Vol.2016-SLP-114 No.18
2016/12/20

英日講義音声翻訳に対する音声認識誤りを考慮したパラレルコーパスの利用

後藤 統興[1,a)]　山本 一公[1,b)]　中川 聖一[1,c)]

概要：自動音声認識と統計的機械翻訳を組み合わせた英語講義音声を日本語文へ翻訳する音声翻訳システムの改善を検討した. 翻訳対象ドメインに近い話し言葉の英語講演音声に対して音声認識を行い, 音声認識誤りを持つ書き起こしを作成し, 正しい書き起こしに対する人手の日本語訳と対にすることで, 音声認識誤り付きのパラレルコーパスを作成した. これを翻訳モデルの学習に利用することで, 音声認識誤りによる翻訳への影響を軽減した. 利用手法として, 機械翻訳の学習に利用する誤りのないコーパスに追加する手法と, 誤りのないコーパスのみで作成したフレーズテーブルに対して誤り付きのコーパスのみで作成したフレーズテーブルを統合する手法を比較した. 性能の比較には自動評価指標 BLEU を用いた. 対象ドメインの音声認識結果を翻訳した結果, 誤り付きのパラレルコーパスを学習コーパスへ加えることにより, 最大で 10.0(BLEU) を得ることができ, ベースラインの 7.7(BLEU) を上回った. また, 学生を対象に, 翻訳対象の講義音声の書き起こしと音声の翻訳実験を行い, 機械による音声認識と翻訳結果との比較を行った.

Utilization of Parallel Corpus with Speech Misrecognition for English to Japanese Lecture Speech Translation

Goto Norioki[1,a)]　Yamamoto Kazumasa[1,b)]　Nakagawa Seiichi[1,c)]

1. はじめに

近年, ウェブ上において利用可能な講義映像が増加している (例 MITOpenCourseWare (MITOCW)[1]). しかし, これらの講義映像は一般的に, その言語が母国語でない学生にとって学習意欲や効率を減少させる外国語である. この問題に対して, 字幕付きの講義映像は有効である [2]. 我々は時間やコストのかかる人手による書き起こしの翻訳に対し, 機械によって少ないコストでそれらを翻訳することに焦点を当てている [3], [4].

本稿で提案する音声翻訳のためのシステムは, 繰り返しやポーズなどを含む英語講義音声を扱う. 講義は旅行アシスト対話システムのようなドメインの翻訳 [5] に比べて広いトピックを持つ傾向があり, 翻訳が困難になる. また, あらかじめ用意されたスピーチや短い発話などに比べて, しばしば会話的で砕けた発話スタイルとなる講義は, タスクに対して複雑さを増加させる.

このことから, 講義音声に対する, 自動音声認識 (ASR) と統計的機械翻訳 (SMT) を組み合わせて作成される音声翻訳 (SLT) システムは困難なものとなる. 音声翻訳システムのもっとも簡単な手法は ASR の 1 ベストの出力を SMT の入力として翻訳することである. これまで ASR と SMT は個々の改善が報告されている [6].

Kolss らはドイツ語-英語講義の翻訳システムにおいて音声翻訳の進捗を報告している [5]. 同一グループの Stuker らの KIT のコンピュータサイエンスの講義ドメインにおけるドイツ語-英語翻訳実験では, リファレンスに対する書き起こしの BLEU[7] が 30.2,ASR 出力の翻訳性能は BLEU が 21.0 であった [8]. 一般的にフランス語やドイツ語といったヨーロッパ圏の言語と英語との翻訳は, 日本語への翻訳よりも容易である [9], [10]. 例えば, 日本語に近い統語構造を持つトルコ語やヘブライ語, ヒンディー語に対して, 英語講演音声の BLEU は 10 以下である [11].

ASR 出力における音声の誤認識は翻訳精度に対して大きな悪影響を持つ. この問題に対処するために,Quan らは ASR の N-best のリランキングを行い, 誤認識による影響を削減した (BLEU で約 1 向上)[12]. また,Tsvetkov らは ASR の認識誤りをシミュレーションし, 疑似的な ASR の出力を生成した [11]. その疑似的な ASR の出力を用いたフレーズテーブルの拡張は 4 つの言語において翻訳性能を改善している (BLEU 値で約 1 向上). また,Segal らは全文の ASR の

[1] 豊橋技術科学大学
TUT, Toyohashi, Aichi 441–8580, Japan
[a)] ngoto@slp.cs.tut.ac.jp
[b)] kyama@slp.cs.tut.ac.jp
[c)] nakagawa@slp.cs.tut.ac.jp

結果を含むパラレルコーパスを用いた翻訳性能の改善を報告している (BLEU 値で約 1 向上)[13].

音声翻訳の性能評価について, 菅谷らは人手による音声翻訳と機械による音声翻訳の性能を比較する実験を行っている [14]. この実験では旅行対話という限られたドメイン内 (パープレキシティ約 50, 1 文平均 8 単語程度)[15] で,TOEIC スコア別に機械と人との日本語-英語の翻訳性能を比較しており, 入力をテキストとし, 翻訳部のみを用いた場合に機械翻訳性能は TOEIC スコア 708 点と同等であるという結果を示したている. また, 音声認識も含めた場合には TOEIC スコア 548 点の人間の翻訳精度と同等であるという結果も示している.

