USE OF GEOGRAPHICAL META-DATA IN ASR LANGUAGE AND ACOUSTIC MODELS

Enrico Bocchieri and Diamantino Caseiro

AT&T Research, 180 Park Ave, Florham Park, New Jersey 07932

ABSTRACT
The query distribution, in the speech recognition applications of directory assistance (DA) and voice-search, depends on the customer's location. This motivates the research on query models conditioned on the user location, here denoted as local models. We describe and test our methods for the estimation of local models with various degrees of spatial "granularity", for the recognition of city-state (sub-task of DA) and for the recognition of business listings, spoken over iPhones in a nation-wide business-listing voice-search service. Our local language models improve the accuracy of city-state by 2.4% absolute (32% relative error reduction), and of voice-search by 2.2% (7% relative).

Index Terms— Local, language, acoustic, model, metadata, ASR

1. INTRODUCTION
The query distribution, in the automatic speech recognition (ASR) applications of directory assistance (DA) and voice-search, depends on the location of the user that often looks for nearby addresses and businesses. This motivates the usage of local models of user queries \( x \), conditioned on the user location \( l \), instead of global models not conditioned on \( l \).

Numerous applications of voice-search for business listings and media data [1, 2, 3, 4, 5] have recently been developed for cell-phones and mobile devices. The user interface has gradually evolved from the so-called two-box configuration, with the user saying the desired business name and place in two separate turns, to one-box, with the user freely uttering business name and place in one sentence [6]. In this work, utilizing the capability of multi-modal phones, the voice-search service Speak4It presents to the user the location of the requested business on a map drawn on the iPhone display. User queries are freely spoken, and the application is aware of the user's location, by means of the iPhone location hardware. Thus queries of categories ("Chinese restaurant"), or of nation-wide businesses ("Walmart"), intended to visualize the location of nearby stores, are common (more than two thirds, in our data). Queries of this type are similar across the nation, and therefore the advantage of local rather than global language models (LM) becomes less evident. Instead, in the traditional business DA, the user answers to prompts like "what city and state ?" and "what business ?" mainly to find out the telephone number of a specific business: local LM's play an intuitively stronger role.

This difference motivates this study of local LM's for the ASR of business listings spoken over iPhone hand-sets in a nation-wide voice-search service (Speak4It, with one box interface), versus the recognition of city-state, spoken in response to the prompt "what city ?" for business DA.

A previous paper [7] is concerned with the recognition of business DA (411) queries, reporting results limited to queries for which there is a matching listing in the city/state provided by the caller. We report on voice-search local models at the level of city, area-code and state, showing better ASR accuracies with the coarsest "granularities", and taking advantage of model interpolation/combination. We also present preliminary results on local acoustic models.

2. THE ASR TASKS
2.1. City-state
The data used for city-state ASR are:
- 46,000 US city/locality names, with respective state.
- 36,000 transcribed city-state utterances for LM and hidden Markov model (HMM) estimation, and 36,000 transcribed business-name (DA) utterances, also used for HMM estimation.
- Development and test sets of 610 and 1433 of city-state utterances, respectively.

The available geographical meta-data are:
- Business counts for all cities/localities.
- Area-code of customer's cell phone, for above sentences.
- Telephone area-codes of US cities/localities.
- Adjacency matrix of US telephone area-codes (non-overlapping, in number of 236 plus overlays), and area-code to state mapping.

2.2. Voice-search
The data used for voice-search ASR are:
- 63,000 transcribed spoken business queries, anonymously recorded over iPhones from users all over the country.
These data were used as part of the baseline global LM, and as a development set for local LM’s.

- 2714 spoken utterances for testing, randomly selected from a week of recordings, excluded from the 63,000 utterances.
- 100 million anonymous typed queries of business names, from the logs of the YellowPages.com website, with the two fields of search term (ST), and location (LO). The search term is either a category or a business name, while the location consists of city and state. The two fields were combined in a single sentence using the five most frequent carrier phrases observed in the iPhone queries: “ST”, “ST LO”, “ST in LO”, “ST near LO”, “LO”. The carrier phrases were weighted according to their relative frequency in the iPhone spoken queries. The typed queries contain numerous spelling errors and business name spelling inconsistencies. These errors are problematic because they dilute the probabilities of search terms among alternative tokens. Mis-spelled results also affect negatively the user experience. To clean the corpus, a set of 15,000 correction rules was developed and applied to the queries.

