音声認識誤りを考慮した英語講義音声の日本語への音声翻訳システムの検討

後藤 統興　　山本 一公　　中川 聖一
豊橋技術科学大学

{ngoto, kyama, nakagawa}@.slp.cs.tut.ac.jp

1 はじめに

音声翻訳を困難にしてる問題点として, 自動音声認識 (ASR) の出力における音声の誤認識があげられる. この問題に対処するために, 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 [1]. 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)[2].

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比較的大規模な TED と少ないインドメインのパラレルコーパスを用いた統計的機械翻訳 (SMT) によって構成されている.

本稿では SMT に対する ASR の認識誤りへの影響の軽減適応を行う. ASR の誤りの特性を考慮し, 認識誤りに適応するために, 実際の ASR の出力を SMT の学習に利用する. また, 書き起こしから疑似的な誤りを伴った ASR 出力を作成し, 同様に学習に利用する. 誤り付きのパラレルコーパスを 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. これらの音声認識誤りに対する適応を行った英日翻訳システムについて, MITOCW(MITOpenCourseWare) の講義の書き起こし, および音声の ASR の出力を翻訳した翻訳性能を報告する.

また, 学生による英語講義への理解度を調査するために, 講義の書き起こしに対する翻訳実験と講義の音声に対する聞き取りとその翻訳実験を行い, 本システムとの比較を行った.

2 システム概要

本研究で用いる英日音声翻訳 (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 の認識誤りをパラレルコーパスに利用する手法によって SMT の改善を報告する.

3 誤り付きパラレルコーパスの作成と利用

3.1 実際の ASR 出力の作成 [5]

本稿では, ベースラインの SMT の学習データには英語講演映像の書き起こし及びその翻訳文をパラレルコーパスとしたものを用いている. 実際の ASR 出力の利用では, 講演映像に対する実際の認識結果を誤り付きの英語コーパスとし, 正しい書き起こしに対する日本語の翻訳文を対応付けることで誤り付きのパラレルコーパスとして利用することで, SLT システム内の SMT への入力となる ASR の誤りに対して適応を行う. 誤りのバリエーションを増加させることによる音声翻訳結果を比較するために 2 種類の言語モデルと 2 種類の音響モデルを作成し, それらの組み合わせとなる計 4 種類の ASR を使用する.

適応を行うにあたり, 音声認識誤りの多すぎる文を用いると正しく学習されるフレーズへ悪影響があると考えられる. 講演映像には観客の拍手などの雑音が多く, 音声認識精度が低くなってしまうため, 認識対象の講演の書き起こし文も加えて言語モデルを作成し, パープレキシティを人為的に下げて, 音声認識率の向上を図り, 利用できる文を増加させる.

利用する誤り付きコーパスの音声認識精度による翻訳の影響を調査するために, 利用する ASR 出力の WER に制限を付けた場合についても実験を行った.

3.2 疑似的な ASR 出力の作成

パラレルコーパスの英語側のパラレルコーパスに対して疑似的な ASR 出力を作成し, 日本語側のコーパスと対応を取ることで認識誤り付きパラレルコーパスを作成する.

疑似的な ASR 出力を作成するために, まず SMT の学習コーパスで用いる英日書き起こしのうち, 英語の書き起こし (単語列) を音素列に変換する. 音素列の変換には, 認識で用いる発音辞書を使用する. 発音辞書に含まれていない語彙が書き起こし中に存在するため, 発音辞書の拡充に Grapheme to Phoneme ツールを用いた.

作成した音素列に対して編集を加えることで, 疑似的な認識誤りを作成する. 音声認識誤りは, 音素の挿入・削除・置換誤りを想定している. 置換誤りは ASR の認識誤り傾向を考

慮するために, 学習済みの GMM-HMM を用いた ASR に含まれる各音素モデルを比較し, 距離の近いものに確率的に置換する. 挿入誤りは, 各音素に対して挿入誤り確率 P_i を定め, 音素を挿入する. また, 削除誤りを行う際も同様に, 各音素に対して削除誤り P_d を定め, 音素を削除する.

編集を行った音素列を, 音声翻訳で実際に用いる発音辞書と N-gram 言語モデルから作成された WFST に入力し, 音素列を単語列に変換することで疑似的な ASR 出力を作成する.