我々のベースラインである英語-日本語の話し言葉翻訳システムは以前報告を行った [3], [4]. この音声翻訳システムは DNN-HMM に基づいた ASR に対してインドメインの講義による追加学習を行ったものと, アウトドメインである TED のパラレルコーパスと少ないインドメインを用いた SMT によって構成されている.

本稿では SMT に対する ASR の認識誤りへの適応を行う. 認識誤りへの適応のために, 我々は実際の原言語の ASR の認識誤りと正しい目的言語への翻訳をパラレルコーパスとして用いる. 認識精度の低すぎる出力は SLT システムに対して悪影響になると思われるため, 一定以上の精度を持つ, 高い信頼性のある認識結果のみを認識誤り付きのコーパスとして用いている. 作成した音声認識誤り付きの書き起こしは,SMT の学習コーパスに対して追加するか, 学習済みのフレーズテーブルに誤り付きのコーパスのみを用いて学習したフレーズテーブルを統合する形で利用する. 本稿では,TED コーパスの音声認識誤り付きの書き起こしを用いることで,MITOCW の講義に対する英語-日本語の機械翻訳システムの改善結果を報告する.

また,TOEIC スコアが 400 から 700 の日本人学生に対して, 専門性や自由度の高い英語が用いられる MITOCW の講義の書き起こしや音声に対してどの程度翻訳ができるのかを調査し, 本システムの性能と比較する.

2. システム概要

本研究で用いる英日 SLT システムは ASR と SMT によって構成されている. ASR は英語の発話から抽出される特徴量パラメータ列 X が与えられたとき, 以下の統計的な音声認識の定式化を用いて最適な単語列 W を探索する.

$$\hat{W} = \arg\max_{W} p(W|X) = \arg\max_{W} p(X|W)p(W) \quad (1)$$

ここで,$p(W)$ は英語の言語モデルによって計算される単語列の出現確率であり,$p(x|w)$ は音響モデルによって計算される観測パターンの出現確率である.

ASR 出力の英語の文である単語列 W が与えられたとき,SMT は以下の統計的な定式化によって最適な目的言語の単語列 Y を探索する.

$$\hat{Y} = \arg\max_{Y} p(Y|W) = \arg\max_{Y} p(W|Y)p(Y) \quad (2)$$

ここで $p(Y)$ は日本語の言語モデルによって計算される単語列の出現確率であり,$p(W|Y)$ は翻訳モデルによって計算される翻訳確率である. 我々の目的は MIT の講義映像に対して,SLT システムを通して SMT の出力である日本語の文を出力することである. 本稿では ASR の認識誤りを利用する手法によって SMT の改善を報告する.

3. ASR システムの適応

我々は追加学習をすることで発話スタイルと話者適応を DNN-HMM の音響モデルに対して行った [3], [4]. 本実験で用いる音響モデルはモデルパラメータの初期値として Wall Street Journal(WSJ) コーパスで学習された DNN を使用している. このモデルに対し, 音声認識対象である MIT 講義のテスト話者を除く 23 名の話者約 4 時間の音声と, テスト話者の約 5 分の音声によって追加学習を行うことでドメインに対する講義の発話スタイルと話者の適応を行った.

4. ASR による認識誤り付き書き起こしの利用

SMT のための学習コーパスは講演の映像とそれに対応する原言語と目的言語の書き起こしが存在している. 発話の単位として, 音声は原言語の書き起こしにあるピリオドごとに音声が区切られる. 原言語の文はその翻訳文にそれぞれ対応付けられている. SMT の ASR の認識誤りに対する適応は, 実際の講演の映像の音声に対する複数認識結果とそれに対応する人手による目的言語への翻訳を用いて行う.

4.1 低 WER 認識結果の作成

誤り付きの書き起こしを作成する際に用いる発話は話し言葉の講演音声であるため, 音楽や観客の声, 拍手といった背景雑音が含まれており, また話題も多岐にわたりパープレキシティーも大きく (約 450), 先行研究で用いていた ASR[3], [4] による音声認識では精度が著しく減少してしまう. 信頼性の低い認識結果を用いると, かえって翻訳モデルに対して悪影響をもたらすことが考えられる. このため, 利用可能な信頼性の高い認識結果を増やすために,ASR の言語モデルの適応を行った. まず,TED の英語書き起こしを約 1000 発話になるように講演毎に分割し, 言語モデルを作成する. その後, 分割したセットごとに全コーパスを用いて作成した言語モデルに対して線形重み付き統合を行い, 各分割したセットに対するパープレキシティを減少させる. 音声認識時には分割した書き起こしに対応する認識モデルを用いてを認識する. この言語モデルは誤り付きのコーパスを作成するための TED 音声に対して非常にクローズなものになっているが, 本研究の評価に用いる対象は MITOCW であり, 評価対象に対してクローズになっていないため翻訳実験に支障はない.

