Vol.2013-NL-212 No.9
2013/7/19

同時通訳データを利用した 同時通訳用機械翻訳システムの構築

清水 宏晃[1,a)]　Graham Neubig[1,b)]　Sakriani Sakti[1]　戸田 智基[1]　中村 哲[1]

概要：本稿では，同時通訳者のような訳文を生成する機械翻訳システム（同時通訳用機械翻訳システム）が構築可能かどうかを調査する．従来，機械翻訳システムに利用される学習データには主に翻訳データが用いられている．しかし，通訳者は同時性を高めるために様々な工夫を訳出に施しており，翻訳データで学習された機械翻訳システムではこの工夫を捉えきれない．このため，本研究ではまず，通訳者が実際に同時通訳を行った音声を収録し，同時通訳データを作成する．そして，その作成したデータを機械翻訳システムの学習データに利用し，同時通訳用機械翻訳システムを構築する．実験の結果，同時通訳データを言語モデルの学習とチューニングデータに利用することで，同時通訳用機械翻訳システムが構築可能であること分かった．

1. はじめに

音声翻訳は，ある言語の音声を異なる言語の音声に翻訳する技術である．最近の技術では，旅行会話で使用される音声翻訳が実用可能なレベルまで到達している．しかし，現在の音声翻訳はいくつかの問題点を抱えている．その中の一つに, 遅延時間の問題が挙げられる．遅延時間とは，発話が始まってから翻訳された音声が合成されるまでの処理時間を示す．通常の音声翻訳は発話された音声を認識した後，認識結果を翻訳し，翻訳された音声を合成する．日常会話に音声翻訳を使用する場合，短い発話が多いため，遅延時間が短い．しかし，講演や講義などの一人が話し続ける独話に音声翻訳を使用する場合，長い発話が多いため，遅延時間は長くなる．すると，合成された音声は，発話に比べ大きく遅れるため，ユーザーはリアルタイムで講演の内容を理解することが困難となる．

ここで，この問題点を解決するために，通訳者の同時通訳に着目する．通訳者は同時通訳をする際に，様々な工夫を施し，遅延時間を短縮しようと試みている．例えば，Jonesは，通訳者が実際の発話をそのまま通訳するのではなく，不必要な情報を省略したり，発話内容を要約したりするなどの工夫を行っていることを報告した [1]．システムも通訳者のような工夫を施した訳文を生成できれば，通訳者と同じように遅延時間を短縮することができると考える．

[1] 奈良先端科学技術大学院大学 情報科学研究科
Nara Institute of Science and Technology
[a)] hiroaki-sh@is.naist.jp
[b)] neubig@is.naist.jp

そのため本稿では，自動的に通訳者に近い訳文を生成するような機械翻訳システム（以降，同時通訳用機械翻訳システム）の構築を試みる．具体的には，同時通訳データの収集を行い，学習データに同時通訳データを利用する．同時通訳データとして，比較のため，通訳経験年数が異なる複数の通訳者が英日同時通訳を行った音声を収録し，その音声を書き起こしたデータを用いる．

まず，この通訳者の工夫が施された同時通訳データと翻訳者が作成した翻訳データの違いを調査する．次に，学習データに同時通訳データを利用することにより，同時通訳用機械翻訳システムが構築可能かどうかを調査する．実験の結果，言語モデルの学習とチューニングに同時通訳データを利用することで，同時通訳用機械翻訳システムが構築可能であることが分かった．最後に，構築した同時通訳用機械翻訳システムの性能と通訳者の訳文の精度を機械翻訳の自動評価尺度を用いて比較する．

2. 同時通訳データ

翻訳と同時通訳について定量的な分析を行うために，同一の内容に関する翻訳データと同時通訳データを収集する．本節では，我々がデータ収集を行った収録環境やデータの詳細について記述する．

