通訳品質評価に関するデータ収集と分析

安田圭志[1] 菅谷史昭[1] 池田美紀子[2] 米澤早紀恵[2] 隅田英一郎[3]
[1]マインドワード株式会社 [2]株式会社インターグループ [3]情報通信研究機構
{ke-yasuda,fsugaya }@mindword.jp
{mikeda,s-yonezawa}@intergroup.co.jp
eiichiro.sumita@nict.go.jp

概要

本論文では、通訳者による翻訳の訳質を評価することを目的とし、実際の通訳者および、通訳学習者による通訳結果のデータ収集した後、実施した主観評価を説明する。次に、主観評価のコストを低減することを視野にいれ、機械翻訳の研究分野で利用される翻訳自動評価法による自動評価値と、主観評価値との関係性についての分析を行なう。

1 はじめに

近年の人工知能技術の発展により、機械翻訳性能が著しく向上し、一般の業務などで機械翻訳システムを利用するユーザが増加している。一方、翻訳誤りが許容されないような状況においては、いまだ人手による翻訳や通訳が行われている。将来的には、このような人手による翻訳および通訳業務に対するニーズは相対的に低くなっていくことが予測されるものの、これらの業務が、完全に機械に置き換えられることは現時点では想定できない。

ここで、人手による通訳について注目すると、今後、期待されるのは、音声翻訳システムよりも優れた訳質を維持することである。通訳の質を担保する上では、通訳結果の評価が重要である。すでに、旅行分野の通訳においては多肢選択問題で構成される全国通訳案内士の資格試験が存在するが、通訳が必要とされる専門分野ごとに多肢選択試験を作成するのは、コスト面の課題がある。

本研究では、このような課題を解決するため、講演音声を通訳評価用の試験問題として利用する方法を採用した。次に、この方法に基づいて、実際の通訳データを収集し、評価指標を定義した上で主観評価を実施した。最後に、主観評価のコストを減らすことを目的とした将来的な評価自動化を見据えて、主観評価値と、各種翻訳自動評価法による自動評価値との関連性についての分析を行なっている。

2 通訳データ収集

ここでは、本研究で収集した同時通訳データおよび、主観評価データについて述べる。

2.1 同時通訳データ

24 発話からなる英語の模擬講演音声を 48 名の通訳学習者に提示し、日本語への同時通訳作業として日本語音声で発話させることによりデータを収集した。

2.2 主観評価データ

主観評価実施にあたり、前述の模擬講演音声と、通訳学習者による日本語音声に対して、人手による書き起こしを行なった。この原言語と目的言語のペアに対して、7 名のプロの通訳者が主観による評価を実施している。ここで、各評価者は、表 1 に示す 5 つの各評価項目に対して、評価を実施している。「的確さ」を除く 4 つの評価項目においては、以下の 4 段階の基準で 0〜4 ポイントの減点評価を行っている。

- スピーカーの意図は過不足なく伝わり、解釈に影響なし (0)
- 不自然な個所や多少の誤用/抜けはあるものの、解釈に大きな影響はない(大筋の理解には影響なし) (-1)
- 解釈が揺れる(文脈から理解は可能だが、聞き手によっては解釈を誤る場合がありえる、微妙なニュアンス) (-2)
- 解釈の妨げになる/一部解釈を誤る/欠落する (-3)
- 解釈を誤る(致命的なミス) (-4)

また、「的確さ」については、加点箇所1か所につき 1 ポイントの加算を行っている。

最小の評価単位については、「訳抜け」の場合のみ、評価の最小単位となるチャンク分割を、評価者に提示している。「訳抜け」以外の評価項目につい

表 1　主観評価における評価項目

評価項目	評価項目詳細	最小評価単位あたりのスコアレンジ	最小評価単位	総評価数	1発話あたりの平均評価数
構文・文法	文法/構文の解釈ミス／使用ミスが原因での誤訳を評価	-4～0	構文の場合は文全体、文法の場合は２～５単語程度（評価者による判断）	7,879	0.977
語彙（情報の正確さ）	スピーカーの述べている情報が誤って伝わる訳となっている	-4～0	単語あるいは３～5単語程度のフレーズ（評価者による判断）	19,053	2.363
語彙（運用の正確さ）	目的言語としてその文脈内での単語・表現の使い方が誤っている	-4～0	単語あるいは３～5単語程度のフレーズ（評価者による判断）	4,056	0.503
訳抜け	原文の意図の伝達に影響をおよぼす情報が抜けている場合のみ減点	-4～0	１問約40単語程度を4～7程度のチャンクに分割（共通のチャンク分けを提示）	21,274	2.638
的確さ	訳語選択、訳処理として優れていると感じた箇所は、1か所につき1点加点	0～1	語彙あるいは文全体（評価者の判断）	333	0.041