3.3 ASR 出力を利用したパラレルコーパス

実際の ASR 出力および疑似的な ASR 出力と, それに対応する正解の日本語の文を対にし, 誤り付きのパラレルコーパスを翻訳モデルの学習に利用する. 誤りのない SMT の学習用パラレルコーパスに誤り付きのパラレルコーパスを追加する手法と, 誤りのないパラレルコーパスのみから学習されたフレーズテーブルに対して, 誤り付きのパラレルコーパスのみから学習されたフレーズテーブルを統合する手法の 2 つの利用手法について実験し, 比較を行う.

4 実験条件

4.1 ASR システム [3]

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 ($MFCC, \Delta MFCC, \Delta\Delta MFCC$) 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 (トライフォンの共有状態数) ユニットの出力層で構成されている.

SLT システムの ASR には WSJ コーパスを用いて学習し, MITOCW のテスト以外の話者 23 名とテスト話者 1 名の発話を合わせたコーパスによって追加学習行った音響モデルを使用する. テスト話者のデータ量は同じ音声を 3 回重複して追加学習データに追加している. 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.

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には 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 の英語言語モデルを作成した.

この ASR を用いてテストデータに対する音声認識を行ったところ, 話者 A に対する音声認識の単語誤り率 (WER) は 26.2%, 話者 B に対する WER は 13.9%, 平均で 21.0% の WER となった.

4.2 SMT ベースラインシステム [3, 4]

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 [6].

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 TEDTalks 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 1 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.

[*1] https://ocw.mit.edu/courses/audio-video-courses/

表 1 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 誤り付きコーパス作成
MIT (パラレル)	54 (話者 A)	MERT(開発用)
	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

4.2.1 実際の ASR 出力

TED の講演音声の認識は, 話し言葉でトピックも多岐にわたるため非常に難しい (WER は約 50%〜60%). 本実験の目的は単語認識誤り率が 10〜20% 程度の認識結果を利用することである. そのため, 認識に用いる言語モデルを認識対象文で適応し, パープレキシティを下げて認識率を人為的に向上させた. これにより, パープレキシティは約 450 から約 40 程度となった

TED コーパスの ASR 出力を得るための音響モデルおよび言語モデルで用いたコーパスを表 2 に示す. 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.

表 2 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.

ASR 識別番号	言語モデル (文数)	音響モデル (発話数)	TED 全文 WER
ASR-1	TED &MIT (145032 &348952)	WSJ&MIT (49180&1780)	29.20%
ASR-2		WSJ+TED &MIT (49180+42012 &1780)	25.52%
ASR-3	WSJ+TED &MIT (49180+145032 &348952)	WSJ&MIT (49180&1780)	28.54%
ASR-4		WSJ+TED &MIT (49180+42012 &1780)	25.40%

4.2.2 疑似的な ASR 出力

疑似的な ASR の誤りは, 音声翻訳システムに使用する翻訳モデルの学習に用いた TED コーパスの英語書き起こしから生成した. 音素の置換確率は 10 または 15% のいずれかで実行しており, 挿入・削除誤りはそれぞれ 0 または 5% で実行している. 音声の置換誤りは, 正解音素に近い 3 つの音素に順に置換誤り確率の 1/2, 1/3, 1/6 の確率で起こるものとしている.

実際の ASR 出力を用いた場合と同様に, 疑似的な ASR 出力全文の, 誤りが大きく翻訳モデルの学習に使用できないものは除外している. また, 誤りのついている文のみを使用するため, WER が 0% の文は除外している.

4.3 学生による英語講義音声翻訳実験

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に対する翻訳と, 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. 被験者は講義内容 (コンピュータプログラミングの基礎) 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 750 程度の者 15 名に対して依頼した. なお, 未知語に対しては辞書引きを許した。また, 音声の書き起こし時には, 3 回の聞き直しを許した。

正確な書き起こし (テキスト入力) の翻訳の実験には, 表 1 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先の 100 文とは別の 20 文のみを選択し, 対応する音声を提示した.

5 実験結果

5.1 実際の ASR 出力の利用

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 3 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, ほとんどの場合でベースラインを上回る翻訳精度を得た. この現象は文献 [2] でも報告されている。これは学習文に誤りが含まれることで, 英語と日本語の正しい単語及びフレーズに対するアライメントが頑健になるためであると考えられる.

