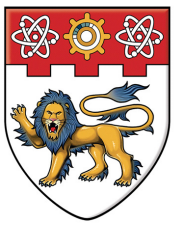
very well written
with depth of analysis
and coherence!
well done!

Domain Adaptation of Language Model for Speech Recognition



NANYANG
TECHNOLOGICAL
UNIVERSITY

A Confirmation Report
Submitted to the School of Computer Science and Engineering
of the Nanyang Technological University

by

Yerbolat Khassanov

for the Confirmation for Admission
to the Degree of Doctor of Philosophy

January 7, 2017

Abstract

Acknowledgments

I would like to express my sincere thanks and appreciation to my supervisor Dr. Chng Eng Siong for his invaluable guidance, support and suggestions. His knowledge, suggestions, and discussions help me to become a capable researcher. His encouragement also helped me to overcome the difficulties encountered in my research.

I also want to thank my colleagues in Rolls-Royce@NTU Corporate lab for their generous help. I want to thank Chong Tze Yuang for his generous help to write my first paper and prepare presentation slides. I also want to thank Benjamin Bigot for introducing me to the speech recognition systems.

I am very grateful to the members of our RT1.1 team. It is a pleasure to collaborate with my team mates, Kyaw Zin Tun and San Linn.

Last but not least, I want to thank my family in Kazakhstan, for their constant love and encouragement.

Contents

List of Figures

List of Tables

List of Abbreviation

AM	Acoustic Model
ASR	Automatic Speech Recognition
BOW	Bag-of-words
CE	Cross Entropy
CED	Cross Entropy Difference
DNN	Deep Neural Networks
fMLLR	Feature Space Maximum Likelihood Linear Regression
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IDF	Inverse Document Frequency
KN	Kneser-Ney
LDA	Latent Dirichlet Allocation
LM	Language Model
LSA	Latent Semantic Analysis
LVCSR	Large Vocabulary Continuous Speech Recognition
MFCC	Mel-Filterbank Cepstral Coefficient
ML	Maximum Likelihood
MLLT	Maximum Likelihood Linear Transform
MT	Machine Translation
NER	Named-entity Recognition
NLP	Natural Language Processing
POS	Part-of-speech
PPL	Perplexity
RBM	Restricted Boltzmann Machine
RNN	Recurrent Neural Network
SLM	Statistical Language Model
SMT	Statistical Machine Translation
sMBR	State-level Minimum Bayes risk
TF	Term Frequency
TM	Translation Model
WER	Word Error Rate
WFST	Weighted Finite State Transducers
WWW	World Wide Web

Chapter 1

Introduction

1.1 Motivation

A brief history of speech recognition systems. Designing a machine that can mimic complex human behaviors such as understanding spoken language and responding accordingly has been envisioned long before advancement of computers. The major step in fulfilling this vision is to develop automatic speech recognition (ASR) systems which have attracted a substantial amount of effort over the last few decades [1].

Given the complexity of human language, the speech recognition technology evolved gradually. The first speech recognition systems focused on simple tasks such as recognizing numbers. For example, in 1952, Bell Laboratories designed *Audrey* [2] which is a first known and documented speech recognizer. *Audrey* could recognize ten digits spoken isolatedly by a single speaker with an accuracy of 97-99%. In 1962, IBM demonstrated *Shoebox*[1], a system which could recognize sixteen words, including ten digits and six arithmetic operations. Over the next decade, speech recognition technology advanced progressively from a simple machine that can recognize a few words to a sophisticated system that can recognize speech with a large vocabulary. Notably, in 1971, DARPA initiated Speech Understanding Research program which was responsible for Carnegie Mellon's *Harpy* [3] system. *Harpy* could recognize speech using a vocabulary of $1,011$ words, approximately the vocabulary of an average three-year-old.

In these large vocabulary systems, however, the complexity of the task had considerably increased, particularly the confusion attributed to homophones. For example, the

[1] www-03.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html

words 'buy', 'bye' and 'by' comprise same phoneme sequence 'B AY' (based on ARPAbet[2] phoneme set). Distinguishing such words was an infeasible task for the early speech recognition systems that mainly relied on acoustic information. Thus, the recognition capability of large vocabulary systems was limited.

Introduction of language models in ASR. The use of only acoustic information proved to be insufficient to achieve human-like performance. Hence, other sources of knowledge were required. Therefore, in 1975, Jelinek et al. [4] proposed to incorporate a grammar structure of the natural language into the speech recognizer. The grammar structure was encoded into a language model (LM) based on statistical principles. The function of the statistical LM was to encapsulate syntactic, semantic, and pragmatic properties of the language considered. In the speech recognition system, the encapsulated knowledge was used to constrain search in a decoder by limiting the number of possible words to follow at any one point. The consequence was faster search and higher recognition accuracy. Since then, the statistical LMs have become an indispensable part of large vocabulary speech recognition systems. We will provide a thorough explanation of the state-of-the-art statistical LMs in chapter 2.

