A Study on Latent Words Language Models for Automatic Speech Recognition

著者	MASUMURA　RYO
学位授与機関	Tohoku University
学位授与番号	11301甲第17192号
URL	http://hdl.handle.net/10097/00096959

ますむら　りょう
氏　　　　　名　増 村　亮
研究科，専攻の名称　東北大学大学院工学研究科（博士課程）通信工学専攻
学 位 論 文 題 目　A Study on Latent Words Language Models
for Automatic Speech Recognition
（音声認識のための潜在語言語モデルに関する研究）
論 文 審 査 委 員　主査　東北大学教授　伊藤　彰則　東北大学教授　大町　真一郎
東北大学教授　乾　健太郎　東北大学准教授　能勢　隆

論文内容要約

This thesis aims to enhance statistical language model (LM) technologies for practical automatic speech recognition (ASR) systems. The LMs define a probability distribution over sequences of words and are essential for ASR systems. Modern LMs can show superior ASR performance if domain-matched training data sets are sufficiently obtained. However, in practical cases such as spontaneous speech tasks, large amounts of domain-matched training data sets are not available. Therefore, LM technologies that can flexibly utilize limited domain-matched data sets or out-of-domain data sets are desired. To utilize the limited domain-matched data set and the out-of-domain data sets, there are two important technologies: a robust modeling technology and a mixture modeling technology for domain adaptation. The robust modeling technology is the most important in language modeling. When an LM is constructed from a limited data set, it is expected to robustly predict the probability of unobserved linguistic phenomena. Thus, an LM constructed from a limited domain-matched data set is required to widely work for target domain. In other words, an LM constructed from a certain domain data set is required to robustly work for unknown domains. The mixture modeling technology is also important in language modeling. In fact, both limited domain-matched data sets and out-of-domain data sets should be utilized smartly to specialize in a certain domain.

Thus, multiple LMs are merged for enhancing a certain domain performance for domain adaptation. These two technologies are closely related because the mixture modeling technology is strongly dependent on the robust modeling technology.

To advance LM technologies, this thesis focuses on latent words LMs (LWLMs) recently proposed in the machine learning area. LWLMs are generative models similar to Bayesian hidden Markov models (HMMs), but they have special latent variables called latent words. While standard Bayesian HMMs set up a latent variable size to a small number, LWLMs have vast latent variable space whose size is equivalent to the

vocabulary size of the training data set. This yields characteristics in which latent variables in LWLMs are represented as specific words in the vocabulary. A latent word is regarded as a representative word behind an observed word, and words similar to the latent word have similar probabilities. Thus, LWLMs can automatically optimize the latent variable modeling without determining the size of latent variable space. The attributes efficiently realize robust modeling. Therefore, it can be expected that LWLMs will robustly cover unknown domains and will be effective as component models in the domain adaptation. In addition, the characteristics that latent variables are represented as specific words yield another important property, which is that multiple LWLMs can share a common latent variable space.

The latent variables in usual latent variable based modeling are model-dependent indices, so each model has a different latent variable space. On the other hand, in LWLMs, a latent variable space mixture modeling can be performed. It can be expected that adequate adaptation performance will be offered by out-of-domain component models. Furthermore, any LWLM can be split into two element models, so each element model can be mixed independently. This concept of mixture modeling yields flexibility in that both components are the intersections of different data sources.

A goal reported in this thesis is to develop LWLM-based technologies that can utilize limited domain-matched data sets or out-of-domain data sets for ASR. Four challenges must be faced in this regard.

