Intent Classification of Voice Queries on Mobile Devices

Subhabrata Mukherjee
IBM India Research Lab
subhabmu@in.ibm.com

Ashish Verma
IBM India Research Lab
vashish@in.ibm.com

Kenneth W. Church
IBM T.J. Watson Research Center
kwchurch@us.ibm.com

ABSTRACT
Mobile query classification faces the usual challenges of encountering short and noisy queries as in web search. However, the task of mobile query classification is made difficult by the presence of more inter-active and personalized queries like *map*, *command and control*, *dialogue*, *joke etc*. Voice queries are made more difficult than typed queries due to the errors introduced by the automatic speech recognizer. This is the first paper, to the best of our knowledge, to bring the complexities of *voice search* and *intent classification* together. In this paper, we propose some novel features for intent classification, like the *url's* of the search engine results for the given query. We also show the effectiveness of other features derived from the *part-of-speech information* of the query and *search engine results*, in proposing a multi-stage classifier for intent classification. We evaluate the classifier using tagged data, collected from a voice search android application, where we achieve an average of 22% f-score improvement per category, over the commonly used bag-of-words baseline.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Algorithms, Experimentation

Keywords
Intent Classification, Mobile Search

1. INTRODUCTION
Mobile queries have been moving away from traditional web search queries, and becoming a lot more personalized and inter-active with the increased usage of smartphones, intelligent agents like *Siri, Google Voice* and navigational assistants like GPS. This has led to new query types requiring local information (*where is the closest restaurant*), command-and-control (*open facebook*), dialogue (*how are you?*), joke (*will you marry me?*) *etc.* The usage of an automatic speech recognizer introduces errors in the text transcript due to incomplete queries and background noise. Coupled with these issues are the standard challenges of short query classification with a lot of ambiguity due to lack of context. A particular challenging aspect of intent classification is to come up with a single appropriate category for a query. The query *find me the nearest restaurant* can be map, command-and-control or a restaurant query with different probabilities. To resolve this confusion we propose a multi-stage classifier where the first stage predicts the top *K* categories for a given query. The second stage combines the first stage prediction with some *additional features* using *regression* to predict the most appropriate query category.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil. ACM 978-1-4503-2038-2/13/05.

Web query classification accuracy is found to be boosted with the usage of a web search engine that helps in query expansion [1]. However, the features commonly used for web query expansion, like the *contents* or *meta-data of a web page*, are less preferable for mobile queries to keep the online data requirement low for a faster response to keep up with the inter-activeness of mobile query applications. We propose a novel method to use *search engine results (only url's*) for intent classification of queries.

In this work, we show that the part-of-speech tag of the query helps in certain categories. For example, command-and-control queries are more likely to start with verbs (*call mom*) than knowledge queries (*what is the capital of India*) which are more likely to start with the POS categories like *WP*, *WRB etc.*

In addition to the before mentioned features, we used other derived features like *url ranking*, *query-url overlap*, *sequence of POS tags etc.* and propose a multi-stage classifier for effective query categorization. Although query length statistics are similar to earlier studies [2][3], the proportion of *music* and *navigational* queries is found to increase a lot in our data (~ 8-10%) with similar decrease in proportion of *news* and *sports* queries. *Joke, dialogue, asr_error, universal_acceptor, endpoint* and *command-and-control* query categories are newly introduced in our work.

2. Query Intent Classification
An automatic speech recognizer is used to obtain the text transcript of voice queries over the mobile. The resultant text is fed into the multi-stage classifier. We have classified the voice queries into *thirty* broad categories. Table 1 shows the query categories along with the impressions per category. An impression of a query denotes the number of times the query was repeated.

offer_suggestions, stock_prices, book, calendar, math, tv, news, request_for_documentation, math, news, joke, eventsearch, non-english, date_time, image_search, food, asr_error, product	**< 1 %**
sports, restaurant, knowledge, endpoint, weather	**2-3%**
dialogue, movies, music	**3-4%**
universal_acceptor	**7%**
map, command-and-control	**11-12%**
navigational, websearch	**22-24%**

Table 1. Query Categories with Percentage of Impressions

In *navigational* queries the user has a *target* website in mind but does not input the full name (example: *starbucks coffee*). Contrast these queries to *websearch* which are informational queries (example: *hidden markov model*). *Endpoints* are queries where the user is cut off midway (example: *go to the closest*) and *universal acceptors* are words picked up by the ASR system accidentally as the user did not intend to talk to the system (example: *or, and*). The first stage of classification involves a *multi-class Naïve Bayes*

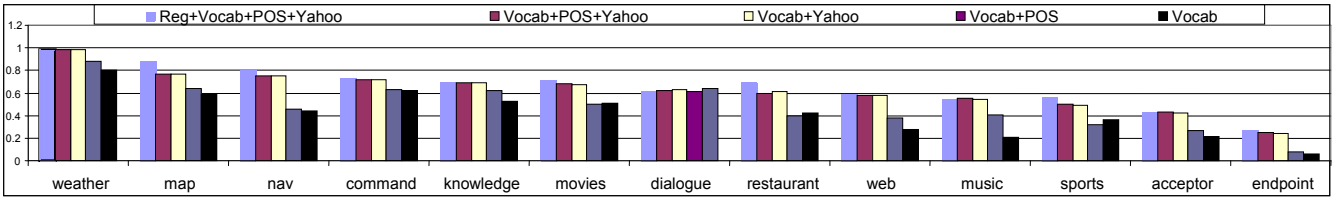

Figure 1. F-Score of Intent Classification Models

classifier using bag-of-words (which forms the vocab), part-of-speech tag information and domain words of the query as features. The domain words of a query are obtained by passing the query through a web search engine and extracting the domains from the url's of top ranked retrieved pages. For example, given the query *the avengers*, the search engine retrieves the domain-words *imdb marvel en.wikipedia youtube imdb tv trailers.apple yahoo marvel.*

