International Journal of Computer & Information Technologies (IJOCIT)
www.ijocit.org & www.ijocit.org
ISSN = 2345-3877

Efficient Incorporation of PLSA and LDA Semantic Knowledge in Statistical Language Model Adaptation for Persian ASR Systems

Seyed Mahdi Hoseini[1], Behrouz Minaei[2]
Computer Department of Shafagh University[1], Computer Department of Iran University of Science & Technology[2]
Tonekabon[1], Tehran[2]
Mahdi_hoseini@shafagh.ac.ir[2], minaeibi@cse.msu.edu 2

Abstract: - Language models (LMs) are important tools for especially ASR systems to improve their efficiency. Development of robust spoken language model ideally relies on the availability of large amounts of data preferably in the target domain and language. However, more often than not, speech developers need to cope with very little or no data, typically obtained from a different target domain. Language models are very brittle when moving from one domain to another. Language model adaptation is achieved by combining a generic LM with a topic-specific model that is more relevant to the target domain. We review a two major topic-based generative language model techniques designed to gain semantic knowledge of text. We show that applying a tf-idf-related per-word confidence metric, and using unigram rescaling rather than linear combinations with N-grams produces a more robust language model which has a significant higher accuracy on FARSDAT test set than a baseline N-gram model.

Keywords: Speech Recognition, Statistical Language Model Adaptation, Corpus,

1. Introduction

Most of the existing automatic speech recognition (ASR) systems generate word hypotheses by computing the probability of a sequence of words as a product of conditional probabilities of words with their context (1). Therefore the probability of a sequence of N words W_1^N is computed as follows:

$$P(W_1^N) = \prod_{n=1}^{N}(w_n|H_n) \quad (1)$$

Considering that w_i is the i-th word of the sequence and Hi is its history. History classes are considered because computing the conditional probabilities for all possible histories is unrealizable. Language models (LMs) provide probability distributions $P(w_i|h_j)$ for each word w_i of the vocabulary and for each history class h_j, including the null history. The probabilities $P(w_i|h_j)$ are computed on a training corpus. In practice their values and their precision depend

on the corpus, its appropriateness to the application domain, and its size. Indeed using a LM computed on a domain D_1 gives poorer results than a LM computed on a different domain D_2 when the application domain is D_2. Furthermore, below a certain size, a corpus doesn't contain enough data to compute correctly a LM. In practice, when developing a new application in a specific domain D_2, the available corpus size could be insufficient. In such case a solution could be based on the utilization of a large amount of data coming from another domain D_1. In order to get a sufficient amount of appropriate data for the new application, it is possible to adapt the data coming from D_1 to the observations coming from D_2. So language model (LM) adaptation is important in speech recognition in order to better deal with a variety of topics and styles

Recently several generative methods in the line of latent semantic analysis have been proposed and used in LM adaptation, such as probabilistic latent semantic analysis (PLSA) [1], and LDA [2]. These existing approaches are based on the "bag of words" model to represent documents, where all the words are treated equally and no relation or association between words is considered. These methods are good at predicting the presence of words in the domain of the text, but not good enough to predict their exact location. The N-gram model complements these models by filling in the missing information – where exactly the content words should go.

In this work, we define an appropriate confidence metric for each word to not only compensate these generative models' weakness (bag of words), but also control their association with N-gram model. The next section illustrates the adaptation framework. Section 3 briefly describes generative methods. Our proposed incorporating method described in Section 4 and 5, followed by experiments results and conclusions in Section 6, 7.

2. SL M Adaptation Framework

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. 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.

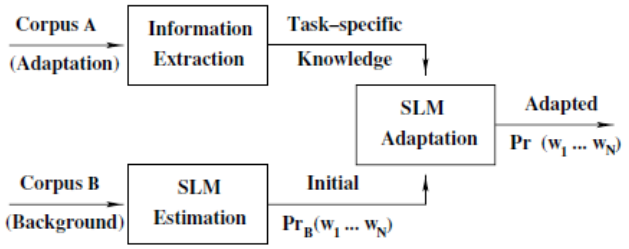

Figure1: General framework for SLM adaptation[3]

The general SLM adaptation framework is depicted in Figure1. Tow text corpora are considered: a (small) adaptation corpus A, relevant to the current recognition task, and a (large) background corpus B, associated with a general task.