The domain mismatch problem. The statistical LMs retain encapsulated knowledge in the form of probability distribution of linguistic units (e.g. words, sentences) learned from textual training data [5]. It is desirable for training data to possess characteristics similar to input utterances submitted to the ASR system. For example, covering similar topics, speaking styles or both. Otherwise, the distribution learned by LM might mismatch with the target domain distribution of input utterances. As a result, the ASR output will be corrupted [5]. For example, a LM trained on industry domain data, but applied to input utterances from the math domain, might confuse the ASR to recognize 'COFACTOR IS' as 'COW FACTORIES' (the Hamming distance between phoneme sequence of these phrases is 1). Therefore, for reliable performance of ASR systems the distribution learned by LM should fit the target domain.

In ASR systems, however, maintaining a LM that fits the distribution of input test data is a challenging task. Specifically, in the cases where input utterances cover several

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

domains changing over the time such as in broadcast news, talk shows and documentary programs. The trivial solution to deal with such heterogeneous inputs is to assemble training data from various domains in order to construct generic LM. The generic LM enables ASR system to handle input utterance from any domains. While the generic LM offers a good coverage, the recognition performance of ASR system will still be sub-optimal due to the distribution mismatch between generic and specific domain data. Particularly, in the generic ASR systems, commonly used terms will 'push aside' domain-specific terms (e.g. law, technical and medical domain terms). For example, technical domain term 'iPad' might be misrecognized as a combination of two commonly used terms such as 'eye' and 'pad'. The domain-specific terms constitute an essential part of utterances that contribute to the context and meaning. Therefore, the correct recognition of such terms is crucial. In this thesis, we will focus on adapting generic LM to better fit the specific domain data. The comprehensive explanation of domain mismatch problems will be provided in chapter 2

Extracting target domain information. To perform LM adaptation, information about target domain is required, such as a list of keywords, a topic of discourse or collection of in-domain documents. This target domain information can be obtained in a supervised or an unsupervised manner. In a supervised manner, the target domain information is manually generated by domain experts, for example, by analyzing initial ASR output (word lattice or 1-best) produced by a generic LM. Whereas, in an unsupervised manner, the domain information is generated automatically, for example, derived from the initial ASR output by employing information retrieval techniques. While the supervised approach provides reliable and adequate information, it is time and cost ineffective. In this work, we will extract domain information in an unsupervised manner from the ASR output.

Although the ASR output is a valuable source of target domain information, it is prone to errors caused by the recognition process. The recognition errors might corrupt the domain information present in the ASR output. Nevertheless, by simulating different levels of word error rate (WER) in the ASR output, Clarkson and Robinson [6] showed that transcripts with high WER can still benefit the adaptation process. In another work,

Lecorvé et al. [7] used only incorrectly recognized parts of the ASR output to perform adaptation. Surprisingly, they obtained more than 10% relative perplexity improvement. They concluded that some misrecognized words are still in-domain words aiding to capture appropriate domain information and others are harmless. Thus, despite the presence of errors, the ASR output contains valuable information which can be effectively utilized by adaptation techniques.

A brief review of existing adaptation techniques. Since the introduction of statistical LMs, several LM domain adaptation techniques have been proposed to alleviate the effect of distribution mismatch [8]. In practice, LMs can be adapted in two different stages of recognition process: online and offline. In online adaptation, the LM is adapted during the decoding process. However, the decoding process itself is a highly complex mechanism involving intensive computations which make the online LM adaptation impractical. In offline adaptation, on the other hand, the generic LM is first applied to produce initial ASR output (word lattice). Then, produced ASR output is utilized to generate target domain information, in supervised or unsupervised manner, which is employed to adapt the generic LM offline. Lastly, the adapted LM is applied to re-decode the input utterances, or to re-score the word lattice (or N-best list).

Given the complexity of decoding process, re-decoding the input utterances is a tedious task. Hence, only a few LM types, for which fast decoding algorithms are available, are eligible for this task such as backoff n-gram models [10, 11]. The backoff n-gram model is a predominant choice for decoding in the state-of-the-art ASR systems due to its effectiveness and simplicity [12] (generic LM is a backoff n-gram model). Whereas, more complex models, such as neural network based [13], are usually employed to re-score the word lattices [9]. While the complex models are expected to have greater prediction power, the efficacy of re-scoring is constrained by the quality of generated word lattice which contains only subset of all possible hypotheses. For example, an inadequate LM used during the decoding stage might discard correct hypotheses, as a result, a deficient word lattice will be produced [14]. Hence, in this work, we will focus on adapting backoff n-gram models which can be effectively applied to re-decode the input utterances.

