多言語音声の同時認識のための統計的翻訳モデル

大村 絵梨　　南條 浩輝*

龍谷大学 理工学部 情報メディア学科

e-mail: ohmura@nlp.i.ryukoku.ac.jp

1 はじめに

国際化に伴いTV番組や国際会議における会議録の動画コンテンツに，複数言語の音声情報および字幕文字情報を付与することが求められており，音声認識技術が注目されている．しかし，話し言葉の音声認識精度は低く，その高精度化が求められている．

我々は，『複数の言語で同じ内容の発話がなされている状況（例えば，TV放送での主音声と副音声）』に着目し，音声認識の精度向上を目的として，主音声とその対訳となる副音声を相互に補完しながら同時に音声認識する『多言語音声の同時認識の枠組み』を提案している [1]．これにより，音声認識誤りや同音異義語（例えば，日本語発話「友達がくるまで待っている」の「くるまで」が「車で」か「来るまで」か）を他の言語の発話の情報から，解決できる可能性がある．

本研究では，この枠組みにおいて重要な役割りを果たす統計的翻訳モデル（TM）に着目し，音声認識の高精度化を図る．具体的には，TMの学習データ量，TMのモデル化手法およびTMスコアの計算方法と認識精度との関係について調べる．

2 多言語音声の同時認識

はじめに，同時音声認識の枠組み [1] について日英音声の日本語の音声認識を例にとり説明する．概要を図1に示す．これは，日本語音声 X と英語音声 Y が与えられたときに，それらを最もよく説明する日本語文字列 $\hat{J}$ を求めるプロセス（式(1)）と表せる．式(1)を変形する [1] と式(2)が得られる．

$$\hat{J} = \underset{J}{\operatorname{argmax}} P(J|X,Y) \tag{1}$$

$$\hat{J} = \underset{J}{\operatorname{argmax}} \Big(\log P(X|J) + \alpha \log P(J) + \beta N + \gamma \log \sum_{m} P(Y|E_m) P(E_m) P(J|E_m) \Big) \tag{2}$$

式(2)は，日本語の（対数）音声認識スコア $\log P(X|J) + \alpha \log P(J) + \beta N$，英語の音声認識スコア $P(Y|E_m)P(E_m)$，TMスコア $P(J|E_m)$ からなることがわかる．ここで，英語音声認識結果の第一候補（1-best）のみを用いて $\hat{J}$ を算出することを考えると，式(2)は式(3)に変形できる．本研究では，式(3)を用いて実験を行う．

$$\hat{J} = \underset{J}{\operatorname{argmax}} \Big(\log P(X|J) + \alpha \log P(J) + \beta N + \gamma \log P(J|E) \Big) \tag{3}$$

この各スコアが正解に対して高くなるようにモデル化を行うことで認識精度の向上が期待できる．本研究では，TMに着目する．

本研究では，TMにIBM-modelを用いる．IBM-modelでは，TMスコアを単語アライメント A を用いて計算する（式(4)）．しかし，この計算方法では計算コストが大きいため，本研究では式(5)で近似を行う．

$$P(J|E) = \sum_{A} P(J,A|E) \tag{4}$$

$$P(J|E) = \max_{A} P(J,A|E) \tag{5}$$

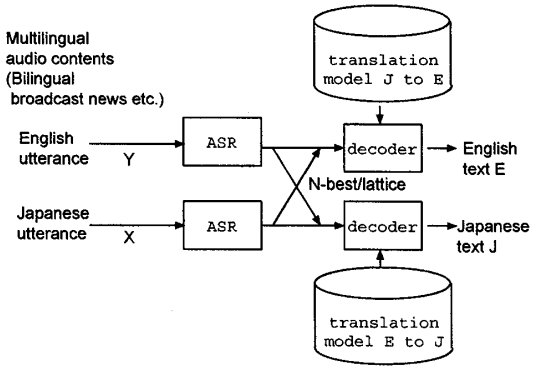

図 1: 同時音声認識の枠組み

3 評価実験

3.1 翻訳モデルと学習データ

TMの学習に用いたデータを表1に示す．本実験では，TMにIBM-model1から3を採用し，GIZA++[2]を用いて学習を行った．学習データには1文100単語以下の文のみを採用した．その際，学習データ内の出

*Statistical Translation Models for Automatic Speech Recognition of Multilingual Audio Contents by E.OHMURA, and H.NANJO, (Ryukoku University)