表 2　評価者間の相関係数

	評価者1	評価者2	評価者3	評価者4	評価者5	評価者6	評価者7
評価者1	1.000						
評価者2	*0.645*	1.000					
評価者3	0.918	*0.538*	1.000				
評価者4	0.931	*0.628*	0.914	1.000			
評価者5	0.903	*0.563*	0.885	0.890	1.000		
評価者6	0.879	*0.517*	0.896	0.884	0.904	1.000	
評価者7	0.902	*0.685*	0.900	0.888	0.870	0.857	1.000

ては、評価の最小単位は、各評価者の判断により決定している。表 1 の総評価数は、全ての評価者が、24 名の通訳者に対して、減点あるいは加点した箇所の合計である。平均評価数は、総評価数を、8,064(＝7 評価者×24 発話×48 通訳者)で割った値であり、1 発話あたりの、減点（あるいは加点）された箇所の平均数である。各発話において、減点（あるいは加点）された箇所が無い場合は 0 のカウントとなるため、評価項目によっては、1 発話あたりの平均評価数が 1 未満の値を取る場合もある。

スコアの集計方法としては、各通訳者の全発話における全評価箇所に対する 5 つのスコアを合計した値を各通訳者の評価値としている。「的確さ」以外の評価項目が減点評価であるため、この合計値は、全通訳者においてマイナスの値となった。

7 名の各評価者は、評価作業を分担するのではなく、各評価者が全評価対象データ（全通訳者 48 名×24 発話）の評価を行っている。このため、各通訳者に対して 7 名による評価結果が得られることになる。表 2 に、7 名の評価者における、通訳者評価結果間の相関係数を示す。表 2 に示す通り、評価者 2 を除き、評価者間で 0.85 以上の相関係数が得られている。評価者 2 の評価結果についてみると、語彙に関する評価において、他の評価者よりも細かい粒度での評価を行なっている傾向があり、このことが相関係数低下の原因であると考えられる。

3 自動評価実験

2.2 で述べた主観評価は、評価実施において非常に大きなコストがかかる。ここでは、評価の自動化についての検討をするため、機械翻訳分野で利用されている翻訳自動評価法による自動評価値と、主観評価値との関連性についての分析を行なう。

3.1 主観評価値と自動評価値の相関

2.2 で述べた通り、各通訳者に対して、7 名の評価者による評価結果が得られているが、ここでの分析は、7 名の評価結果の合計値を主観評価値とし、自動評価値との比較を行っている。

翻訳自動評価法としては、BLEU[1], REOUGE[2] (REOUGE-1, REOUGE-2, REOUGE-L)、BERT Score (Precision, Recall, F1)[3]を分析の対象とした。自動評価で必要となる参照訳は、各発話に対して 7 名分作成し評価に用いた。

表 3　自動評価値と主観評価値間の相関

	構文	語彙 （情報の正確さ）	語彙 （運用の正確さ）	訳抜け	的確さ
構文	1.000				
語彙 （情報の正確さ）	***0.685***	1.000			
語彙 （運用の正確さ）	-0.012	0.222	1.000		
訳抜け	0.205	0.275	-0.297	1.000	
的確さ	0.458	0.520	0.153	0.427	1.000
ROUGE-1	0.529	0.720	0.171	0.616	0.494
ROUGE-2	0.446	0.630	0.061	0.626	0.443
ROUGE_L	0.544	**0.724**	0.159	0.640	0.523
BLEU	0.538	0.646	0.057	0.421	0.275
BERT score (Precision)	0.599	0.703	**0.249**	0.342	0.418
BERT score (Recall)	0.530	0.579	-0.109	**0.873**	**0.589**
BERT score (F1)	**0.620**	0.705	0.079	0.648	0.541

表 3 に各自動評価値と、主観評価値との相関行列を示す。ここでは、表 2 の場合と異なり、各通訳者の 5 つの個別の評価項目と自動評価値との相関を求めている。主観による評価項目間の相関は比較的低く、最も高い場合でも、文法 - 語彙（情報の正確さ）間の 0.685 であった。一方、主観評価値と自動評価値間の相関係数は、比較的高く、最も高い場合で、訳抜け - BERT Score (Recall)　間の 0.873 であった。

