情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-SLP-84 No.1
2010/12/20

カテゴリ推定に基づく動的な言語モデル適応

山本 仁[†1] 花沢 健[†1]
三木清一[†1] 篠田浩一[†2]

音声入力インタフェースを備える情報検索システムにおいて、検索キーワードの音声認識精度を高めるための言語モデル適応手法を提案する。提案手法は、検索発話に現れるキーワードに対して動的に言語モデルを適応させる。まず、カテゴリ識別処理によって、入力された発話に含まれるキーワードのカテゴリとその区間を推定する。次に、この推定結果に基づき、発話内の区間ごとに異なる重みを用いて、カテゴリごとに用意した言語モデルを混合しながら音声認識を行う。テレビ番組検索タスクにおいて、音声認識実験による評価を行い、キーワード誤り率の 22.0%の削減を確認した。

Dynamic Language Model Adaptation using Keyword Category Classification

Hitoshi Yamamoto,[†1] Ken Hanazawa,[†1]
Kiyokazu Miki[†1] and Koichi Shinoda[†2]

This paper describes a language model adaptation method for improving speech recognition of keywords in spoken queries occurring in information retrieval tasks. The method dynamically adapts language models to keyword categories within a single utterance; it first estimates keyword categories and their positions in an input query utterance and then dynamically changes the weights for language models designed for individual keyword categories on the basis of the estimation results. The method has been evaluated in speech recognition experiments on television program retrieval tasks and has demonstrated a 22.0% reduction in keyword error rates.

†1 日本電気株式会社
NEC Corporation
†2 東京工業大学
Tokyo Institute of Technology

1. はじめに

音声入力インタフェースを備える情報検索システムにおいて、検索キーワードの音声認識精度を高めるための言語モデル適応手法を提案する。

近年の自動音声認識技術の進展により、モバイル端末や情報家電などの機器で情報検索のようなアプリケーションを操作するための入力手段として、音声が用いられるようになってきた。本稿では、テレビ番組やレストランなどのデータベースを対象とした情報検索タスクにおける音声入力の精度向上をはかる。

この種の情報検索の音声インタフェースは、使用者が自発的に話した、さまざまなカテゴリのキーワード（名前・場所・時間など）を含む発話を受理できることが望ましい。本稿では、このための音声認識処理として、文法規則ではなく、統計的言語モデルに基づく大語彙連続音声認識を採用する。その上で、キーワードの音声認識精度を高めるための言語モデル適応手法として、カテゴリごとに構築された言語モデルを混合させて用いる。

従来、複数の言語モデルを混合する言語モデル適応手法として、トピック依存言語モデルを用いる手法が知られている[1),2)]。これらの手法の多くは、トピックの重み推定のために文単位の履歴情報を要するが、本稿で扱うような情報検索タスクの場合、入力されるのはひとつの短い発話のみであり、そのような履歴情報は得られない。また、検索発話は複数の異なるカテゴリのキーワードを含むことがあるため、発話内においても重みを変更できることが望ましい。このような条件に対して、トピック依存言語モデルを用いることは難しい。

また、発話内の履歴情報を用いて、クラスの出現確率や混合重みを推定する方法が提案されている[3),4)]。たとえば、クラス 3-gram による手法[3)] は、先行する 2 単語を履歴情報として次の単語のクラスを推定する。この手法は、ひとつの発話内においても重みを動的に推定すると言えるが、発話全体にわたる言語的な特徴を扱っていない。検索発話のキーワードは、そのカテゴリによって、発話内での出現傾向が異なる。たとえば、「△△チャンで○○さんが出演しているトーク番組」という発話からは、人名のキーワードは、「出演する」などの動詞に先行する、放送局名のキーワードに後続する、などの現象が観察できる。このような発話全体にわたる言語的な特徴を扱うことによって、カテゴリの推定がより正確に行えると期待できる。

本稿では、カテゴリ識別に基づく動的な言語モデル適応手法を提案する。提案手法は、二段階処理によって、検索発話に現れるキーワードに対して動的に言語モデルを適応させる。まず、入力発話に対し一旦音声認識を行い、その認識結果の単語境界を用いて発話を複数の

