PERSONALIZED LANGUAGE MODELING BY CROWD SOURCING WITH SOCIAL NETWORK DATA FOR VOICE ACCESS OF CLOUD APPLICATIONS

Tsung-Hsien Wen[1]*, Hung-Yi Lee*[2]*, Tai-Yuan Chen*[2]*, and Lin-Shan Lee*[1.2]

[1]Graduate Institute of Electrical Engineering,
[2]Graduate Institute of Communication Engineering,
National Taiwan University, Taipei, Taiwan

r00921033@ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

Voice access of cloud applications via smartphones is very attractive today, specifically because a smartphones is used by a single user, so personalized acoustic/language models become feasible. In particular, huge quantities of texts are available within the social networks over the Internet with known authors and given relationships, it is possible to train personalized language models because it is reasonable to assume users with those relationships may share some common subject topics, wording habits and linguistic patterns. In this paper, we propose an adaptation framework for building a robust personalized language model by incorporating the texts the target user and other users had posted on the social networks over the Internet to take care of the linguistic mismatch across different users. Experiments on Facebook dataset showed encouraging improvements in terms of both model perplexity and recognition accuracy with proposed approaches considering relationships among users, similarity based on latent topics, and random walk over a user graph.

Index Terms: Language Model Adaptation, Social Network, Personalized Language Model, Speech Mobile Interface.

1. INTRODUCTION

With the mass production and rapid proliferation of smartphones in recent years, voice access becomes a dream for many cloud applications [1]. It is definitely attractive for many applications if the user can enter his input directly by voice, given the smartphone itself is a voice-operated device. On the other hand, social networks over the Internet have been very popular among almost all people for sharing information, ideas, interests and experiences, as well as interacting with each other in different ways. It turns out that smartphone applications, special voice access of social networks over the Internet, offer some advantages for speech recognition. First, a smartphone is usually used by a single user, so only speaker dependent speech recognizers are needed. Second, large quantities of texts are available over the social networks with known authors and given relations among the authors. So it is possible to train personalized language models, because it may be reasonable to assume that users with close relationships may share some common subject topics, wording habits and linguistic patterns. In addition, most texts on social networks are relatively casual with slightly higher tolerance for recognition errors. This paper is focused on the personalized language modeling problem mentioned above.

N-gram-based language models including various adaptation techniques have been proven to work very well in many applications. For the problem considered here, however, since different users tend to post messages about many relatively disjoint topics on the social networks with significantly different n-gram statistics, the language model trained to work reasonably well for a large group of users may not perform as well for individual users. This can be considered as a cross-individual linguistic mismatch problem, which may have been ignored when the cross-domain linguistic mismatch problem was considered for conventional language model adaptation [2]. Probably because of the lack of large enough personal corpora in early days, it was hard to realize the concept of personalized language model, and therefore we have to aggregate the corpora produced by many different individuals but on similar domains to perform domain-oriented language model adaptation. However, as mentioned above, as the social media blossoms today and given the fact that each user is a part of the social media, the huge quantities of texts left on the network by large number of users with known relationships are handy and therefore the social networks become a very valuable linguistic data resource for language model adaptation. This leads to the fact that the idea of personalization, which has been intensively studied in several other tasks such as personalized search [3, 4] and speaker adaptation [5, 6, 7], is now possible for language modeling.

Unlike most previous works on language model adaptation [8, 9] focusing on the problem of domain mismatch, in this paper we propose a different concept of personalized language modeling with a goal to serve a single user for voice access of cloud applications, in particular social networks, via smartphones. The basic assumption here is that the text messages a user has posted on the social networks are the best source to predict what this user would like to say in the social networks in the future. But, compared with the vast amount of training corpus needed for n-gram language model training, the text messages each user has left on the social network are very limited for the purpose. A nice situation, however, is that it may be reasonable to assume that users with close relationships may share some common subject topics, wording habits and sentence patterns, and the relationships among users are actually given over the social networks. We therefore propose in this paper a crowd-sourcing [10] method which exploits the resource of social network media to develop personalized language models. The experiments showed that by considering of the user's friends over the social networks, the recognition accuracy can be significantly improved. It is also possible to find different weighting schemes to emphasize different aspects of linguistic similarities between the target user and other users to boost the recognition performance further.

