INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Using Knowledge Graph And Search Query Click Logs in Statistical Language Model For Speech Recognition

Weiwu Zhu

Microsoft, 555 110th Ave NE, Bellevue, WA, United States

weiwuzhu@microsoft.com

Abstract

This paper demonstrates how Knowledge Graph (KG) and Search Query Click Logs (SQCL) can be leveraged in statistical language models to improve named entity recognition for online speech recognition systems. Due to the missing in the training data, some named entities may be recognized as other common words that have the similar pronunciation. KG and SQCL cover comprehensive and fresh named entities and queries that can be used to mitigate the wrong recognition. First, all the entities located in the same area in KG are clustered together, and the queries that contain the entity names are selected from SQCL as the training data of a geographical statistical language model for each entity cluster. These geographical language models make the unseen named entities less likely to occur during the model training, and can be dynamically switched according to the user location in the recognition phase. Second, if any named entities are identified in the previous utterances within a conversational dialog, the probability of the n-best word sequence paths that contain their related entities will be increased for the current utterance by utilizing the entity relationships from KG and SQCL. This way can leverage the long-term contexts within the dialog. Experiments for the proposed approach on voice queries from a spoken dialog system yielded a 12.5% relative perplexity reduction in the language model measurement, and a 1.1% absolute word error rate reduction in the speech recognition measurement.

Index Terms: knowledge graph, search query click log, statistical language model, speech recognition, named entity recognition

1. Introduction

Statistical language model is a classical model designed for speech recognition to estimate the prior probabilities of word strings [1]. This paper focuses on statistical language models for online large vocabulary continuous speech recognition (LVCSR), especially spoken dialog systems on mobile devices. The language model applied in these speech recognition systems usually operate with a very large but finite training data. Lots of named entities, such as restaurants, persons, organizations, and events, especially the noteless ones, are usually not covered in the finite training data, and this kind of data missing often results in recognition performance loss. Although modern commercial speech recognition systems try their best to build a sufficient and large training data that includes lots of named entities, it is not possible to collect all the necessary ones due to the training performance and the model size, as well as the data collection cost. However, named entities are often present in the speech utterances in spoken dialog systems, and these entities are very likely to be recognized as other common words with the similar pronunciation. Once this recognition issue can be well addressed, the performance of spoken dialog systems will be improved significantly.

There are already a few approaches proposed to identify the unknown named entities for speech recognition. Class-based n-gram language model is a common solution that handles unseen words by assigning similar words to classes [2], and a hierarchical language model with conventional word and word-class n-grams is proposed in [3]. Some researches for the specific scenarios like directory assistance use geographical data to improve the recognition of business listing names [4, 5]. The context among the adjacent utterances within a dialog is another important factor frequently used in spoken dialog systems [6, 7, 8, 9, 10, 11, 12]. Information retrieval (IR) technologies are also very hot in language models, such as relevance based long-term context information [13], search engine query logs [14], and IR based model adaptation [15]. Other attempts, including transcript normalization [16], incorporating speech recognition confidence features [17], and applying LVCSR word lattices [18], have also shown improvements.

Instead, this paper innovatively takes the approach of using the comprehensive entities and their relationships from KG and the query clicks from SQCL to improve the statistical language model for speech recognition. KG comprises a large collection of entities, their facts, and a rich set of relationships among them in the form of subject-predicate-object expression [19]. Figure 1 shows a piece of entities and their relationships in KG. A typical commercial KG contains hundreds of millions of entities and billions of their facts, and is a critical component to understand the web more intelligently in the large scale commercial search engines. Meanwhile, search engines such as Google or Bing log more than hundreds of millions of user queries every day, and each query in the log has an associated set of URLs that are clicked after the users issued the query [20, 21]. This user click data could be used to find queries that are highly related to the contents of the clicked URLs, as well as queries that are related to each other.