Vol.2010-SLP-84 No.1
2010/12/20

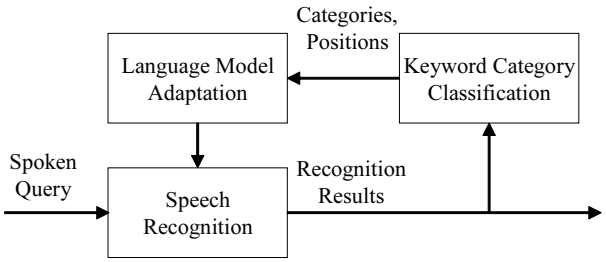

図 1 提案手法の構成図

区間に分割する。次に、カテゴリ識別処理によってそれぞれの区間の属するカテゴリを推定する。さらに、推定結果に基づいて定めた重みによって、カテゴリごとに構築した言語モデルを区間ごとに混合し、音声認識を行う。カテゴリ識別では、発話全体にわたるコンテキスト情報を特徴として扱う識別モデルを用いる。また、発話の区間ごとにカテゴリを推定するため、重みは 1 つの発話内で変動する。そのため、検索発話に複数の異なるカテゴリのキーワードが含まれる場合であっても、それぞれのキーワードに対して動的に言語モデルを適応できる。

以下、第 2 節で提案手法の枠組み、第 3 節でカテゴリ識別、第 4 節で言語モデルの動的適応について述べる。第 5 節では、テレビ番組検索発話の音声認識実験による評価結果を報告する。

2. 提 案 手 法

提案手法は、図 1 に示すように、3 つの処理から構成される。音声認識処理（Speech Recognition）は、入力された検索発話を 2 回処理する。第 1 パスの音声認識では、従来の言語モデルを用いて、音声認識結果を出力し、その認識結果の単語境界を用いて発話を複数の区間に分割する。カテゴリ識別処理（Keyword Category Classification）は、区間ごとにそのカテゴリを推定する。言語モデル適応処理（Language Model Adaptation）は、推定結果に基づき求めた重みで、カテゴリごとの言語モデルを混合する。第 2 パスの音声認識は、この区間ごとに混合重みが異なる言語モデルを用いて発話を再び認識する。

3. カテゴリ識別

カテゴリ識別部は、入力された検索発話に対して、その区間ごとに適切なカテゴリを推定する。第 1 パスの音声認識によって、入力発話に対応する単語列を求め、その単語列の各単語をカテゴリに分類する。

この自動分類を行うため、本稿では CRF（Conditional Random Fields）[5] を用いる。CRF は、単語列などの系列に対するラベリングに適した識別モデルであり、入力された単語列の各単語に対して、多種の素性関数により特徴を抽出し、対応するラベルとその事後確率とを出力することができる。CRF を次式（1）のように用いることで、区間ごとのカテゴリ推定を一度の識別処理によって行うことができる。

$$P(\boldsymbol{C}|\boldsymbol{W}) = \frac{\exp(\boldsymbol{\Lambda}\cdot\boldsymbol{\Phi}(\boldsymbol{C},\boldsymbol{W}))}{\sum_{\tilde{\boldsymbol{C}}}\exp(\boldsymbol{\Lambda}\cdot\boldsymbol{\Phi}(\tilde{\boldsymbol{C}},\boldsymbol{W}))} \tag{1}$$

ここで、$\boldsymbol{W}$ は第 1 パスの音声認識結果として出力される単語列、$\boldsymbol{C}$ は $\boldsymbol{W}$ に対応するカテゴリ列である。本稿では、あらかじめ定めたキーワードのカテゴリに加えて、設定したカテゴリに属さない単語も「非キーワード」というひとつのカテゴリとして扱う。$\boldsymbol{\Phi}$ は $\boldsymbol{W}$ と $\boldsymbol{C}$ から特徴を抽出する素性関数からなるベクトルである。$\boldsymbol{\Lambda}$ は $\boldsymbol{\Phi}$ の要素それぞれの重みからなるベクトルで、CRF のパラメタである。このパラメタは、カテゴリ列と対応づけられた単語列からなる学習データに対して最尤になるように最適化される。与えられた単語列 $\boldsymbol{W}$ に対して、単語列 × 出力候補カテゴリからなるラティス上の探索により、最適なカテゴリ列 $\hat{\boldsymbol{C}}$ を得る。

$$\hat{\boldsymbol{C}} = \underset{\boldsymbol{C}}{\operatorname{argmax}} \log P(\boldsymbol{C}|\boldsymbol{W}) \tag{2}$$

また、各単語に対応するすべてのカテゴリの事後確率は、ラティス上の前向き後向き計算によって求める。

$$P(c_i = K_j|\boldsymbol{W}) = \frac{\alpha(c_i = K_j)\beta(c_i = K_j)}{Z} \tag{3}$$

