1J3-OS-22a-5

話し言葉の書き起こし文章の話者性の変換

Transforming Individuality of Transcriptions of Speech

水上 雅博 Masahiro MIZUKAMI　Graham NEUBIG　Sakriani SAKTI　戸田 智基 Tomoki TODA　中村 哲 Satoshi NAKAMURA

奈良先端科学技術大学院大学 情報科学研究科
Nara Institute of Science and Technology, Graduate School of Information Science

In text and speech, there are various features that reflect the individuality of the writer or speaker. These features play an important role in ensuring smooth communication. Our research aims to create a method for text translation from the texts with one writer's individuality to the texts with another writer's individuality. At first, we introduce related research about translation from spoken language to written language and about authorship attribution using writers language models. Next, we describe a text translation system that translates text with one writer's individuality to another writer's individuality by using a language model. Experimental results demonstrate that the proposed methods effectively translate from the individuality to another individuality.

1. まえがき

言語情報のみでなく，誰が話しているかといった話者性などの非言語情報も伝達することは，円滑な意思疎通を行う上で有効である．例を挙げれば，敬体よりも常体を用いることで，親しみやすい雰囲気を伝えることができ，ユーザが子供であれば柔らかい話し方を，大人であれば硬い話し方を用いることで，より円滑な対話を実現できると考えられる．そのため，対話システムにおいて，ユーザや場面に応じて，対話エージェントの豊かな話者性と雰囲気を表現することは重要であると考えられ，対話システムの応答を適切に制御する技術の構築が望まれる．

対話システムの応答生成に関連する技術の一つである音声合成においては，非言語情報に着目した研究が盛んに行われている．特に，話者性に関しては，声質の変換 [Abe 88] や個人性を考慮した音声合成 [Yamagishi 10] の研究が進んでおり，その発展として，話者の声質を考慮した音声翻訳システム [Qian 13] なども実現されている．その一方で，話し言葉の書き起こし文章の持つ話者ごとの特徴（以下，話し言葉における話者性と呼ぶ）までも制御する研究は多くなされていない．類似の研究として性格や礼儀正しさなどの心理学的要因を考慮し，話者性を考慮したルールベースの文生成を行う研究がある [Mairesse 11]．対話システムにおいて，統一的な話し方をする応答を生成するよりも，音声合成で得られる声質と合致する話者性を話し言葉に与えたほうが，より自然である．

本稿では，話し言葉における話者性を自由に変換する技術の実現を目指し，翻訳辞書と言語モデルを用いた話し言葉の話者性変換手法を提案する．提案法は，変換対象の話者によらず，変換先の目標の話者性を持った言語モデルのみで変換可能で，話者の特徴推定を必要としない．変換は雑音のある通信路モデルに基づいており，自動で構築された置き換え可能な機能語を集めた翻訳辞書と，変換先の目標となる話者の話者性を捉えた言語モデルを事前に用意する．変換時には，与えられた文章に対して翻訳辞書による機能語の置き換えを行い，置き換えを行った候補の中で言語モデル確率が最大となる文章を選択する．客観的および主観的な評価結果から，提案手法の有効性を示す．

連絡先: 水上雅博, 奈良先端科学技術大学院大学 情報科学研究科 知能コミュニケーション研究室,masahiro-mi@is.naist.jp

2. 言語モデルに基づく話者判別

本研究の最終目標は話者性の変換であるが，それを実現するためには話者性を定量的に評価可能な手法を確立する必要がある．話者性を定量的に評価することで，その評価を指標とした話者性の変換が可能となる．これに関連する先行研究として，著者判別の研究がある．言語モデルに基づく著者判別法 [Juola 05] では，著者ごとの単語の使用傾向や言い回しを話者性とみなし，著者が既知の文章からそれらの特徴をモデル化することで，与えられた文章の著者を識別する．学習時には，著者 S によって書かれた文章に基づいて単語系列 W の発生確率 $P(W|S)$ を表す言語モデルを学習する．識別時には，識別する文章の生成確率 $P(W|S)$ が最大となる著者 $\hat{S}$ を当該著者とする．

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, P(W|S) \tag{1}$$

また，単語数が異なる文章間での比較を行う際には，単語数で正規化された尺度として，式 2 に示すエントロピーを使用することができる．

$$H(W|S) = \frac{-\log_2 P(W|S)}{|W|} \tag{2}$$

この言語モデルに基づく著者判別法は，オランダ語において有効性が示されている [Juola 05]．ただし，言語モデルの構築法は言語に依存せず利用可能であるため，日本語の話し言葉における話者判定においても同様の有効性が期待される．

