音声出力を考慮した同時音声翻訳のための評価尺度

三重野 隆史，Graham Neubig，Sakriani Sakti，戸田 智基，中村 哲
奈良先端科学技術大学院大学 情報科学研究科
{mieno.takashi.mh1,neubig,ssakti,tomoki,s-nakamura}@is.naist.jp

1 はじめに

音声翻訳は，入力音声を異なる言語に翻訳する技術であり，長年の研究によりその性能は改善しつつある．しかし，文単位で翻訳する従来の音声翻訳 [9] が講演のような発話が長い場面に使用される場合，発話開始から翻訳開始までの時間（以降，遅延時間）が長くなる．同時音声翻訳 [1, 12, 13] は文の途中で翻訳を開始し，遅延時間を短縮する．同時音声翻訳で重要となるのは，翻訳精度をできるだけ維持しつつ，遅延時間を短縮することであり，従来の研究では遅延時間を短縮する様々な手法が提案されてきた．

しかしながら，遅延時間を減らせば減らすほど，翻訳に利用できる文脈情報も減るため，遅延時間を短縮すると翻訳精度が劣化することも知られている [5]．この中で我々は同時音声翻訳における翻訳精度と遅延時間を同時に考慮した評価尺度の提案をしている [15]．具体的には，同一の入力動画に対して，精度の異なる複数の翻訳結果を作成し，動画に話者の実際の発話より遅れて提示する．こうすることにより，同一の動画に対して様々な翻訳精度と遅延時間を持った動画が得られ，これらを被験者に見せてランク形式で評価を行ってもらう．そして，遅延時間と翻訳精度を入力として人手評価結果を推定するランキング学習の問題として定式化し，評価関数を学習する．

文献 [15] では，翻訳文の提示に字幕データを用いており，音声からテキストへの翻訳（以降，S2T 翻訳）を想定している．しかし，音声から音声への翻訳（以降，S2S 翻訳）においても同様の傾向が見られるか明らかではない．そこで本研究では，音声データを用いて文献 [15] の手法に従い，S2S 翻訳のための翻訳精度と遅延時間を同時に考慮した評価尺度の作成方法を提案し，S2T 翻訳との違いを調査する．

実験では，評価の対象に TED 講演 [1] を用い，英日方向で翻訳された結果の読み上げ音声を被験者に提示した．評価データの翻訳精度の素性として人手評価（5 段階評価）と自動評価を用いる．比較検証を行った結果，遅延時間及び各評価尺度を素性に用いた場合，分類精度がチャンスレートを上回ることが確認された．また，音声データは字幕データを用いた場合と比べて翻訳精度が遅延時間よりも重要であることが示唆された．

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

2 評価関数

同時音声翻訳における翻訳精度と遅延時間を同時に考慮した評価を行うために，任意の同時音声翻訳結果が与えられたとき，主観評価と相関のある評価スコアを返す評価関数 s を式 (1) のように定義する．

$$s = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{1}$$

ここで，$\boldsymbol{x}$ は提示された動画を示し，$\boldsymbol{\phi}$ は $\boldsymbol{x}$ から同時音声翻訳の評価に有用な素性を計算する関数である．本稿で $\boldsymbol{\phi}(\boldsymbol{x})$ を遅延時間と翻訳精度という 2 つの値を計算し，ベクトルとして返す関数とする．[2] $\boldsymbol{w}$ はこの素性の相対的な重要度を表す重みベクトルである．本研究の目標は，この重みベクトルをデータに基づいて推定することで，同時音声翻訳において遅延時間と翻訳精度が聞き手の主観に与える影響を明らかにすることであり，次節以降にその具体的な手続きを説明する．

3 評価データの収集法

本節では，前節で述べた評価関数を推定するための，同時音声翻訳の翻訳精度と遅延時間を同時に考慮した人手評価データの収集法を記述する．

3.1 評価データの形式

2 節の自動評価関数は動画 $\boldsymbol{x}$ を受け取り，スコア s を返す．この関数を学習するデータを作成する方法として，まず，評価者に動画を視聴してもらい，スコア s を直接 5 段階評価などで評価付ける方法が考えられる．しかし，翻訳精度と遅延時間を総合的に評価する人手評価指標は確立しておらず，その設計が容易ではない．