ここで c_i は単語列 $\boldsymbol{W}$ 中の単語 w_i に対応するカテゴリ $(i = 1, 2, \cdots, M)$、K_j は第 j のカテゴリを表す $(j = 1, 2, \cdots, N)$。また $\alpha()$ と $\beta()$ はそれぞれ前向き・後向きスコア、Z は正規化項である。

カテゴリ識別に用いる素性関数について述べる。素性関数 $\phi_k(\boldsymbol{W})$（$\boldsymbol{\Phi}$ の要素）は、条件 k にあてはまる単語数を計数して返す。条件 k は $\{f, v, i\}$ の 3 つの要素からなる。この f

表 1 カテゴリ識別に用いる素性関数 $\phi_{\{f,v,i\}}$ の一覧。$i:0$ は対象区間の単語自身を示す。

Feature (f)	Value(v)	Relative Position (i)
(a) Word ID	Integer	$[-7,-2], -1, +1, [+2,+7]$
(b) Confidence score	Real ($[0,1]$)	$[-7,-2], -1, 0, +1, [+2,+7]$
(c) # of syllables	Integer	0
(d) Preceding pause	Binary	0

と v はそれぞれ特徴の種類とインスタンス、i は対象区間との相対的な位置関係を示す。例えば、素性関数 $\phi_{\{\mathrm{word,the},-1\}}(\boldsymbol{W})$ は、直前の区間（$i:-1$）の単語（$f:$ word）の値（v）が “the” であるときにその数 “1” を返す。素性関数の一覧を表 1 に示す。提案手法では、認識結果の各単語の特徴として、単語 ID（Word ID）、認識信頼度（Confidence score）、音節数（# of syllables）、先行無音の有無（Preceding pause）の 4 種類を用いる。

素性関数（a）はカテゴリ識別で発話内の共起単語を扱うために使用する。その際、共起単語は位置関係によって区別する。たとえば、素性関数 $\phi_{\{(a),N,[+2,+7]\}}$ は、2 単語から 7 単語離れて後続する単語のうち、単語 ID が N であるものの数を返す。このように、前後関係や距離を反映した特徴を用いることで、キーワードの出現傾向を学習できる。また、自発的な発話にしばしば観察されるフィラーや言い淀みが存在する場合でも、ローカルな履歴のみを用いるクラス n-gram と比べて、頑健な識別処理を期待できる。

また、音声認識結果に誤りが含まれることを考慮し、音声認識処理の過程で得られるその他の情報も特徴として用いる。認識結果の確からしさを表す特徴を抽出する素性関数として（b）と（c）、また、キーワードの存在しやすさを表す特徴を抽出する素性関数として（d）を用いる。これらの特徴により、第 1 パスの認識結果に誤りに対しても頑健にカテゴリが推定されると期待する。

4. 言語モデルの動的適応

言語モデルの動的適応は、入力発話に対して、前段で得られた区間ごとのカテゴリ推定結果を用いて、言語モデルを適応する。すなわち、発話の区間ごとに推定された各カテゴリの事後確率を用いて、区間ごとの各カテゴリの混合重みを推定し、それを用いて言語モデルを線形補間する。

提案手法では、言語モデルの単語出力確率を次式のように表わす。

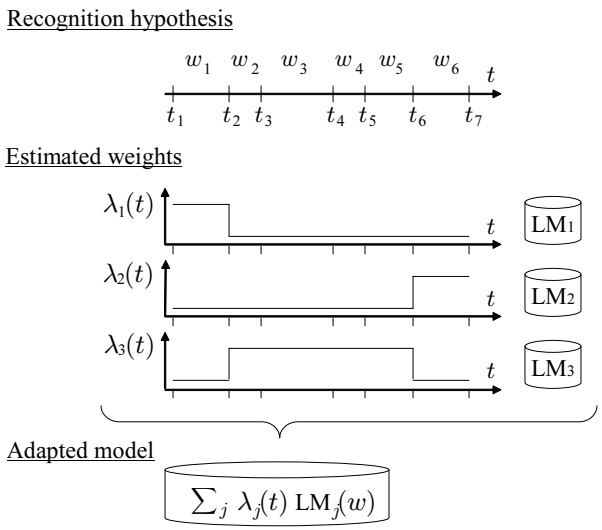

図 2 提案手法の模式図。発話を第 1 パスの音声認識結果の単語列 $w_1, \cdots, w_6$ の境界時刻 $t_1, \cdots, t_7$ で分割する（例: 単語 w_3 は区間 $[t_3, t_4]$ に相当）。カテゴリ識別部は、各区間ですべてのカテゴリ（$\lambda_1, \lambda_2, \lambda_3$）の事後確率を求める。言語モデル適応部はこれらの確率値を各フレームにおける混合重みとして扱う。たとえば、区間 $[t_3,t_4]$ での第 1 カテゴリの重みは $P(c_3 = K_1|\boldsymbol{W})$ として与える。この区間では、第 3 カテゴリの重みが大きいことを示している。