One simple way to leverage KG is to add all the entity names into the language model vocabulary and training data. However, just adding entity names into the training data may result in tiny or even zero probability of the transition n-gram that consists of a part of the entity name and the adjacent words. Instead, this work utilizes much more capabilities of KG and SQCL. First, the entity names and other important facts in KG and the queries from SQCL are used to improve the named entity coverage in the language model training data. Meanwhile, the language model can be adapted based on the user location to perform better on the entities close to the user, given that these nearby entities are likely to be asked if the user issues a query with the local intent. Second, the entity relationships in KG and the entity relevance mined from the user click data in SQCL can be utilized to boost the probability of the word sequence paths that contain the related entities in a conversational dialog. Experiments conducted on this new approach show promising

http://dx.doi.org/10.21437/Interspeech.2017-1790

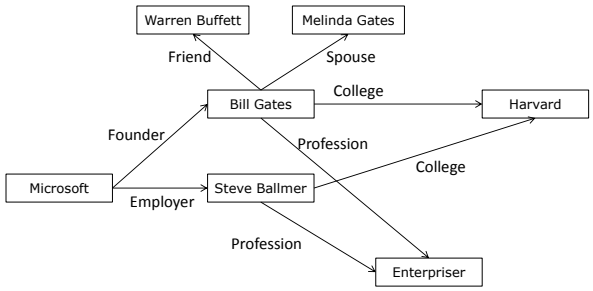

Figure 1: *An example of entities and their relationships in knowledge graph*

improvements on the performance of both the language model and the speech recognition.

This paper is organized as follows. Section 2 explains two concrete components of the proposed approach that leverage KG and SQCL to improve the language model and the speech recognition. Section 3 shows experiments for the language model performance and the speech recognition word error rate (WER). Finally conclusions are drawn in section 4.

2. Language Modeling Using Knowledge Graph And Search Query Click Logs

This work seeks to utilize two important concepts from KG and SQCL to enhance statistical language model, including the facts of the entities, and the relationships among them. The entity facts, together with the queries, are collected to train a geographical adaptive language model; the entity relationships are designed for the context to rescore the n-best hypotheses in the lattice. Here a concrete description of a n-gram language model augmented by KG and SQCL is given to demonstrate how to use them.

2.1. Geographical adaptive statistical language model

Named entities often appear in the voice queries in spoken dialog systems, but some noteless ones are likely to be recognized as other common words that have the same or similar pronunciation. For example, it is very common to ask a spoken dialog system how to go to a particular restaurant, or the facts of a local business like phone numbers. If the entity in the speech utterance is not well known, speech recognizer often fails to recognize it. Please notice that the noteless entities in the voice queries are usually local businesses and have strong relationship with the city where the user is located. Meanwhile, it is much more likely for the user to query a nearby entity than the one that is hundreds of miles away. If the language model can cover these entities in the training data and boost the weights of the nearby entities according to the user location in the recognition phase, the issue can be resolved significantly. Fortunately, lots of entities in KG have the location data. For example, a restaurant owns an address, a person was born or lives in a city, and an event happened in a particular location. This large scale entity location data and entity names along with aliases in KG can be utilized to resolve the issue of unknown entities.

Figure 2 demonstrates how the geographical adaptive statistical language model is built and used. First, all the entities in KG are clustered by location. Here the location is not a specific address but a broader region, like a city group or a state. For example, all restaurants in Seattle and several adjacent cities are gathered into Seattle cluster. Second, in each entity cluster SQCL is used to extract all the queries that contain any names or aliases of the involved entities to build one location based training set. Ideally this location based training set should contain all the entities located in it. In order to make the smoothing better, a generic but small training set can be added to the location based training set. A n-gram language model is then trained on it for each entity cluster. Here a kind of crowdsource is utilized by leveraging the user generated data in search engines, which has a large number of users who keep sending various of queries every day, and the sentence structure of these queries are very similar to the voice queries sent to spoken dialog systems. Each location based training set can be defined as:

$$Set_i \quad = \quad \{Q_j \mid E_k \in Q_j,\ E_k \in C_i\} \tag{1}$$

where Q_j is a query from SQCL, and E_k and C_i are an entity and its corresponding entity cluster respectively.

In online recognition part, one of the n-gram language models that has the same location as the user will be selected as the final geographical language model, and then the static generic n-gram language model trained on a large training set is linearly interpolated with this selected geographical language model. The new geographical adaptive n-gram language model can be represented as:

$$p(w_i|w_{i-n+1}^{i-1}) = \lambda p_s(w_i|w_{i-n+1}^{i-1}) + (1-\lambda) p_g(w_i|w_{i-n+1}^{i-1}) \tag{2}$$

where w_{i-n+1}^{i-1} are the last $n-1$ words before the current word w_i, p_s is the static generic n-gram language model, p_g is the selected geographical n-gram language model based on the user location, and λ is the interpolation coefficient. In this adaptive language model, the selected geographical language model can boost the weights of the nearby entities, which are more likely to be mentioned in the queries than the ones far away. It is easy to get the user location data from GPS, and it can be sent to the speech recognizer along with the voice query without any extra latency.