4.2 認識結果の選択と利用

精度の低すぎる認識結果では SMT に対してかえって悪影響を及ぼしてしまうと考えられるため, 一定の単語誤り率 (WER) 以下の発話を認識誤り付きのコーパスとして採用した. WER の値を基準にし, 各発話ごとに ASR による音声認識結果を採用するかどうかの選択を行っている. これらを以下の二つの方法で利用する

- 選択された ASR の認識誤りを持つ発話文集合コーパスを SMT の学習コーパスに追加する
- 認識誤り付きのコーパスをフレーズテーブルの作成に

使用し, そのフレーズテーブルを認識誤りを含まない元のパラレルコーパスによって作成されたフレーズテーブルと統合を行う

4.3 複数 ASR による認識結果の作成

信頼性の高い音声認識結果を用いる場合, その誤りは発話ごとに部分的なものになると考えられる. 誤りのバリエーションを増やすことによる誤りそのものに対する翻訳や, 誤りの周辺にある正しい書き起こしの翻訳に対する性能への影響を調査するため, 学習コーパスの異なる複数の音響モデル・言語モデルを作成し, 誤りの異なる認識結果の影響や複数システムの統合の効果を検討した. すべての ASR システムにおいて, 上述の 4.2 節に示した言語モデルの適応を行っている.

5. 人手による音声翻訳実験

本研究の音声翻訳システムが, 使用している評価基準においてどの程度学生の理解に近いかを調査するために, 日本人学生に対して本稿で対象とする MITOCW の翻訳実験を行った. 翻訳実験では講義の正確な書き起こしへの翻訳と, 学生自身による講義音声の書き起こしとその翻訳を行った.

正確な書き起こし (テキスト入力) の翻訳には講義映像が提供されている Web サイトから講義の書き起こしを取得し, それを提示して被験者による日本語訳を行う.

講義音声 (音声入力) に対する翻訳では, まず講義映像が提供されている Web サイトから動画を取得し, 音声へと変換したのちに被験者に提示し, その英語書き起こしを実施した. その後, 被験者自身が書き起こした英語文に対する日本語への翻訳を実施した. 得られた学生による英語音声の書き起こし結果 (認識結果) は ASR による認識結果と WER を用いて比較する.

テキスト入力・音声入力の両実験の際には実際の講義を聞き取る環境に近づけるためにいくつかの条件を設けた (6.4 節参照).

6. 実験条件

6.1 テストデータに対する ASR システム

6.1.1 音響モデル

本稿で用いるテストデータを認識する ASR システムの音響モデルには,WSJ コーパスから得られる 129 話者のアメリカ英語 49190 発話を用いて学習された DNN-HMM を用いた. また, 音素には 39 音素の CMU 音素語彙を用いた. モデルの詳細を表 1 に示す. DNN-HMM の教師データとして, 同じコーパスによって学習された GMM-HMM による自動アライメントされた状態ラベルファイルを使用している. 特徴量パラメータは平均 0, 分散 1 に正規化している. ネットワークは 7 層で,429 ユニット (39 次元の特徴量× 11 フレームコンテキスト) の入力層,4096 ユニットをそれぞれ持つ 5 つの隠れ層,2001 ユニットの出力層で構成されている. 学習プロセスの速度向上のため, 事前学習を行わず, 一般的に学習に用いられているシグモイド関数の代わりにリクティファイド・リニア関数を用いている. また先行研究同様に, 発話スタイルと話者の特徴を DNN-HMM 音響モデルを追加学習によって適応している [3], [4]. モデルパラメータの初期値として WSJ コーパスを用いて DNN モデルを学習し, その後にテスト話者を除く 23 名分の MITOCW の話者の音声と, テスト話者 1 名分の音声を合わせて追加学習用のコーパスとして, 発話スタイルの適応と話者適応を行っている. 上記の話者適応のためのテスト話者のデータ量は同じ音声を 3 回重複している. 音声のデータベースを表 2 に示す. テストデータのパープレキシティは 110.5, 平均文長は 22 単語であり, 部分文や複文を含んでいる.

表 1 DNN-HMM 構成

特徴量	12 MFCCs+Δ+ΔΔ+energy+Δenergy+ΔΔenergy
入力層	11 フレームコンテキスト [429 ノード]
隠れ層	7 層 [各 2048 ノード]
出力層	2001 ノード
目的関数	クロスエントロピー
活性化関数	リクティファイドリニア

表 2 ASR データベース

コーパス	話者	発話時間	発話数	用途
WSJ	129	85 時間	49190	DNN モデル学習 英語言語モデル学習
MIT	-		300000	英語言語モデル学習
	23	4 時間	1672	発話スタイル適応
	2 (A,B)	10 分	125	話者適応
		10 分	159	認識精度評価

6.1.2 言語モデル

テストデータ認識用の ASR システムに使用する言語モデルのために,1987 年から 1989 年の WSJ コーパス 85445 文書 (36754891 単語) と MITOCW の講義書き起こしの PDF800 ファイルから得られた 300000 文を用いて 3gram の言語モデルを作成した. モデルの作成にはオープンソースの SRI 言語モデルツールキットを用いている [16]. WSJ と MIT コーパスを組み合わせた辞書を用いることで, テストデータに対するパープレキシティは適応前である 305.0 から 110.5 に減少し, 未知語率は 3.96%から 0.20%に減少している. 音素辞書には WSJ コーパスのモノに加え,MITOCW のコーパスを組み込んだ約 40000 の語彙を持つ辞書を使用している [3], [4].

