International Journal of Computer & Information Technologies (IJOCIT)
www.ijocit.ir & www.ijocit.org

ISSN = 2345-3877

Statistical Language Model Adaptation for Persian Speech Recognition

Seyed Mahdi Hoseini[1], Ahmad Akbari Azirani[2]
Computer Department of Shafagh University[1], Computer Department of Iran University of Science and Technology[2]
Tonekabon[1], Tehran[2]
Iran[1,2]
Mahdi_hoseini@shafagh.ac.ir[1], akbari@iust.ac.ir[2]

Abstract:- Language models are important in various applications especially in speech recognition. Extracting n-gram statistics is a prevalent approach for statistical language modeling. But Traditional n-gram language models suffer from insufficient long-distance information and have crucial dependency on the training corpus. The aim of language model adaptation is to exploit specific, albeit limited, knowledge about the recognition task to compensate for this mismatch. This paper presents an overview of the major adaptation approaches proposed to deal with this issue and we implement these approaches for Persian continuous speech recognition.

Keywords: Speech Recognition, Statistical Language Model Adaptation, Corpus

1. Introduction

Statistical language models have long been used to estimate probabilities for the next word given the preceding word history. A language model plays an important role in automatic speech recognition, both for improving word accuracy and decreasing search costs by constraining the search space and/or allowing more aggressive hypothesis pruning. The task of a language model can be understood as calculating the probability $P(w_i \mid h_i)$, where w_i is the i-th word in the given text and h_i is the history of the word w_i.

Language models represent prior knowledge of word usage in a task domain and, in the case of a stochastic model, assign different probability estimates to different word sequences. Basically, its function is to encapsulate as much as possible of the syntactic, semantic, and pragmatic characteristics. Most of speech recognition systems use a stochastic language model. The most common type of language models used in speech recognition is n-gram language model. This model is simple and fast to use while giving good results. The n-gram is quite powerful but it also has some drawbacks, including: no consideration of long distance dependencies, no understanding the semantics, exponential growth in parameters as a function of n, which increases the storage and compute costs of the recognition search the problem of sparse data for parameter estimation. In large vocabulary speech recognition, most n-gram models use $n \leq 3$. Even with such a small n and a training corpus of several million words, there are still many unobserved n-grams. Smoothing and back-off techniques are therefore introduced to better estimate the probabilities for rare or unseen events. Also, language modeling becomes more problematic when it related to languages with low resource availability.

In this work, we endeavor to build outstanding adaptation methods for Persian speech recognition. The next section poses the adaptation problem. It covers adaptation methods, including Maximum entropy, LSA, pLSA, LDA. LM adaptation experiments are described in Section 3, followed by conclusions Section 4.

2. Language Model Adaptation

Two text corpora are considered: very large and general domain corpus, called background corpus, and small adaptation corpus that pertinent to the current recognition task.

For any upcoming word, we have two distinct sources of information. A general task language model that appropriate for initializing and support the speech recognition, which may helpful for unseen words in current task. And the adaptation language model, which is extracts some specific information relevant to the current task. As we will see in next section, this information may take the form of Maxent constraints, topic identity, etc.

The general idea is to dynamically modify the background statistical language model estimate on the basis of what information can be extracted from adaptation corpus. The adaptation information is incorporated in background language model. However; the adaptation procedure depends critically on the quality of the existing adaptation corpus. In what follows, we concentrate on how to accomplish this adaptation procedure.

2.1. Maximum Entropy Adaptation

Maximum entropy (ME) modeling is a framework that has been used in a wide zone of natural language processing tasks. A conditional ME model has the following form:

$$p(w|h) = \frac{\exp\{\sum_i \lambda_i f_i(w,h)\}}{Z(h)} \qquad (1)$$

Where *w* is a word, and *h* is the word history and:

$$Z(h) = \exp\{\sum_j \lambda_j f_j(w_j, h)\} \qquad (2)$$

The functions f_i are (typically binary) feature functions. During ME training, the optimal weights $\lambda_i$corresponding to features$f_i(w,h)$ are learned. More precisely, finding the ME model is equal to finding weights that maximize the log-likelihood $L(X;\Lambda)$ of the training data Λ. The weights are learned via improved iterative scaling algorithm or some of its modern fast counterparts (e.g., conjugate gradient descent).