3.2 重回帰分析

各自動評価値が、どの主観評価項目に重点をおいた評価になっているのかを明らかにするため、自動評価値を目的変数、主観評価における 5 つの評価項目を説明変数とした重回帰分析を実施した。

表 4 に重回帰分析の結果を示す。ここでは、主観評価結果として、全評価者を用いた場合と、他の評価者との相関が低かった評価者 2 を除く 6 名の評価者を用いた場合で分析を行っている。

7 名の評価者を用いた場合において、重相関係数は、BERT Score (Recall)が最も高い値をとっており、0.961 であった。この場合の重回帰係数は、「訳抜け」に対する値が最も大きくなっている。BERT Score (Recall)と同様に、参照訳を正解とした Recall をもとにスコア付けを行う REOUGE-L においても同様の傾向が見られている。これらの Recall に基づく評価は、参照訳からの情報の欠落を評価していることから、「訳抜け」に対する重回帰係数が高くなっていると考えられる。BERT Score 間で比較すると、Precision においては、語彙(情報の正確さ)に対する重回帰係数が最も高くなっている。F1 においては、Precision と Recall の中間的な重みの配分なっている。

評価者 2 を除く 6 名の評価者を用いた結果について見ると、前述の Recall に基づく自動評価指標においては、重相関係数および、重回帰係数の大きな変化は見られなかった。一方、BERT Score (Precision)および、BLEU といった Precision に基づく自動評価への影響がでている。Precision に基づく評価は、評価対象の翻訳に、どの程度正しい情報が含まれているかを評価しており、主観評価項目「語彙(情報の正確さ)」との関連性が高いと考えられる。評価者２の評価結果をみると、特に「語彙(情報の正確さ)」の評価項目に対する評価を多くつける傾向が強かったことから、評価者 2 を除くことが、Precision に基づく自動評価に対する重回帰分析結果に大きな影響を与えたと考えられる。

表 4　重回帰分析の結果

分析に利用した評価者数	目的変数（スコア自動評価値）	決定係数	重相関係数	重回帰係数				
				構文	語彙（情報の正確さ）	語彙（運用の正確さ）	訳抜け	的確さ
全7名の評価者	BLEU	0.709	0.842	0.248	**0.538**	0.094	0.335	*-0.163*
	ROUGE-L	0.796	0.892	0.222	0.337	0.191	**0.597**	*-0.048*
	BERT score (Precision)	0.744	0.863	0.260	**0.561**	0.159	0.203	*-0.046*
	BERT score (Recall)	**0.923**	**0.961**	0.262	0.047	0.114	**0.806**	*0.018*
	BERT score (F1)	0.835	0.914	0.294	0.343	0.147	**0.529**	*-0.022*
評価者2を除く6名の評価者	BLEU	0.579	0.761	0.338	0.374	0.152	**0.408**	*-0.273*
	ROUGE-L	0.792	0.890	0.263	0.317	0.273	**0.627**	*-0.136*
	BERT score (Precision)	0.624	0.790	**0.458**	0.279	0.249	0.273	*-0.135*
	BERT score (Recall)	**0.926**	**0.962**	0.290	0.086	0.087	**0.779**	*0.037*
	BERT score (F1)	0.795	0.892	0.416	0.204	0.189	**0.560**	*-0.061*

4 まとめと今後の検討課題

人手による通訳の評価に関して、5 つの評価項目を定義し、実際の通訳学習者による通訳データを収集し、主観評価を実施した。次に、翻訳自動評価指標との関連性を明らかにするため、相関分析と重回帰分析を実施した。

実験の結果、BERT Score (Recall)に対する重相関係数が 0.9 以上の高い値となった。また、BERT Score (Recall)は、主観評価における評価項目「訳抜け」に重点をおいた評価となっていることが明らかとなった。

今回は、講演を単位とした評価を実施したが、今後は、文や発話といったより細かい単位での評価に対する分析を実施したい。また、今後の評価の自動化のために、今回実施した重回帰分析とは逆に、主観評価値を目的変数とし、種々の自動評価値から主観評価結果を予測する方法についても、今後検討していきたい。

謝辞

本件は、総務省の「ICT 重点技術の研究開発プロジェクト (JPMI00316)」における「多言語翻訳技術の高度化に関する研究開発」による委託を受けて実施した研究開発による成果です。

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318.

[2] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Association for Computational Linguistics.

[3] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberge and Yoav Artz. 2020. BERTSCORE: EVALUATING TEXT GENERATION WITH BERT. In Proceedings of the 8th International Conference on Learning Representations