統計的機械翻訳の枠組みを用いた話し言葉の整形

Graham Neubig　　秋田 祐哉　　森 信介　　河原 達也
京都大学 情報学研究科

1　はじめに

従来の自動音声認識（ASR）は音響信号 X から忠実な発話内容 V を求めるように定式化され，統計的モデルを構築して事後確率 $P(V|X)$ が最大となる $\hat{V}$ を探索する過程により実現される．しかし，忠実な発話 V には言いよどみや冗長な表現，口語的表現，脱落された単語等が多く含まれており，完全に発話内容が復元できても記録文書としてふさわしくない．このため，会議録・講演録作成のための音声認識システムでは，発話内容を整形し，文書体に近づける必要がある．

これらの現象を自動的に整形する手法はいくつか研究されており，非流暢現象（フィラー・言い直し等）の削除と句読点の挿入を扱う研究は特に多い [2, 4]．しかし，人間の速記者が会議録・講演録を作成する場合，非流暢現象の削除と句読点の挿入以外にも，口語的表現の書き言葉への言い換えや脱落された単語の挿入などの訂正も行う．講演会や議会などのフォーマルな場では，言いよどみや言い直しが比較的少なく，文体の整った会議録・講演録が求められるので，このような現象を扱うことは特に重要である．

このような様々な現象を扱う先行研究としては，統計的機械翻訳（SMT）の技術を利用し，忠実な書き起こしと正式の会議録・講演録を異なる言語とみなして書き起こしから文書体へと “翻訳” するアプローチがある．下岡ら [9] は雑音のある通信路モデルを用いて，忠実な書き起こしを整形する方法を提案した．我々は，このモデルをさらに拡張し，対数線形モデルに様々な素性を導入して重み付き有限状態トランスデューサ（WFST）で実装を行った [5].

本稿では SMT に基づく話し言葉の整形の精度を向上させるために，2 つの手法を提案する．まず，音声翻訳で用いられている同時確率モデル [1] を条件付き確率に変換することで，文脈を考慮した翻訳モデルを構築する手法について述べる．これに加えて対数線形モデルを用いて同時確率モデルと条件付き確率モデルを組み合わせることで，頻度の高い翻訳パターンが優先的に利用されるようにする．

国会審議の会議録作成と「日本語話し言葉コーパス」の講演録作成で実験を行い，提案手法の有効性を検証する．具体的には，音声認識結果及び人手による忠実な書き起こしを原言語とし，会議録・講演録を目的言語として，これらの “対訳コーパス” を学習データとする．確率的モデルを重み付き有限状態トランスデューサーで実装し，会議録・講演録を正解とみなして評価を行う．

2　話し言葉の整形のモデル化

2.1　雑音のある通信路モデル

SMT に基づいた話し言葉の整形は，認識結果（または忠実な書き起こし）V を会議録・講演録 W へ変換する．具体的には $P(W|V)$ を計算する統計的モデルを構築し，ある V に対して $P(W|V)$ を最大化する $\hat{W}$ を探索する．モデルのパラメターを推定するために，V と W の対訳コーパスを学習データとして利用する．V も W も揃っている対訳コーパスのサイズは W のみの会議録・講演録コーパスより小さい場合が多いため，ベイズ則を用いて $P(W|V)$ を翻訳モデル確率 $P_t(V|W)$ と言語モデル確率 $P_l(W)$ に分解する[1].

$$\hat{W} = \underset{W}{\operatorname{argmax}} P_t(V|W)P_l(W). \tag{1}$$

翻訳モデルの学習に対訳コーパスが必要であるのに対し，言語モデルの推定には V の必要がなく，W のみが存在する “単言語” コーパスが利用できる．このようなモデルは “雑音のある通信路” モデルと呼び，従来の話し言葉の整形の研究で用いられている [2, 9].