In addition to the geographical meta-data for city-state ASR, we have also at our disposal the city (with respective area-code and state) nearest to the customer, at the time he/she is uttering the business queries.

3. "LOCAL" LANGUAGE MODELS

Data scarcity makes it impossible to directly estimate local models \( p(x|l) \) by partitioning the data according to the user location \( l \). Therefore, we design the local models by mixtures of p.d.f.’s as in (1).

\[
p(x|l) = w_1 p_1(x) + w_{l1} p_{l1}(x) + w_{l2} p_{l2}(x) + w_u p_u(x) \quad (1)
\]

The weight \( w_1, w_{l1}, w_{l2}, w_u \) are to be interpreted as the probabilities that someone located in \( l \) asks for any item (any city or any business) in \( l \) and in topologically concentric locations \( l_1, l_2, u \). The probabilities of items within the concentric locations are denoted by \( p_1(x), p_{l1}(x), p_{l2}(x), p_u(x) \), respectively. This model design allows for nation-wide coverage, and for calibration to the user’s locality by the weight estimation on (relatively few) data.

3.1. City-state

We have implemented the following global LM’s.

- **CS_USA_UNW.** Same probability to all the 46,000 localities.
- **CS_USA_BCW.** City-state query probabilities assumed proportional to the city business count.
- **CS_USA_BCCE.** To estimate from data the relation between business-count and city probability, we cluster the 46,000 cities according to their business counts. Then, the cluster probabilities (and, trivially, the cities’) are estimated from the 36,000 transcribed utterances.

Since CS_USA_BCW and CS_USA_BCCE have similar performance (Section 4.1), the local models below assume that the probabilities of cities in a certain area are proportional to the business counts.

**CS_AREA_STATE** has three mixture components, the first component probability \( p_1(x) \) in Figure above (1)) is of the cities within the area-code of the user’s cell-phone, the second \( p_{l1}(x) \) is of cities in the other area codes of the same state, and then \( p_u(x) \) is of cities in the rest of the US. Mixture component \( p_{l2} \) of (1) is missing.

**CS_AREA_ADJ** uses the US area-code adjacency matrix to define the components of (1): \( p_1(x), p_{l1}(x), p_{l2}(x) \) and \( p_u(x) \) denote the probabilities of the cities in the area-code \( a \) of the user’s cell phone, in the area-codes adjacent to \( l \), in the area-codes adjacent to \( l_1 \) (but not to \( l \)), and in the rest of the US, respectively.

The mixture weights of CS_AREA_ADJ in (1) were estimated from the counts of the 36,000 training city-state queries as

\[
w_1 = 0.632, \ w_{l1} = 0.172, \ w_{l2} = 0.023, \ w_u = 0.135 \quad (2)
\]

This greatly increases the estimated probabilities of the cities in the user’s area code \( l \): consider that the user’s area code weight \( w_l = 0.632 \) is 150 times larger than the equal weighting \( 1/236 \) for the 236 area-codes.

3.2. Voice-search

Our baseline location-independent language model (VS_USA) is a mixture language model. We start by clustering the 100M web queries in 10 clusters using the k-means algorithm and the cosine distance to measure the similarity between two queries. Then, 3-gram models built from these clusters are linearly interpolated with a model built from the complete data set and with models trained with the 63k transcriptions. The interpolation weights are optimized on a held-out subset of the transcriptions.
We do not use the 63,000 tagged transcriptions to build local model due to their small quantity. Instead, we train local language models using the 100M web queries. This constrains their resolution to the city level. In this work, we consider three different granularities:

VS_CITY. For every city, we estimate a 3-gram LM on the web query data that refer to that city. For the recognition of user queries we select the LM of the city nearest to the customer, when he/she is uttering the business query.

VS_STATE. Here we build a model for each US state using the respective subset of web queries.

VS_AREA. To build models of intermediate spacial granularity, we interpolate models trained from the cities contained in the user area code and the cities in the two adjacent rings as shown in equation (1). However we do not include the US global component that, instead, will be combined with the local models as in Section 4.2. The interpolation weights \( w_1, w_2, w_3 \) are estimated by optimizing the perplexity on iPhone queries from the respective area code. Although we do not have access to the phone number code of our iPhone users, the area-code is still a natural choice to base our intermediate granularity models because it reflects the population density throughout the country, to a large extent.