2. APPLICATION SCENARIO

The application scenario of the proposed approach is shown in Fig 1, which is actually physically implemented for the purpose of this project. A speech recognition module including a recognition engine is available over the cloud. The smartphone users can utilize

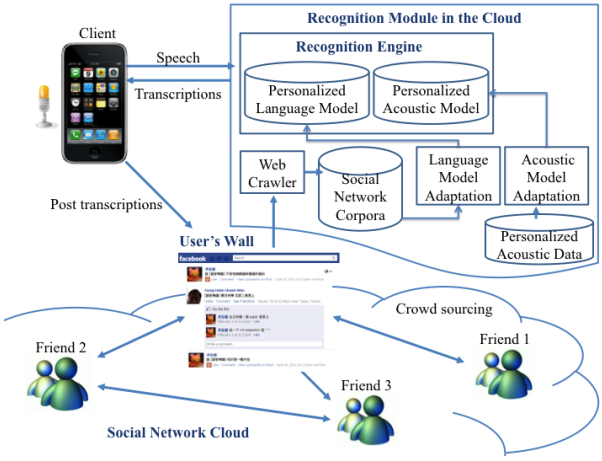

Fig. 1. *The application scenario for a mobile user to post messages over social network by speech. A recognition module with a recognition engine in the cloud is needed, which includes a web crawler to collect the texts for personalized language modeling.*

the speech recognition service to post text to the social network by voice. The space for posting the user's message is referred to as the *Wall* of the user. The recognition module maintains a pair of personalized language and acoustic models for each user. The recognition module transcribes the utterances produced by the user and sends the transcriptions back. When the transcriptions are shown on the screen, the user can decide whether to post it on the *Wall* or not. If the transcriptions are not accurate, the user can correct it manually.

A web crawler is implemented in the recognition module over the cloud for collecting adaptation corpora from the social network. Because of privacy issue, only those information granted by the user is accessible to the crawler, but the crawler is also able to access all public data made available by all other users of the social network. The personalized acoustic utterances produced by each individual user are also collected for acoustic model adaptation, although it is out of the scope of this paper. Below in this paper we focused on personalized language modeling, primarily how to select and weight the corpora posted by different users to better model the language of an individual user.

3. PERSONALIZED LANGUAGE MODELING

The basic framework for language model adaption takes two text corpora into consideration: the adaptation corpus A, which is in-domain or updated with respect to the target recognition task, but probably small and insufficient to train a robust stand-alone language model; the other is a large background corpus B which may not be sufficiently related to the target task or perhaps out-of-dated. Given a word sequence of length N, $\{w_q : 1 \leq q \leq N\}$, that is somehow consistent with the corpus A, the goal is to estimate the probability

$$P(w_1, ..., w_N) = \prod_{q=1}^{N} P(w_q|h_q) \quad (1)$$

where h_q is the history available at time q. The most common approach of doing this is to train two different language models with corpora A and B respectively and adopt a linear interpolation strategy on the model level,

$$P(w_q|h_q) = (1-\alpha)P_A(w_q|h_q) + \alpha P_B(w_q|h_q) \quad (2)$$

where the estimates of $P(w_q|h_q)$ based on the language models trained with corpora A and B are respectively denoted as $P_A(w_q|h_q)$ and $P_B(w_q|h_q)$. $0 \leq \alpha \leq 1$ is the interpolation weight, which is typically estimated on a held-out data from a subset of A under a maximum likelihood criterion using standard EM algorithm.

When applying the above general language model adaption framework to the personalized language model adaptation task considered here, large background corpus B is again assumed available and is used to train a robust background model. The difference here is that the web crawler in Fig. 1 is able to pick up a number of small personal corpora A_i for each user i and the personal corpus A_u of the target user u. With the background estimates $P_B(w_q|h_q)$ based on corpus B, the original personal estimates for the target user u $P_u(w_q|h_q)$ based on corpus A_u, and a set of estimates $P_i(w_q|h_q)$ for user i based on corpus A_i, it is straightforward to rewrite Eq. (2) for a better language model $P^{(u)}(w_q|h_q)$ for the target user u as

$$\begin{aligned} P^{(u)}(w_q|h_q) &= \alpha^{(u)} P_u(w_q|h_q) + \\ &\quad \beta^{(u)} \Sigma_{A_i \in H} \lambda_i^{(u)} P_{A_i}(w_q|h_q) + \\ &\quad (1-\alpha^{(u)} - \beta^{(u)}) P_B(w_q|h_q) \end{aligned} \quad (3)$$

where $H = \{A_i; i = 1, 2...\}$ is the set of all available personal corpora picked up from the social network by the crawler but not including A_u, $\alpha^{(u)}$ and $\beta^{(u)}$ are the weights for the original personal estimates and background estimates respectively given the target user u, $\lambda_i^{(u)}$ is the weight for the estimates based on the i-th personal corpus A_i except the target user u itself, and $\sum_{A_i \in H} \lambda_i^{(u)} = 1$.

