Vol.2009-SLP-79 No.17
2009/12/22

文脈を考慮した確率的モデルによる話し言葉の整形

Graham Neubig [†1]　秋田祐哉[†1]
森信介[†1]　河原達也[†1]

自動音声認識（ASR）の結果には認識誤りのみならず，言いよどみや口語的表現など，会議録にふさわしくない現象が多く含まれている．これらの現象を整形し，自然な会議録を作成するために，認識結果（または忠実な書き起こし）と会議録を異なる言語とみなし，統計的機械翻訳を用いて認識結果から会議録へと“翻訳”する．本研究では，この枠組みの中で 2 つの手法を提案する．まず，文脈情報を考慮した翻訳モデルを導入し，システムのさらなる精度向上を目指す．また，翻訳モデルの条件付き確率と同時確率の対数線形補間を行うことで，高頻度の翻訳パターンを優先的に利用することを可能とする．有限状態トランスデューサー（WFST）による実装を行い，国会会議録と音声認識結果を用いた評価実験を行った．

Context-sensitive Statistical Models for Speaking-style Transformation

Graham Neubig ,[†1] Yuya Akita ,[†1] Shinsuke Mori [†1]
and Tatsuya Kawahara[†1]

Automatic speech recognition (ASR) results contain not only recognition errors, but also disfluencies and colloquial expressions that are not appropriate for inclusion in official transcripts. In order to correct these phenomena and create natural transcripts, we treat ASR results (or faithful transcripts) and official transcripts as different languages and use techniques from statistical machine translation (SMT) to “translate” between the two. In this paper, we present two novel methods in this framework. First, we introduce a technique to create context-sensitive translation models, improving the modeling accuracy. Second, we use log-linear interpolation to combine the translation model’s joint and conditional probabilities, allowing for frequently observed patterns to be given higher priority. A system containing these improvements was implemented using weighted finite state transducers, and an evaluation was performed on transcripts from meetings of the Japanese Diet (national congress).

1. まえがき

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[1)–3)]．しかし，人間の速記者が会議録を作成する場合，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を異なる言語とみなして書き起こしから会議録へと“翻訳”するアプローチが多い．下岡ら[4)] 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，SMT に基づく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，重み付き有限状態トランスデューサー（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，以前発表した文脈に依存する同時確率モデル[6)] 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のタスクで評価実験を行い，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，国会の会議録を目的言

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

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 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．W も V も揃っている対訳コーパスのサイズは W のみの会議録コーパスよりはるかに小さいため，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主に用いられている[1),3),4)]．

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，それぞれの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それぞれ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 $P(w|\epsilon)$ 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に，単語分割の境界が曖昧となる．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．

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．

$$\begin{aligned}\hat{W} &= \underset{W}{\operatorname{argmax}} P(V|W)P(W) \\ &= \underset{W}{\operatorname{argmax}} P(V,W)\end{aligned}$$

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[7)] 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．

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

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，一般性を失わずに，単語の条件付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．

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

⋆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 は会議録 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．

Vol.2009-SLP-79 No.17
2009/12/22

式 (4) をさらに変形すると以下が得られる．

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

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．

$$P_t(w_i|\gamma_{i-n+1},\ldots,\gamma_{i-1}) = \sum_{\gamma_j \in \{\tilde{\gamma}:\tilde{w}=w_i\}} P_t(\gamma_j|\gamma_{i-n+1},\ldots,\gamma_{i-1}). \tag{6}$$

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補間

条件付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補間[8] を利用し，言語モデル確率・翻訳モデル条件付確率・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．

$$M(W,V) = \lambda_1 \log P_l(W) + \lambda_2 \log P_t(V|W) + \lambda_3 \log P_t(W,V) \tag{7}$$

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，$\lambda_2 = 0$ にして第 1・第 3 項のみを対数線形補間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，このように補間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補間を行う際に，前節で導入した条件付確率モデルは必要不可欠である．

表 1 評価用データの詳細，人手による書き起こしと音声認識結果に対する整形パターン数
Table 1 Size of the test set, and number of transformations necessary for faithful transcripts and ASR results

ターン数			1,023
単語数			300,059
読点数			20,629
句点数			7,196
		人手	ASR
削除	フィラー	22,520	19,468
	その他	24,450	42,105
置換		4,954	28,503
挿入		4,584	11,332

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[9] を用いてシステムを学習し，国会の公式な会議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．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．音声認識は国会審議用の音声認識システム[9] によって行われ，単語誤り率 (WER) は 17.1%であった．テストセットは 2007 年 10 月に開かれた衆議院審議から構成されており、詳細は表 1 の通りである．