6.2 SMT システム

テストデータを翻訳する SMT システムの翻訳モデルは単語アライメントツールキット GIZA++を伴った MOSES デコーダのツールを用いて作成している [17]. また, 言語モデルには SRI 言語モデルツールキットを用いて 3gram の日本語の言語モデルを作成した.

6.2.1 パラレルコーパス

本研究の SMT システムの学習には英語と日本語の文対応がとられたパラレルコーパスが必要となる. Web サイト上から得られる TED Talks のページから英語の書き起こしと, それに対する日本語訳を収集した. 得られた英日の TED Talks コーパスに対して, 時間的な対応を取り, 翻訳に適さないタグデータ等の文を取り除く処理を行い, 約 140000 文のパラレルコーパスを作成した. 表 3 にパラレルコーパスの詳細を示す. MIT 講義の英語書き起こしは MITOCW の Web サイトから参照することができる [1]. ASR と SMT のシステムに対する開発・評価データには” コンピュータプログラミング” に関する講義を行う 2 話者の 284 発話を選択した. 選択したデータのうち 159 発話を評価データとし,125

発話を話者適応及び開発データとして切り分けた.SMT の翻訳モデルパラメータ調整には 125 発話のうち, 話者 A の発話である 54 発話を使用している. 評価データと開発データは MITOCW のサイトから取得した英語の書き起こしに対して, プロの翻訳家に依頼し, 日本語データを作成した.

6.2.2 パラメータ調整

翻訳モデルの言語モデル重み等のパラメータを調整するために, 開発データ 54 文を用いて Moses に搭載されている最少エラー率学習 (Minimum Error Rate Training; MERT) を行った. MERT の際に, 開発データの少なさからパラメータに対する偏りを軽減させるため, 同じモデルに対して 5 度乱数を用いた調整を行い, そのパラメータの平均値を用いている.

表 3 SMT データベース

コーパス	文数	用途
TED Talk (パラレル)	140,000	翻訳モデル学習 誤り付きコーパスの作成
MIT (パラレル)	54 (話者 A)	翻訳モデルパラメータ調整
	159 (話者 A:65, 話者 B:94)	翻訳精度評価
TED Talk (日本語)	140,000	日本語言語モデル学習

6.3 ASR による認識誤り付き書き起こしの利用

誤認識付きの書き起こしを作成するために,SMT の学習に用いる TED パラレルコーパスに対応する音声 140,000 発話に対し, 音声認識を行った. 音声認識には 2 種類の音響モデルと 2 種類の言語モデルの組み合わせ, 計 4 種類の ASR システムを作成した. それぞれの認識結果, またはすべての認識結果を, テストデータを翻訳するための SMT システムに対するパラレルコーパスに追加する学習データとして使用した場合の翻訳結果を比較・調査した. また, 学習データとして使用する際に, 認識結果の信頼性とその文数について調査を行うため,WER が 0%より大きく 10%以下, 20%以下, 30%以下のみの認識結果を用いる場合と全文を用いる場合について比較を行った. 表 4 に使用した音響モデルおよび言語モデルのデータを示す. なお, 表中の+は該当するコーパス同士を合わせて一つの学習コーパスで用いたものであり,&は左項のコーパスによって学習済みの言語モデルまたは音響モデルに対して, 右項のコーパスを用いて適応または追加学習を行ったモデルである. また,ASR 識別番号は便宜上 ASR を区別するために用いる識別用の番号である. 音響モデルの追加学習に用いる TED コーパスは ASR-1 で認識した場合の WER が 20%以下 (WER = 0 も含む) のものを使用している. これらの認識結果である誤り付きの英語文は, 正しい書き起こしの日本語翻訳と対応付けられ, 誤認識付きのパラレルコーパスとなる.

オリジナルのコーパスを使用して作成されたフレーズテーブルに認識誤り付きのフレーズテーブルを統合する場合, オリジナルのフレーズテーブルの重みは誤認識付きのフレーズテーブルよりも,5 倍大きく設定した.

6.4 人手による音声翻訳実験

正確な書き起こし (テキスト入力) の翻訳の実験には, 表 3 で示した翻訳精度評価のためのデータのうち,100 文を選択し, 被験者に対してランダムに提示した. また, 音声の書き起こし翻訳 (音声入力) の実験には同様に翻訳精度評価データから被験者の負担を軽減するために 20 文のみを選択し, 対応する音声を提示した.

表 4 誤り付きコーパス抽出のための ASR に使用したデータベース.+は左項と右項を合わせて一つのデータベースとして学習に用いたモデル.&は左項で学習したモデルに対して右項のデータを用いて適応を行ったモデル

言語モデル (文数)	音響モデル (発話数)	ASR 識別番号
TED&MIT (145032&348952)	WSJ&MIT (49180&1780)	ASR-1
	WSJ+TED&MIT (49180+42012&MIT)	ASR-2
WSJ+TED&MIT (49180+145032&348952)	WSJ&MIT (49180&1780)	ASR-3
	WSJ+TED&MIT (49180+42012&MIT)	ASR-4

人手による MITOCW の講義書き起こし及び講義音声への日本語訳を被験者に依頼する際に, 実際の英語講義を閲覧する環境に近づけるため条件を設けた. 英語の音声を再生する場合は 3 回まで全体の再生を可能とし, それまでに聞き取ることのできた英語文の書き起こしを入力する条件とした. この際,1 回目の音声再生時に同じ個所を 2 度聞くといったような時間的に遡る動作を禁止とした. 一時再生・中断は許可している.