3. 言語モデルと翻訳辞書を用いた話者性変換

話者性変換のように，同言語間において特定の性質を変換する先行研究として，話し言葉から書き言葉への変換があげられる [Neubig 12]．この話し言葉の変換では，話し言葉から書き言葉への変換を統計的機械翻訳の問題とみなす．話し言葉の書き起こしテキスト V と整形された書き言葉のテキスト W を異なる言語とみなし，V から W への統計的機械翻訳を雑音のある通信路モデルを用いて行う．利用可能なコーパスの分量を考慮し，大量に確保することが難しい対訳コーパスを用い

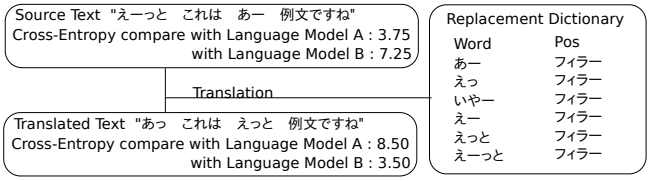

図 1: 個人性変換手法のイメージ図

る翻訳モデル確率 $P(V|W)$ と大量に確保可能な出力側コーパスを用いる言語モデル確率 $P(W)$ の二つに分解し，事後確率 $P(W|V)$ を以下のようにモデル化する．

$$P(W|V) = \frac{P(V|W)P(W)}{P(V)} \tag{3}$$

与えられた V に対して $P(W|V)$ が最大となる $\hat{W}$ を探索する．$P(V)$ は W の選択によらず変動しないため，以下のように表せる．

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(V|W)P(W) \tag{4}$$

本研究ではこの雑音のある通信路モデルを話者性変換へ適用する．話者性変換を，変換の対象となる話し言葉のテキスト V と変換先の目標となる話者性を持つテキスト W の二つの異なる言語間での翻訳とみなす．この時，2 節で紹介した言語モデル確率 $P(W)$ はそのまま話者性変換に適用可能である．一方で，翻訳モデル確率 $P(V|W)$ は変換の対象となる話し言葉のテキスト V と変換先の目標となる話者性を持つテキスト W の対訳コーパスを用いた学習が必要となる．しかしながら，同じ意味を持ちながら異なる個人性を持った対訳コーパスを大量に収集することは困難であり，翻訳モデルの学習は困難である．

これを解決するために，本研究では翻訳辞書による翻訳モデルを構築する．シソーラスや辞書を用いた言い換えを行う手法は提案されているが，同義語や類義語であっても意味や用法に何らかの差がある場合がほとんどで，無条件で置換できる語のペアは必ずしも多くない [乾 04]．そこで，本研究では変換対象を話し言葉としている点を利用し，文章の意味を保持しながら交換可能な機能語であるフィラーと感動詞を抽出し，翻訳辞書を構築する．翻訳辞書に登録された機能語の翻訳確率は，語によらず一様であると仮定することで，翻訳モデル確率 $P(V|W)$ を定義する．結果，事後確率 $P(W|V)$ を最大化する $\hat{W}$ の探索処理は，翻訳辞書の置き換えパターンで生成されるテキストの中から $P(W|S)$ を最大化する，すなわち，クロスエントロピー $H(W|S)$ を最小とする $\hat{W}$ を選択する処理として表される．変換のイメージを図 1 に示す．

4. 評価的実験

4.1 実験条件

本研究の題材として二つの日本語話し言葉コーパスを用いる．

カメラ販売対話コーパス 3 名の店員と 19 名の客の 1 対 1 でのカメラ販売に関する対話をまとめたコーパスである [平岡 13]．店員 3 名のコーパスをそれぞれ話者ごとに 2 つに分け，一方を学習データ，もう一方を評価データとする．

医療対話コーパス 2 名の話者が医者役，患者役に分かれ医療行為を目的として対話を行うコーパスである．それぞれ話者と役割の 4 種に分割し，さらにそれぞれを 2 つに分け，各種の一方を学習データ，もう一方を評価データとする．各コーパスの緒元を表 1 および 2 に示す．