The three popular backoff n-gram model adaptation techniques applied to ASR systems are cache-based, topic-mixture and query-based. These techniques employ domain-specific information to tune distribution of generic LM so that it better matches the target domain. For example, the cache-based techniques [6, 15–18] are based on the hypothesis that a word used in a recent past is more likely to be used again. Hence, the probabilities of recognized words are increased within the generic LM. In the topic-mixture techniques [6, 18–21], the generic LM is dismantled into several sub-domain (or sub-topic) LMs interpolated together. Here, the domain of the final interpolated LM can be controlled by tuning interpolation weights of the sub-topic LMs. Hence, the ASR output is used to find the 'closest' sub-topic LMs to increase their weights. The query-based techniques [7, 22–24], on the other hand, use ASR output to generate queries which are submitted to external sources, such as world wide web (WWW), to retrieve 'similar' data. The retrieved data is then used to update parameters of the generic LM. For example, by training new 'pseudo' in-domain LM from the retrieved data and interpolating it with the generic LM. These LM adaptation techniques have been shown effective to improve recognition performance of ASR systems. The complete review of these techniques will be given in chapter 2.

The proposed adaptation approach based on data selection. The existing adaptation techniques typically adjust the distribution learned by generic LM to match the target domain distribution. This adjustment is performed by directly changing parameters of the generic LM, for instance, by increasing or decreasing probabilities of individual words (or n-grams). Changing parameters of LM might help to achieve desired distribution, however, the adapted LM most probably won't represent a distribution corresponding to the natural text produced by human. Consequently, the encapsulated knowledge might be corrupted. Thus, in this work, rather than directly updating parameters of the generic LM, we will examine other adaptation methods that preserves the 'natural' distribution of linguistic units.

In particular, we propose to manipulate the training data used to build generic LM. As was mentioned previously, the training data consist of text assembled from various domains. Hence, we will employ data selection techniques [25] to select a subset of

training data 'similar' to the ASR output (broad overview of data selection techniques will be exposed in chapter 3). As a result, out-of-domain sentences will be discarded, leaving only in-domain sentences. The in-domain sentences are then used to train new LM which is expected to better converge with the target domain distribution. In addition, the new LM will represent adapted version of generic LM, since it was build from the same, but pruned data. More importantly, the adapted LM produced in this way will encapsulate appropriate linguistic knowledge which complies with the regularities of natural language. To evaluate the effectiveness of proposed approach we conducted several experiments on TED-LIUM speech corpus which will be described in chapter 4.

1.2 Contributions

In this thesis, we proposed unsupervised LM adaptation framework to address domain mismatch problem inherent in generic ASR systems. The proposed framework is based on data selection technique which customizes generic background corpus to produce domain-specific LM. The novelties of the proposed framework are listed below:

1) Existing LM adaptation techniques aim to tune parameters of the generic model to diverge its focus towards target domain. Different from them, the proposed approach employs ASR output and data selection techniques to perform adaptation at the data level. This work shows that a LM adapted in this way possesses a strong discriminative ability that results in substantial WER reduction.

2) Although the generic background corpus is sufficiently large and contains data from various domains, several adaptation techniques (e.g. query-based) still require in-domain data retrieved from external sources such as WWW. Unlike these approaches, our method efficiently utilizes available background corpus by intelligently selecting in-domain sentences. Hence, the proposed method doesn't rely on any external source which might be unavailable for some tasks involving private corporation or military domains.

Experiments performed on TED-LIUM speech corpus show that proposed adaptation framework can produce domain-specific LM that achieves up to 10% relative WER reduction. When we adapted LM to a more specific domain the WER reduction up to 12% was observed. Moreover, we compared our approach against standard adaptation

method based on linear interpolation which directly updates parameters of a LM, and observed better WER.

The work on unsupervised LM adaptation by data selection was accepted by ACIIDS conference [26].

1.3 Report Organization

The report is organized as follows:
In Chapter 2, we provide background information on ASR systems, statistical LMs, and domain mismatch problem. We describe general LM adaptation framework, followed by a review of popular LM adaptation techniques applied to ASR systems.
In Chapter 3, we provide an overview of the current state-of-the-art data selection techniques including linguistic features used to represent data and similarity metrics. We briefly review other natural language processing (NLP) applications where data selection has been employed.
In Chapter 4, we propose data selection based unsupervised LM adaptation framework for ASR systems. We explain experiment setup and data. Lastly, obtained results are discussed.
Chapter 5 concludes the report and lists future research directions.