN	Transcription
SearchTerm	61.0%	**63.0%**	95.1%
LocationTerm	88.5%	88.4%	97.4%

Table 1: Concept Extraction Accuracy

4. CONCLUSIONS

This paper described a statistical approach to voice query parsing. We demonstrated the effectiveness of this approach on a mobile search application.

5. REFERENCES

[1] J.Feng, S. Bangalore, and M.Gilbert. Role of natural language understanding in voice local search. In *INTERSPEECH*, 2009.

[2] A. S. L. Mangu, E. Brill. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computation and Language*, 14(4):273–400, October 2000.

[3] B. T. F. Peng. Unsupervised query segmentation using generative language models and wikipedia. In *Proceedings of WWW-2008*, 2008.