従来の雑音のある通信路モデルでは，文の翻訳確率 $P_t(V|W)$ をモデル化するために，単語の翻訳確率は独立であると仮定し，文の翻訳確率をそれぞれの単語翻訳確率で近似していた．

$$P_t(V|W) \approx \prod_i P_t(v_i|w_i). \tag{2}$$

ここで単語翻訳確率は最尤推定によって求められる．ただし，挿入と削除を扱うために，空文字列を表す記号 ϵ を語彙に入れ，確率 $P_t(v|\epsilon)$ と $P_t(\epsilon|w)$ を推定する．一対多や多対一の変換（「いろんな」→「いろいろ な」）を扱うために，頻繁に変換される単語列も 1 つの単語として語彙に含める．このような単語が存在するため，単語分割の境界が曖昧になる．この問題を解決するために，単語境界確率を 1-gram の分割モデルで推定する．

[1] ここで P_t は対訳コーパスを用いて推定される確率，P_l は W のみが存在するコーパスを用いて推定される確率をさす．

2.2 同時確率モデル

前節では，単語翻訳確率は独立であるという仮定に基づいて文翻訳確率を推定したが，多くの場合，この文脈非依存性は必ずしも成り立たない．特に「ですね」や「と」のような，変換するか否かが文脈に依存する現象については，翻訳モデルに直接文脈を取り入れた方が有効であると考えられる．

文脈に依存する翻訳モデルを作成するために，雑音のある通信路モデルでモデルの分解を行わずに，対訳コーパスのみを用いて同時確率 $P(V,W)$ を直接モデル化する方法がある．

$$\hat{W} = \underset{W}{\mathrm{argmax}} P(V,W)$$

同時確率をモデル化する具体的な方法はいくつか提案されているが，本研究では GIATI[1] と呼ばれる方法を採用する．アライメントされた V と W の組を表す $\Gamma = \gamma_1, \gamma_2, ..., \gamma_k$（$\gamma_i = \langle v_i, w_i \rangle$）を用いて，平滑化された n-gram モデルを構築する．

$$P_t(V,W) = P_t(\Gamma) \approx \prod_i^k P_t(\gamma_i | \gamma_{i-n+1}, \ldots, \gamma_{i-1}) \quad (3)$$

2.3 文脈に依存する翻訳モデル

同時確率をモデル化することで文脈を利用することはできるが，言語モデル $P_l(W)$ との併用は容易ではなく，単言語のデータを利用することが困難である．本研究では，GIATI の同時確率を書き換え，文脈に依存する条件付き確率を得る手法を提案する．

まず，翻訳モデル確率 $P_t(V|W)$ は単語の条件付き確率の積として表現できることに着目する．

$$\begin{aligned} P_t(V|W) &= \prod_{i=1}^k P_t(v_i | v_1, \ldots, v_{i-1}, w_1, \ldots, w_k,) \\ &= \prod_{i=1}^k P_t(v_i | \gamma_1, \ldots, \gamma_{i-1}, w_i, \ldots, w_k) \end{aligned}$$

また，v_i の確率は w_i 以降の単語に依存しないと仮定する．

$$P_t(V|W) \approx \prod_{i=1}^k P_t(v_i | \gamma_1, \ldots, \gamma_{i-1}, w_i)$$

さらに，翻訳モデルが依存する単語履歴の長さをオーダ n のマルコフモデルで制限する．

$$P_t(V|W) \approx \prod_{i=1}^k P_t(v_i | \gamma_{i-n+1}, \ldots, \gamma_{i-1}, w_i) \quad (4)$$

式 (4) をさらに変形すると以下が得られる．

$$P_t(V|W) \approx \prod_{i=1}^k \frac{P_t(\gamma_i | \gamma_{i-n+1}, \ldots, \gamma_{i-1})}{P_t(w_i | \gamma_{i-n+1}, \ldots, \gamma_{i-1})} \quad (5)$$

式 (5) の分子は式 (3) の n-gram 確率と同一である．また，分母は以下のように確率を周辺化することにより，式 (3) の確率から得ることができる．

$$\begin{aligned} &P_t(w_i | \gamma_{i-n+1}, \ldots, \gamma_{i-1}) = \\ &\qquad \sum_{\gamma_j \in \{\tilde{\gamma}: \tilde{w} = w_i\}} P_t(\gamma_j | \gamma_{i-n+1}, \ldots, \gamma_{i-1}). \end{aligned} \quad (6)$$