表 1: カメラ販売対話コーパスの緒元

用途	話者	対話数	発話数	単語数
学習	店員 A	5	171	931
	店員 B	6	162	1068
	店員 C	6	155	844
評価	店員 A	5	128	879
	店員 B	6	153	922
	店員 C	6	156	843

表 2: 医療対話コーパスの緒元

用途	話者	対話数	発話数	単語数
学習	医者 A	1	11	78
	医者 B	1	16	82
	患者 A	1	14	33
	患者 B	1	10	41
評価	医者 A	1	11	71
	医者 B	1	14	92
	患者 A	1	14	42
	患者 B	1	10	35

日本語話し言葉におけるクロスエントロピーに基づく著者判別法の評価を行うために，上記のコーパスに対して著者判別実験を行う．形態素解析を MeCab[Kudo 04] で行い，言語モデルの学習，評価を京都言語モデルツールキット（Kylm）[Neubig] で行う．各コーパスごとに語彙を揃え，平滑化に Kneser-Ney 法を利用し，2-gram の言語モデルを構築する *1．これらの学習データと評価データに対して言語モデルに基づく著者判別法を適用し，著者を判別し，クロスエントロピーを示す．

次に，提案法を人手で評価する．カメラ販売対話コーパスを用い，対話中の一部を 5〜10 発話程度で，どのような話題，どのようなタスクについて話していたかわかるように切り出し，その中からある店員の発話一文を削除する．削除された発話と同様の意味を持つ他の店員の発話を探し，変換の対象とする．変換対象の発話に対して言語モデルと翻訳辞書を用いた話者性の変換を行う．翻訳辞書は MeCab での形態素解析結果に基づいてフィラーと感動語を抽出し，自動構築する．話者性の変換の際は話者判別と同様の言語モデルを利用する．また，同発話に対してそれぞれ翻訳辞書を用いてランダムに置き換えた発話を 4 つ作成する．被験者には削除部分に対して，これら 5 つの発話のどれが最も適切といえるかを質問し，さらに選択した発話がその対話の流れに対して適切か，対話中の話者と同一の話者性を持っているかの 2 項目について 1 から 5 の 5 段階で評価をし，提案法が選択される割合と話者性をもって評価値とする．被験者は 5 名で，それぞれ 12 対話に対して評価を行う．

また，5 つの発話がそれぞれどのような割合で個人性の持つか評価するため，式 (4) の事後確率を利用する．事後確率は式 (4) に対して周辺化を行うことで式 (5) として容易に求められ

*1 2-gram を利用した理由は，医療対話コーパスを話者，役割で分割した場合のコーパスサイズが非常に小さいため，より長い文脈を使ったモデルの利用に効果がないためである．

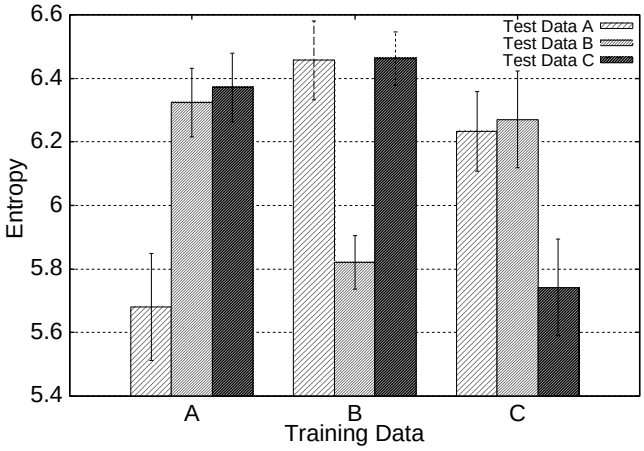

図 2: カメラ販売対話コーパスにおける各話者のクロスエントロピー評価

表 3: 医療対話におけるクロスエントロピー評価

		学習データ			
評価データ		A	A	B	B
話者	役割	医者	患者	医者	患者
A	医者	3.03	3.31	4.32	3.33
A	患者	4.48	3.31	5.37	3.93
B	医者	3.95	3.03	3.03	3.26
B	患者	4.59	3.93	4.68	3.00

る．なお，$\mathcal{W}$ は選択肢として選ばれた 5 つの発話を示す．

$$P(W|V) = \frac{P(V|W)P(W)}{\sum_{\tilde{W}\in\mathcal{W}} P(V|\tilde{W})P(\tilde{W})} \quad (5)$$