Estimating optimal feature weights for language model scan take prohibitively long time if done straightforwardly: in each iteration of the estimation algorithm, one has to calculate normalization factors $Z(h)$ for all observed contexts in the training data. For each context; this requires looping overall words in the vocabulary – also those that didn't occur in a given context. However, [1] proposed a technique that greatly decreases the complexity of calculating normalization factors when features are nested and not overlapping, e.g., n-gram features.

2.2. Latent Semantic Analysis

LSA is one of a growing number of corpus-based techniques that employ statistical machine learning in text analysis. This is stands on the concept of a document, i.e., a ''bag-of-words'' entity forming a semantically homogeneous unit. LSA analyzes relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. The resulting semantic knowledge is encapsulated in a continuous vector space (LSA space) of comparatively low dimension, where all words and documents in the training data are mapped. This mapping is derived through singular value decomposition (SVD) of the co-occurrence matrix between words and documents. Thereafter, any new word and/or document is itself mapped into a point in the LSA space, and then compared to other words/documents in the space using a simple similarity measure.

This framework is very effective at reducing the underlying dimensionality of the discourse, and thus offers promise in tracking semantic changes.

In [2, 3], the LSA framework was embedded within the conventional n-gram formalism, so as to combine the local constraints provided by n-grams with the global constraints of LSA. The outcome is an integrated SLM probability of the form [4]:

$$p(w|h,h') = \frac{p(w|h)\rho(w,h')}{Z(h,h')} \qquad (3)$$

Where *h'* represents the global (''bag-of-words'') document history, $\rho(w,h')$is a measure of the correlation between the current word and this global LSA history [4], and$Z(h,h')$ ensures appropriate normalization. The language model $p(w|h,h')$ represents, in effect, a modified n-gram SLM incorporating large-span semantic information derived through LSA.

2.3. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) is a statistical latent class model or aspect model [5, 6]. The model is fitted to a training corpus by the Expectation Maximization (EM) algorithm [7]. It assigns probability distributions over classes to words and documents and thereby allows them to belong to more than one class, and not to only one class as is true of most other classification methods. PLSA represents the joint probability of a document *d* and a word *w* based on a latent class variable *z*:

$$p(d,w) = p(d)\sum_z p(w|z)p(z|d) \qquad \mathbf{(4)}$$

PLSA has the following view of how a document is generated: first a document *d* ∈*D* (i.e., its dummy label) is chosen with probability $p(d)$. For each word in document *d*, a latent topic *z* ∈*Z* is chosen with probability $p(z|d)$, which in turn is used to choose the word *w* ∈*W* with probability $p(w|z)$.

A model is fitted to a document collection *D* by maximizing the log-likelihood function *L*:

$$L = \sum_{d\in D}\sum_{w\in d} f(d,w)\log p(d,w) \quad (5)$$

The *E*-step in the EM-algorithm is:

$$p(z|d,w) = \frac{p(z)p(d|z)p(w|z)}{\sum_{z'} p(z')p(d|z')p(w|z')} \qquad (6)$$

And the *M*-step consists of:

$$p(w|z) = \frac{\sum_d f(d,w)p(z|d,w)}{\sum_{d,w'} f(d,w')p(z|d,w')} \qquad (7)$$

$$p(d|z) = \frac{\sum_w f(d,w)p(z|d,w)}{\sum_{d',w} f(d',w)p(z|d',w)} \qquad (8)$$

$$p(z) = \frac{\sum_{d,w} f(d,w)p(z|d,w)}{\sum_{d,w} f(d,w)} \qquad (9)$$

The parameters are either randomly initialized or according to some prior knowledge. The parameters $p(w|z)$ obtained in the training process are used to calculate $p(w|d')$for new documents d'with the folding-in process. Folding-in uses Expectation-Maximization as in the training process; the *E*-step is identical, the *M*-step keeps all the $p(w|z)$ constant and re-computes $p(w|d')$. Usually, a very small number of iterations are adequate for folding-in.

