Vol.2011-SLP-89 No.1
2011/12/19

予稿の話し言葉変換に基づく言語モデルによる講演音声認識

渡邉真人[†1]　秋田祐哉[†1]　河原達也[†1]

講演のような話し言葉の音声認識では，言語モデルがドメインに関連する表現とフィラーや口語表現などの話し言葉特有の表現の両方をカバーすることが求められる．本研究では，単語・構文などの情報に基づくルールベースの話し言葉テキスト変換と，N-gram の統計的話し言葉変換を組み合わせて，書き言葉スタイルの予稿テキストから話し言葉スタイルの言語モデルを構築する手法を提案する．学会講演音声を対象とした評価実験において，提案手法の効果の評価を行った．

Automatic Transcription of Lecture Speech using Language Model based on Speaking-Style Transformation of Proceeding Texts

Makoto Watanabe,[†1] Yuya Akita[†1]
and Tatsuya Kawahara[†1]

For automatic speech recognition of spontaneous lecture speech, language models need to cover spoken-style expressions such as fillers and colloquial expressions, as well as domain-dependent topic words. We propose an approach to make a spoken-style language model from written-style texts by combining two transformation methods: a rule-based text transformation using lexical and syntactic information, and statistical transformation of N-gram entries. Experiments over academic presentations were conducted to evaluate the proposed approach.

†1 京都大学 情報学研究科
Graduate School of Informatics, Kyoto University

1. はじめに

近年，TED[⋆1] のような講演アーカイブ，またオープンコースウェア[⋆2]のような講義アーカイブの公開が広がってきている．講演・講義は長時間におよぶことから，効率的に視聴するには書き起こしなどを用いてインデックスを付与し，また検索を実現することが重要である．一方，聴覚障害者や高齢者，非ネイティブの視聴者のためには字幕を付与することが望ましいが，人手による字幕付与は人的，時間的コストがかかるうえ，全ての内容をリアルタイムに人手で書き起こすことは不可能である．そこで，これらに音声認識を活用する研究が進められている．[2)3)]

講演や講義の音声の特徴として，フィラーや口語表現など話し言葉特有の表現に加えて，専門用語のように分野（ドメイン）に強く依存する単語が出現することが挙げられる．したがって，講演や講義を対象とした音声認識システムを実現するには，ドメインに関連する表現と話し言葉特有の表現の両方を言語モデルがカバーすることが求められる．しかし，そのような２つの特徴を持ち合わせた学習コーパスをドメインごとに準備するのは容易でない．ドメインに関連する表現を含んだコーパスとしては，講演であれば予稿集や論文集，講義であれば教科書が考えられるが，これらは文体が書き言葉であり，フィラーなどは含まれない．そのため，一般的にはドメインに関連するコーパスと何らかの話し言葉コーパスを混合して言語モデルを作成することが行われる．たとえば，会議や講演の音声認識[4)5)] では，ニュースステキストや Web テキストに電話会話音声の書き起こし（Switchboard・Fisher コーパス）を組み合わせて言語モデルを構築している．また，講義の音声認識[6)7)] では，講義スライドや Web テキストと電話会話音声の書き起こし（Switchboard コーパス）を用いて言語モデルを作成している．このようなコーパス混合により一定の音声認識精度は得られるが，ドメインに関連のない表現が言語モデルに含まれたり，単語間の接続関係に不整合が発生したりするなどの問題が起きうる．

これに対して，我々は言語モデルの話し言葉へのスタイル変換[8)] というアプローチを提案している．本手法は，統計的な変換モデルを用いて話し言葉の N-gram の生成と統計量の推定をインドメインのコーパスのみから行うため，コーパス混合のような問題は発生しない．本手法は国会の会議録テキストから，国会審議の言語モデルを構築するのに適用されて