Even though that EM algorithm can be applied here to estimate $\alpha^{(u)}$, $\beta^{(u)}$, and $\lambda_i^{(u)}$ under the maximum likelihood criterion, it is not able to offer a good solution to the problem here because of the two reasons below:

I. The number of users on the social cloud is very large, while the data of a single user that can be observed by the system as described may be very limited. With small quantity of training data but large number of parameters to be estimated, the EM algorithm tends to overfit to the training data with poor generalization capabilities.

II. The messages from a social network user tend to cover many completely disjoint topics especially when the messages are posted across different time spans. As a result, the training set and the held-out set used to estimate the model may be quite different from the test set, and can't reflect the true statistics of the language of the user.

In this paper, a set of approaches based on the above framework in Eq. 3 for personalized language modeling is proposed to mitigate the problem, as will be discussed in the following subsections.

3.1. Model Adaptation Framework

The proposed framework for personalized language modeling is shown in Fig. 2. When a target user u wishes to have a personalized language model of his own, he first logs in the social network website and activate the web crawler in the recognition module to start collecting the data, as shown in the left part of Fig. 2. This yields the set $H = \{A_i, i = 1, 2...\}$ of many personal corpora A_i for user i, as well as the personal corpus of the target user u. The personal corpus of the target user u is then divided into two parts: training set A_u and development set D_u. These are also shown in the left part of Fig. 2. The training set A_u for the target user u as well as other personal corpora $H = \{A_i, i = 1, 2...\}$ collected by the crawler are then used to generate one or more intermediate language model(s) based

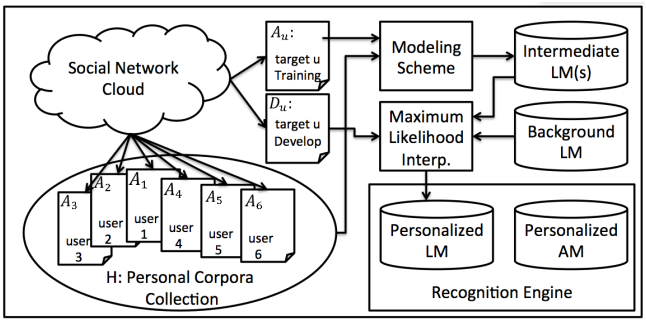

Fig. 2. *Framework for the proposed Personalized Language Modeling using Social Network data.*

on some modeling schemes. Details of these modeling schemes will be discussed later on in the subsections 3.2 and 3.3. This is shown in the upper right corner of Fig. 2. The intermediate LM(s) are then interpolated with the background LM by a set of weights tuned under the maximum likelihood criterion evaluated on the development set D_u for the target user, as in the middle right of Fig. 2. This gives the desired personalized LM to be used together with the personalized acoustic models in the recognition engines at the lower right corner of Fig. 2.

3.2. Personal Corpora Weighting

In this section we are going present a series of approaches for integrating the set of available personal corpora $H = \{A_i, i = 1, 2...\}$ with the personal corpus A_u of the target user u into a single intermediate LM as shown in the upper right corner of Fig. 2. We wish to use Eq. (3) directly by properly estimating the weight $\lambda_i^{(u)}$ for the target user u and all other users i in different ways based on social network relations (Subsection 3.2.1), latent topic similarity (Subsection 3.2.2), and random walk over a user graph (Subsection 3.2.3).

3.2.1. Social Network Relationships (REL)

As several sociolinguistic studies suggested [11, 12], social network relations or contacts among people in a society are recognized as the principle vehicle of language exchange. We follow this idea and assume that in a given social network, if two users are linked together (e.g. friends or relatives) in the network, one is more likely to share some similar linguistic patterns with the other. If there are more frequent interactions between the two, the relation is assumed to be stronger with a higher probability for the two to use same similar linguistic patterns.