また, 英語書き起こしの翻訳や, 被験者自身の書き起こしの翻訳には一文当り 2 分 30 秒の制限時間を設け, その時間を過ぎた場合に速やかに翻訳を終了する様指示を行った. 翻訳の際に, 被験者にとって意味の分からない単語が出現した場合には, 大規模辞書を持つ機械の優位性の差を避けるため, 紙・電子辞書等の使用を許可した. ただし, 辞書上の例文検索機能や Web 上の翻訳サイトを用いる等の使用は禁止した.

被験者は講義内容 (コンピュータプログラミング) が理解できる情報系の学生を対象とし,TOEIC スコアが 400 から 700 の者 12 名に依頼した (大学生の全国平均は 568 点, 大学院生の全国平均は 605 点 [*1]).

6.5 評価基準

テストデータの翻訳精度の評価には, 翻訳で広く用いられている以下の BLEU を用いた.

$$BLEU = BP \times exp(\sum_{n=1}^{N} \frac{1}{N} logP_n) \quad (3)$$

ここで BP は参照訳に対する機械翻訳文の単語数のペナルティであり, 機械翻訳文が短いほどペナルティが大きくなる. また,P_n は ngram の一致率である. 本稿では 1 から 4gram の BLEU の幾何平均を用いた. この評価値は値が高くなるほど, 参照した正解の日本語訳に対して近い質の翻訳結果であることを示している.

書き起こしの音声認識率を求める際, 冠詞の有無 (an, a, the の種類・挿入・脱落) や過去形 (ed 等), 複数形や三人称単数現在形 (s 等), 単語がわかる程度のスペルミスの誤りは日本語への翻訳に大きな影響を及ぼさないと考え, それらの認識誤りを訂正した場合の WER も比較対象としている.

[*1] http://www.toeic.or.jp/library/toeic_data/toeic/pdf/data/DAA.pdf

7. 実験結果

7.1 テストデータに対する ASR

MIT の発話スタイルおよび話者への適応なしでは, ベースラインの WER は 29.7%であった. 発話スタイルと話者適応をベースラインに対して同時に実行することにより, テストセットに対して 21.0%の認識精度を得た [3], [4].

7.2 認識誤り付きの書き起こしデータの利用

7.2.1 認識誤り付きデータ量

表 5 に認識する TED の講演音声 140000 に対する認識結果と各 WER の条件下での使用可能な文数を示す. ここで,ASR-1〜4 は 4 つの誤り付きコーパス抽出用の ASR の認識結果をすべて合わせた場合である. 各 ASR において, 全文の数が誤りなしのコーパスの全文よりも減少しているが, 大きすぎる認識誤り (WER1000%以上) のものやデコーダによって出力が得られなかった発話があり, システム的なエラーにより翻訳モデルの学習を行うことができなくなるため, そういった文を除外しているからである. なお, 言語モデルの適応なしの場合は, パープレキシティは約 450 程であり,WER は約 50%と本目的に使用するには精度が低すぎた. そのため,4 節で述べた言語モデルの人為的な適応を行うことにより, パープレキシティは平均約 40 になった.

表 5 TED140000 発話に対する各誤り付きコーパス導出用の ASR の認識結果

ASR	文数				TED 全文 WER
	全文	30% 以下	20% 以下	10% 以下	
誤りなし	145032				-
ASR-1	125328	60112	39439	13605	29.20%
ASR-2	117733	65091	44924	17319	25.52%
ASR-3	121905	62080	41457	14924	28.54%
ASR-4	117871	64764	44607	17210	25.40%
ASR-1〜4	482837	252047	170427	63058	-

7.2.2 認識誤り付き書き起こしの学習コーパスへの追加

認識誤り付きの書き起こしを学習コーパスへ加えて SMT の学習を行い, テストデータの書き起こしを翻訳した結果 (テキスト入力) を表 6 に示す. ベースラインには誤りコーパスを利用しない場合の翻訳モデルを用いた結果を示している. また, 重複倍数の列には元のコーパスを重複して用いた回数を示している. テキストデータを入力した場合は, 誤りコーパスを加えると翻訳精度の低下を招くのではと考えられたが, ベースラインと同等か, それ以上の翻訳精度を示すものが多くあった. 下線はベースラインと比べて 0.5 以上の BLEU の減少があった場合を示している. WER の使用条件として全文を用いた場合でも 0.5 以上の BLEU の減少は見られなかった. テキスト入力の翻訳において,ASR-1 と ASR-3 に良い性能が見られることが多いという傾向がある. ASR-1 と ASR-3 には音響モデルに TED コーパスを使用していないという共通点があげられる.

また, テストデータに対する ASR を用いて認識を行った結果 (音声入力) を翻訳した結果を表 7 に示す. 全ての場合においてベースラインの BLEU を上回っており, 認識誤りによる翻訳性能への悪影響を軽減できていることがわかる (BLEU で約 2.0 の向上). 傾向として重複回数が少ない場合において翻訳性能の向上が高いことがあげられる. 最高性能は ASR-1 を用いた場合であり,WER10%以下の誤り付きコーパスを用いた場合にテキスト翻訳の結果とほぼ同等の精度である 10.0 を得た. ASR-1〜4 を用いることで誤りのバリエーションが増加し, もっともよい精度が出ることを期待したが, そうではなかった.