$$p(w|h) = \sum_{j=1}^{N} \lambda_j(t_w) p_j(w|h), \quad \sum_{j=1}^{N} \lambda_j(t_w) = 1, \tag{4}$$

ここで、$p(w|h)$ は履歴 h を条件とする単語 w の出力確率、N はカテゴリの総数、$p_j()$ は第 j カテゴリ用言語モデルの出力確率 ($j = 1 \cdots N$)、$\lambda_j()$ は第 j カテゴリの混合重みである。この混合重み λ_j は、発話内で単語 w が出現した時刻（フレーム）t_w に依存する関数であり、次式のように、式 (3) で求められる各カテゴリの事後確率を用いる。

$$\lambda_j(t) = P(c_i = K_j|\boldsymbol{W}),\ \tau_s(i) \le t \le \tau_e(i), \tag{5}$$

ここで c_i は第 1 パスの認識結果 $\boldsymbol{W}$ の単語 w_i に相当する区間のカテゴリ、$\tau_s(i)$ と $\tau_e(i)$ はそれぞれその区間の始端と終端の時刻である。式（5）において、時刻 t における第 j カテゴリの混合重みは、カテゴリ c_i が K_j であるときの事後確率で与えられる。このようにして、カテゴリ推定結果を言語モデルに反映し、発話内の区間に応じて出現しやすいカテゴリに偏りを与えることができる。以上の手続きの模式図を図 2 に示す。

5. 評価実験

5.1 実験条件

提案手法の評価のため、テレビ番組検索タスクにおいて音声認識実験を行った。評価音声には我々が収集した日本語のテレビ番組検索発話を用いた。一般に募集した話者に自発的な発話を促すようにし、音声入力インタフェースを備える検索システムにより音声を収集した。本実験では、キーワードのカテゴリ 3 種類（人名・番組名・放送局名）と、前記のいずれにも属さない単語を表わす「非キーワード」の計 4 種類のカテゴリを設定した。

収集した話者 210 名の発話のうち、話者 40 名の 5,841 発話を評価データとした。評価データ中のキーワード数は 2,107、カテゴリごとのキーワード数は、人名 : 706、番組名 : 520、放送局名 : 881 であった。また、キーワードを含む発話数は 1,941 であり、そのうち 153 発話が複数のキーワードを含むものであった。発話あたりの平均単語数は 4.23（最少 : 1、最多 : 27）であった。発話例を以下に示す。

- 「＜人＞の出演しているドラマを見たい」
- 「先週の＜番組＞を教えて」
- 「＜放送局＞で夜九時からやっている番組」

カテゴリ識別では、表 1 に示す素性関数を用いた。音声認識結果の信頼度には単語事後確率[6] を用いた。学習データは、収集データから評価データを除いたもののうち、設定した 3 種類のカテゴリのキーワードを含む 8,117 発話とした。学習に用いるカテゴリ列と認識結果単語列の組は、学習データの発話の認識結果単語列について、正解単語列との対応付けをとり、正解単語のカテゴリを付与することで作成した。CRF の学習には CRF++[⋆1] を用いた。

音声認識には我々の開発した大語彙連続音声認識デコーダ[7] を用いた。提案方法を実行するため、音声の分析フレームごとに所定の重みで複数の言語モデルを線形補間する仕組みをデコーダに実装した。フレーム t_w の言語スコアは式（5）により得た。このデコーダは 2 パスのフレーム同期ビームサーチ方式をとっており、第 1 パスの単語終端、第 1 パスの先読み（ファクタリング）、第 2 パスの単語終端の 3 箇所で、提案手法により言語スコアを取得した。

音響モデルは評価データを含まない 600 時間の音声で学習した性別依存の不特定話者トライフォン HMMs（状態数 4,000）とし、40 次元の音響特徴量（12 次元の MFCC・Δ・ΔΔ、対数パワーの Δ・ΔΔ、ピッチとその Δ）をフレームごとに抽出した。