Following the above concept, we define a set of social relationship features $f_j(u, i), j = 1, 2...$ for the target user u and all other user i which can be extracted from the records of past interactions of u and i. Good examples of $f_j(u, i)$ include number of common friends between u and i, number of comments sent from u to i and i to u, number of commonly joined groups by u and i, and so on. These features are listed in Table. 1 We can then compute a relevance score $R(u, i)$ for each user i with respect to the target user u:

$$R(u, i) = \sum_j b_j log(f_j(u, i)) \tag{4}$$

where $b = \{b_j, j = 1, 2...\}$ is a weighting vector that is tuned by a development set. We take the *log* over each feature $f_j(u, i)$ so the differences of the feature values can be expanded when small and compressed when large. The value of $\lambda_i^{(u)}$ can then be defined by smoothing and normalizing $R(u, i)$ as

$$\lambda_i^{(u)} = \frac{R(u, i) + \epsilon}{\sum_i [R(u, i) + \epsilon]} \tag{5}$$

where ϵ is a smoothing constant.

Table 1. The components of feature $f_j(u, i)$.

j	Description
1	# of common friends between target u and user i
2	# of comments sent from target u to user i
3	# of comments received by target u from user i
4	# of "Like!" sent from target u to user i
5	# of "Like!" received by target u from user i
6	# of commonly joined groups of target u and user i
7	# of commonly subscribed pages of target u and user i

3.2.2. Latent Topic Similarity by Latent Dirichlet Allocation (LDA)

Instead of considering the social network relationships as described above, another approach is to consider the similarity between the texts posted by the users based on the latent topics addressed in the texts [13]. Various approaches have been proposed to consider the word co-occurrence relations among documents and classify the words and documents into soft clusters called latent topics[14, 15, 16]. Latent Dirichlet Allocation (LDA) [15] has been widely used along this direction, and some language model adaptation works also used it to identify the relevant in-domain corpora addressing similar latent topics to the target recognition tasks [8, 9, 17].

LDA is used in this work for latent topic modeling. Several existing algorithms [15, 18, 19, 20] can be adopted to inference the parameters θ and ϕ, the topic distribution over documents and the word distribution over topics respectively. The collapsed Gibbs sampler [13, 18] is chosen here for parameter estimation. Different from the conventional LDA model based on the concept of "documents", for the problem considered here the unit within each personal corpus over the social network is usually a sentence rather than a document. But LDA modeling on sentence level is difficult because of the sparseness of word observations in each sentence. This is why we train the LDA model on the corpus level. That is, we treat each personal corpus as a "document" in LDA modeling and try to discover the word co-occurrence relations between users. This way of LDA modeling makes sense for the propose here, because it considers the word co-occurrence relationships across different users, which may help in the personalized language modeling. The weighting parameter $\lambda_i^{(u)}$ in Eq. (3) can then be defined as the cosine similarity between $\theta^{(u)}$ and $\theta^{(i)}$, the topic distribution vectors for personal corpora A_u and A_i,

$$\lambda_i^{(u)} = sim(u, i) = \frac{\theta^{(u)} \cdot \theta^{(i)}}{|\theta^{(u)}| \times |\theta^{(i)}|}, \tag{6}$$

which is symmetric for u and i.

3.2.3. Random Walk Over a User Graph (RW)

Random Walk over graphs has been shown to be effective in different tasks, including video search[21], spoken term detection[22], and speech summarization[23]. Since our weights $\lambda_i^{(u)}$ obtained above in Eqs. (5)(6) only take the pairwise relationships of users into consideration, use of a user graph to consider the global relationship

structure among all users may be helpful. In this approach, we construct a directed user graph for the target user u, in which every other user i is a node and the edges describe the relationships between the other users i and j.

We first construct two directed edges between each pair of nodes i and j, both weighted by the same similarity between them,

$$sim(i,j) = \frac{\theta^{(i)} \cdot \theta^{(j)}}{|\theta^{(i)}| \times |\theta^{(j)}|} \tag{7}$$

which is actually the same as Eq. (6), with $\theta^{(i)}$, $\theta^{(j)}$ being the LDA topic distribution vectors. We then prune the edges by keeping only the top K outgoing edges with the highest weights for each node. The similarity in Eq. (7) is further normalized,

$$\rho(i,j) = \frac{sim(i,j)}{\sum_{j \in O_i} sim(i,j)} \tag{8}$$

where O_i is the set of the top K neighbors outgoing from node i. Now the score of node i at time t for the target user u is denoted as $\nu_t^{(u)}(i)$, and the iterative updating formula for $\nu_t^{(u)}(i)$ is then