表 6 原コーパスへの認識誤り付きコーパス追加後の SMT 翻訳結果 (テキスト入力)

使用 ASR	重複回数	BLEU [使用 WER 条件]			
		[全文]	[30%]	[20%]	[10%]
ベースライン	-	10.3			
ASR-1	1	10.8	**11.0**	10.9	9.3
	2	10.1	9.0	9.9	**11.5**
	5	11.2	10.2	10.6	10.7
ASR-2	1	10.7	10.2	10.9	10.8
	2	10.9	10.3	10.6	9.9
	5	10.3	9.6	10.1	10.5
ASR-3	1	10.7	10.7	10.1	10.5
	2	10.5	10.7	9.4	11.1
	5	**12.2**	10.8	10.5	10.5
ASR-4	1	10.6	10.1	10.9	10.3
	2	10.1	10.7	10.8	9.9
	5	10.9	10.2	9.9	11.2
ASR-1〜4	1	10.9	10.2	10.1	9.3
	2	10.8	9.7	10.2	9.9
	5	10.1	10.3	**11.5**	10.5

表 7 原コーパスへの認識誤り付きコーパス追加後の SMT 翻訳結果 (音声認識結果入力)

使用 ASR	重複回数	BLEU [使用 WER 条件]			
		[全文]	[30%]	[20%]	[10%]
ベースライン	-	7.7			
ASR-1	1	8.9	9.1	8.6	8.1
	2	8.2	7.8	8.5	**10.0**
	5	8.8	8.7	8.5	8.7
ASR-2	1	9.5	8.4	8.8	8.9
	2	8.6	8.9	8.6	8.6
	5	8.8	8.8	8.3	9.3
ASR-3	1	**9.6**	9.1	8.8	8.5
	2	8.1	**9.5**	8.6	8.8
	5	9.5	8.8	8.6	9.3
ASR-4	1	9.0	8.8	8.7	8.8
	2	8.3	8.7	8.6	8.6
	5	8.3	8.7	8.6	9.3
ASR-1〜4	1	9.3	8.6	8.2	8.6
	2	9.0	8.6	8.5	8.0
	5	9.0	8.8	**9.6**	8.8

7.2.3 認識誤り付き書き起こしのフレーズテーブルの追加

誤りのないパラレルコーパスによって学習されたフレーズテーブルに対して, 誤りのあるパラレルコーパスのみを用いて学習されたフレーズテーブルを統合し, テストデータの書き起こしを翻訳した結果 (テキスト入力) を表 8 に示す. また, テストデータに対する ASR を用いて認識を行った結果 (音声入力) を翻訳した際の結果を表 9 に示す. テキスト入力の場合は精度が向上するといった場合が存在しなかった. これは, 誤りコーパスを利用する場合に, 誤りを含むコーパスのみを用いてフレーズテーブルを作成したことで正しいフレーズ対応を取ることができず, 本来正しく対応の取れているフレーズに対して悪影響を与えているものと考えられる. 誤りを含むフレーズを利用する場合には正しいパラレルコーパスと共に用いなければ, 認識誤りの起

表 8 フレーズテーブルへの認識誤り付きコーパス追加後の SMT 翻訳結果 (テキスト入力)

使用 ASR	BLEU [使用 WER 条件]			
	[全文]	[30%]	[20%]	[10%]
ベースライン	10.3			
ASR-1	9.9	10.4	9.7	10.6
ASR-2	10.3	10.2	9.7	10.5
ASR-3	9.9	10.2	10.3	10.1
ASR-4	10.1	10.3	9.9	9.3
ASR-1〜4	8.1	9.7	8.7	10.3

表 9 フレーズテーブルへの認識誤り付きコーパス追加後の SMT 翻訳結果 (音声認識結果入力)

使用 ASR	BLEU [使用 WER 条件]			
	[全文]	[30%]	[20%]	[10%]
ベースライン	7.7			
ASR-1	**8.7**	**8.6**	**8.6**	8.0
ASR-2	8.3	7.8	8.4	7.5
ASR-3	8.4	8.0	7.9	8.1
ASR-4	8.2	7.9	6.8	7.4
ASR-1〜4	6.8	8.4	7.3	**8.5**

こった周辺の学習ができないと思われる.

一方, 音声認識結果入力 (音声入力) の場合はパラレルコーパスへの追加と比べて, 改善率は小さいが, 一定の効果は認められる. 傾向としては学習コーパスに追加したとき同様,ASR-1 と ASR-3 の翻訳精度が良かった. 全ての ASR による認識結果を用いた場合 (ASR1〜4, 全文) に翻訳精度が減少していることから, 誤りパターンの登録を多くしすぎると悪影響を及ぼすと考えられる.

7.3 人手による音声翻訳実験

MITOCW の講義書き起こし (テキスト入力)100 文に対する学生の翻訳結果と, 我々のシステムを用いた場合の翻訳結果を表 10 に示す. 正確な人手による書き起こしを入力する場合 (テキスト入力), 人の手による翻訳は機械翻訳を上回っている場合が多いが, 機械翻訳の性能は一部の被験者 (4 名) と同等であった.