4.2 実験結果

4.2.1 言語モデルに基づく話者判別法の評価

まず，カメラ販売対話コーパスに対して言語モデルに基づく話者判別法を適用する．学習データと評価データのそれぞれの組み合わせにおけるクロスエントロピーを図 2 に示す．学習データと評価データが同一の話者のものであればクロスエントロピーは異なる場合に比べ明らかに小さくなることが確認できる．

次に，医療対話コーパスに対してクロスエントロピーに基づく話者判別法を適用する．表 3 にそれぞれの話者，役割を与えた評価データと学習データの組み合わせにおけるクロスエントロピーを，表 4 にそれぞれの話者と役割の組み合わせにおける平均クロスエントロピーを示す．データが小規模にも関わらずこちらもクロスエントロピーは同一話者かつ同一役割において最も小さい値を示し，続いて同一話者，相違役割の場合，相違話者，同一役割の場合，相違話者，相違役割の場合と続く．これらのことから言語モデルに基づく著者判別法は日本語話し言葉の書き起こしテキストに対しても十分に有効であると評価できる．

また，医療対話においては役割によって話者間のクロスエントロピーがわずかではあるが変化する傾向がみられた．医者，患者役において話者を変えた際のクロスエントロピーの差をまとめたものを表 5 に示す．同一話者と相違話者のクロスエントロピーを比較する際，医者役を与えた場合より患者役を与えたほうがクロスエントロピー差がわずかに小さいことがわかる．これらのことから，言語モデルに基づく話者性の評価には，話者に依存する言語表現のみでなく，役割に依存する言語表現の両方の影響を受けていると考えられる．

表 4: 各対話設定におけるクロスエントロピーの平均

話者	役割	クロスエントロピー
同一	同一	3.15
同一	相違	3.92
相違	同一	4.00
相違	相違	4.08

表 5: 役割による話者間のクロスエントロピー差分の変化

役割	同一話者	相違話者	差分
医者	3.15	4.14	0.99
患者	3.16	3.93	0.77

4.2.2 話者性変換法の評価

次に，本研究で提案した話者性変換法の評価を行う．まず，被験者に「5 つの候補の中から最もその話者らしい発話」を選ばせた結果，提案手法によって生成された発話であった割合（正解率）は 51.7%であった．5 つの候補の中から文をランダムに選択した際の正解率である 20.0%を大幅に上回っているため，提案手法は明らかに話者性を変換できているといえる．

また，変換モデルの事後確率が話者性の評価として適切であるかを調べる．それぞれの発話の事後確率とその発話が選択された割合（選択率）を図 3 に示す．発話の事後確率と選択率に対して線形回帰分析を行ったところ，相関係数 $R^2 = 0.30$ となり，緩やかな相関があることがわかった．発話のエントロピーと発話の話者性の評価値との関係を図 4 に示す．発話のエントロピーと話者性の平均評価値に対して線形回帰分析を行ったところ，相関係数は $R^2 = 0.12$ となり，緩やかな相関があることがわかった．これらのことから，言語モデルに基づく話者性の評価と人間による話者性の評価には相関があるといえる．

次に，提案手法による変換で発話の持つ対話の文脈的な意味が変化していないことを調べる．変換後の発話が対話の流れに対して適切か（適切性）を評価した結果，平均適切性は 3.81 であった．このことから，提案手法による文脈的意味の著しい変化や損失はないことがわかった．

これらの結果から，提案手法は話者性を十分変換することが可能であることがわかる．しかしながら，より正確に話者性を変換するために言語モデルと翻訳モデルをさらに洗練させる必要がある．特に，前後の文脈やフィラーと感動詞の発生位置，その発生数まで考慮して話者性を評価可能な言語モデルの構築が必要であろう．より人間の評価に近い評価値を提案することは今後の重要な課題である．

5. まとめ

本稿では，言語モデルに基づく話者判別法の日本語話し言葉における有効性の評価と，それを利用した翻訳辞書による話者性変換手法を提案し，実験的評価を用いて有効性を示した．今後の課題として，話者判別および話者性の評価手法については，より人間の評価に近い言語モデルを用いた評価方法の提案