言語モデルは単語トライグラムとし、評価データを含まないテレビ番組検索発話コーパス 27,431 文（221,069 語）を学習データとした。第 1 パスの音声認識で用いる適応なし言語モデルと「非キーワード」カテゴリの言語モデルは、このコーパス全体で学習した。このとき、認識辞書の語彙サイズは 28,504 語、評価データのキーワードの未知語率は 3.6%であった。また、カテゴリごとの言語モデルは、このコーパス中のそのカテゴリの単語を含む文のみを用いて学習した。また、比較対象としてクラス 3-gram を用意した。本実験で設定した 3 種類のキーワードのカテゴリのみをクラスとして扱い、前述のコーパス全体で学習した。

提案手法では、言語モデルの混合重みをカテゴリ識別結果にしたがって定めるが、参考実験として、カテゴリとその区間を既知とした条件（カテゴリ識別誤りがない場合に相当）でも同様の認識実験を行った。混合重みは、正しいカテゴリに 0.9（それ以外のカテゴリに 0.0）、「非キーワード」カテゴリに 0.1 を与えた。また、区間は、発話と正解単語列とのアライメントによって与えた。

評価尺度は、第 2 パスの音声認識結果のキーワード誤り率（KER: Keyword Error Rate）とした。

表 2 表 1 の特徴を用いた CRF によるカテゴリ識別率（%）。3 種類のカテゴリの平均値。全単語（All）と第 1 パスで誤認識された単語（Error）の場合。

Features	All	Error
Word IDs; (a)	49.3	40.2
All features; (a) to (d)	57.2	42.2

5.2 実験結果

カテゴリ識別の正解率を表 2 に示す。表 1 の 4 種類の素性関数を用いたとき、キーワードのカテゴリ 3 種類での平均は 57.2%だった。また、第 1 パスで誤認識されたキーワードについても 42.2% が正しいカテゴリに識別された。また、4 種類の特徴のうち、発話内で共起する単語の ID（a）の寄与が最も大きく、キーワードのカテゴリそれぞれについて、識別に寄与する共起単語のグループが観察された。たとえば、人名に関しては、先行する時間表現や放送局名の単語、後続する接尾辞や動詞、ジャンル表現などであった。

提案手法により動的に適応した言語モデル（Proposed）と、適応なしの言語モデル（Baseline）を用いたときのキーワード誤り率を表 3 に示す。提案手法のキーワード誤り率は他の

⋆1 http://crfpp.sourceforge.net/

表 3 音声認識実験結果のキーワード誤り率 (%)。Baseline は適応なしの言語モデル（第 1 パス）、Class はクラス 3-gram、Proposed は提案法で動的に適応した言語モデル（第 2 パス）、Known は提案法で正しいカテゴリと区間を与えたときの結果。N は発話中のキーワードの単語数。かっこ内はキーワードの挿入誤りの単語数。Non-KW は非キーワードの誤り率。

	Keywords				Non-
	all	$N \geq 2$	$N = 1$	$N = 0$	KW
Baseline	14.8	14.9	12.9	(32)	24.8
Class	13.8	13.9	11.9	(32)	24.6
Proposed	13.1	11.8	10.0	(59)	26.8
known	*12.0*	*11.5*	*9.8*	*(41)*	*24.7*

方法よりも低い値（13.1%）を示した。一方で非キーワード（Non-KW）の単語誤り率は Baseline よりも高かった。この結果は、提案法は検索発話のうち特にキーワードの認識精度を高めることに効果があったことを示す。また、発話中のキーワードの単語数（N）によって分類すると、提案法は、$N \geq 2$ のとき Baseline に対してキーワード誤り率の 20.8%（3.1pts.）の削減を示した。この結果は、検索発話に複数のキーワードが含まれる場合でも、提案法が有効であったことを示す。なお、キーワード誤り率の削減は $N = 1$ のときは 22.5%（2.9 pts.）、キーワードを含む発話全体（$N \geq 1$）では 22.0%（2.9 pts.）の削減を達成した。この認識精度は、カテゴリとその区間を既知とした条件（Known）と同程度であった。

クラス 3-gram（Class）に対しても、提案法は $N \geq 1$ において従来法を上回る結果を示した。本実験で用いたクラス 3-gram は、先行する 2 単語の履歴情報のみからクラスを推定したのに対し、CRF によるカテゴリ識別では、その履歴情報を包含するさらに広いコンテキスト情報に基づいて推定したためと考えられる。第 2 パスの言語モデルの動的適応では、CRF の事後確率を混合重みとして単語 3-gram を線形補間するため、クラス 3-gram よりも良好に働いたと考えられる。