MITOCW の講義 20 発話の音声 (音声入力) に対する学生の書き起こし精度と, その翻訳結果を表 11 に示す. WER* は日本語への翻訳に影響がないような認識誤りを許容した場合 (5 節参照) の認識精度を示している. 人による認識精度は非常に悪く, 多くが単語誤り率は 50%を超えており, 日本人学生が講義音声を聞き取ることの困難さを示している. 実験条件ではある程度の反復した聴講を許可しているため, オンラインで講義を聴取するような現実的な場面では, より精度が低くなると思われ, 機械による音声認識の優位性が確認できる. また, 書き起こし正解率 (認識精度) が低いため翻訳の BLEU は機械よりも非常に悪く, その翻訳文もほぼ意味をなしていない結果となった. 聞き取りにおいて専門用語等を聞き逃してしまうと内容の理解ができなくなってしまうため, 認識精度に頑健で, 専門用語を理解することのできる音声翻訳システムが有用であると考えられる.

表 10 英語講義音声書き起こし 100 文に対する人手による翻訳精度 (テキスト入力)

被験者 ID	TOEIC スコア	BLEU
01	495	14.3
02	635	16.3
03	450	13.1
04	640	14.8
05	630	10.8
06	570	12.5
07	475	10.0
08	475	11.7
09	565	11.1
10	475	10.2
11	480	13.0
12	560	10.5
被験者平均	537.5	12.4
ベースライン (機械)	-	10.1
ASR-1〜4 学習コーパス追加 (5 回重複 ; WER <= 20%)	-	10.6

表 11 英語講義音声 20 発話に対する人手による書き起こしと翻訳精度 (音声入力)

被験者 ID	TOEIC スコア	WER	WER*	BLEU
01	495	72.1	71.8	1.4
02	635	46.2	35.5	5.3
03	450	54.7	46.1	4.7
04	640	60.2	51.4	7.5
05	630	56.9	46.5	2.6
06	570	69.3	64.5	5.3
07	475	70.7	65.4	3.7
08	475	64.3	55.3	4.5
09	565	72.6	66.5	4.3
10	475	78.2	69.5	0.0
11	480	67.0	62.3	2.7
12	560	65.1	57.3	5.5
被験者平均	537.5	64.8	57.7	4.0
ベースライン (機械)	-	23.8	19.5	7.5
ASR-1〜4 学習コーパス追加 (5 回重複 ; WER <= 20%)	-	23.8	19.5	7.3

表 12 に書き起こしに対する, 人間の翻訳結果, 機械翻訳による翻訳結果の例を示す (テキスト入力). また, 表 13 に音声入力に対する, 人間の書き起こしと翻訳結果, 機械による認識結果と翻訳結果の例を示す (音声入力).

テキスト入力の英語言語モデルに対するパープレキシティは 122.6, 平均文長は 16 単語であり, 音声入力の正解書き起こし文のパープレキシティは 130.3, 平均文長は 18 単語である. 今回提案した音声認識誤りに頑健な SMT システムを用いた場合, 音声翻訳実験に用いた 20 文のテストセット全体ではベースラインに比べ,BLEU の劣化 (7.5→7.3) が見られたが, 表 13 に示すように, 日本語としてある程度自然な文となっており, 内容の理解の助けになる翻訳が出現していることがわかる.

8. まとめ

本稿では MITOpenCourseWare の Web 上の講義映像に対する自動音声翻訳システムの改善について検討した. 認識誤りの適応のために,TED Talk の音声に対して 4 つの ASR を用いて認識誤り付きの書き起こしを作成し, 翻訳モデルの学習へ利用した. 一文内の誤りを小さいものに限定するため,WER の大きさ毎に使用の可否を決定し, それぞれ使用した場合の比較を行った. 結果として, 書き起こしを入力した際のベースラインの BLEU10.3 に対し, 最大で 12.2 の BLEU を得ることができた. また, 音声認識結果を入力した際のベースラインの BLEU 7.7 に対して最大で 10.0 の

表 12　テキスト入力に対する翻訳例

翻訳者	翻訳結果	BLEU
書き起こし文	The semantics was what caused the problem, because the operator was expecting a particular kind of structure there.	-
正解文 (プロ)	問題を引き起こした原因は意味です演算子がそこに特定の構造を期待していたからです	100
被験者 01	セマンティックはオペレータが特定の構造の種類を期待していたために問題の原因となったものでした	25.7
被験者 02	意味論は問題を起こすことでしたなぜなら演算子はそこで個々の構造の種類を無視するからです	0.0
被験者 03	意味はプログラムで何が起こったかすなわちオペレーターは特定の種類の構造を期待しているということです	23.1
被験者 04	オペレーターは特定の種類の構造をここでは想定していたため意味論はこのような問題を引き起こしました	17.4
ベースライン	何を期待してたのは特殊な構造を引き起こしたのでしたからですの SEMANTICS のオペレータの問題があります	17.7
ASR-1	問題の原因は何を期待してたのはある特定の種の構造が SEMANTICS のオペレータでしたからです	17.8