$$\nu_{t+1}^{(u)}(i) = (1-\gamma)\nu_t^{(u)}(i) + \gamma \sum_{j \in I_i} \rho(j,i)\nu_t^{(u)}(j), \tag{9}$$

where γ is the trade-off parameter, I_i is the set of neighbors connected to node i via incoming edges. The first term on the right hand side of Eq. (9) is the initial score for node i at time t, and the second term describes the score propagation over the graph at time t. So those nodes strongly connected to more nodes with higher scores will have higher scores as well. For $t = 0$, the initial score $\nu_0^{(u)}(i)$ can be simply $\nu_0^{(u)}(i) = \lambda_i^{(u)} / \sum_i \lambda_i^{(u)}$, or $\lambda_i^{(u)}$ of Eqs. (5) or (6) but normalized over all users i. The random walk theory guarantees the score propagation converges after a certain number of iterations N, and the final scores $\nu_N^{(u)}(i)$ are taken as the new weights for $\lambda_i^{(u)}$.

3.3. Sentence Clustering (CLU)

Another possible approach is to cluster all the sentences in all corpora of the target user u and other users i, $\{A_u, A_i, i = 1, 2...\}$ into L sets and estimate an intermediate LM $P_k(w_q|h_q)$ for each of them, where $1 \le k \le L$. We cluster a sentence s based on its topic distribution vector evaluated by a pre-trained L-topic LDA model as described in Sec. 3.2.2:

$$C_k = \arg \max_{1 \le k \le L} \theta_k^{(s)} \tag{10}$$

where C_k is the k-th cluster, $\theta_k^{(s)}$ is the k-th component of topic distribution vector $\theta^{(s)}$ of sentence s as defined for Eq. (6). Given the background estimates $P_B(w_q|h_q)$ and each cluster estimates $P_k(w_q|h_q)$, the personalized estimates $P^{(u)}(w_q|h_q)$ can then be defined as:

$$P^{(u)}(w_q|h_q) = \lambda_B^{(u)} P_B(w_q|h_q) + \sum_{k=1}^{L} \lambda_k^{(u)} P_k(w_q|h_q) \tag{11}$$

where $\lambda_B^{(u)}$, $\lambda_k^{(u)}$ are again weights such that $\lambda_B^{(u)} + \sum_k \lambda_k^{(u)} = 1$. $\lambda_k^{(u)}$ can be either treated as the topic distribution vector computed by sampling latent topics over the development set D_u from the pre-trained LDA model (*CLU1*), or simply optimized by EM algorithm under the maximum likelihood criterion on the background LM and all the intermediate LM(s) based on the development set D_u (*CLU2*).

4. EXPERIMENTS

The recognition module and cloud application scenario as shown in Fig. 1 was implemented with personalized acoustic/language models for registered users constructed, serving as the experimental platform for this research. We chose Facebook as the source of personal corpora for experiments. A detailed analysis of the Facebook corpora is discussed in Sec. 4.1, followed by the experimental setup in Sec. 4.2. The experimental results are then in Sec. 4.3.

4.1. Data Analysis

There are a total of 21 users who logged in and authorized this project to collecting their messages and basic information for the purpose of academic research. These 21 users were treated as our target users u and we tried to build a personalized language model for each of them. Furthermore, with their consents, the public data that are observable to these 21 target users are also available to our system. Through this process, besides the personal corpora for the 21 target users, we also had a whole set of publicly observable data, which are primarily data for those individuals linked with the 21 users on the network. In this way, we had a total of 12635 anonymous personal corpora collected, denoted as users i in the above. With careful survey, we noted the following phenomena in the Facebook dataset: (I) Usually the users only posted part of their thoughts on the wall, so the topics addressed were not continous in time and very often switched very frequently from time to time, which may be the primary reason why the conventional ML criterion is not adequate here. (II) Almost every user had his/her own recurrent usage of some words or expressions. Most of the time it was a signature phrase of the person. Language model estimation may benefit from giving those repeatedly appearing words or expressions higher probability estimates. Some statistics of the Facebook dataset are summarized below: with 21 target users and 12635 other users. We collected a total of 450,000 sentences. After preprocessing and filtering, the total number of sentences used in the work was approximately 280,000. The number of sentences for each user range from 1 to 2943 with mean 23.12, with 12.02 words (Chinese or English or mixed) per sentence in average. On the network, each user was linked with an average of 250 others. This number may become much smaller if only those having real interactions (*comments, likes, common groups, common pages*) with the user were considered.