2.1 収集材料

同時通訳データの収録材料には TED 講演[*1]を利用した．TED 講演は学術・エンターテイメントなど様々な分野の

[*1] http://www.ted.com/

表 1 通訳者のプロフィールデータ

通訳経験年数	ランク	通訳者数
15 年	S ランク	1 名
4 年	A ランク	1 名
1 年	B ランク	1 名

人物が英語でプレゼンテーションを行っている．TED 講演を選んだ理由の一つは，ボランティアの翻訳者が作成した字幕データを無料で利用できるためである．我々は，その字幕データと収録した同時通訳データを収集し，翻訳と同時通訳の比較を行った．収録した TED 講演数は S ランクが 46 講演，A ランクと B ランクがそれぞれ 34 講演である．TED 講演の通訳音声は合計約 24 時間である．通訳者のランクに関しては次節で記述する．

2.2 通訳者

通訳経験年数が異なる複数の通訳者に同時通訳を行ってもらった．これは，通訳経験年数が異なる通訳者のデータを用意することにより，それぞれの通訳の特徴を比較するためである．表 1 に各通訳者のプロフィールデータを示す．表中のランクは通訳者の熟練度を表している．これらのランクは通訳経験年数によって決定され，通訳の熟練度が高い順に S, A, B ランクとなる．また，全ての通訳者はプロとして活躍しており，全員日本語母語話者である．

2.3 収録環境

通常，通訳者が講演を同時通訳する際には，通訳者は事前にその講演の内容を理解した上で同時通訳を行う．収録では，普段行われる同時通訳と同じ条件に揃えるため，事前に通訳者へ講演に関する資料を配布した．その資料には，講演の要約と講演中に出現する専門用語が記載されている．通訳者はその資料を読み，講演内容を理解した上で同時通訳を行った．また，通訳者は TED 講演の動画を視聴しながら同時通訳を行った．動画を見せた理由は，通訳者にとって，講演者の発話内容や音声だけでなく，表情やプレゼンテーションに用いられるスライドなどの視覚情報も重要なためである．

2.4 通訳音声の書き起こし

通訳者が同時通訳を行った通訳音声は，日本語話し言葉コーパスの書き起こし基準 [2] に準拠し，書き起こした．図 1 に通訳音声を書き起こしたデータの一例を示す．書き起こしたデータは，通訳音声の書き起こし以外に，フィラーや言い間違いなどのタグ情報も付与されている．また，通訳音声の発話単位は通訳者の音声を 0.5 秒以上の無音区間（ポーズ）で分割している．書き起こしたデータには，発話単位ごとに発話番号，発声の開始時刻と終了時刻も付与されている．

```
0001 - 00:44:107 - 00:45:043
本日は<H>
0002 - 00:45:552 - 00:49:206
みなさまに(F え)難しい話題についてお話したいと思います。
0003 - 00:49:995 - 00:52:792
(F え)みなさんにとっても以外と身近な話題です。
```

図 1 通訳音声の書き起こし例

表 2 翻訳データと同時通訳データ

データ		行数	形態素数（英）	形態素数（日）
翻訳データ	T1	167	3.11k	4.58k
	T2			4.64k
同時通訳データ	I1			4.44k
	I2			3.67k

3. 翻訳データと同時通訳データの比較

従来使用されている翻訳データと今回収録した同時通訳データの違いを調査する．翻訳者が翻訳データを人手で作成する際は，時間の制約が存在しない．しかし，通訳者が同時通訳を行う際は，時間の制約が存在する．そのため，翻訳データと同時通訳データは本質的に異なると考えられる．本節では，これらのデータの違いを調査する．

3.1 実験条件

比較対象として，この実験では，TED 講演を翻訳した 2 種類のデータと，同時通訳を行った 2 種類のデータの計 4 種類を使用する．データの詳細を表 2 に示す．T1 は翻訳者による翻訳結果であり，T2 は TED の字幕である．I1 は S ランク，I2 は A ランクの同時通訳データである．