表 13　音声入力に対する翻訳例

書き起こし/翻訳者	書き起こし/翻訳結果	WER[%]	BLEU
正解 書き起こし	Because if you did what I suggested with the list, the time to look up the key would be linear in the length of the list.	0	100
正解 翻訳 (プロ)	なぜならもし皆さんがリストについて私の指示する通りにしたならキーを探す時間はリストの長さに比例するでしょう		
被験者 01 書き起こし	BECAUSE I DID I SUGGESTED WITH THE LIST THE TIME TO LOOK UP THE KEY WOULD BE LINEAR IN LINE TO THE LISTS	26.9	11.5
被験者 01 翻訳	なぜなら私はリストで示唆しましたキーを検索する時間はリストの列に線形であると		
被験者 02 書き起こし	BE LOOK AT THIS JUST WAY TO THE LIST THE TIME TO LOOK AT KEY THE LIST	65.4	0.0
被験者 02 翻訳	これを見てくださいリストのキーを見つける時間の方法です		
被験者 03 書き起こし	BECAUSE I DID I SAY JUST WITH THE LIST THE TIME TO LOOK UP THE KEY WOULD BE LINEAR	42.3	0.0
被験者 03 翻訳	ちょうどリストといったので時間が線形になります		
音声認識結果	COULD THAT YOU DIDN'T WORK RIGHT SUGGESTED WITH THE LIST THE TIME TO LOOK UP THE KEY WOULD BE LINEAR IN THE LENGTH OF THE LIST	19.2	-
ベースライン翻訳	あなたができるのリストを調べ鍵となるでしょうリストの長さで時間の仕事をしていませんでした正しい線形		11.2
ASR-1 翻訳	あなたが正しいことを提案していませんでしたが鍵となるでしょうリストの長さで線形時間のリストを調べたのです		10.1

BLEU(テキスト入力のベースラインと同等) を得ることができ, 認識誤りによる翻訳への悪影響を軽減することを示すことができた.

また, 本稿でテスト対象となる MITOCW の書き起こし及び音声に対する学生による翻訳実験を行った. 書き起こしの実験では人手による翻訳性能が機械翻訳を上回ったが, 音声に対する翻訳では人の認識精度・翻訳精度が非常に悪く, 音声翻訳システムの利便性の可能性を示した.

謝辞　本研究は JSPS 科研費 25280062 の助成を受けたものです。

参考文献

[1] Abelson, H.: The creation of opencourseware at MIT, *Jour. Science Education and Technology*, Vol. 17, No. 2, pp. 164–174 (2008).
[2] Ferdiansyah, V. and S.Nakagawa: English to Japanese Spoken Language Translation System for Classroom Lectures, *Proc. ICAICTA*, pp. 34–38 (2014).
[3] Goto, N., Yamamoto, K. and Nakagawa, S.: English to Japanese Spoken Lecture Translation System by Using DNN-HMM and Phrase-based SMT, *Proc. ICAICTA* (2015).
[4] 後藤統興，山本一公，中川聖一：対象ドメイン内高頻出句の対訳作成による講義音声翻訳の検討，日本音響学会講演論文集，pp. 201–204 (2016).
[5] Kolss, M., Wolfel, M., Kraft, F., Niehues, J., Paulik, M. and Waibel, A.: Simultaneous German-English Lecture Translation, *Proc. IWSLT*, pp. 174–181 (2008).
[6] Matusov, E., Kanthak, S. and Ney, H.: Integrating Speech Recognition and Machine Translation: Where Do We Stand?, *Proc. ICASSP*, pp. 1217–1220 (2006).
[7] Papineni, K., Roukos, S., Ward, T. and Zhu, W.: BLEU: A method for automatic evaluation of machine translation, *Proc. ACL*, pp. 311–318 (2002).
[8] St$'$uker, S., Kraft, F., Mohr, C., Herrmann, T., Cho, E. and Waibel, A.: The KIT lecture corpus for speech translation, *Proc. LREC*, pp. 3409–3414 (2012).
[9] Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L. and Federico, M.: Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014, *Proc. IWSLT*, pp. 2–17 (2014).
[10] Eck, M. and Hori, C.: Overview of the IWSLT 2005 evaluation campaign, *Proc. IWSLT*, p. 22 (2005).
[11] Tsvetkov, Y., Metze, F. and Dyer, C.: Augmenting Translation Models with Simulated Acoustic Confusions for Improved Spoken Language Translation, *Proc. EACL*, pp. 616–625 (2014).
[12] Quan, V., Federico, M. and Cettolo, M.: Integrated n-best reranking for spoken language translation, *Proc. EuroSpeech* (2005).
[13] Segal, N. and et al.: LIMSI English-French Speech Translation System, *Proc. IWSLT*, pp. 106–112 (2014).
[14] 菅谷史昭，竹沢寿幸，横尾昭男，山本誠一：音声翻訳システムと人間との比較による音声翻訳能力評価手法の提案と比較実験，信学誌，Vol. J84-D-, No. 11, pp. 2362–2370 (2001).
[15] 松田 繁樹他：多言語音声翻訳システム"VoiceTra" の構築と実運用による大規模実証実験，信学誌，Vol. J86-D, No. 10, pp. 2549–2561 (2001).
[16] Stolcke, A.: SRILM – An Extensible Language Modeling Toolkit, *Proc. ICSLP*, pp. 901–904 (2002).
[17] Koehn, P. and et al.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. ACL*, pp. 177–180 (2007).