言語モデルの学習データとして 1999 年 1 月〜2007 年 8 月の間に開かれた審議の 1.58 億語を用いた．また 2003 年 1 月〜2006 年 12 月の間に開かれた審議の 232 万語の対訳コーパスを翻訳モデルの学習に用いた．対数線形モデルの重みは 6.63 万語のヘルドアウトデータで推定した．

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．GIATI 法の n-gram も Kneser-Ney 法で平滑化され，1-gram，2-gram，3-gram のそれぞれについて試みた[⋆1]．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．最小化・決定化などで，必要な記憶量を削減し，探索が効率的となる．本研究ではこれらのアルゴリズムを実装したライブラリ OpenFst[10] を利用した．合成されたモデル上で全探索を行うことは計算量的に現実的でないため，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[11]．

3.3 翻訳モデルの効果

表 2 に，式 (5) の雑音のある通信路モデル，対数線形重みのかけられた雑音のある通信路モデル（$\lambda_3 = 0$ の式 (7)），式 (3) の同時確率モデル，式 (7) の対数線形補間されたモデルを比較する．1-gram の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と同一であるため，それをベースラインとする．

雑音のある通信路と同時確率の線形補間された 3-gram モデルは忠実な書き起こしの入力に対して 4.05%，音声認識結果の入力に対して 20.03%の単語誤り率となっており，もっとも精度が高かった．これは 2 群比率の差検定 (有意確率 $P = 0.01$) でベースラインの誤り率（それぞれ 6.51%，21.83%）と比べて有意な改善である．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この改善に大きく貢献した．

同時確率のみを用いるモデルは，人手による書き起こしでは雑音のある通信路モデルを上回ったものの，音声認識結果ではベースラインも下回って，もっとも低い精度となった[⋆4]．これは認識結果のばらつきにより，学習データがスパースとなったからである．これに対して，大量のデータで学習された言語モデルを用いる雑音のある通信路モデルは対訳コーパスのスパースさに比較的頑健といえる．

雑音のある通信路と同時確率の対数線形補間されたモデルではさらに大きな精度の向上が見られた．特に，2-gram や 3-gram では，同時確率から与えられる頻度情報は低頻度の翻訳パターンによる誤った変換を抑えることができた．

表 2 各モデルの評価結果．LL は対数線形モデルを指す．右側は n-gram のオーダー．斜字はベースライン（アンダーライン）との有意な差を指す．
Table 2 Each model, whether it is log-linear (LL), and its WER for each TM order. Italics are statistically significantly different from the baseline (underlined).

人手による書き起こし（整形なしは 18.62%）

モデル	LL	1-gram	2-gram	3-gram
雑音のある通信路 (Noisy)		6.51%	*5.33%*	*5.32%*
雑音のある通信路 (Noisy LL)	○	*5.99%*	*5.15%*	*5.13%*
同時確率 (Joint)		9.89%	*4.70%*	*4.60%*
雑音のある通信路+同時確率 (Noisy+Joint LL)	○	*5.81%*	*4.12%*	***4.05%***

認識結果（整形なしは 36.10%）

モデル	LL	1-gram	2-gram	3-gram
雑音のある通信路 (Noisy)		21.83%	*21.00%*	*21.09%*
雑音のある通信路 (Noisy LL)	○	21.63%	*20.97%*	*21.09%*
同時確率 (Joint)		28.61%	22.62%	21.98%
雑音のある通信路+同時確率 (Noisy+Joint LL)	○	*21.32%*	*20.04%*	***20.03%***

3.4 コーパスサイズの影響

モデル学習に必要なデータサイズを調べるべく，対訳コーパスのサイズを変動させて，整形の精度を調べた．その結果は**図 1** と**図 2** の通りである．

この実験からいくつかの傾向が見られる．まず，雑音のある通信路モデルは対訳コーパスが小さい場合でも言語モデルの情報が利用できるため，小さい対訳コーパスでも比較的頑健である．その一方，同時確率を用いたモデルはデータが多くなるとともに精度が向上する傾向があり，図 1 の人手による書き起こしを入力とする実験では，雑音のある通信路モデルの精度を上回ることもある．雑音のある通信路と同時確率を補間したモデルは両方のモデルの長所をあわせもつ．小さいデータにも頑健であり，データが増えるほど性能がよくなるため，すべてのデータサイズにおいてもっとも高い精度となった．

また，同時確率を用いるモデルでは，232 万語の対訳コーパスを用いても飽和現象が見られず，さらにデータを増やせばまだ改善する余地がある．音声認識結果を入力とするシステ

⋆1 予備実験で 4-gram の精度は 3-gram の精度を下回ったため，4-gram は用いなかった．
⋆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 %ほど高くなった．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，そして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だと考えられる．
⋆3 http://www.phontron.com/kyfd/
⋆4 文脈をまったく考慮しない 1-gram モデルを除く．