4.2. Experimental Setup

We build a personalized language model for each of the 21 Facebook target users. For each target user, 3/5 of his corpus is taken as the training set (A_u), 1/5 as the development set (D_u), and the rest 1/5 as testing data for computing the model perplexity. For the background language model, 250M sentences were collected from another popular social network site called Plurk. There were both Chinese and English words in the Plurk data with mixing rate 9:1. Modified Kneser-Ney algorithm [24] was used for language model smoothing. 10K Chinese words and the most frequent 5K English words appearing in the Plurk dataset were selected to form the lexicon. The SRILM [25] toolkit was used for language model training and adaptation. For recognition experiments, 1000 longest sentences selected from the 21 test sets for the 21 target users were read by two male speakers in the preliminary experiments via a smartphone. Recognition accuracy reported below is the accuracy averaged over the 1000 sentences. The Mandarin tri-phone acoustic models were trained on the ASTMIC corpus with 37 Chinese phone set, while the English tri-phone acoustic models were trained on the Sinica Taiwan English corpus with 35 English phone set, both training sets

including hundreds of speakers. The decode weights of 5.0 and 0.5 for language and acoustic models respectively were set empirically. The recognition beamwidth was set to 100. MLLR speaker adaptation was also adopted here.

4.3. Experimental Results

The experimental results are divided into two parts: the preliminary perplexity experiments in Sec. 4.3.1 and the recognition results discussed in Sec. 4.3.2.

4.3.1. Perplexity Analysis

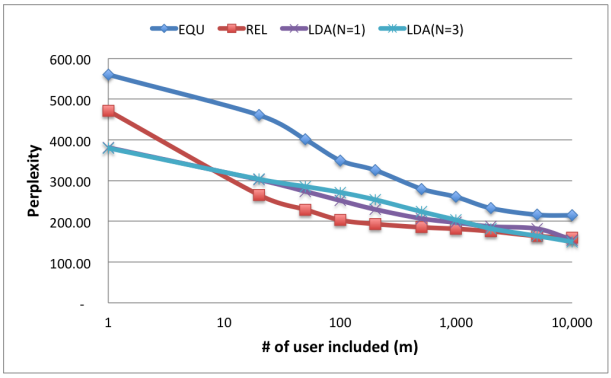

Fig. 3. *The perplexity obtained with different number (m) of personal corpora (A_i of user i) and different modeling approaches. The perplexity was averaged over 21 target users. The number of topics in LDA was set to $L = 50$, with $N = 1$ and 3 indicating that the LDA trained and interpolated on unigram and trigram corpora respectively.*

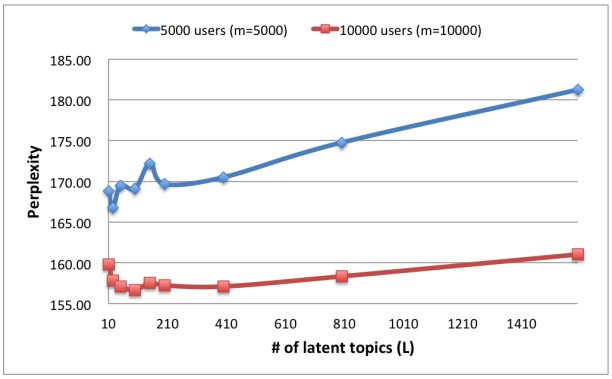

Fig. 4. *The perplexity with respect to the number of latent topics(L) in LDA trained on unigram corpora (N=1) for m=5000 and 10000 other users.*