データ同士の違いは類似度を用いて評価する．具体的には，類似度の計算には機械翻訳の自動評価尺度である BLEU [3] と RIBES [4] を用いる．BLEU，RIBES は非対称であるため，類似度は，通常のスコアと正解文と翻訳文を入れ替えたスコアの平均で計算される．つまり，BLEU(R, E) を，正解文 R に対して翻訳文 E の BLEU を計算する関数と定義すると，類似度は以下の式で表すことができる．

$$\frac{1}{2}\{\mathrm{BLEU}(R, E) + \mathrm{BLEU}(E, R)\} \quad (1)$$

また，RIBES を用いた類似度の計算も BLEU と同様に計算される．

3.2 実験結果

実験結果を図 2 に示す．まず，翻訳データ同士に着目する．T1-T2 (T1 と T2 の類似度) は，BLEU と RIBES がそれぞれ 19.18 と 71.39 であり，この中で最も高い．そのため，翻訳データ同士は全ての組み合わせの中で，最も類似度が高いことが分かる．

次に，同時通訳データ同士に着目する．I1-I2 は，BLEU と RIBES がそれぞれ 10.44 と 52.51 であり，T1-T2 より類似度が低い．この類似度が低い原因の一つは，I2 の訳出

表 3　翻訳データと同時通訳データの例

	例文
原言語文	and the disasters around the world have been increasing at an absolutely extraordinary and unprecedented rate
T1	様々な災害は今までにない異常なペースで増えています
T2	そして世界中で大災害がこれまでに例を見ない率で増えているのです
I1	世界中の自然災害は急速に最近増加しておりまして
I2	異常にこれまでにない例で

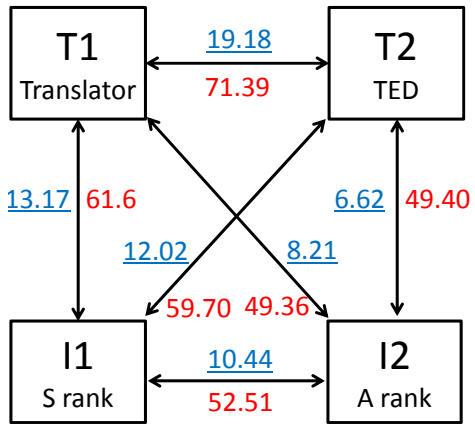

図 2　翻訳データと通訳データの比較実験結果
(下線部は BLEU，他方は RIBES を示す)

の長さである．表 3 に，翻訳データと同時通訳データの一例を示す．この表を見ると，I2 の訳出は I1 より短縮していることが分かる．また，同様に，表 2 の日本語の単語数に着目しても，I2 の単語数は他のデータと比較して少ないことが分かる．すなわち，通訳者は時間の制約に縛られているため，I1 が通訳できた文でも，I2 は上手く通訳ができなかった場合がある．つまり，I1-I2 が T1-T2 より類似度が低い理由は，通訳経験年数による通訳精度の差が現れたと考えられる．一方，翻訳者は時間の制約がないため，T1-T2 のように翻訳結果が個人によって比較的左右されないことが分かる．

最後に，翻訳データと同時通訳データに着目する．T1-I1, T2-I1, T1-I2, T2-I2 の類似度は，T1-T2 には及ばない．そのため，翻訳データ同士の方が翻訳データと同時通訳データより類似していることが分かる．また，T1-I1 と T2-I1 は類似度が高く，T1-I2 と T2-I2 は類似度が低い．この原因は，I1-I2 と同様に，通訳経験年数による I1 と I2 の通訳精度の差であると考えられる．

上記の翻訳データと同時通訳データの違いの結果から，これらのデータは本質的に異なることが分かった．そのため，翻訳者のような翻訳結果が，必ずしも熟練した通訳者に近い訳出を実現できているとは限らない．すなわち，同時通訳の精度を計測するには，同時通訳データを正解文に用いる必要がある．