評価データに対する重み推定の一例を図 3 に示す。図に示す通り、CRF によりカテゴリを正しく推定したことにより、第 2 パスの音声認識において、第 1 パスでキーワードを誤った区間であっても、キーワードを正確に認識できたことが確認された。なお、評価データの一部に、キーワード単独の発話や、第 1 パスでほとんど誤認識された発話が存在した。このような場合では、発話内の共起単語の特徴を扱うことが困難であるため、提案手法による効果は小さかった。

検索発話が本実験で設定したキーワードを一つも含まない場合（$N = 0$）、提案手法により、キーワード挿入誤りが増加した。本実験で用いた CRF は、キーワードのない発話でも、

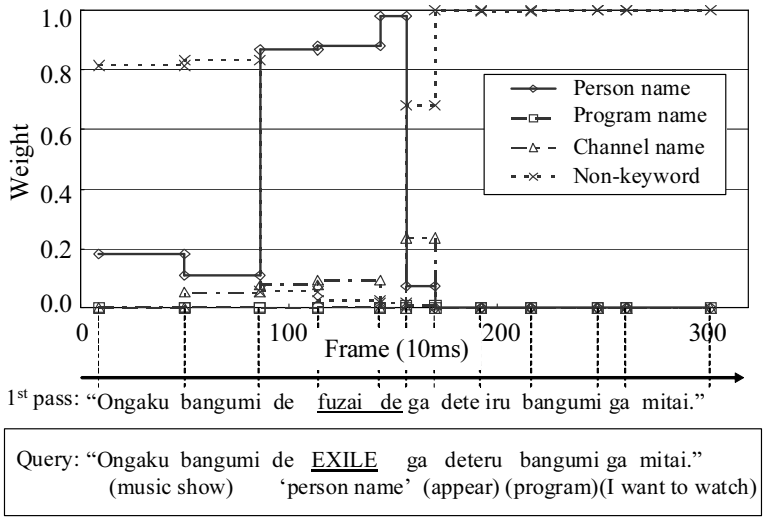

図 3 評価データの重み推定結果の例。キーワード「EXILE」が第 1 パスで「不在で」に誤認識されたが、その区間で「人名」のカテゴリが高く推定された。

キーワードと含む発話と類似した表現があるときに、非キーワードの区間をキーワードのカテゴリとして推定したと考えられる。カテゴリの誤推定がない条件（Known）では、そのような挿入誤りの増加は比較的少なかった。このことから、カテゴリ識別精度を高めることによって、キーワード認識率をさらに向上できると期待される。また、キーワードを含まない発話を検出して再認識しないようにしたり、認識結果から信頼度の低い単語を除いたりすることにより、このような挿入誤りを低減する方法も考えられる。

6. ま と め

本稿では、検索発話のキーワードの認識精度を高める方法として、カテゴリ識別に基づく言語モデルの動的適応手法を提案した。提案方法は、検索発話の音声認識結果に対して CRF によるカテゴリ識別処理を行って求めた事後確率を重みとして、発話の区間ごとにその重みを変えながら、カテゴリごとに構築された言語モデルを混合することにより、検索発話に現れるキーワードに動的に言語モデルを適応する。提案手法の効果をテレビ番組検索タスクにおける音声認識実験によって評価したところ、キーワードを含む発話に関して、適応

Vol.2010-SLP-84 No.1
2010/12/20

なしの言語モデルを用いた場合と比べて、提案法による 22.0% のキーワード誤り率の削減を確認した。

今後の課題として、誤認識を含む区間でのカテゴリ識別の強化や、言語モデルの混合重みの最適化方法の検討、情報検索のタスク達成率の評価などが挙げられる。

参 考 文 献

1) R. M. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures versus dynamic cache models," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 1, pp. 30–39, 1999.
2) D.Gildea and T.Hofmann, "Topic-based language modeling using EM," *Proc. of Eurospeech*, pp. 2167–2170, 1999.
3) P.F.Brown *et al.*, "Class-based n-gram models of natural language," *Computer Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.
4) B-J. Hsu, "Generalized linear interpolation of language models," *Proc. of ASRU*, pp. 136–140, 2007.
5) J.Lafferty *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proc. of ICML*, pp. 288–298, 2001.
6) F. Wessel *et al.*, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 288–298, 2001.
7) R. Isotani *et al.*, "An automatic speech translation system on PDAs for travel conversation," *Proc. of ICMI*, pp. 211–216, 2002.