The perplexity averaged over 21 target users when different number (m) of personal corpora were added for all users (target user u and other users i) equally weighted (EQU), weighted with social network relationships in Sec. 3.2.1 (REL) or LDA topic similarity in Sec. 3.2.2 (LDA) with unigram (N=1) or trigram (N=3) are shown in Fig. 3. Very significant reduction in perplexity appears in the early phase as the number of personal corpora used glowed regardless of the modeling approach used. This indicates that the background language model maintains a significant mismatch to the target corpora even though the data sources Plurk and Facebook are similar in some sense. Obviously carefully selecting the weight for each user is better than giving all users equal weights (*REL, LDA vs EQU*). The advantage of taking the social network relations (*REL*) into consideration is apparent in the beginning ($m < 100$ or so), because there were only less than 100 or so close-related users for most of the 21 target users. The latent topic similarity (*LDA*) based on linguistic information starts to exhibit its superiority in being able to continuously reducing the perplexity when the number of added personal corpora become so large ($m > 200$ or so) that the social network relationships were not able to capture the relationships anymore. However, regardless of the modeling approach it is clear that more data gave better performance. Furthermore, the selection of the number of latent topics (L) for LDA is a key issue. The perplexity results of *LDA(N=1)* for 5000 or 10000 other users ($m = 5000$ and 10000) are shown in Fig. 4. The results suggest that the number of latent topics (L) doesn't affect the perplexity too much if it is confined in a reasonable range, say 30-200, regardless of the number of personal corpora involved. In Fig. 5 we show the detailed perplexity results for each of the 21 target users for equal weights (*EQU*), weighted by social network relationships (*REL*), by LDA(N=3) plus random walk in Sec. 3.2.3 (*LDA(N=3)+RW*), and sentence clustering with EM estimate for weights λ_k in Sec. 3.3 (*CLU2*) for $L = 50$ and $m = 10000$. We see that the perplexity differs from user to user, high perplexity users tend to remain high across different modeling approaches, and a better approach did keep the perplexity low for all the 21 target users.

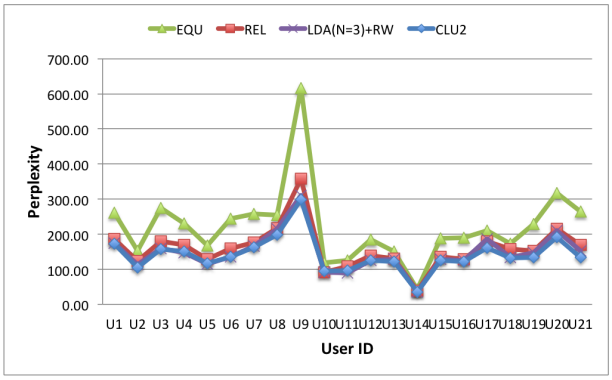

Fig. 5. *The perplexity for each of the 21 target users (U1-U21) for EQL, REL, LDA(N=3) plus random walk(LDA(N=3)+RW) and sentence clustering with EM estimate for weights λ_k (CLU2) for $L = 50$ and $m = 10000$*

4.3.2. Recognition Results

The recognition results with the standard acoustic models (AM) and those adapted with MLLR (AM+MLLR) are shown in Fig. 6. Three baselines are shown in the leftmost section of Fig. 6 using background LM only (*BACK*), the background LM interpolated with the target user's personal corpus by a well-tuned weight (*SELF*), and with all the personal corpora with equal weighting for each user (*EQU*). One can immediate find out that the personal corpus helps significantly (*SELF vs BACK*), offering an approximate 5.6% premium regardless of whether the acoustic model were adapted. With all the personal corpora added with equal weight, the performance can be further improved (*EQU vs SELF*). The middle-left section of Fig. 6 is the family of proposed approaches by personal corpora weighting, with the first two based on social network relations (*REL*), and latent topic similarity by (LDA *LDA(N=1)*) respectively. Both of them provided performance improvements over the base-

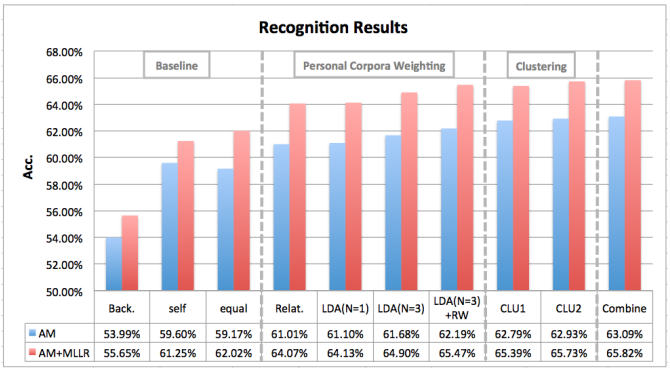

Fig. 6. *The recognition accuracies for baselines and two different proposed adaptation framework: personal corpora weighting and sentence clustering, with/without speaker adaptation respectively.*