2.4. Latent Dirichlet Allocation

LDA extends the PLSA model by treating the latent topic of each document as a random variable. The number of parameters is controlled even through the size of training documents is considerably increased. Different from PLSA, LDA model is capable of computing likelihood function of unseen documents. Typically, LDA is a generative probabilistic model for documents in text corpus. The documents are represented by the random latent topics, which are characterized by the distributions over words. The graphical representation of LDA is shown in Figure 1. The LDA parameters consist of $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$ where $\boldsymbol{\alpha} = \{\alpha_1 \dots, \alpha_M\}$ denotes the Dirichlet parameters of *M* latent topic mixtures, and $\boldsymbol{\beta}$ is a matrix with multinomial entry $\beta_{i,w} = p(w|z_i)$.Using LDA, the probability of an *n*-word document $\boldsymbol{w} = \{w_1, \dots, w_n\}$is calculated by the following procedure.

First, a topic mixture vector $\boldsymbol{\theta}$is drawn from the Dirichlet distribution with parameter $\boldsymbol{\alpha}$. The corresponding topic sequence $\boldsymbol{z} = [z_1 \dots, z_N]$ is generated based on the multinomial distribution with parameter $\boldsymbol{\theta}$in document level. Each word $w_n$is generated by the distribution $p(w_n|z_n,\boldsymbol{\beta})$. The joint probability of $\boldsymbol{\theta}$, topic assignment **c** and document **w** is given by

$$p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha})\prod_{n=1}^{N} p(z_n|\boldsymbol{\theta})p(w_n|z_n,\boldsymbol{\beta}) \quad (10)$$

By integrating (10) over $\boldsymbol{\theta}$and summing it over **z**, we obtain the marginal probability of document **w** by

$$p(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\alpha})\prod_{n=1}^{N}\sum_{z_n=1}^{M} p(z_n|\boldsymbol{\theta})p(w_n|z_n,\boldsymbol{\beta})\,d\boldsymbol{\theta}. \quad (11)$$

The LDA parameters $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$ are estimated by maximizing the marginal likelihood of training documents. The parameter estimation was solved by approximate inference algorithms including Laplace approximation, variational inference, and resampling method [8]. Using variational inference, the variational parameters were adopted for calculating the lower bound of marginal likelihood. The LDA parameters were estimated by maximizing the lower bound, or equivalently minimizing the Kullback-Leibler divergence between the variational distribution and the true posterior $p(\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta})$.

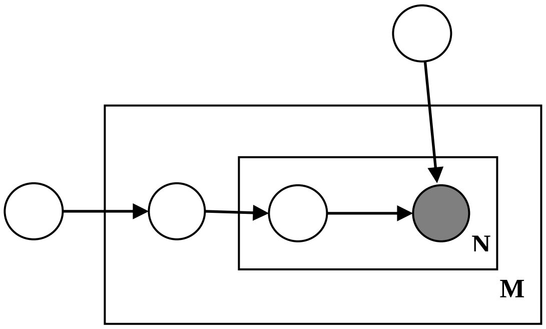

Figure1: Graphical model of LDA

3. Experimental Results

We evaluated our speech recognition system on TFarsdat, a database of Persian conversational telephone speech [9]. This database consists of 320 audio files spoken by 64 different speakers. Speakers have a wide variety of genders, ages and educations. They also cover 10 different Persian dialects. Number of different phones in this database considered as 30.

The training set was adopted for HMM estimation with 52 dimensional PLP acoustic features. The HTK toolkit was used for context-dependent (triphone) acoustic model training. In our setting, each phoneme was represented by a simple left-to-right 3 state HMM with 64 Gaussian mixtures per state. Our train set for ASR, has 25k words and 4k sentences, and our test set has 11k words and 2k sentences.