文献 [15] では，s を直接付与する絶対評価ではなく，複数の候補を比較して評価する相対評価を用いることでこの問題を回避しており，本研究でもこの手法を用いる．具体的には，同一の動画に対して，複数の異なった翻訳精度と遅延時間を持った翻訳結果を評価者に見せ，理解のしやすい順にランク付けを行ってもらう方法を用いる．

1 つの動画を作成するために，まず平均文数 4〜5 文程度となるように動画の一部を選択し，切り出す．[3] 文数は原文である英文のピリオドを基準に算出する．選択する基準は，なるべくそれ以前の内容に依存せず，発話開始のタイミングが明確であることを重視する．

[2] つまり，線形モデルに限定される．

[3] 4〜5 文を利用する理由は，評価文数が多すぎる場合，被験者に負担がかかりすぎて評価が曖昧になることを回避するためである．

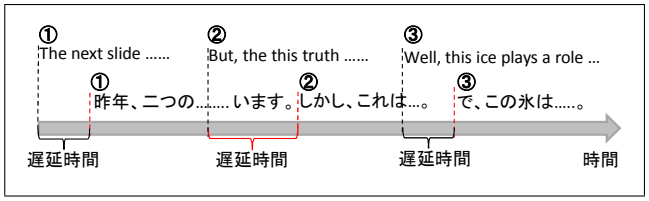

図 1: 遅延時間の例

3.2 翻訳結果の提示

【音声】

次に，実際に評価者に見せる動画を作成する．文献 [15] では，翻訳結果の出力を字幕データとして提示していたが，本研究では，読み上げ音声を用いる．

翻訳結果の提示方法として，収録音声（読み上げ音声）と合成音声を用いる方法が考えられる．収録音声と合成音声にはそれぞれ以下の特徴がある．

収録音声：自然性が高く，聞き取りやすく理解しやすいが，話速や声色，イントネーションなどを一定に保ちにくい．

合成音声：話速や声色，イントネーションなどを一定にすることが可能であるが，自然性が収録音声に比べ低く聞き取りにくい．

本研究では，翻訳精度と遅延時間が重要な要素であるため，聞き取りやすさを重視した収録音声を用いる．

音声の収録は，文のピリオド単位で行う．その後，収録音声と対応する原言語の発話タイミングで連結し，評価動画に合成する．その際，通常の人間による吹き替え翻訳に習い [4]，原言語の音声は評価音声の聞き取りを阻害しない程度まで音量を落とす．

【参照訳の提示】

翻訳精度評価の際に，事前にテキストベースで参照訳（正解訳）の提示を行う．被験者にはまず参照訳を一読してもらい，参照訳をもとに評価音声の翻訳精度を評価してもらう．参照訳を事前に提示した理由は，原言語の音声は評価の際，音量が十分に小さいため提示された翻訳音声の内容を正しいか否かを判断できない問題を回避するためである．

【遅延時間】

3.1 項で作成した翻訳結果の提示には，無作為に選択された遅延時間を付加し，切り出した動画の開始時点を遅延 0 秒として提示する．本研究において遅延時間とは，講演者の発話開始から翻訳データの提示開始までに要した時間とすることに注意されたい．具体例を図 1 に示す．図から分かるように，仮に 20 秒の動画を選択した場合，翻訳結果の提示に遅延時間を 5 秒設けると，その動画は合計 25 秒の動画となる．ただし，この場合，伸びた表示の時間だけ提示する動画の長さを伸ばすこととする．また，翻訳文によっては，収録音声が原文の発話時間を超過する場合が存在する．この場合は一つ前の翻訳文の提示が終了した直後に提示することとする（図 1:②）．

3.3 動画の評価

動画の評価には，理解のしやすい順にランクを付ける方法を用いる．具体的には，ひとつの画面に異なる翻訳精度と遅延時間を持つ同一の動画を複数提示し，被験者に任意のタイミングで視聴してもらいランク付けを行う．このとき，正確な評価データを得るために同一の動画に関しては何度でも視聴し比較することは可能とする．ただし，翻訳精度と遅延時間を評価対象とし，音声の話速や声色，イントネーションなどは評価対象外とする．