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 4 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. 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. 最良の条件で BLEU が 1.9 向上し, テキスト入力のベースラインに近い BLEU が得られた (9.6 vs 10.3).

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 5 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 6 に示す. 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.

表 3 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 [WER 条件] [全文]	[≥ 20%]	[≥ 10%]
ベースライン	-	10.3		
ASR-1	1	10.8	10.9	9.3
	5	11.2	10.6	10.7
ASR-2	1	10.7	10.9	10.8
	5	10.3	10.1	10.5
ASR-3	1	10.7	10.1	10.5
	5	**12.2**	10.5	10.5
ASR-4	1	10.6	10.9	10.3
	5	10.9	9.9	11.2
ASR-1〜4	1	10.9	10.1	9.3
	5	10.1	**11.5**	10.5

表 4 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 [WER 条件] [全文]	[≥ 20%]	[≥ 10%]
ベースライン	-	7.7		
ASR-1	1	8.9	8.6	8.1
	5	8.8	8.5	8.7
ASR-2	1	9.5	8.8	8.9
	5	8.8	8.3	9.3
ASR-3	1	**9.6**	8.8	8.5
	5	9.5	8.6	9.3
ASR-4	1	9.0	8.7	8.8
	5	8.3	8.6	9.3
ASR-1〜4	1	9.3	8.2	8.6
	5	9.0	**9.6**	8.8

表 5 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 [WER 条件] [全文]	[≥ 20%]	[≥ 10%]
ベースライン	10.3		
ASR-1	9.9	9.7	10.6
ASR-2	10.3	9.7	10.5
ASR-3	9.9	10.3	10.1
ASR-4	10.1	9.9	9.3
ASR-1〜4	8.1	8.7	10.3

表 6 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 [WER 条件] [全文]	[≥ 20%]	[≥ 10%]
ベースライン	7.7		
ASR-1	**8.7**	**8.6**	8.0
ASR-2	8.3	8.4	7.5
ASR-3	8.4	7.9	8.1
ASR-4	8.2	6.8	7.4
ASR-1〜4	6.8	7.3	8.5

5.2 疑似的な ASR 出力の利用

TED の書き起こしに対する疑似的な ASR 出力を作成し, 学習コーパスに加えたモデルについて, テキスト入力の翻訳結果を表 7 に, 音声入力の結果を 8 に示す.

音声認識で劣化した翻訳精度が改善され, 音声翻訳システムとしての性能が向上した. このことから, 疑似的な ASR 出力を用いる場合には, 実際の音声認識結果を用いる際と同様に学習コーパスに加えることが有効であることがわかった. 本手法では実際の学習用パラレルコーパスに学習データがなくても, あるいは実際の音声認識をしなくても認識誤り付きのパラレルコーパスを利用できる利点がある.

5.3 学生による英語講義翻訳実験

学生に対する英語講義の書き起こしの翻訳実験の結果を 9 に示す. 学生の翻訳結果の平均は BLEU 12.9 となっており,

機械翻訳に対して上回った結果となった．書き起こしが存在する場合には, 人手による翻訳は精度が高い結果となった．また, BLEU の低い学生であっても, 機械翻訳の翻訳文に比べて日本語として流暢な翻訳文になっているケースが多く, 語順等がバラバラになってしまう機械翻訳の翻訳結果よりも優れたものが多くなった．テキスト文の機械翻訳を授業の理解に用いる場合にはより高い精度を得る必要がある．

学生に対する講義音声への書き起こしとその翻訳を依頼した結果について表 10 に示す．WER*は冠詞や単数形の誤りなどの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の認識精度を示している．学生による音声認識は精度が非常に悪く, 聞き取りが困難なことから翻訳精度も大幅に減少し, 授業理解が困難になることがわかった．650 未満の TOEIC スコアを持つ学生は聞き取った内容がまるで異なっている発話もあり, その聞き取りに対する翻訳結果は, 人が確かめた場合にも大きくずれたものになっている．また, 高い TOEIC スコアを持つ学生でも翻訳の精度は低く, 機械翻訳の精度を上回る学生は 15 名中 1 名のみとなった．授業を受ける際に音声翻訳システムを用いることで, 手助けになることが期待できる．

表 7　疑似的な ASR 出力を学習コーパスに加えたモデルの翻訳結果 (テキスト入力)