line methods *SELF* and *EQU* by 2% or more. *LDA(N=3)* was similar to *LDA(N=1)* but the topic models were trained on unigram, bigram, and trigram corpus respectively, and the intermediate LM was interpolated with different n-gram weights given by the three models. The accuracy can be further improved by approximately 0.6% in this way (*LDA(N=3) vs LDA(N=1)*). Random walk over the user graph offered another additional 0.5% (*LDA(N=3)+RW vs LDA(N=3)*). The middle-right section are the results of the sentence clustering based on LDA with two different estimate approaches for weights λ_k, sampling (*CLU1*) and EM (*CLU2*) respectively as described in Sec. 3.2.2. Although there's no significant accuracy differences between the last three approaches (*LDA(N=3)+RW vs CLU1 vs CLU2*), the performance of *CLU2* was slightly superior than all other approaches, probably due to the larger degree of freedom when applying EM. Finally, the three approaches (*REL, LDA(N=3)+RW, and CLUS2*) that stress different aspect of similarities among people were integrated together in a way that all the intermediate LMs generated by the three were interpolated with the background simultaneously via EM under maximum likelihood criterion again. The combined approach (*Combine*) outperformed all others approaches, with roughly 0.1% to 0.15% improvements over the best (*CLUS2*).

5. CONCLUSION

In this paper, we explore a novel research area of building a personalized LM for a specific user for voice access of cloud applications using the personal corpora of the users left on the social network. Different adaptation frameworks are proposed and different resources are exploited for the purpose. The experiments yielded very encouraging results with the Facebook data. However, this is just a very beginning attempt in this direction, and certainly much more approaches and directions remain to be explored in the future.

6. REFERENCES

[1] D. Hakkani-Tur, G. Tur, and L. Heck, "Research challenges and opportunities in mobile applications [dsp education]," *Signal Processing Magazine, IEEE*, 2011.

[2] Jerome R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication*, 2004.

[3] Geoffrey Zweig and Chang Shuang yu, "Personalizing model m for voice-search," in *Interspeech*, 2011.

[4] M. Speretta and S. Gauch, "Personalized search based on user search histories," in *Web Intelligence, 2005*, 2005.

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," 1995.

[6] P. C. Woodland, "Speaker adaptation for continuous density hmms: A review," 2001.

[7] Jean luc Gauvain and Lee Chin Hui, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," 1994.

[8] Aaron Heidel and Lee Lin shan, "Robust topic inference for latent semantic language model adaptation," in *ASRU*, 2007.

[9] Hsu Bo-June and James Glass, "Style & topic language model adaptation using hmm-lda," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*.

[10] Munro and Robert, "Crowdsourcing and language studies: the new generation of linguistic data," in *NAACL 2010*. Association for Computational Linguistics.

[11] John Paolillo, "The virtual speech community: Social network and language variation on irc," *Journal of Computer-Mediated Communication*, 1999.

[12] Devan Rosen and Margaret Corbit, "Social network analysis in virtual environments," in *ACM on Hypertext*, 2009.

[13] Thomas L. Griffiths and Mark Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, 2004.

[14] Thomas Hofmann, "Probabilistic latent semantic indexing," in *ACM SIGIR*, 1999.

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, 2003.

[16] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth, "The author-topic model for authors and documents," in *UAI*, 2004.

[17] Tam Yik-Cheung and Tanja Schultz, "Unsupervised language model adaptation using latent semantic marginals," in *INTERSPEECH*, 2006.

[18] Gregor Heinrich, "Parameter estimation for text analysis," Tech. Rep., 2004.

[19] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *KDD*. 2008, ACM.

[20] Teh Yee Whye, David Newman, and Max Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *NIPS*, 2006.

[21] Hsu Winston H., Lyndon S. Kennedy, and Chang Shih-Fu, "Video search reranking through random walk over document-level context graph," in *MULTIMEDIA*. 2007, ACM.

[22] Chen Yun-Nung, Chen Chia-Ping, Lee Hung-Yi, Chan Chun-An, and Lee Lin-Shan, "Improved spoken term detection with graph-based re-ranking in feature space," in *ICASSP*, 2011.

[23] Chen Yun-Nung, Huang Yu, Yeh Ching-Feng, and Lee Lin-Shan, "Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms," in *Interspeech*, 2011.

[24] Frankie James, "Modified kneser-ney smoothing of n-gram models modified kneser-ney smoothing of n-gram models," Tech. Rep., 2000.

[25] Andreas Stolcke, "Srilm - an extensible language modeling toolkit," 2002.