4 ランキング学習による重みの推定

前節で述べたデータを用いて重みベクトルを推定する．重みベクトルの推定にはランキング学習を使用する．ランキング学習の目的は，提示された動画から抽出された素性ベクトル（本稿では翻訳精度と遅延時間）に基づき，各動画に対するランキングを出力することである．ランキング学習の学習データは，動画から抽出された素性ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ と評価者により判定されたランク $y_i \in \{1, 2, ...\}$ のペア集合 $\{(\boldsymbol{\phi}(\boldsymbol{x}_i), y_i)\}_{i=1}^{m}$ により構成される．ランキング学習では，素性ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ のランクが高い（つまり，数字が低い）ほど大きな値を出力する関数 f を作成することが目標となる．

関数 f を $f(\boldsymbol{\phi}(\boldsymbol{x})) = \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})$ とすると，ランキング学習は各インスタンスのペア (i, j), $\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}(\boldsymbol{x}_j)$ に関して，

$$\begin{aligned} y_i < y_j \quad &\Leftrightarrow \quad f(\boldsymbol{\phi}(\boldsymbol{x}_i)) > f(\boldsymbol{\phi}(\boldsymbol{x}_j)) \\ &\Leftrightarrow \quad \boldsymbol{w}^T(\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\phi}(\boldsymbol{x}_j)) > 0 \end{aligned}$$

となる重みベクトル $\boldsymbol{w}$ を求めることになる．このようなベクトルを適切に学習するため，各素性ベクトルのペアを考え，新たに $\{(\boldsymbol{\phi}(\boldsymbol{x}_{i1}) - \boldsymbol{\phi}(\boldsymbol{x}_{i2})), z_i\}_{i=1}^{n}$ を作成する．ここで，

$$z_i = \begin{cases} +1 & y_{i1} < y_{i2} \\ -1 & y_{i1} > y_{i2} \end{cases}$$

であり，n は全ての可能なペア数を表す．この新たなデータを学習データとして 2 クラス分類問題を解くことによって上記の大小関係を満たす関数を学習することができる．その際に，同じペアで順序を入れ替えただけのペアは境界からの距離が等しいという特徴を利用し，$z = +1$ のペアのみ学習に利用する．

5 実験的評価

本節では，実験設定および実験結果について記述する．

5.1 実験設定

【評価データ】

評価データには TED 講演を使用し，被験者は同一の動画に対して英日方向に翻訳された結果を付与した

表 1: 評価データの各翻訳精度，TED の字幕データ (TED)，通訳者 S-rank の書起しデータ (S), 通訳者 A-rank の書起しデータ (A),Travatar の翻訳結果 (Trav)，Moses の翻訳結果 (Mos)，5 つの参照訳の適用 (multi)

	TED	S	A	Trav	Mos
BLEU+1	0.38	0.14	0.11	0.20	0.16
BLEU+1(multi)	0.57	0.25	0.19	0.44	0.36
RIBES	0.82	0.59	0.53	0.67	0.59
RIBES(multi)	0.86	0.68	0.63	0.79	0.72
人手評価	0.89	0.57	0.45	0.48	0.38

3 つの動画を視聴しながら内容の理解しやすさを基準に 1 から 3 のランク付けを行ってもらった．ランク付けを被験者 15 人が行い，被験者は全て日本語を母国語とする．

TED 講演を選んだ理由は，2 つ挙げられる．1 つ目は，リアルタイム性の高い動画であるからである．動画には講演者の身振りやスライドが含まれており，遅延時間が長くなると翻訳内容と講演者の身振りやスライドなどの表示内容とのずれが生じ易いため，同時音声翻訳の評価タスクに適している．2 つ目は，TED 講演が機械翻訳の性能を評価する際のテストセットとして頻繁に使用されているからである．

3.1 項のもと，20 秒から 30 秒程度の動画を 20 種類用意し，無作為に選ばれた翻訳結果と遅延時間を付加した字幕データを与えた．今回の評価データでは，1 動画の平均文数は約 4.45 文となった．動画にはスライドを含むもの（翻訳データと表示内容の遅延が分かりやすいもの）と，スライドを含まないものを 10 種類ずつ用いた．また，女性講演者の音声は女性が，男性講演者の音声は男性が収録を行い，同数ずつ用いた．