⋆1 www.ted.com
これを用いた研究[1)] も行われている．
⋆2 ocw.kyoto-u.ac.jp

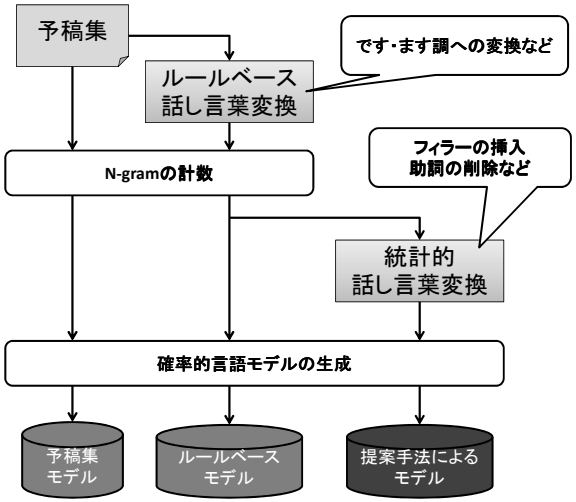

図 1 言語モデル構築の流れ
Fig. 1 Flow of language model training

いる[9]．ただし，会議録は話し言葉を書き起こしたものであり，予稿のような完全な書き言葉テキストと話し言葉の書き起こしを対応付けて変換モデルを学習することは容易でない．そこで本稿では，完全な書き言葉コーパスから話し言葉スタイルの言語モデルを構築するために，言語モデルの統計的話し言葉変換に加えて，書き言葉テキストに対するルールベースの話し言葉変換[10] を用いる手法を提案する．

ルールベースの話し言葉変換では，あらかじめ定めた変換規則に従って，書き言葉表現から話し言葉表現への置換や，文末のです・ます調への変換などを行う．一方，統計的枠組みを利用した話し言葉変換では，あらかじめ学習した変換パターンとその確率に従って，フィラーの挿入，助詞の削除，口語表現の置換などを行う．本稿では予稿がある講演の音声認識に適用し評価を行う．

2. 話し言葉スタイル変換に基づく言語モデル

2.1 言語モデル構築のあらまし

本研究では**図 1** の手順で言語モデルを作成する．書き言葉コーパスとしては，講演と関連の深い学会の過去の予稿集や論文集を想定する．まず前処理として，英語論文の削除，表記の揺らぎの統一，参考文献記述箇所の除去などの整形を行う．次にルールベース話し言葉変換を適用し，予稿集テキストの文体の変換を行う．その出力に対して，統計的話し言葉変換を適用してフィラーなどの話し言葉表現を含む N-gram 統計量を推定する．最後に N-gram 統計量から言語モデルの確率を計算する．なお中間的な言語モデルとして，話し言葉変換を適用せずに作ったもの（予稿集モデル），ルールベース話し言葉変換のみ適用したもの（ルールベースモデル）も作成可能である．

2.2 ルールベース話し言葉変換

ルールベースの話し言葉変換手法[10] は，人手で用意した変換規則に基づいて書き言葉テキストから話し言葉テキストへの変換を行う．もともと自然な音声合成のために開発されたものであるが，本研究では言語モデル構築に適用する．書き言葉テキストから，自然でわかりやすい話し言葉表現を生成するという観点から，文献 10) の変換システムでは以下に分類された書き言葉表現を対象としている．

- **書き言葉特有の表現**
  普通体：聞き手を想定しない脱待遇の表現
  （例）吾輩は猫 である．
  文語調：堅苦しく古めかしい表現
  （例）本日 午前十時 より 会議室 にて ミーティングを行う．
  漢語調：難解な漢字を使った表現
  （例）酸欠 のために，養鰻場 で 曝気 をする．

- **複雑な構造を持つ表現**
  名詞化された用言を含む表現：主にサ変名詞や形容詞・副詞の語幹などを含む複合語からなり，格関係が明示されていない表現
  （例）車両点検後，始業開始前点呼 を行う．

本研究では以上の書き言葉表現のうち，変換精度の比較的高かった普通体と文語調の表現を対象として話し言葉変換を適用する．書き言葉特有の表現の中で，普通体と文語調の表現は主に機能的表現であり，テキスト中に出現する頻度は高いが，その種類は多くない．そのため，それぞれの表現に対して口語調表現への変換規則を作成することで対応することがで