The first challenge is to introduce the LWLMs to ASR because it is impractical to rigorously compute a generative probability of words using the LWLMs. This thesis introduces two methods that can achieve reasonable implementation. One is an n-gram approximation method in which an LM with a back-off n-gram structure is trained from words randomly sampled on the LWLM. This makes one-pass ASR decoding possible. The other is a Viterbi approximation method that simultaneously decodes a recognition hypothesis and its latent word sequence. Chapter 3 proposed an n-gram approximation method for introducing LWLMs to one-pass ASR decoding. Experimental results revealed that random sampling based on LWLM can generate various linguistic phenomena, and a smoothed n-gram LM constructed from the generated data performs robustly in not only in-domain tasks but also out-of-domain tasks. In addition, an interpolation of the approximated LWLM and a standard n-gram LM effectively improved ASR performance. Although a lot of data was needed to adequately approximate LWLM to the back-off n-gram structure, an entropy pruning was useful in reducing constructed model size efficiently. Chapter 4 proposed a Viterbi approximation method that directly takes account of the latent words assignment. The Viterbi approximation was implemented as a two-pass process in which several recognition hypotheses are initially decoded using the standard n-gram LM; these hypotheses are then rescored using the joint probability between the recognition hypothesis and the latent word assignment. Experiments showed that the Viterbi approximation was effective when it was combined with the first pass results. Moreover, the combination of the n-gram approximation method and Viterbi approximation method improved ASR performance.

The second challenge is to advance a model structure of LWLMs for further domain robustness to various ASR tasks. This thesis presents two novel model structures: latent word recurrent neural network LMs (LWRNNLMs) and hierarchical LWLMs (h-LWLMs). The LWRNNLMs have a soft class structure based on a latent word space as well as LWLMs, where the latent word space is modeled using an RNN structure. The h-LWLMs can be regarded as a generalized form of the standard LWLMs. The key advance is introducing a multiple latent variable space with a hierarchical structure that can flexibly take account of linguistic phenomena not present in a training data set. Chapter 5 proposed LWRNNLMs by combining an RNNLM structure and an LWLM structure. The LWRNNLMs can capture long range relationships in the latent word space while standard LWLMs can only take small context information into consideration. Experiments showed that LWRNNLM, RNNLM and LWLM complement each other and their combinations achieve performance improvement in both n-gram approximation and Viterbi approximation. Chapter 6 proposed hierarchical LWLMs that have a hierarchical latent word space. Experiments showed that h-LWLM offers improved robustness for out-of-domain tasks; an n-gram approximation of h-LWLM is also superior to a standard LWLM in terms of PPL and WER. Furthermore, the proposed approach is significantly superior to the smoothed n-gram LMs or the RNNLMs in out-of-domain tasks.

The third challenge is to establish mixture modeling technologies that can flexibly integrate multiple LWLMs. This thesis presents latent word space mixture modeling methods, i.e., LWLM mixture modeling and LWLM cross-mixture modeling. The latent word space mixture modeling can be expected to efficiently utilize out-of-domain data sets in domain adaptation. For the domain adaptation, this thesis also presents methods to optimize mixture weights using a validation data set. Chapter 7 displayed LWLM mixture modeling and LWLM cross-mixture modeling to utilize out-of-domain data sets including partially matched data sets. The proposed methods perform latent word space mixture that can mitigate a domain mismatch between a target domain and training data sets. Detailed experiments showed that LWLM mixture modeling outperformed n-gram mixture modeling. In addition, a combination of LWLM cross-mixture model and standard LWLM mixture models yielded performance improvements, while using an LWLM cross-mixture model by itself offers little benefit.

The fourth challenge is to reveal relationships between various LM technologies including LWLMs. It is unclear whether a combination of the LWLMs and other important LM technologies is effective or not in practical ASR tasks. Therefore, this thesis examines various combination settings in which the applicable scope of each LM technology is considered. The examinations employ not only manual transcriptions of a certain domain but also external text resources. In addition, unsupervised LM adaptation based on multi-pass decoding and rescoring methods such as discriminative LMs are also added to the combination. The examination presented in Chapter 8 employed major LM technologies while taking their applicable scope into consideration. Experiments demonstrated that significant performance improvements were possible by combining various technologies, compared to using each technology in isolation. The investigations revealed several remarkable facts: the power of a back-off n-gram modeling with combining technologies for direct decoding including vocabulary expansion, the relationship between RNNLM rescoring or unsupervised adaptation and other technologies, and the uniqueness of DLM.

This thesis will show these four challenges can provide ASR performance improvement and beneficial knowledge to language modeling in practical ASR systems.