表 1: 翻訳モデル（TM）学習データ一覧

TM 学習データの種類	評価データとのドメインの一致度	文対数	単語数	
			日本語	英語
会話	×	63K	567K	484K
記事	○	56K	1.6M	1.3M
新聞	○	179K	5.3M	4.9M
会話+記事+新聞	△	299K	7.5M	6.7M

表 2: 評価データ

テキスト	日本語，英語各 50 文
単語数	711（日本語），476（英語）
音声	5 名（日本語母語話者）のテキスト読み上げ（計 250 発話）

現頻度 2 以下の単語を未知語として学習を行った．学習データには，ATR の SLDB，会話表現および対話データベース（会話）とロイター日英記事対訳コーパス [3]（記事）と日英新聞記事対応付けデータ [3]（新聞）の 3 種類を用いた．

3.2 評価データ

評価データを表 2 に示す．日本語とその対訳となる英語は，『The NEWS HOUR リスニング』から，双方の単語数が 60 以下のものを 50 ペア抽出した．日本語音声は日本語母語話者 5 名にテキストを読み上げてもらったデータ（計 250 発話）である．本研究では，英語音声認識誤りの影響がない状態で TM の効果を見るため，英語音声を用いずに正解文を与えて実験を行った．

3.3 結果

本研究では，式 (3) の重みを $\beta = 6$，$\gamma = 1$ に固定して実験を行った．式 (3) を実現するために，日本語音声を音声認識して上位 300 件を出力し，TM スコアでリスコアした．

はじめに，TM 学習データと認識対象とのドメインの一致度の影響，学習データ量の影響および TM の違いの影響を調べた．正確な比較のために TM の語彙は全てのデータで設定したものとした（日本語 36K，英語 33K）．結果を表 3 に示す．学習データと評価データのドメインが異なるデータ（会話）で TM を学習した場合の実験では，TM を用いる効果が見られなかった．ドメインが一致しているデータ（記事）で TM を学習した場合は，日本語音声認識のみの結果より WER が 9.3%（12.58%→ 11.65%）改善した．ドメインが最も一致しているデータ（新聞）で TM を学習した場合は 19.6%改善した．これは，学習データ量の違いの影響とも考えられる．全てのデータを用いて TM を学習した場合は，新聞のみで学習した場合と比べて差はみられなかった．これは，ドメインが大きく異なるデータ（会話）を含んでいるためと考えられる．これらのことはドメインが一致したデータを大量に用いて学習することが重要であることを示している．

表 3: TM の学習データと音声認識結果

TM 学習データの種類	TM	WER
なし（ASR のみ）	なし	12.58%
会話	IBM-model3	13.19%
記事		11.65%
新聞		10.62%
会話 + 記事 + 新聞	IBM-model1	11.81%
	IBM-model2	11.47%
	IBM-model3	10.82%

表 4: TM スコア計算方法と認識結果

TM 学習データの種類	TM スコア計算方法	TM	WER
記事	SUM	IBM-model3	12.21%
	MAX		11.73%

IBM-model1 から 3 の比較実験では，IBM-model2 は IBM-model1 より適しており（WER 3.4%改善），IBM-model3 は IBM-model2 より適している（WER 6.5%改善）ことがわかった．

最後に，TM スコア計算方法の違いによる認識結果の比較を行った．結果を表 4 に示す．近似を行ったほうが速度だけでなく精度も高かった．なおここでは，記事データのみで語彙を設定しており，語彙が異なるため，表 3 の結果と WER が一致しない．

4 おわりに

音声認識の高精度化を目的として，多言語音声の同時認識枠組みにおける TM について研究を行った．具体的には，TM の学習データ量，TM のモデル化方法および TM の計算方法と認識精度との関係について調べた．モデル化手法として，IBM-model3 が有効であること，ドメインと一致した学習データが多く必要であることを確認した．また，TM スコアの計算時に，総和計算を最大値計算で代用しても十分有効であり，計算時間の短縮が可能であることがわかった．

今後の課題として，1）学習データおよび評価データの量を増やして実験，2）本研究で扱わなかった種類の TM（IBM-model4，IBM-model5 など）で実験，3）実際の英語音声を用いた実験，4）TM の統合重みと単語数重みの最適値の検討が挙げられる．

参考文献

[1] 南條浩輝. 多言語音声の同時認識枠組みの提案. 情報処理学会論文誌, 第 49 巻, pp. 4044–4048, 2008.

[2] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. Vol. 29, pp. 19–51, 2003.

[3] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning japanese-english news articles and sentences. In *ACL-2003*, pp. 72–79, 2003.