【です・ます調への変換】
走らない。 ⇒ 走りません。
ねじ曲げたのだ。⇒ ねじ曲げたのです。

【文語表現の変換】
・・・のみならず ⇒ ・・・だけではなく
・・・せざるを得なかった ⇒ ・・・しなくてはいけなかった

図 2 ルールベース話し言葉変換の例
Fig. 2 Examples of rule-based style-transformation

きる．

普通体は相手の存在を前提としない脱待遇の表現であり，話し言葉で用いられることは少ない．そこで，公の場での発話に一般的に用いられる丁寧体（です・ます調と呼ばれる表現）を用いることで待遇表現を補う．文語調の古めかしい言い回しは，その表現と意味的に対応する日常的に用いられる表現へと言い換える．これらの変換例を**図 2** に示す．これらの変換規則は，変換箇所の前後の文脈の形態素情報，特に品詞や活用形，係り受けなどの情報を参照しているため，様々な単語の組み合わせに対しても柔軟に適用することができる．

2.3 統計的話し言葉変換

自然発話にはフィラーや発話の怠けなどが存在するため，話し言葉の特性をモデル化するには，ルールベース話し言葉変換による文体の変換だけでは不十分である．そこで，統計的機械翻訳（Statistical Machine Translation; SMT）の枠組みに基づく話し言葉変換[8] を導入する．

SMT では，翻訳元の言語における文 X と翻訳先の言語の文 Y について，事後確率 $P(Y|X)$ が最大となる Y を翻訳文として出力する．ベイズ則に基づき $P(Y|X)$ は (1) 式で与えられる．

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

ここで，$P(X)$ と $P(Y)$ はそれぞれ，翻訳元・翻訳先の言語の言語モデルである．$P(X|Y)$ は翻訳モデルと呼ばれ，両言語の対応関係を規定している．実際の SMT では，$P(X)$ は Y の選択に寄与しないことから無視される．

SMT に基づく統計的話し言葉変換では，書き言葉スタイル (X) と話し言葉スタイル (Y) をそれぞれ別の言語としてとらえ，書き言葉スタイルの言語モデル $P(X)$ から話し言葉スタイルの言語モデル $P(Y)$ を生成する．変換の概念図を**図 3** に示す．(1) 式より，$P(Y)$ は (2) 式で表される．

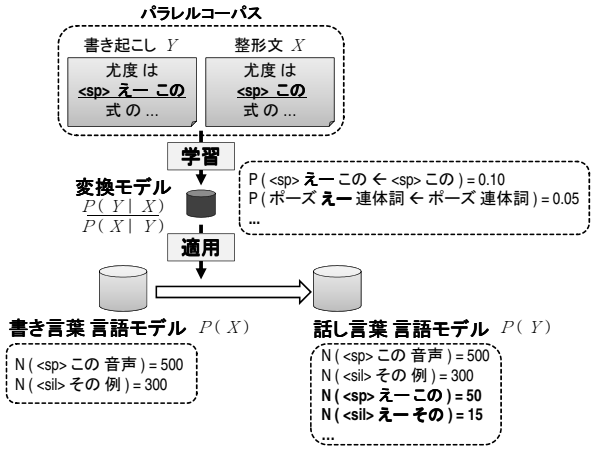

図 3 統計的な話し言葉変換の流れ
Fig. 3 Flow of statistical speaking-style transformation of language model

$$P(Y) = P(X)\frac{P(Y|X)}{P(X|Y)} \tag{2}$$

ここで条件付き確率 $P(Y|X)$・$P(X|Y)$ すなわち変換モデルは，音声の書き起こしと，それを書き言葉スタイルに編集したテキストによるパラレルコーパスを用いて推定される．

本研究では，まず CSJ「日本語話し言葉コーパス」のコア（177 講演）の書き起こしとそれに対応する整形文をパラレルコーパスとして用意し，その差異から変換モデルを推定した．そして，これをルールベース話し言葉変換の出力に対して適用した．変換モデルには，少量の学習データからでも多様な単語の接続関係に柔軟に対応できるようにするため，単語の一致だけでなく，品詞の一致による変換も含まれている．この変換によって，主にフィラーの挿入，助詞の脱落，口語表現の置換などが行われる．