誤り確率 (%) 置換	削除	挿入	WER (%)	BLEU[WER 条件] [全文]	[≥ 20%]	[≥ 10%]
ベースライン			-	10.3		
10	0	0	10.3	10.2	10.8	10.9
10	0	5	12.4	11.3	10.8	11.5
10	5	0	12.0	11.2	10.9	11.0
10	5	5	12.5	11.2	10.4	11.2
15	0	0	10.9	10.8	11.3	11.5
15	0	5	13.5	10.4	10.7	10.9
15	5	0	12.9	10.6	11.0	11.2
15	5	5	14.5	11.1	10.6	10.7

表 8　疑似的な ASR 出力を学習コーパスに加えたモデルの翻訳結果 (音声入力)

誤り確率 (%) 置換	削除	挿入	WER (%)	BLEU[WER 条件] [全文]	[≥ 20%]	[≥ 10%]
ベースライン			-	7.7		
10	0	0	10.3	8.4	9.4	8.8
10	0	5	12.4	9.8	8.9	9.1
10	5	0	12.0	8.9	8.8	8.3
10	5	5	12.5	9.0	7.8	8.9
15	0	0	10.9	8.8	9.3	8.9
15	0	5	13.5	8.7	8.8	8.8
15	5	0	12.9	8.8	8.4	8.7
15	5	5	14.5	8.6	9.0	9.3

表 9　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	TOEIC	BLEU
TOEIC400〜550 程度の学生の平均 (6 名)	475.0	12.1
TOEIC550〜650 程度の学生の平均 (6 名)	600.0	12.7
TOEIC750 程度の学生の平均 (3 名)	761.7	15.3
被験者 (15 名) 平均	582.3	12.9
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
疑似的な ASR；学習コーパス追加 (置換 10%, 削除 0%, 挿入 5%, $WER \geq 30$)	-	11.9

表 10　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	WER	WER*	BLEU
TOEIC400〜550 程度の学生の平均 (6 名)	67.8%	61.7%	2.8
TOEIC550〜650 程度の学生の平均 (6 名)	61.7%	53.6%	5.1
TOEIC750 程度の学生の平均 (3 名)	44.1%	36.8%	7.8
被験者 (15 名) 平均	57.9%	53.5%	4.7
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%)			7.3
疑似的な ASR；学習コーパス追加 (置換 15%, 削除 5%, 挿入 0%, $WER \geq 30$)			7.4

6　おわりに

本研究では, 学習に用いる講演映像の実際の認識誤りまたは疑似的な認識誤りを作成し, 正しい日本語訳とのパラレルコーパスと対応させ, 誤り付きのパラレルコーパスを作成した．実際の認識誤りのパラレルコーパスを SMT の学習コーパスに加えることで, テキスト入力による翻訳結果がベースラインの BLEU 10.3 から 12.2 へ改善した。また, 音声入力による場合は BLEU 7.7 から 9.6 へ改善した．また, 疑似的な誤りを学習コーパスに加えることで, テストデータのテキスト入力に対して最大で BLEU 11.5, 音声認識結果に対して 9.8 の翻訳精度を得た．疑似的な誤りコーパスは音声を使用せずに書き起こしのみから作成することができるため, 音声がない場合においても利用できるため, 様々なドメインへの利用が期待できる．また, 本システムの音声翻訳性能は TOEIC450〜650 程度の学生による音声翻訳性能を大きく上回った．

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] Y. Tsvetkov, F. Metze, and C. Dyer. Augmenting Translation Models with Simulated Acoustic Confusions for Improved Spoken Language Translation. In *Proc. EACL*, pp. 616–625, 2014.

[2] N. Segal and et al. LIMSI English-French Speech Translation System. In *Proc. IWSLT*, pp. 106–112, 2014.

[3] N. Goto, K. Yamamoto, and S. Nakagawa. English to Japanese Spoken Lecture Translation System by Using DNN-HMM and Phrase-based SMT. In *Proc. ICAICTA*, 2015.

[4] V. Ferdiansyah and S.Nakagawa. English to Japanese Spoken Language Translation System for Classroom Lectures. In *Proc. ICAICTA*, pp. 34–38, 2014.

[5] 後藤統興, 山本一公, 中川聖一. 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. 情報処理学会論文集 (SLP), Vol. 2016-SLP-114, No. 18, pp. 1–7, 1016.

[6] P. Koehn and et al. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. ACL*, pp. 177–180, 2007.