【翻訳データ】

翻訳データには，TED の字幕データ，S ランク（通訳経験年数 15 年）及び A ランク（通訳経験年数 4 年）[14] の同時通訳者が同時通訳を行なった際の書起しデータ，機械翻訳システム Travatar[10] 及び Moses [7] の翻訳データの 5 種類を用いた．

翻訳精度には，自動評価尺度 BLEU+1[8] 及び RIBES[6]，複数の参照訳を用いた BLEU+1 及び RIBES，人手評価の 5 つを用いた．参照訳は 5 人の翻訳者がそれぞれ翻訳を行ったものである．複数の参照訳を用いた理由は，自動評価における精度を改善するためである．人手評価には忠実性を 5 段階評価 [2] で被験者 3 人に評価を行ってもらい，その結果を加算平均し，0 から 1 の間になるように正規化する．自動評価を計算する際に，日本語の単語分割には KyTea[11] を使用した．参照訳には TED の字幕データとは異なる翻訳者の翻訳結果を用いた．各評価データの翻訳精度を表 1 に示す．

表 2: 音声提示とテキスト提示の結果得られた分類精度

素性	評価尺度	音声	テキスト [15]
遅延	-	0.54	0.67
精度	BLEU+1	0.55	0.50
	BLEU+1(multi)	0.53	-
	RIBES	0.61	0.44
	RIBES(multi)	0.57	-
	人手評価	0.70	0.71
遅延＋精度	BLEU+1	0.60	0.66
	BLEU+1(multi)	0.56	-
	RIBES	0.61	0.67
	RIBES(multi)	0.57	-
	人手評価	0.72	0.81

表 3: 音声提示時の重み $\boldsymbol{w}$ と重み $\boldsymbol{w}$ の比

評価尺度	$\boldsymbol{w}$		$\boldsymbol{w}$ の比	
	遅延	翻訳精度	平均	分散
BLEU+1	-0.013	2.031	154.8	540.9
BLEU+1(multi)	-0.017	0.598	35.7	43.6
RIBES	-0.018	1.508	86.6	170.6
RIBES(multi)	-0.018	1.32	65.6	175.4
人手評価	-0.018	1.988	114.2	324.7

【遅延時間】

遅延時間は秒単位で，$\boldsymbol{D} = \{0, 1, 2, 3, 5, 7, 10\}$ の 7 種類で与えた．3.2 項で示したように，今回の評価データにおいて遅延時間は発話開始からの時間とした．

【学習・評価】

学習器には LIBLINEAR[3] による線形 SVM を用いた．学習器の諸設定はデフォルトのままとした．

5.2 実験結果

ランキング学習の結果，得られた各分類精度を表 2，翻訳精度と遅延時間の重みとその比を 3 に示す．ここで，分類精度はランクの正解率を示しており，チャンスレートは 0.5 である．また，遅延時間及び翻訳精度の重みは，各動画の平均値を表しており，重み $\boldsymbol{w}$ の比は翻訳精度の重みを遅延時間の重みで割ったものである．なお，字幕提示により行われた文献 [15] の実験結果も表 2 に示すが，音声とテキストの違い以外にも参照訳の提示の有無や翻訳システムの種類も異なることに注意されたい．

まず，表 2 の結果より，どの素性を用いても分類精度がチャンスレートを上回ることが分かる．更に，音声データの実験おいても字幕データ [15] と同様に人手評価が素性として最も有効である事が確認された．BLEU+1 と人手評価を素性として用いた場合，遅延時間と翻訳精度を同時に素性とすることにより分類精度が若干上昇することが確認された．また，音声デー