3. 評 価 実 験

3.1 評価タスクと言語モデル

提案手法の評価を講演音声の認識タスクで行った．テストセットは，音声ドキュメント処

理ワークショップの講演（2007 年〜2009 年）から選択した 10 講演である．主要なトピックは言語モデル・検索・機械翻訳などの言語処理で，1 講演あたり平均 20 分，単語数は平均 4.4K である．

評価で用いる言語モデルの仕様とテストセットにおけるパープレキシティ（PP）・未知語率（OOV）を**表 1** に示す．言語モデルの構築に利用した形態素解析器は茶筌 Ver2.4.4+UNIDIC-1.3.9 である．書き言葉コーパスとして，言語処理学会年次大会の 2004 年から 2009 年までの 5 年分の予稿集を用意し，話し言葉変換を適用しないモデル（NLP），ルールベース話し言葉変換のみ適用したモデル（NLP-rule），統計的変換まで適用したモデル（NLP-stat）の 3 つを作成した．比較対象として CSJ の模擬講演（1715 講演）の書き起こしを用いて作成した trigram 言語モデル（CSJ_E）を用意した．CSJ_E と NLP のモデル学習に用いた整形済み学習テキストの総単語数はそれぞれ 4.1M, 2.7M である．さらに CSJ_E に NLP の 3 種のモデルをそれぞれ混合したモデル（CSJ_E+NLP，CSJ_E+NLP-rule，CSJ_E+NLP-stat）も用意した．混合比は事前実験により 0.5:0.5 と定めた．

表 1 から，NLP-stat では，ドメインに関連する単語と話し言葉表現の両方をカバーすることができたため，CSJ_E より低い未知語率および NLP より低いパープレキシティを実現できている．ルールベース話し言葉変換は話し言葉の N-gram を十分に補うものではないため，パープレキシティの削減は大きくないが，これに続いて統計的話し言葉変換を行うことで，パープレキシティが大きく削減できている．また CSJ_E と NLP の混合モデルでは，それぞれが単独では補えなかった語彙が追加されたため，さらに未知語率が減少した．ただし，NLP 単独の場合に比べて語彙サイズが 2 倍以上になっている．

表 1 のモデルでは，当該の講演の予稿は用いられていない．一方，実際の場面では当該講演の予稿を事前に入手して言語モデルに反映させることも考えられるため，テストセットの講演の予稿を混合した言語モデルも講演ごとに作成した（**表 2**）．予稿 1 つあたりの平均単語数は 4.1K である．preprint-rule と preprint-stat は NLP の場合と同様に，講演予稿にそれぞれルールベース話し言葉変換のみ適用したモデル，それに加えて統計的話し言葉変換も適用したモデルである．事前実験により，NLP のモデルと予稿によるモデルの混合比は 0.9:0.1，CSJ_E と予稿によるモデルの混合比は 0.7:0.3，CSJ_E と NLP のモデルと予稿によるモデルの混合比は 0.45:0.45:0.1 とした．全般的に，予稿がない場合に比べてより発話内容をカバーできたため，パープレキシティや未知語率は低くなっている．

3.2 音声認識による評価

以上の言語モデルを用いて，テストセットに対して音声認識を行い単語正解精度を求め

表 1 言語モデル（予稿なし）の仕様
Table 1 Specifications of language models

言語モデル	1-gram エントリ数	2-gram エントリ数	3-gram エントリ数	PP	OOV
NLP	9.94K	124K	283K	245	2.56%
NLP-rule	9.97K	124K	280K	227	2.55%
NLP-stat	10.0K	304K	333K	95.5	2.14%
CSJ_E	19.9K	190K	335K	210	4.49%
CSJ_E+NLP	24.0K	282K	592K	109	1.09%
CSJ_E+NLP-rule	24.0K	282K	589K	108	1.08%
CSJ_E+NLP-stat	24.0K	439K	630K	92.7	1.08%

表 2 言語モデル（予稿あり）の仕様
Table 2 Specifications of language models using preprint of the lecture