3. Generative Models

A generative model for documents is based on simple probabilistic sampling rules that describe how words in documents might be generated on the basis of latent (random) variables. When fitting a generative model, the goal is to find the

best set of latent variables that can explain the observed data (i.e., observed words in documents), assuming that the model actually generated the data. Given the observed words in a set of documents, we would like to know what topic model is most likely to have generated the data. This involves inferring the probability distribution over words associated with each topic, the distribution over topics for each document, and, often, the topic responsible for generating each word.

A variety of probabilistic topic models have been used to analyze the content of documents and the meaning of words [2, 4 - 8]. These models all use the same fundamental idea – that a document is a mixture of topics – but make slightly different statistical assumptions. Hofmann [7, 8] introduced the probabilistic topic approach to document modeling in his Probabilistic Latent Semantic Indexing method (pLSI; also known as the aspect model). The pLSI model does not make any assumptions about how the mixture weights are generated, making it difficult to test the generalizability of the model to new documents. Blei extended this model by introducing a Dirichlet prior on mixture weights, calling the resulting generative model Latent Dirichlet Allocation (LDA) [2]. As a conjugate prior for the multinomial, the Dirichlet distribution is a convenient choice as prior, simplifying the problem of statistical inference.

PLSA and LDA represent two basic formulations of topic language models. Since they were proposed, many extensions of them have been made. See [9] for more information about some of their extensions.

4. TF-IDF-Related Confidence Metric

A drawback with PLSA and LDA language models is that compared to the N-gram it is a weak predictor of function words, and other common words with uniform distribution over contexts. But they perform well at predicting the occurrence of content words which are specific to a context, even if they have not occurred yet in the document. We propose a tf-idf-related confidence metric associated with each word that helps determine to what degree these models are effective at predicting that word. In order to integrate n-gram and PLSA or LDA probabilities, we introduced a tf-idf-related confidence measure for the PLSA or LDA component, based on the observation that words that occur in many different contexts cannot well be predicted by PLSA or LDA.

tf-idf, term frequency–inverse document frequency, is a numerical statistic that reflects how prominent a word is to a document in a set or corpus.

Thus, the confidence for term t is calculated by

$$Confidence_t = \frac{tf(t,d) \times idf(t,D)}{z} \qquad (2)$$

Where Z is for normalization. Several ways for clarifying the exact values of both statistics exist [10]. We use augmented frequency for $tf(t,d)$, to avoid a bias towards longer documents, e.g. raw frequency divided by the maximum raw frequency of any term in the document:

$$tf(t,d) = 0.5 + \frac{0.5 \times freq(t,d)}{max\{freq(w,d) : w \in d\}} \qquad (3)$$

Where $freq(t,d)$is the number of term t in the document d. For reverse document frequency is a measure of whether the term is widespread or rare across all documents. It is found by dividing the total number of documents by the number of documents covering the term, and then taking the logarithm of that quotient.

$$idf(t,D) = log \frac{|D|}{1+|\{d \in D : t \in d\}|} \qquad (4)$$

With $|D|$ is cardinality of D or the total number of documents in the corpus and $|\{d \in D : t \in d\}|$is number of documents where the term t appears. Mathematically the base of the log function does not matter and constitutes a

constant multiplicative factor towards the overall result.

Words that are very casual, occurring in many documents without regard to their content (uniform distribution) will get a very low confidence value, while words that are less casual (more discriminative), usually occurring with the same group of words, will get a higher confidence value.

5. Incorporate PLSA or LDA Semantic Knowledge to N-grams

While the N-gram model is a worthy predictor of words, we want to assure that it contributes at least half the probability mass to predicting the next word. Therefore, we divide the confidence in half. Thus, for words that PLSA or LDA is very confident about predicting, high confidence, the PLSA or LDA and N-gram models will be about equally considered.

$$\mu_i = \frac{Confidence_i}{2} \qquad (5)$$

We found simple linear mixture to be insufficient (Equation 6), partially because the PLSA or LDA estimator often predicts words that are syntactically forbidden.

$$p(w_i|h) = \frac{P_A(w_i|h_A)\times\mu_i + P_B(w_i|h_B)\times(1-\mu_i)}{\sum_{j=1}^{N}\left(P_A(w_j|h_A)\times\mu_j + P_B(w_j|h_B)\times(1-\mu_j)\right)}$$

(6)

We need a non-linear combination function that gives a much higher probability when the two models agree — that is, when the predicted word is both syntactically and semantically likely — and gives a low probability if either estimator believes a word unlikely. The combination scheme we favor is based on an intuition from maximum entropy model fitting by Iterated Proportional Scaling [11]. When the PLSA or LDA confidence is high, this function makes the N-gram and PLSA or LDA model to accept a word is likely in order to get a high resulting probability. When the PLSA or LDA confidence is low, the need for acceptance is reduced.

$$p(w_i|h) = \frac{1}{Z}\left(\frac{P_A(w_i|h_A)^{\mu_i} \times P_B(w_i|h_B)^{(1-\mu_i)}}{p(w_i)}\right)$$

(7)

Where Z is a normalization parameter.