Background corpus is the FARSDAT [10] text corpus with about 38K words and 4.5k sentences was used to train baseline Kneser-Ney back-off trigram. And adaptation corpus is the text of 1k words and 0.1k sentences of acoustic train set. We trained our adaptation models (LSA etc.) on adaptation corpus and combined them with our baseline model by linear interpolation technique. The table1 shows that adaptation models improve baseline system significantly.

TABLE 1: RECOGNITION RESULTS

	λ for sentence Acc	λ for word Acc	Acc in Sentence	Acc in words
Baseline (Trigram)	-	-	11.70	51.81
ME	-	-	15.25	65.72
Trigram + Trigram	0.4	0.5	20.81	67.95
Trigram + LSA	0.4	0.4	18.73	67.36
Trigram + PLSA	0.5	0.5	21.28	67.28
Trigram + LDA	0.3	0.5	22.31	69.41

4. Conclusions

An adaptive language model seeks to maintain an adequate representation of the domain under changing conditions involving potential variations in vocabulary, syntax, content, and style. This involves gathering up-to-date information about the current recognition task, whether a priori or possibly during the recognition process itself, and dynamically modifying the language model statistics according to this information.

In this paper we carried out outstanding adaptation techniques for Persian Speech recognition. In the experiments on continuous speech recognition, we obtained desirable improvements of recognition accuracy. In future works, we will adopt some other large-scale tasks to examine the Persian speech recognition and explore the approaches to increase accuracy.

References

[1] Wu, J., Khudanpur, S., (2002), “Building a topic-dependent maximumentropy model for very large corpora”. IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing 2002 (ICASSP-2002), 13th-17th May, Orlando, Florida, USA, pp: 777-780.

[2] Bellegarda, J.R. (1998). "Exploiting both local and global constraints for multi-span statistical language modeling".IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing 1998 (ICASSP-1998), 15th May, Vol. 2. Seattle, Washington, pp: 677–680.

[3] Bellegarda, J.R., "Largevocabulary speech recognitionwith multi-span statistical language models". IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 1, 2000, p: 76–84.

[4] Bellegarda, J.R., "Exploiting latent semantic informationin statistical language modeling". IEEE Transactions on ASSP, Vol. 88, No. 8, 2000, p: 1279–1296.

[5] Hofmann T. (1999). "Probabilistic latent semantic analysis". Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 30th Jul-1st Aug, Stockholm, Sweden, pp: 289–296.

[6] Hofmann T. (2000). "Probabilistic latent semantic indexing". Proceedings of SIGIR-99, 15th-19 Aug, Berkeley, Califonia, pp: 35–44.

[7] Dempster, AP., Laird, NM., Rubin, DB., "Maximum likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, Vol. 39, No. 1, 1977, pp:1–21.

[8] Jordan, M., "Learning in Graphical Models". MIT Press,Cambridge, MA, 1999.

[9] Bijankhan, M., Sheykhzadegan, J., Roohani, M., Zarrintare, R., Ghasemi, S. Z., Ghasedi, M. E. (2003). “TFarsdat– the telephone farsispeech database”, Proceedings European Conferenceon Speech Communication and Technology (Eurospeech-2003), 1st-4th Sep ,Geneva, Switzerland, pp: 1525-1528.

[10] Bijankhan, M., et al. (1994). "FARSDAT – The Speech database Of Farsi Spoken Language" Proceedings Australian Conference On Speech Science and Technology (SST'94), Perth, Australia. pp: 826-830.

Authors Profile:

Seyed Mahdi Hoseini is Instructor at Computer Department of Shafagh University. He teaches Algorithm Design, Data structures, Expert systems and etc. at Shafagh University. His present research interests are Speech recognition, Data mining.

Ahmad Akbari is an Associate Professor at the Faculty of Department of Computer at University of Science and Technology of Iran. He also is on the Board of Computer Society of Iran, Head of Iran University of Science and Technology Computer Center. His present research interests are Speech recognition, Computer Networks.