したがって，$P_t(V|W)$ は GIATI 法によって得られる n-gram 確率から推定することが可能である．この条件付き確率を言語モデル確率 $P_l(W)$ と組み合わせて式 (1) の雑音のある通信路モデルで利用することができる．これにより，文脈を考慮した翻訳確率を利用しながら，大量のテキストを用いた言語モデルも利用することができる．また，$n = 1$ の場合，式 (5) と式 (2) は同値となる．このため，文脈に依存する翻訳モデルを用いた雑音のある通信路モデルは，従来の雑音のある通信路モデルの一般化になっている．

2.4 同時確率との対数線形結合

条件付き確率によって言語モデル確率 $P_l(W)$ との併用が可能となる一方，変換パターンの頻度情報が失われる問題点もある．具体例として，パターン γ_x の頻度がそれぞれ $c_t(\gamma_x) = 100, c_t(w_x) = 1000$，パターン γ_y の頻度がそれぞれ $c_t(\gamma_y) = 1, c_t(w_y) = 10$ のとき，変換パターンの条件付き確率は両方とも 0.1 となる．

$$P_t(v_x | w_x) = P_t(v_y | w_y) = 0.1$$

しかし，低頻度の γ_y はスパースなデータに起因する例外的なパターンである可能性もあり，高頻度の γ_x の方が信頼性が高いと思われる．特に，認識誤りを含む音声認識結果を対象とする場合，低頻度の変換パターンは信頼できない場合が多い．

$P_t(v_x|w_x)$ と $P_t(v_y|w_y)$ が同値となるのに対し，同時確率の $P_t(\gamma_x)$ は $P_t(\gamma_y)$ の 100 倍である．したがって，条件付き確率にない頻度情報は，同時確率を用いることで補完できる．本研究では，同時確率の導入法として対数線形結合 [6] を利用し，言語モデル確率・翻訳モデル条件付き確率・翻訳モデル同時確率をすべて組み合わせたモデル $M(W,V)$ を構築する．

$$\begin{aligned} &M(W,V) = \\ &\quad \lambda_1 \log P_l(W) + \lambda_2 \log P_t(V|W) + \lambda_3 \log P_t(W,V) \end{aligned} \quad (7)$$

ここで $\lambda_3 = 0$ にすると，$M(W,V)$ は式 (1) の雑音のある通信路モデルの拡張とみなせるが，$\lambda_2 = 0$ にして

表 1: 各コーパスの単語数と誤り率

コーパスタイプ		国会	CSJ
言語モデル学習		158M	181k
翻訳モデル学習		2.31M	181k
重み学習		66.3k	21.5k
テストセット		300k	11.4k
音声認識誤り率		17.10%	19.43%
整形前誤り率	書き起こし	18.62%	27.70%
	認識結果	36.10%	36.49%

第 1・第 3 項のみを対数線形結合することは理論的にも実用的にも不適切である．理論的な観点からは，$P_l(W)$ と $P_t(W,V)$ を組み合わせても，求めたい $P(W|V)$ の事後確率を得ることは不可能であるため，適切ではないといえる．また，実用的な立場からは，このように結合されたモデルは単語を過剰に削除する傾向があり，精度は単純な同時確率モデルを上回ることはない．このため，同時確率モデルと対数線形結合を行う際にも，前節で導入した条件付き確率モデルは必要不可欠である．

3 評価実験

3.1 実験の設定

提案手法の有効性を検証するために，衆議院審議コーパス [8]（以下「国会」）と「日本語話し言葉コーパス」[7]（以下「CSJ」）に対して実験を行った．それぞれのコーパスを用いてシステムを学習し，会議録・講演録を正解とみなして，人手による忠実な書き起こしと音声認識結果をそれぞれ入力とする 2 つの実験を行った．ただし CSJ の講演録は本研究室で外注により作成したものである．学習データに含まれていないテストセットを準備し，システム出力と正解の間の単語誤り率（WER）を評価尺度とした．