6. Experimental Results

We evaluated our speech recognition system on TFarsdat, a database of Persian conversational telephone speech [12]. 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 [13] 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 (PLSA & LDA) on adaptation corpus and combined them with our baseline model by proposed tf-idf-related interpolation technique.

TABLE 1. PERSIAN SPEECH RECOGNITION ACCARACY RESULTS

	tf-idf-related linear mixture	tf-idf-related non-linear mixture	linear interpolation	baseline (Trigram)
Trigram + Trigram	-	-	67.95	51.81
Trigram +PLSA	68.54	67.91	67.28	51.81
Trigram + LDA	71.63	70.06	69.41	51.81

Table I shows Persian speech recognition accuracy rates. The baseline system is made by Kneser-Ney back-off trigram model trained on background corpus. All trigrams in table I are Kneser-Ney back-off. Semantic information extracted by PLSA and LDA models from adaptation corpus is integrated to background trigram. Proposed incorporation methods and simple interpolation results have shown in table I for better comparison. Accuracy of linear interpolation of local adaptation information (adaptation corpus trigram) and general background model is added to show the importance of inserting Semantic nonlocal information (Semantic). Table I shows that tf-idf-related nonlinear incorporation method outperforms others.

7. Conclusion

An adaptive language model seeks to retain a sufficient representation of the domain under varying conditions involving potential variations in vocabulary, syntax, content, and style. This involves collecting 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 obtain our semantic knowledge from popular generative models (PLSA & LDA). We have shown efficient tf-idf-related incorporating method that can significantly increase the performance of a language model with semantic information. Our semantic confidence measure improved performance by accurately predicting when a semantic model would be beneficial. Finally, a maximum entropy model fitting combination of evidence favors situations where the two orthogonal models agree.
In the experiments on continuous speech recognition, we obtained desirable improvements of recognition accuracy.

[1] D. Gildea and T. Hofmann, "Topic-based language modelsusing EM," in Proc. of Eurospeech, 1999.
[2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
[3] Bellegarda, J. R., "Statistical Language Model Adaptation: Review and Perspectives" Speech Communication, vol. 42, 2004, pp. 93-108.
[4] Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In Proceedings of the 24th Annual Conference of the Cognitive Science Society.
[5] Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In Neural information processing systems 15. Cambridge, MA: MIT Press.
[6] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Science, 101, pp. 5228-5235.
[7] Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.
[8] Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning Journal, 42(1),pp.177-196.
[9] C. Zhai. Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies, 1(1):93–97, 2009.
[10] Manning, C. D.; Raghavan, P.; Schutze, H. (2008). "Scoring, term weighting, and the vector space model". Introduction to Information Retrieval. p. 100. doi:10.1017/CBO9780511809071.007. ISBN 9780511809071.
[11] J. Darroch and D. Ratcliff. Generalized iterative scaling for log–linear models. Ann. Math. Statist. , 43:1470–1480, 1972.
[12] M. Bijankhan, J. Sheykhzadegan, M. Roohani, R.Zarrintare, S. Z. Ghasemi, M. E. Ghasedi, "TFarsdat– the telephone farsispeech database", *Proc. EuropeanConferenceon Speech*

Communication and Technology, pp. 1525-1528,2003.

[13] Mahmood Bijankhan" et al. (1994):FARSDAT – The Speech database Of Farsi Spoken Language" Proceedings Of The 5 Australian Conference On Speech Science and Technology" VOL 2 pp.826-830 " Perth"Australia"2.

Editors 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.

Behrouz Minaei is an Assistant Professor at the Faculty of Department of Computer at University of Science and Technology of Iran. He also is the Director of National Foundation of computer games and Research Senior of Technical High Council Information. His research interests are Data Mining, NLP and machine learning.