言語モデル	1-gram エントリ数	2-gram エントリ数	3-gram エントリ数	PP	OOV
NLP+preprint	9.99K	124K	284K	184	1.99%
NLP-rule+preprint-rule	10.0K	124K	282K	168	1.98%
NLP-stat+preprint-stat	10.1K	305K	358K	72.0	1.57%
CSJ_E+preprint	20.0K	191K	338K	78.8	1.44%
CSJ_E+preprint-rule	20.0K	191K	338K	78.4	1.45%
CSJ_E+preprint-stat	20.0K	196K	368K	74.7	1.45%
CSJ_E+NLP+preprint	24.0K	283K	594K	75.5	0.75%
CSJ_E+NLP-rule+preprint-rule	24.0K	282K	590K	74.1	0.75%
CSJ_E+NLP-stat+preprint-stat	24.0K	440K	655K	65.0	0.75%

た．予稿を用いない場合の 10 講演の平均の精度を**図 4**，用いた場合の精度を**図 5** に示す．なお，音声認識における音響モデルは，CSJ の学会講演（257 時間）による，3000 状態・16 混合の triphone HMM である．CMN, CVN, MPE, VTLN を適用し，MLLR による教師なし話者適応を行っている．特徴量には，MFCC およびその 1 次・2 次差分各 12 次元とエネルギーの 1 次・2 次差分の計 38 次元を用いた．デコーダは Julius rev.4.1.5 である．

予稿を用いない場合は，CSJ_E のみによる認識精度は 73.50%であり，NLP は 73.89%，NLP-rule は 74.13%で改善は小さい．これに対して，NLP-stat は 82.64%と高い精度を実現した．CSJ_E と NLP のモデルの混合モデルではさらに高い精度となり，CSJ_E+NLP で 83.92%，CSJ_E+NLP-rule で 84.01%，CSJ_E+NLP-stat で 84.26%で，CSJ モデルと提案法によるモデルの混合モデルが最も高い精度を得た．話し言葉変換の効果は小さくなったが，これは CSJ_E との混合によって，CSJ_E に含まれる話し言葉表現と NLP に含まれるドメインに関連する単語の両方がカバーされたためであると考えられる．

一方予稿を用いた場合，CSJ_E+preprint は 6.18%，NLP+preprint は 1.32%の精度の改善が見られた．NLP+preprint の改善幅が小さいのは，予稿に含まれるドメインに関連する表現がすでに言語処理学会年次大会の論文集でカバーされていたためである．全て混合した CSJ_E+NLP+preprint では，85.56%とさらに精度が向上した．次に，preprint-rule を混合した場合は，preprint の混合と比べて精度の向上はほとんどみられず，preprint-stat でも同様であった．これは，予稿のサイズが小さすぎるため話し言葉変換で十分な統計量が得られなかったこと，CSJ_E や NLP-stat にはすでに話し言葉表現が含まれていることが理由と考えられる．

4. おわりに

本稿では，ルールベース話し言葉変換と統計的話し言葉変換を用いて，書き言葉スタイルのテキストから話し言葉スタイルの言語モデルを作成する手法を提案した．ルールベース変換により，書き言葉テキストの文末表現や文語表現などを言い換える．これに統計的変換を適用することで，フィラーなどを含む話し言葉の言語モデルを構築する．講演音声の音声認識において評価を行い，話し言葉のコーパスを用いなくてもインドメインの語彙のみで高い認識精度が得られることが示された．

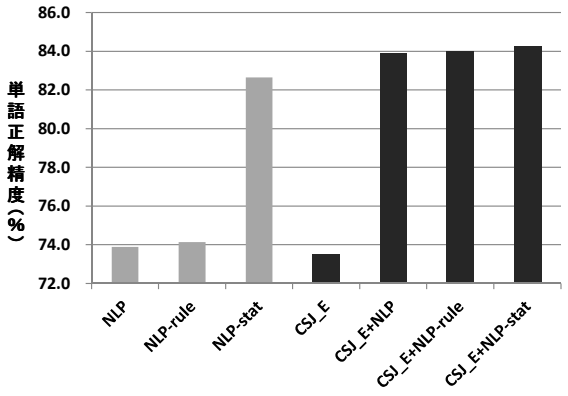

図 4 単語正解精度（予稿なし）
Fig. 4 Word accuracy (LM without preprint of the lecture)