国会のデータサイズは CSJ を大幅に上回り，言語モデル用のデータは CSJ の 872 倍，翻訳モデル用のデータは CSJ の 12.7 倍であった（表 1 参照）．したがって，国会では CSJ より高い精度が期待できる．

句読点は翻訳モデルの中で通常の単語として扱っているが，音声認識結果を入力とする場合，200ms 以上のポーズがあった箇所にポーズを表す記号が入れられ，句読点挿入の手がかりとして用いた．CSJ の忠実な書き起こしにはポーズ情報が付与されており，同じくポーズがある箇所に記号を挿入した．国会の場合，忠実な書き起こしにはポーズ情報は付与されていなかったため，会議録に句読点があった箇所に記号を入れた．表 1 の忠実な書き起こしにおける CSJ と国会の整形前誤り率の差は主にこの句読点の異なった扱いによるものである．

表 2: 翻訳モデルのオーダと誤り率の関係

タスク		1-gram	2-gram	3-gram
国会	書き起こし	6.51%	5.33%	**5.32%**
	認識結果	21.84%	**20.99%**	21.09%
CSJ	書き起こし	15.06%	14.70%	**14.63%**
	認識結果	25.98%	**25.30%**	25.53%

3.2 学習・探索

言語モデルは Kneser-Ney 法で平滑化された 3-gram モデルを使用した．この言語モデルはすべての雑音のある通信路モデルで用いられた．

忠実な書き起こしを入力とするシステムの学習には，忠実な書き起こしと会議録の対訳コーパスを翻訳モデルの学習データとした．音声認識結果を入力とするシステムには，認識結果と会議録の対訳コーパスを利用した[2]．

各モデルを WFST で表現して，それらを合成することで大規模な 1 つのモデルを構築した．探索には WFST のビームサーチデコーダ Kyfd[3] を用いる．対数線形モデルの重みの学習には SMT ツールキット Moses に含まれているツールを利用した [3].

3.3 文脈を利用した翻訳モデルの効果

まず，雑音のある通信路（式 (5)）の枠組みにおいて翻訳モデルのオーダ n を 1〜3 の間で変動させ，翻訳モデルに文脈を利用する効果を調べた（表 2 参照）．その結果，CSJ と国会の両方で文脈を利用した翻訳モデルは 1-gram の翻訳モデルを上回った．特に，比較的学習データの多い国会では文脈を考慮したモデルの効果は顕著であった．

書き起こしの場合は 3-gram 翻訳モデルにより最も高い精度となり，認識結果の場合は 2-gram 翻訳モデルと 3-gram 翻訳モデルがほぼ同等の精度となった．認識誤りを含む認識結果では 3-gram の統計量を安定して抽出することが困難と考えられる．これ以降の実験では，忠実な書き起こしで 3-gram，認識結果で 2-gram を翻訳モデルとして用いる．

3.4 翻訳モデルの種類の影響

次に，本研究で提案した 3 種類の文脈を考慮した翻訳モデルを比較した．

[2] 予備実験で人手による書き起こしと会議録の対訳コーパスを用いた翻訳モデルも比較したが，単語誤り率は 3 %ほど悪くなった．これは主に，書き起こしを学習に用いたシステムが句読点を正しく挿入できなかったため，及び認識結果を用いたシステムが頻出する認識誤りを正しく訂正できたためと考えられる．

[3] http://www.phontron.com/kyfd/

表 3: 各モデルの評価結果．斜字は **Baseline** に比べて有意な改善を示す．

タスク		整形前	Baseline	Joint	Moses	Noisy	Noisy LL	Noisy+Joint
国会	書き起こし	18.62%	6.51%	*4.59%*	*5.45%*	*5.32%*	*5.13%*	***4.05%***
	認識結果	36.10%	21.84%	22.61%	*20.97%*	*20.99%*	*20.97%*	***20.04%***
CSJ	書き起こし	27.70%	15.06%	14.55%	14.72%	14.63%	14.49%	***13.54%***
	認識結果	36.49%	25.98%	25.37%	25.88%	25.30%	24.92%	***23.67%***