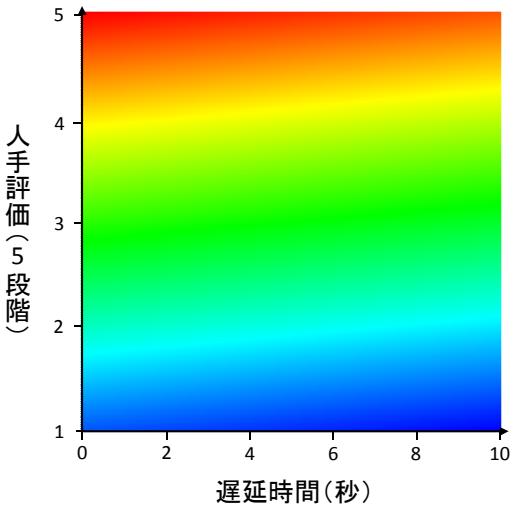

図 2: 評価関数によって得られたヒートマップ

タと字幕データの結果を比較すると，音声を用いた実験の方が翻訳精度の重みが大きく，相対的に重要であることが示唆された．

その一方，複数の参照文の結果では，期待していた自動評価における精度改善が確認されなかった．

表 3 より，$\boldsymbol{w}$ の比から人手評価の翻訳精度が 1 段階上がることは同じ翻訳システムに 28.55 秒の遅延時間を加えることと同じであると示している．今回の評価実験により得られた人手評価の翻訳精度を素性として導いた評価関数 s をヒートマップとして図 2 に示す．ヒートマップの左上に行くに従い評価スコアが高くなり，右下に向かうほど評価スコアが低くなっていることが分かる．先行研究 [15] と比較すると，ヒートマップの傾斜が緩やかである．これは，音声情報と視覚情報による理解過程の差異によるものである可能性が高いが，提示方法以外の条件を揃えた実験が今後の課題である．

6 おわりに

本研究では，同時音声翻訳システムの評価手法として，S2S 翻訳のための翻訳精度と遅延時間を同時に考慮した評価方法を提案した．その結果，遅延時間及び各評価尺度を素性に用いた場合，どの素性を用いても分類精度がチャンスレートを上回ることが確認された．また，先行研究 [15] と比べて，音声データを用いた場合，字幕データを用いた場合と比べて翻訳精度が遅延時間よりも重要であることが示唆された．今後の課題としては，音声とテキストを同等の条件での実験，話速や声色，イントネーションなどの影響を調査，自動評価における精度改善，非線形なモデルへの適用などが挙げられる．

7 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] S. Bangalore, V. K. R. Sridhar, P. K. L. Golipour, and A. Jimenez. Real-time incremental speech-to-speech translation of dialogs. In *Proc. NAACL*, 2012.

[2] DARPA. Linguistic Data Annotation Specification:Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations. 2002.

[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9, 2008.

[4] E. Franco, A. Matamala, and P. Orero. Voice-over translation. Peter Lang Pub Inc, 2013.

[5] T. Fujita, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. Simple, lexicalized choice of translation timing for simultaneous speech translation. In *Proc. 14th InterSpeech*, 2013.

[6] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. EMNLP*, pp. 944–952, 2010.

[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, pp. 177–180, 2007.

[8] C.-Y. Lin and F. J. Och. A method for evaluating automatic evaluation metrics for machine translation. In *Proc. COLING*, pp. 501–507, 2004.

[9] E. Matusov, A. Mauser, and H. Ney. Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proc. IWSLT*, pp. 158–165, 2006.

[10] G. Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proc. ACL*, pp. 91–96, 2013.

[11] G. Neubig, Y. Nakata, and S. Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. ACL*, pp. 529–533, 2011.

[12] Y. Oda, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. Optimizing segmentation strategies for simultaneous speech translation. In *Proc. ACL*, 2014.

[13] K. Ryu, A. Mizuno, S. Matsubara, and Y. Inagaki. Incremental Japanese spoken language generation in simultaneous machine interpretation. In *In Proc. Asian Symposium on Natural Language Processing to Overcome language Barriers*, 2004.

[14] H. Shimizu, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. Constructing a speech translation system using simultaneous interpretation data. In *Proc. IWSLT*, 2013.

[15] 三重野, G. Neubig, S. Sakti, 戸田, 中村. 同時音声翻訳における翻訳精度と遅延時間を同時に考慮した評価尺度. 情報処理学会 第 219 回自然言語処理研究会 (SIG-NL), 東京, 12 2014.