2.2. Rescoring n-best hypotheses using entity relationships

The motivation of leveraging the entity relationships is that the entities related to the ones from previous utterances are more likely to be mentioned in the coming voice query than the irrelevant ones within a conversational dialog. For example, if a query "American restaurants near me" was asked and a new voice query is issued in the same dialog, the new query is likely to be about a particular American restaurant, like "show me the phone number of Cheesecake Factory". When the particular entity is not famous or has some special words in its name, like "ayutthaya thai restaurant", it becomes very challenging for the spoken dialog system to recognize it. Although the geographical adaptive statistical language model described in 2.1 can be used to address it, the probability of the entity name might be still very small due to the data sparseness issue. In order to overcome this challenge, the entity relationships from KG and SQCL are utilized to boost the scores of the n-best lists that contain the entities related to the ones in the previous utterances.

In order to represent the entity relationships, an entity relevance model is mined from both KG and SQCL. The relevance of any two entities E_i and E_j can be defined as:

$$R(E_i, E_j) = \hat{\lambda} R_{kg}(E_i, E_j) + (1 - \hat{\lambda}) R_{sqcl}(E_i, E_j) \tag{3}$$

where $R_{kg}(E_i, E_j)$ is the relevance that is mined from KG, $R_{sqcl}(E_i, E_j)$ is the relevance computed from SQCL, and $\hat{\lambda}$

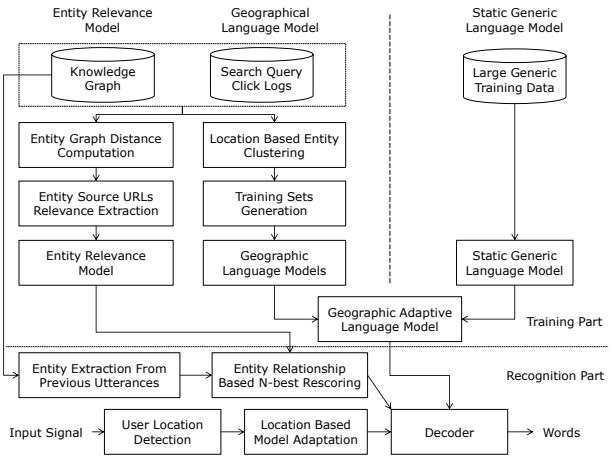

Figure 2: *Architecture of the statistical language model using knowledge graph and search query click logs*

is the linear parameter that can be estimated by the logistic regression. Given that the related entities are connected with each other in KG, the relevance from KG is designed based on the shortest distance $D(E_i, E_j)$ between the two entities in the graph,

$$R_{kg}(E_i, E_j) = \frac{1}{1 + log_2 D(E_i, E_j)} \tag{4}$$

sometimes the path connecting the two entities in KG may be too long, or they may not be connected at all due to the data issue, but they may still have some kinds of relationships. In order to improve the entity relationship coverage, SQCL is leveraged to mine more relationships. Typically every entity in KG has source URLs that identify which webpages the entity facts are extracted from, so once the connection between two source URLs can be detected, the relationship of the two corresponding entities will be found. The following source URLs may have relationships: (a) the ones co-clicked in the same search session, (b) the ones both present in the search results for the same query, and (c) the ones whose webpage contents contain links to each other. The relevance from SQCL is defined as:

$$R_{sqcl}(E_i, E_j) = w_1 C_a(E_i, E_j) + w_2 C_b(E_i, E_j) + w_3 B(E_i, E_j) \tag{5}$$

where $C(E_i, E_j)$ is the correlation of the source URLs of the two entities in set (a) or set (b), $B(E_i, E_j)$ is a boolean value that indicates whether the source URLs have links to each other or not, and w_i is the linear weight that obeys $\sum w_i = 1$. The correlation of the source URLs of any two entities is defined as:

$$C(E_i, E_j) = \frac{2N(E_i, E_j)}{N(E_i) + N(E_j)} \tag{6}$$

where $N(E_i, E_j)$ is the count of the source URLs of the two entities that are both clicked in the same search session or both present in the search results for the same query, and $N(E_i)$ is the count of the source URLs that are clicked or present for one single entity. Assume that one dialog consists of multiple utterences and only the last L previously processed ones have potential connections with the current utterance, and the kth utterance to the last one contains M_k named entities E_j, whose distance to the current utterance is k, and E_i is one of the N named entities in one n-best hypothesis in the lattice for the current utterance, then the entity relevance model is defined as the average relevance with the utterance distance penalty of all the possible entity pairs, which have one entity from the current utterance and the other from the previous utterances. The model can be described as:

$$R(\pi) = \frac{1}{NL} \sum_{i=1}^{N} \sum_{k=1}^{L} \{ \frac{1}{kM_k} \sum_{j=1}^{M_k} R(E_i, E_j) \} \tag{7}$$

where π is one path in the lattice, and typically in most dialogs N and M_k are usually 1 or 0, and L is usually up to 2. If any one of them are 0, the entire score will be 0.

Figure 2 shows how this entity relevance model is used in online recognition part. Once the n-best hypotheses are generated from the lattice, their scores will be linearly interpolated with the entity relevance model. The final likelyhood of a n-best hypothesis can be defined as:

$$H(\pi) = \bar{\lambda} \{ p(\mathbf{Y}|\pi) \}^{\alpha} P_{LM}(\pi) + (1 - \bar{\lambda}) R(\pi) \tag{8}$$

where π is a path in the lattice, $P_{LM}(\pi)$ is the language model probability of the word sequence path π, $p(\mathbf{Y}|\pi)$ is the acoustic likelihood of the observation sequence $\mathbf{Y}$ for the utterance, α is the acoustic scaling factor, $R(\pi)$ is the entity relevance score, and $\bar{\lambda}$ is the interpolation coefficient, which can be estimated by using EM algorithm to minimize the WER.

3. Experiments

Experiments were conducted on voice queries collected from Microsoft Cortana spoken dialog system to evaluate the language model perplexity and the speech recognition WER. Totally the test set collected around 100K sentences that contain 21K named entities and more than 87K unique words from 40K dialogs. All the recognition experiments used a 72-mixture triphone acoustic model trained on 2K hours of Switchboard and Fisher data, and all the language models are trained by using SRILM [1]. The baseline language model was a 4-gram language model with Kneser-Ney smoothing trained on a mixture of voice queries, telephone conversations, broadcast news, lectures, and common named entities, and the vocabulary contained around 103K words. This language model was also used as the static generic language model in the proposed geographical adaptive language model. Since SQCL were used to train this proposed language model, for the sake of fairness, they were also further added into the baseline training set and vocabulary to train an enhanced 4-gram language model as another reference. Microsoft knowledge repository cataloging more than 500M of entities was used as the KG. Note that there are other free knowledge graph products with open access, like Freebase [22]. Bing query click logs over one year that contain more than 1.8B queries were collected to train the geographical adaptive language model and the entity relevance model.

Table 1 shows the comparative results for various language models. Generic (4-gram) is the generic 4-gram language model as the baseline, generic + cache is the cache-based language model [8] that was interpolated with the generic 4-gram language model, enhanced 4-gram is another 4-gram language model that collected both the generic training data and the queries from SQCL in its training set, and generic + geographical adaptive is the proposed geographical adaptive language model. The proposed geographical adaptive language model achieved a perplexity of 77.4, which were 12.5% and 6.4% relative improvements over the generic 4-gram baseline

Table 1: *Perplexity of various language models*

Model	Perplexity
Generic (4-gram)	88.5
Generic + Cache	85.1
Enhanced 4-gram	82.7
Generic + Geographical Adaptive	**77.4**

Table 2: *Comparative WERs of various language models*

Model	WER
Generic (4-gram)	16.2
Generic + Cache	15.9
Enhanced 4-gram	15.6
Generic + Geographical Adaptive	15.4
Entity Relationship based n-best Rescoring	15.8
Proposed	**15.1**

model and the enhanced 4-gram model respectively. This significant improvement demonstrated KG and SQCL are superior data sources for language modeling.