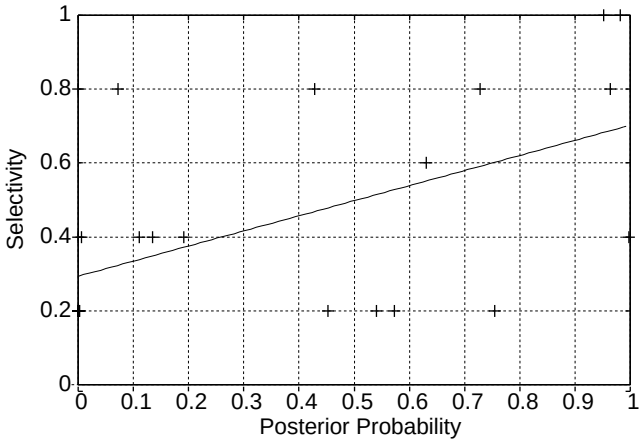

図 3: 事後確率と選択率の相関

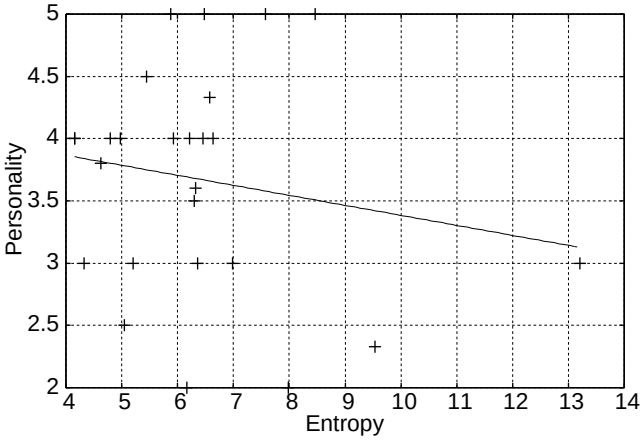

図 4: エントロピーと話者性評価の相関

と，文法・文脈的要素に対応した話者性変換手法の提案する．また，話者性変換については言語学の知見 [Teshigawara 12] を考慮したより正確な個人性変換の翻訳モデルを構築すること，シソーラスの利用も含めた翻訳辞書の拡張を進める．

参考文献

[Abe 88] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H.: Voice conversion through vector quantization, in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 655–658IEEE (1988)

[Juola 05] Juola, P. and Baayen, R. H.: A controlled-corpus experiment in authorship identification by cross-entropy, *Literary and Linguistic Computing*, Vol. 20, No. Suppl, pp. 59–67 (2005)

[Kudo 04] Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis, in *Proceedings of EMNLP*, pp. 230–237 (2004)

[Mairesse 11] Mairesse, F. and Walker, M. A.: Controlling user perceptions of linguistic style: Trainable generation of personality traits, *Computational Linguistics*, Vol. 37, No. 3, pp. 455–488 (2011)

[Neubig] Neubig, G.: Kylm - The Kyoto Language Modeling Toolkit, `http://www.phontron.com/kylm/`: Accessed: 2013-04-8

[Neubig 12] Neubig, G., Akita, Y., Mori, S., and Kawahara, T.: A monotonic statistical machine translation approach to speaking style transformation, *Computer Speech & Language* (2012)

[Qian 13] Qian, Y., Soong, F. K., and Yan, Z.-J.: A Unified Trajectory Tiling Approach to High Quality Speech Rendering, *IEEE Transactions on Audio, Speech & Language Processing*, Vol. 21, No. 2, pp. 280–290 (2013)

[Teshigawara 12] Teshigawara, M. and Kinsui, S.: Modern Japanese ‘ Role Language ’ (Yakuwarigo): fictionalised orality in Japanese literature and popular culture, in *Sociolinguistic Studies Vol 5-1*, Sheffield: Equinox Publishing (2012)

[Yamagishi 10] Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Guan, Y., Hu, R., Oura, K., Wu, Y.-J., et al.: Thousands of voices for HMM-based speech synthesis–Analysis and application of TTS systems built on various ASR corpora, *IEEE Transactions on, Audio, Speech, and Language Processing,*, Vol. 18, No. 5, pp. 984–1004 (2010)

[乾 04] 乾 健太郎, 藤田 篤 : 言い換え技術に関する研究動向, 自然言語処理, Vol. 11, No. 5, pp. 151–198 (2004)

[平岡 13] 平岡 拓也, Sakti, S., Neubig, G., 戸田 智基, 中村 哲 : 説得対話システム構築のための対話コーパス分析, 日本音響学会 2013 年春季研究発表会 (ASJ), 東京 (2013)