4. 同時通訳データの利用

本節では，学習データに同時通訳データを利用し，同時通訳用機械翻訳システムを構築できるかどうか調査する．具体的には，翻訳モデルの学習，言語モデルの学習とチューニングに TED 講演の同時通訳データを利用する．また，同時通訳用機械翻訳システムの訳文と通訳者の訳文の性能を比較する．

表 4　実験データ

データ		行数	形態素数 (英)	形態素数 (日)
翻訳データ	トレーニング	95.9k	1.57M	2.24M
	チューニング	595	11.9k	18.1k
	テスト	523	9.78k	13.8k
同時通訳データ	トレーニング	2.05k	29.7k	33.9k
	チューニング	595	11.9k	14.6k
	テスト	523	9.78k	12.9k

4.1 同時通訳用機械翻訳システムの構築

実験に使用したデータの詳細を表 4 に示す．同時通訳データには通訳経験年数が最も長い S ランクのデータを利用した．また，チューニングとテストは，翻訳データと同時通訳データ共に同じ講演データを利用した．つまり，原言語（英語）は同じデータで，目的言語（日本語）は翻訳データと同時通訳データとで異なっている．

同時通訳用機械翻訳システムには，通常のフレーズベース機械翻訳 [5] を使用した．以下に使用したツールの一覧を示す．

- 統計的機械翻訳エンジン：Moses [6]
- 単語アライメント：GIZA++ [7]
- 言語モデル：SRILM [8]
- チューニング：MERT [9]
- 日本語の単語分割：KyTea [10]

2.4 節で示した通り，同時通訳データの発話単位は無音区間で定めているため，文アライメントはとれていない．そのため，トレーニングデータの文アライメントは，Champollion Toolkit (CTK) [11] によって，自動的に作成された．ただし，文アライメント誤りによる実験の結果，アライメント精度が 45.6 %であった．そのため，チューニングとテストで使用する同時通訳データの文アライメントは人手によって作成された．翻訳精度の評価は BLEU と RIBES を用いた．

表 5 同時通訳データを利用した実験結果

	TM	LM	Tuning	BLEU				RIBES			
				Dev		Test		Dev		Test	
				T	I	T	I	T	I	T	I
Baseline	T	T	T	12.68	6.10	11.14	7.78	53.27	44.63	52.56	46.40
Tu	T	T	I	10.70	6.88	10.85	7.49	51.95	42.42	52.33	44.60
LM+Tu	T	T, I	I	11.12	6.92	10.77	**8.10**	53.15	44.40	53.39	**48.22**
TM+LM+Tu	T, I	T, I	I	9.74	7.27	9.28	6.69	53.12	45.17	52.04	44.30

表 6 同時通訳データを利用した実験結果の例

	例文
原言語文	and the dad used to use those bulls to tell the boy stories about that civilization and their work
正解文	そして父親からその古代の文明について話を聞きました
Baseline	父とを bulls これを使っているのですが彼らの文明について話して取り組んでいます
LM+Tu	父は bulls を使って話を文明の仕事をしていますが

実験結果を表 5 に示す．TM，LM，Tuning はそれぞれ翻訳モデル，言語モデル，チューニングを示している．T と I はそれぞれ翻訳データと同時通訳データを示している．Dev と Test は正解文がそれぞれチューニングデータとテストデータであることを示している．Baseline は全ての学習データに翻訳データを利用した．Tu はチューニングデータのみに同時通訳データを利用した．LM+Tu は言語モデルに翻訳データと同時通訳データを使用し，線形補間を行った．線形補間のパラメータは，同時通訳データを用いて最適化を行った．TM+LM+Tu は，fill-up 法 [12] を用いてフレーズテーブルを作成した．その際，in-domain に同時通訳データ，out-of-domain に翻訳データを用いた．