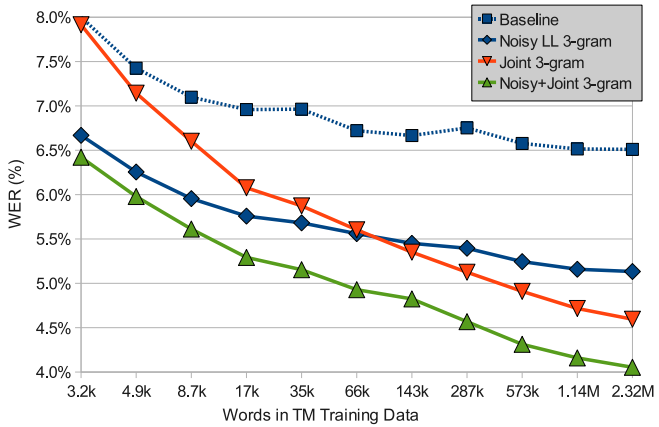

図 1 コーパスサイズの影響（人手による書き起こし）
Fig. 1 Effect of corpus size (manual transcripts)

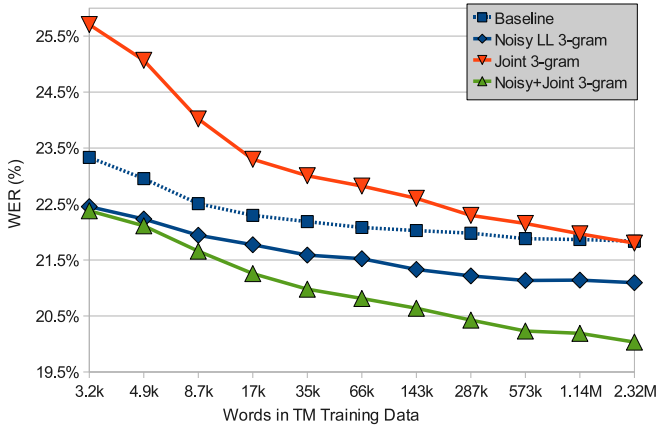

図 2 コーパスサイズの影響（音声認識結果）
Fig. 2 Effect of corpus size (ASR results)

ムでは学習データを増やすのが容易であるので，今後さらに増やしていきたい．

最後に，ほとんどのモデルで，1.7 万語程度のデータサイズで精度が急激に改善し，それ以降では改善幅が緩やかになった．これは頻出するフィラーなどの変換パターンが 1.7 万語までに学習され，それ以降は文脈に依存する変換パターンなどが学習されたためであろう．

4. むすび

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の対数線形補間する手法を提案した．両方の提案手法を用いたシステムは，雑音のある通信路モデルのベースラインに比べて精度を有意に改善できた．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．また，本研究で提案したモデルを 5) で導入した特徴量と組み合わせて，それぞれの相乗効果を調べる予定である．

参 考 文 献

1) Honal, M. and Schultz, T.: Correction of Disfluencies in Spontaneous Speech using a Noisy-Channel Approach, *Proc. EuroSpeech2003*, pp.2781–2784 (2003).
2) Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M. and Harper, M.: Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.14, No.5, pp.1526–1540 (2006).
3) Maskey, S., Zhou, B. and Gao, Y.: A phrase-level machine translation approach for disfluency detection using weighted finite state transducers, *Proc. InterSpeech2006*, pp.749–752 (2006).
4) 下岡和也，南條浩輝，河原達也：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， Vol.11, No.2, pp.67–83 (2004).
5) Neubig, G., Mori, S. and Kawahara, T.: A WFST-based Log-linear Framework for Speaking-style Transformation, *Proc. InterSpeech2009*, pp.1495–1498 (2009).
6) Neubig, G.，森 信介，河原達也：重み付き有限状態トランスデューサーと対数線形モデルを用いた話し言葉の整形，情処研報，SLP-77-21 (2009).
7) Casacuberta, F. and Vidal, E.: Machine Translation with Inferred Stochastic Finite-State Transducers, *Computational Linguistics*, Vol. 30, No. 2, pp. 205–225

Vol.2009-SLP-79 No.17
2009/12/22

(2004).
8) Och, F.J. and Ney, H.: Discriminative training and maximum entropy models for statistical machine translation, *Proc. ACL02*, pp.295–302 (2002).
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，情処研報，SLP-74-21 (2008).
10) Allauzen, C., Riley, M., Schalkwyk, J., Skut, W. and Mohri, M.: OpenFst: a general and efficient weighted finite-state transducer library, *Proc. CIAA '07*, pp.11–23 (2007).
11) Koehn, P. et al.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. ACL07*, pp.177–180 (2007).