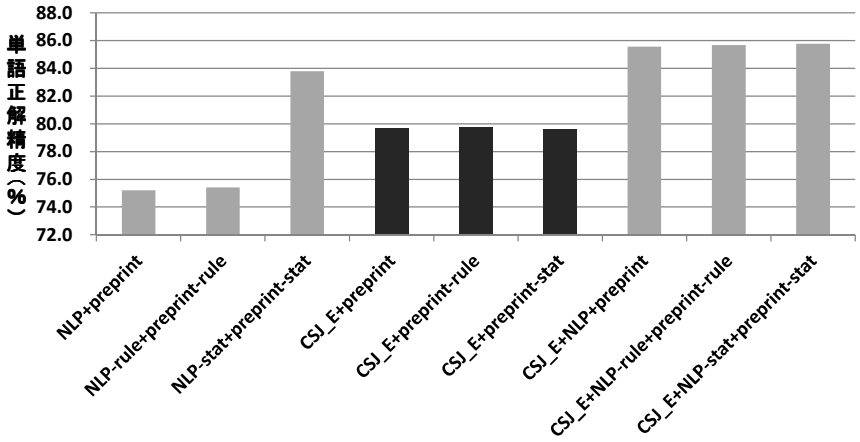

図 5 単語正解精度（予稿あり）
Fig. 5 Word accuracy (LM using preprint of the lecture)

Vol.2011-SLP-89 No.1
2011/12/19

謝　　辞

本研究を進めるにあたり，ルールベース話し言葉変換のプログラムをご提供いただいた京都大学情報学研究科の黒橋禎夫教授，柴田知秀助教に深く感謝いたします.

参 考 文 献

1) J. Lopes, I. Trancoso, and A. Abad. A Nativeness Classifier for TED Talks. In *Proc. ICASSP*, pp. 5672-5675, 2011.
2) 松井淳, 本間真一, 小早川健, 尾上和穂, 佐藤庄衛, 今井亨, 安藤彰男. 言い換えを利用したリスピーク方式によるスポーツ中継のリアルタイム字幕制作. 電子情報通信学会論文誌, Vol. 87, No. 2, pp. 427-435, 2004.
3) 勝丸徳浩, 河原達也, 秋田祐哉, 森信介, 山田篤. 講義音声認識に基づくノートテイクシステム. 電子情報通信学会技術研究報告, SP2009-53, pp. 25-30, 2009.
4) C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan. An Audio Indexing System for Election Video Material. In *Proc. ICASSP*, pp. 4873-4876, 2009.
5) S. Kombrink, T. Mikolov, M. Karafiat, and L. Burget. Recurrent Neural Network based Language Modeling in Meeting Recognition. In *Proc. INTERSPEECH*, pp. 2877-2880, 2011.
6) J. Glass, T .Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. INTERSPEECH*, pp. 2553-2556, 2007.
7) C. Munteanu, G. Penn, and R. Baecker. Web-Based Language Modelling for Automatic Lecture Transcription. In *Proc. INTERSPEECH*, pp. 2353-2356, 2007.
8) Y. Akita and T. Kawahara. Statistical Transformation of Language and Pronunciation Models for Spontaneous Speech Recognition. In *IEEE Trans. Audio, Speech and Language Process*, Vol. 18, No. 6, pp. 1539-1549, 2010.
9) 秋田祐哉, 三村正人, 河原達也. 会議録作成支援のための国会審議の音声認識システム. 電子情報通信学会論文誌, Vol. J93-D, No. 9, pp. 1736-1744, 2010.
10) 黒橋禎夫, 大泉敏貴, 柴田知秀, 鍛治伸裕, 河原大輔, 岡本雅史, 西田豊明. 会話型知識プロセスのための言語情報のメディア変換. 社会技術研究論文集, Vol. 2, pp. 173-180, 2004.