表 5 より，LM+Tu は，BLEU が 8.10，RIBES が 48.22 であり，Baseline より，BLEU が 0.32，RIBES が 1.82 改善している．表 6 に同時通訳データを利用した実験結果の一例を示す．LM+Tu は，同時通訳データでチューニングされているため，Baseline と比較すると，訳文が短縮されており，余計な単語が生成されていない．つまり，全ての学習データに翻訳データを利用するより，言語モデルの学習とチューニングに同時通訳データを利用することで，同時通訳の訳文に近づくことが分かった．

Tu は，Baseline と比較すると，精度が劣化している．Dev の I の BLEU を Baseline と比較すると，0.78 改善し，Test では 0.29 劣化した．この原因を調べるため，目的言語の形態素数を原言語の形態素数で割った値を比較した．結果，Dev は 1.22，Test は 1.31 となり，Test が Dev より値が大きいため，原言語の 1 形態素あたりの目的言語の形態素数が多いことが分かる．つまり，翻訳文の長さが正解文より短縮されており，簡潔ペナルティが影響したことが原因の 1 つである．

また，TM+LM+Tu も Baseline と比較すると，精度が劣化している．この原因の 1 つとして翻訳モデルの学習が挙げられる．fill-up に使用した同時通訳データの文アライメントの精度は 45.6 %である．文アライメントの精度の悪さゆえに，フレーズの対応が上手くとれなかったことが原因の 1 つである．

表 7 同時通訳者との比較実験結果

	BLEU	RIBES	形態素数（日）
A ランク	8.51	49.96	10.8k
B ランク	9.58	49.52	10.7k
LM + Tu	8.10	48.22	12.7k

4.2 同時通訳者との比較

表 5 より，最も精度が高かった同時通訳用機械翻訳システム (LM+Tu) と，実際の同時通訳者 (A ランクと B ランク) の比較を行う．使用した TED 講演は，523 文である．また，S ランクは通訳経験年数が最も長いため，同時通訳データの正解データとして利用した．

実験結果を表 7 に示す．同時通訳者と同時通訳用機械翻訳システムの比較実験では，LM+Tu は A ランクと B ランク共に及ばなかった．そのため，翻訳精度の観点からは，構築したシステムは通訳経験年数が 1 年未満のシステムであることが分かる．

A ランクと B ランクに関しては，BLEU は B ランクの方が高く，RIBES は A ランクの方が高い．このため，自動評価尺度の観点からは，比較的差が見られない．

5. 関連研究

同時通訳に関する研究はいくつかなされている．同時通訳のデータ収集に関する研究として，名古屋大学同時通訳データベース[*2](SIDB) は 5 年間に渡り同時通訳データを収集を行った [13][14][15]．SIDB は独話の通訳音声を含め，全体で約 182 時間の音声を収録した．しかし，SIDB は通訳者の同時通訳データのみを収集しており，翻訳データは作成していない．そのため，SIDB のデータだけでは，

[*2] http://slp.itc.nagoya-u.ac.jp/projects/sidb

翻訳データと同時通訳データの違いを比較することができない．この分析を行うため，本研究では 1 つの講演データに対し，同時通訳データと翻訳データを作成し，比較を行った．

同時通訳データを利用したシステム構築に関する研究も行われている．Ryu らは，構文ルールを利用した同時通訳システムを構築した [16]．この研究は，学習データに SIDB の同時通訳コーパス [13][14][15] を利用しているが，ルールベース機械翻訳を用いているため，多言語に対応することが容易ではない．他にも，Paulik らは統計的機械翻訳に同時通訳データを使用した [17]．しかし，同時通訳データを翻訳データの代替データとして使用できるかという点に注目しており，同時通訳の訳出に近いかどうかは評価を行っていない．そのため，本研究では，同時通訳データを正解文に用いて，機械翻訳用同時通訳システムの訳文を評価した．

