様式：課程博士用

24年　　8月　　31日

電子・情報工学専攻	学籍番号	099311
申請者 氏名	小林彰夫	

指導 教員	中川聖一 秋葉友良

論文要旨(博士)

論文題目	放送番組における音声認識のための識別的言語モデルの研究

(要旨　1,200字程度)

本論文では，音声認識の性能改善を目的とした識別学習に基づく識別的言語モデルについて論ずる．識別的言語モデルでは，音声認識結果に含まれる認識誤りに着目して，その言語的な誤り傾向を統計的にモデル化して認識率の改善を行う．

本論文ではまず，音声データとその書き起こしが与えられた場合の教師あり識別的言語モデルついて論述する．提案する識別的言語モデルでは，文仮説に対して従来の統計的音響モデル・言語モデルによるスコアに加えて，仮説(正解候補となる文)の誤り傾向に応じたペナルティが与えられる．識別的言語モデルの学習では，仮説の誤り傾向をモデルに反映するために識別的な学習方法を採用し，教師あり学習として定式化を行う．識別的言語モデルは，言語的な文脈により活性化する素性関数(例えば，仮説に含まれる単語列の数)とその重みにより表現され，これらが文仮説に対するペナルティとして機能することで音声認識の性能改善を行う．放送ニュースを対象とした音声認識実験により，本論文が提案する教師あり識別的言語モデルは音声認識性能を統計的に有意に改善した．

次に，正解単語列の付与されていないラベルなしデータを用いた識別的言語モデルについて論述する．教師あり識別的言語モデルでは正解ラベルを人手により作成するため，学習データを大量に用意することはコストの面から困難である．したがって，教師ありモデルでは少量のデータから学習せざるを得ず，頑健性が失われることになる．そこで，大量の放送番組の音声認識結果を用いて，教師なしで識別的言語モデルを学習する枠組みを検討する．識別的言語モデルの教師なし学習では，教師あり学習の目的関数を拡張して，ラベルなしデータ全体に対するリスク(期待される誤り)を最小化するように定式化される．放送ニュースを対象とした音声認識実験により，本論文が提案する教師なし識別的言語モデルは音声認識性能を有意に改善した．

最後に，ラベルあり・ラベルなしデータを併用した半教師あり識別的言語モデルについて論述する．識別的言語モデルの頑健性を改善するには，ラベルありデータだけではなく，ラベルなしデータを併用することが有効である．そこで，教師あり・教師なしの識別的言語モデルの学習手法を組み合わせることを目的として，多目的最適化手法に基づく半教師あり識別的言語モデルを検討する．識別的言語モデルの半教師あり学習では，教師あり・教師なし学習の定式化で用いた目的関数を多目的最適化手法で統合する．多目的最適化手法では，2つの目的関数を統合した関数の最適解を求める代わりに，妥協解の集合を求めることにより，評価対象となるデータに適合したモデルを推定する．放送番組を対象とした音声認識実験により，半教師あり学習に基づく識別的言語モデルは，統計的に頑健かつ有意な音声認識性能の向上を示した．

year month day
2012 8 31

Department	Electrical and Electronic Information Engineering	ID	099331
Name	Akio Kobayashi		

Supervisor	Seiichi Nakagawa Tomoyoshi Akiba

Abstract

Title	A Study on Discriminative Language Modeling for Automatic Speech Recognition in Broadcast Programs

(800 words)

This doctoral dissertation describes discriminative language modeling for transcribing broadcast programs. This study focuses on statistical language modeling utilizing tendencies of word errors in automatic speech recognition (ASR) transcriptions with a goal of improvement in performance.

The recent progress in the field of ASR-based language processing has led to its successful application in the real world. For example, NHK has developed a system for closed-captioning broadcast news using real-time ASR. ASR technology also plays a crucial role in the development of a broadcast archiving system, which serves as a basis for spoken document processing applications. The availability of these applications strongly depends on the accuracy of ASR, and recently there has been many interest in applying discriminative acoustic or language models for improvement. Although these models typically require a large amount of manually transcribed (labeled) data, there are only limited resources available in reality. Information from unlabeled data such as ASR transcriptions could, therefore, be useful for increasing the robustness of the models. In this dissertation, then, a novel semi-supervised language modeling method is proposed. The dissertation has three parts: a study on supervised discriminative language modeling, a study on unsupervised discriminative language modeling and a study on semi-supervised discriminative language modeling.

First, discriminative language modeling in a supervised manner is studied under the condition that manually transcribed training data are given. The discriminative language model is formed as a log-linear model, which employs a set of linguistic feature functions and weighting factors. These feature functions are typically activated by linguistic contexts such as word/phoneme sequences. The model assigns the weighting factors as penalty scores to sentence hypotheses according to tendencies of word errors along with conventional scores derived from statistical acoustic/language models. To be obtained such error-sensitive penalties, the weighting factors of the model are discriminatively trained. The training method is based on the minimization of word errors, which is performed on training lattices, more specifically, on the sets of sentence hypotheses. Conventionally, discriminative language modeling has been performed on the basis of maximization of conditional log-likelihood of references, which does not reflect word errors but the scores of references. The results of transcribing Japanese broadcast news showed supervised discriminative language modeling achieved statistically significant performance compared with conventional discriminative models.

Next, an unsupervised version of discriminative language modeling is proposed. In supervised training, a large amount of transcribed data is required for statistically robust modeling. However, such data are often limited from a cost viewpoint. In an unsupervised manner, a large amount of ASR transcriptions are often utilized as training data instead of manually transcribed data. In the perspective of language modeling, unsupervised training is typically conducted by a model-based linear interpolation method. However, this method does not always perform the best in terms of word error rates (WERs), since the model are estimated by using ASR transcriptions containing misrecognized incorrect words. Obviously, the model is needed to reduce the influence of these erroneous words, or reflects information of the errors through discriminative training. In unsupervised discriminative language modeling, the training procedure is performed with ASR transcriptions (lattices) without references. It minimizes the whole risk of training lattices to yield a log-linear model, which is defined as a generalized version of supervised language modeling. Experimental results obtained in transcribing Japanese broadcast news showed significant word error rate reduction for unsupervised discriminative language modeling, while the conventional linear interpolation method achieved larger improvements. The advantageous effect of unsupervised discriminative language modeling is to provide supplementary improvements because of error-sensitive scores.

Finally, the dissertation describes a method for semi-supervised language modeling, which was designed to improve the robustness of a discriminative language model. When extending supervised discriminative language modeling to its semi-supervised version, we have two key issues to be solved. One issue is how to design objective functions for labeled and unlabeled training data. For the maximum use of information from different types of data, the objectives should be required to be compatible, i.e., designed with a consistent criteria. Then, risk-based objectives would be worthy of being utilized in semi-supervised discriminative language modeling. The remaining issue is how the contribution of unlabeled training data is reflected in the supervised models. Although the semi-supervised modeling is formulated as an optimization problem consisting of two independent objectives, it would be difficult to find the optimum that minimizes both objectives simultaneously. To address this issue, a semi-supervised modeling approach based on "multi-objective optimization programming" (MOP) is proposed. In transcribing Japanese broadcast programs, the proposed semi-supervised discriminative language models reduced WERs significantly compared with both supervised and unsupervised discriminative models.

This dissertation concludes that the discriminative language model, which is estimated from the risk-based objective in a supervised manner, could reduce word errors significantly when a sufficient amount of labeled training data is available. It additionally revealed that the discriminative model, which is estimated by semi-supervised discriminative language modeling, could similarly reduce word errors with the assistance of a large amount of unlabeled data even when there was a limited amount of labeled data.