Let $\mathbf{Q} = \{q_i\}$ be a query with the part-of-speech tagset $\mathbf{T} = \{t_i\}$, where t_i is the *POS tag* of q_i. Let **D** be the set of search engine url's for **Q**, where $\{d_j\}$ are domain-words. The feature vector of **Q** is formed by $\mathbf{F} = \{\mathbf{Q}, \mathbf{T}, \mathbf{D}\}$. Probability of **F** belonging to class **C** is given by, $P(C|F)=argmax_c\ \Pi_i\ P(q_i|t_i,C)\ P(t_i|t_{i-1},C)\ \Pi_j\ P(d_j|C)$. Here, certain independence assumptions have been made. A word is taken to be dependant on only its POS tag, and the POS tag of a word depends on the POS tag of the previous word. The domain-words are taken to be dependant on only the class, to deal with the sparsity of the feature space. The *top K ranked classes* are taken along with some additional features (Table 2) which are used in a *logistic regression classifier* in *stage two* of classification.

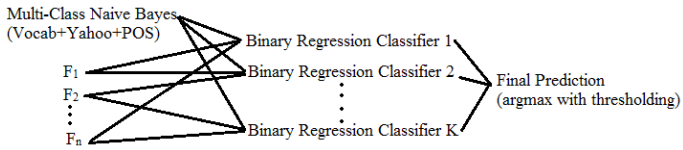

Figure 2. Multi-Stage Classifier for Intent Classification

Navigational Indicators: Query as a prefix of a domain word, query as a complete domain word
Music Indicators: Presence of the substrings *music*, *song*, *lyrics*, *pandora* in the query or domain words
Map Indicators: Presence of the substrings *find*, *close*, *direction*, *near*, *where* in the query
Movie Indicators: Presence of *IMDB* as the topmost domain word
Command-and-Control Indicators: Query starting with a Verb
Websearch Indicators: Presence of *Wikipedia* in the topmost two domain words
Knowledge Indicators: Presence of *ask*, *answer* in the query or domain words; presence of start POS tags as *WRB* and *WP*
Weather Indicators: Presence of the substrings *weather, rain, forecast* in the query or domain words
Sport Indicators: Presence of *sport, goal, nba, espn, play* in query or domain words; overlap with a seed list of Sport terms
Restaurant Indicators: Presence of the substrings *restaurant, food, yelp, eat* in query or domain words
Endpoint or **Universal_Acceptor** Indicators: Query ending with POS tags *VB, JJ, IN, TO, RP, DT;* query Length

Table 2. Snapshot of Features used in Regression Model

The features in Table 2 involve *substring matching* and *domain-word ranking and overlap with query* which is not used in stage one of the classifier. The features for each class are designed to resolve the confusion between it and the classes it is frequently confused with. During training, the regression classifier assigns *positive weights* to the features that support the class and *negative weights* to the features of the conflicting classes. The *substrings* in Table 2 are chosen as the ones having maximum *information gain* from a *heldout* training data. Let $P_c(Q)$ be the probability of query Q belonging to class C, as assigned by the regression classifier. Let Θ_c be the threshold for class C, which is learnt from a split of the training data. The final class of the query is given by,

$$C^* = \arg\max_c \frac{(P_c(Q) - \theta_c)}{\theta_c}, s.t. (P_c(Q) - \theta_c > 0)$$

3. Experimentation

52,282 unique queries, having a total of 1,04,950 impressions, are collected from an android voice search application and manually tagged, of which 11,675 queries are frequent ones (average impression per query > 20). The average number of words per query is 2.3. The average word error rate (WER) of the ASR engine is 20%. *Stanford POS-tagger* [4] and *Yahoo BOSS* [5] *search engine* are used. *Logistic regression* classifier is trained *with impressions,* in the second level of classification. The data is split in the ratio 30-30-20-20 to train the *Naïve Bayes*, *Regression Model*, obtain *threshold* for each class (to maximize *f-score*) and the last split is used for testing the multi-stage classifier. Figure 1 shows the performance of the classifier using features like *vocab*, *part-of-speech information*, *Yahoo BOSS domain words*, and derived features for regression. The categories shown above have per-category impression > 1% of the total impressions. The figure depicts the usefulness of domain words as very powerful features for classification by incorporating external knowledge in the classifier. The POS information proves beneficial in most cases which detects the query pattern (especially *start* and *end* tags). The final regression classifier with all features is the best one.

4. Conclusions

In this work we have shown the usefulness of *domain words*, *part-of-speech-information* and other derived features for intent classification of voice queries over mobile. We have evaluated the proposed approach against manually tagged data and achieved significant improvements over the baseline. External knowledge incorporated by the domain words and the part-of-speech pattern for certain query classes prove effective in distinguishing between close query categories, to predict the most appropriate one.

5. References

[1] Govaerts, S., Corthaut, N., Duval, E. 2009. Using search engine for classification: Does it still work? In: Proc. IEEE Int. Symp. Multimedia.
[2] Yi, J., Maghoul, F., Pendersen, J. 2008. Deciphering Mobile Search Patterns: A Study of Yahoo! Mobile Search Queries. WWW 2008
[3] Kamvar, M., Baluja, S. 2006. A Large Scale Study of Wireless Search Behavior: Google Mobile Search. CHI 2006
[4] http://nlp.stanford.edu/software/tagger.shtml. 2012. Website
[5] http://developer.yahoo.com/boss/search/. 2012. Website