The language models were further evaluated for online LVCSR. Table 2 shows WERs on the evaluation set by running a single-pass real-time non-adapted speech recognition. The proposed approach achieved the best WER of 15.1%, which out-performed the 4-gram baseline and the enhanced 4-gram model by 1.1% absolute and 0.5% absolute respectively. When only one of the two concepts in KG is applied, the geographical adaptive language model and the entity relationship based n-best rescoring decreased the WER by 0.8% and 0.4% absolute respectively. Particularly, in the comparison between the different transcripts generated by the baseline and the proposed approach over the same utterances, around 84% of the difference were caused by the recognition improvement of the named entities. For example, the park named "Weowna Park" in Bellevue Washington can be recognized correctly by the proposed approach, but was recognized as "We gonna park" or "We own a park" by the baseline. Another example is recognizing "Nibbana thai restaurant" as "nearby thai restaurant". Experiments showed the WER reductions for the named entities were statistically significant.

4. Conclusions

This paper has demonstrated that the language model for speech recognition can be significantly enhanced by leveraging KG and SQCL. The proposed approach utilizes the comprehensive entity facts and their relationships to mitigate the false recognition of the named entities. Experiments showed that the language model perplexity was reduced by 12.5% on the voice queries from a spoken dialog system, and WER was reduced by 1.1% absolute over the 4-gram baseline when the proposed approach was applied. Overall, it has been demonstrated that exploiting KG and SQCL for statistical language models in speech recognition is indeed beneficial.

5. References

[1] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Interspeech*, 2002.

[2] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[3] T. Tomita, Y. Okimoto, H. Yamamoto, and Y. Sagisaka, "Speech recognition of a named entity," *ICASSP, IEEE*, pp. 1057–1060, 2005.

[4] A. Stent, I. Zeljkovi, D. Caseiro, and J. Wilpon, "Geo-centric language models for local business voice search," *Computational linguistics*, pp. 389–396, 2009.

[5] E. Bocchieri and D. Caseiro, "Use of geographical meta-data in asr language and acoustic models," *ICASSP, IEEE*, pp. 5118–5121, 2010.

[6] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: A maximum entropy approach," in *ICASSP, IEEE*, vol. 2. IEEE, 1993, pp. 45–48.

[7] R. Sarikaya, Y. Gao, H. Erdogan, and M. Picheny, "Turn-based language modeling for spoken dialog systems," in *ICASSP, IEEE*, 2002, pp. I–781.

[8] R. Kuhn and R. De Mori, "A cache-based natural language model for speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 6, pp. 570–583, 1990.

[9] S. Ikbal, O. Deshmukh, K. Visweswariah, and A. Verma, "Utilizing relationships between named entities to improve speech recognition in dialog systems," in *Spoken Language Technology Workshop, IEEE*, 2010.

[10] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss, "A dynamic language model for speech recognition," in *HLT*, vol. 91, 1991, pp. 293–295.

[11] P. R. Clarkson and A. J. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *ICASSP, IEEE*, vol. 2. IEEE, 1997, pp. 799–802.

[12] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," *SLT*, vol. 12, pp. 234–239, 2012.

[13] M. Mahajan, D. Beeferman, and X. Huang, "Improved topic-dependent language modeling using information retrieval techniques," in *ICASSP, IEEE*, vol. 1. IEEE, 1999, pp. 541–544.

[14] E. Brill, G. Kacmarcik, and C. Brockett, "Automatically harvesting katakana-english term pairs from search engine query logs," in *NLPRS*, vol. 2001, 2001, pp. 393–399.

[15] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker, "Using information retrieval methods for language model adaptation," in *INTERSPEECH*, 2001, pp. 255–258.

[16] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *ICASSP, IEEE*, 2009.

[17] K. Sudoh, H. Tsukada, and H. Isozaki, "Incorporating speech recognition confidence into discriminative named entity recognition of speech data." Association for Computational Linguistics, 2006.

[18] J. Horlock and S. King, "Named entity extraction from word lattices," *International Speech Communication Association*, 2003.

[19] A. Singhal, "Introducing the knowledge graph: things, not strings," *Official google blog*, 2012.

[20] D. Hakkani-Tur, L. Heck, and G. Tur, "Using a knowledge graph and query click logs for unsupervised learning of relation detection," *ICASSP, IEEE*, pp. 8327–8331, 2013.

[21] A. El-Kahky, X. Liu, R. Sarikaya, G. Tur, D. Hakkani-Tur, and L. Heck, "Extending domain coverage of language understanding systems via intent transfer between domains using knowledge graphs and search query click logs," in *ICASSP, IEEE*. IEEE, 2014, pp. 4067–4071.

[22] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 2008, pp. 1247–1250.