### 4. LANGUAGE MODEL EXPERIMENTS

These experiments were performed with the AT&T Watson recognizer, with a triphonic HMM, discriminatively trained on 2,000 hours of telephone data, and adapted on the transcribed in-domain data of Sections 2.1 and 2.2, respectively for city-state and voice-search ASR.

#### 4.1. City-state

To test the local LM’s, the customer’s location \( l \) is approximated by the cell-phone area-code even if he/she may be calling in a different part of the country. Our results, on real application data, reflect this possibility.

Table 1 shows the ASR results of the LM’s of Section 3.1. Perplexities are measured on the development set with no out-of-vocabulary words. The city-state or string accuracies (misrecognition of either city or state counts as an error) are measured on the test set, with development set accuracy in parentheses. Most of the difference between the test and development accuracies are due to out-of-vocabulary words (Canadian cities).

The global CS_USA_BC and CS_USA_BCC models (city probabilities based on business counts) give much better accuracies (7% improvement) than the "flat" model CS_USA_UNW. Under the assumption that LM CS_USA_BC yield additional accuracy improvements (2% and 2.4%), with better performance from the more detailed model CS_AREA_ADJ. The precise values of the weights (2) are not critical: they can vary by more than a factor of two without significant changes of the city-state error rate.

#### 4.2. Voice-search

The performance of the language models described in Section 3.2 is summarized in Table 2. The first observation is that all local models perform worse on their own than the global model. This is not unexpected as the global model is a strong baseline that is adapted to the iPhone utterances. Other contributing factors are the fact that it was trained on a larger data set and it has wider coverage. We also observe that local models with wider coverage perform better. After these results, we conducted experiments combining the local and global models.

Interpolating the global and local models would result in a large number of multi-gigabyte models, prohibiting any future public deployment. Consequently, we decided not to use interpolation, instead, we use the union of two or more models, as it can be dynamically built by the decoder for each utterance. Equation 3 formalizes the union of language models \( p_l(x) \) and \( p_g(y) \). The denominator is not computed because it does not influence decoding.

Table 3 shows the results of combining the local and global models, here we see that the local models complement the global models, with absolute improvements ranging from 0.5% to 2.2%. The state seems to be the best granularity level for this application, being 0.9% better than the area code. Adding the city or the area code does not bring improvements over using the state.

\[
p(x) = \frac{\max_y [p_l(x), p_g(y)]}{\sum_y \max [p_l(y), p_g(y)]}
\]

### 5. "LOCAL" ACOUSTIC MODELS FOR CITY-STATE

We try to capture dialectal variations of American English by adaptation of a general US HMM to speech from differ-

<table>
<thead>
<tr>
<th>Language model</th>
<th>Perplexity</th>
<th>City-state accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS_USA_UNW</td>
<td>46,000</td>
<td>85.1 (86.2)</td>
</tr>
<tr>
<td>CS_USA_BC</td>
<td>3,130</td>
<td>92.3 (93.3)</td>
</tr>
<tr>
<td>CS_USA_BCC</td>
<td>2,950</td>
<td>92.4 (93.3)</td>
</tr>
<tr>
<td>CS_AREA_STATE</td>
<td>249</td>
<td>94.4 (95.1)</td>
</tr>
<tr>
<td>CS_AREA_ADJ</td>
<td>103</td>
<td>94.8 (95.7)</td>
</tr>
</tbody>
</table>

Table 1. Performance of the city-state LM’s underweights cities with few businesses, we introduced CS_USA_BCCE with, however, no significant improvement.

Both local models CS_AREA_STATE and CS_AREA_ADJ yield additional accuracy improvements (2% and 2.4%), with better performance from the more detailed model CS_AREA_ADJ. The precise values of the weights (2) are not critical: they can vary by more than a factor of two without significant changes of the city-state error rate.
The current trend in voice-search applications is to allow the user to “say anything at any time”, providing convenience and flexibility. However, in the trade-off between convenience and accuracy, there may be space for less general interfaces (e.g. in “vertical” search) designed to take better advantage of application constraints, as those provided by local models based on geographical meta-data.

7. REFERENCES