Noisy: 式 (5) の雑音のある通信路モデル

Noisy LL: 対数線形重みを用いた雑音のある通信路モデル（式 (7) で $\lambda_3 = 0$）

Noisy+Joint: 式 (7) の対数線形結合モデル

そして，既存の 3 つの手法を比較対象とした．

Baseline: 式 (2) の文脈を考慮しない雑音のある通信路モデル

Joint: 式 (3) の同時確率モデル

Moses: アライメントテンプレート法を用いたフレーズベース機械翻訳器 Moses [3]

それぞれのシステムの単語誤り率を表 3 に示す．実験の結果，**Noisy+Joint** はすべてのタスクにおいて最も高い精度となった．これは 2 群比率の差検定 (有意確率 $P = 0.01$) で **Baseline** の誤り率と比べて有意な改善である．すなわち，本研究で提案した文脈を考慮した翻訳モデルは音声認識結果の場合にも忠実な書き起こしの場合にも有効であった．

Joint モデルは **Noisy** より単純であり，頻度情報が含まれているため，データが比較的安定している書き起こしでは高い精度が得られた．しかし，認識誤りを含む入力の場合には，出力が整っていることを保証する言語モデルが利用可能な **Noisy** の方がより高い精度を実現した．さらに，両方のモデルを対数線形結合した **Noisy+Joint** が両方を大きく上回ったことから，両方のモデルには相乗効果があることが確認された．

Moses は対数線形モデルやアライメントテンプレート，クラスによる平滑化など，多くの工夫により言語間の翻訳で高い精度をあげている．しかし，本研究の提案手法とは異なり，各単語の翻訳確率は直接文脈により変動することはない[4]．実験の結果，**Moses** はすべてのタスクにおいて **Baseline** を上回ったが，対数線形結合を利用しない **Noisy** とほぼ同等，または低い精度となった．これは **Moses** の工夫は主に単語の正しい並べ替えを狙ったもので，並べ替えの少ない話し言葉の整形では必ずしも精度の向上につながらないからであろう．

[4]翻訳確率はフレーズ内の位置により変動するが，周辺の単語の表層による直接な変動はない．

4 おわりに

本論文では話し言葉の整形のための文脈を考慮した翻訳モデルの構成について述べた．また，翻訳パターンの頻度を反映する方法として，雑音のある通信路モデルと同時確率モデルを対数線形結合する手法を提案した．両方の手法を用いたシステムは，雑音のある通信路モデルやフレーズベース翻訳のベースラインに比べて高い精度を実現した．今後の課題として，WFST の音声認識デコーダとの統合を行い，音響特徴量 X から直接最適な W を探索することなどがある．

参考文献

[1] F. Casacuberta and E. Vidal. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225, 2004.

[2] M. Honal and T. Schultz. Correction of disfluencies in spontaneous speech using a noisy-channel approach. In *Proc. EuroSpeech2003*, pp. 2781–2784, 2003.

[3] P. Koehn, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL07*, pp. 177–180, 2007.

[4] Y. Liu, et al. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540, 2006.

[5] G. Neubig, S. Mori, and T. Kawahara. A WFST-based log-linear framework for speaking-style transformation. In *Proc. InterSpeech2009*, pp. 1495–1498, 2009.

[6] F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. ACL02*, pp. 295–302, 2002.

[7] 前川, 籠宮, 小磯, 小椋, 菊池．日本語話し言葉コーパスの設計．4(2):51–61, 8 2000.

[8] 秋田, 三村, 河原．会議録作成支援のための国会審議の音声認識システム．情処研報, SLP-74-21, 2008.

[9] 下岡, 南條, 河原．講演の書き起こしに対する統計的手法を用いた文体の整形．自然言語処理, 11(2):67–83, 2004.