6. おわりに

本稿では，同時通訳用機械翻訳システムの構築を行った．システム構築に必要な同時通訳データは，実際に通訳者が同時通訳を行った音声をもとに，同時通訳データを作成した．翻訳データと同時通訳データの比較では，データ間の類似度を計測し，互いのデータが本質的に異なることを示した．同時通訳データの利用では，言語モデルの学習とチューニングに同時通訳データを加えることで，同時通訳の訳出に近づくことが確認できた．

本稿では，機械翻訳部に着目したため，翻訳精度の観点から評価を行った．この研究を発展させるためには，評価方法を改善する必要がある．例えば，この同時通訳用機械翻訳システムに，発話の分割位置を自動的に推定する手法 [18][19] を組み込むことが考えられる．そうすれば，発話された音声が翻訳されるまでの遅延時間の評価も行うことができる．遅延時間と翻訳精度の観点から同時通訳システムの性能を調査する必要がある．

謝辞 本研究の一部は，JSPS 科研費 24240032 の助成を受け実施したものである．

参考文献

[1] Roderick Jones. *Conference Interpreting Explained (Translation Practices Explained)*. St. Jerome Publishing, 2002.

[2] 小磯花絵, 間淵洋子, 西川賢哉, 斎藤美紀, and 前川喜久雄. 『日本語話し言葉コーパス』の設計の概要と書き起こし基準について. In **平成 *15* 年度国立国語研究所公開研究発表会講演予稿集**, 2003.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318, Philadelphia, USA, 2002.

[4] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. EMNLP*, pages 944–952, 2010.

[5] Phillip Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proc. HLT*, pages 48–54, Edmonton, Canada, 2003.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, pages 177–180, Prague, Czech Republic, 2007.

[7] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[8] Andreas Stolcke. SRILM - an extensible language modeling toolkit. In *Proc. 7th International Conference on Speech and Language Processing (ICSLP)*, 2002.

[9] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proc. ACL*, 2003.

[10] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. ACL*, pages 529–533, Portland, USA, June 2011.

[11] Xiaoyi Ma. Champollion: A robust parallel text sentence aligner. In *Proc. LREC*, 2006.

[12] Arianna Bisazza, Nick Ruiz, and Marcello Federico. Fill-up versus interpolation methods for phrase-based smt adaptation. In *Proc. IWSLT*, pages 136–143, 2011.

[13] Yasuyuki Aizawa, Shigeki Matsubara, Nobuo Kawaguchi, Katsuhiko Toyama, and Yasuyoshi Inagaki. Spoken language corpus for machine interpretation research. In *International Conference on Speech and Language Processing (ICSLP)*, pages vol. III, page 398–401, 2000.

[14] Shigeki Matsubara, Akira Takagi, Nobuo Kawaguchi, and Yasuyoshi Inagaki. Bilingual spoken language corpus for simultaneous machine interpretation research. In *Proc. LREC*, pages vol. I, page 153–159, 2002.

[15] Hitomi Toyama, Shigeki Matsubara, Koichiro Ryu, Nobuo Kawaguchi, and Yasuyoshi Inagaki. Ciair simultaneous interpretation corpus. In *Proc. Oriental COCOSDA*, 2004.

[16] Koichiro Ryu, Atsushi Mizuno, Shigeki Matsubara, and Yasuyoshi Inagaki. Incremental japanese spoken language generation in simultaneous machine interpretation. In *Proc. Asian Symposium on Natural Language Processing to Overcome language Barriers*, 2004.

[17] Matthias Paulik and Alex Waibel. Automatic translation from parallel speech: Simultaneous interpretation as mt training data. In *Proc. IEEE Workshop on ASRU*, 2009.

[18] Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan Ladan Golipour, and Aura Jimenez. Real-time incremental speech-to-speech translation of dialogs. In *Proc. NAACL12*, 2012.

[19] Tomoki Fujita, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Simple, lexicalized choice of translation timing for simultaneous speech translation. In *Proc. 14th InterSpeech*, 2013.