自動評価法におけるメタ評価と特許データを用いた性能評価

―WMT2018 評価タスクと WAT2017 データに基づく自動評価法の現状について―

Meta-evaluation and experiments using patent data in automatic evaluation metrics

北海学園大学大学院工学研究科教授

越前谷　博

1996 年北海学園大学大学院工学研究科修士課程修了。博士（工学）。2013 年～現在北海学園大学大学院工学研究科教授。機械翻訳の研究に従事。アジア太平洋機械翻訳協会（AAMT）/Japio 特許翻訳研究会委員。

echi@lst.hokkai-s-u.ac.jp　TEL 011-841-1161（内線 :7863）

1 はじめに

機械翻訳研究はこれまでいくつかのブレイクスルーを経て現在に至っている。統計的機械翻訳およびニューラル機械翻訳は機械翻訳研究に大きな転換をもたらした。一方、このようなブレイクスルーを支える技術として自動評価法は不可欠であり、機械翻訳技術の進展と共に発展し続けている。本稿では機械翻訳研究に必須である自動評価法が現在どの程度の性能を有しているのか を WMT2018（Third Conference on Machine Translation）の評価タスク[1] に基づいて概観すると共に、特許データにおける自動評価法の有効性につい て WAT2017（The 4th Workshop on Asian Translation）[2] データに基づき述べる。

2 WMT2018 評価タスクにおける自動評価法のメタ評価

2.1 メタ評価の方法と結果

WMT2018 の評価タスクに基づく自動評価法のメタ評価の方法と結果について述べる。自動評価法の多くは、機械翻訳システムの訳文と人手による正解訳である参照訳を比較することによりスコアを算出する。そして、そのスコアが機械翻訳システムの訳文に対する評価結果となる。WMT では自動評価法のメタ評価は自動評価法が算出したスコアと人手評価値との間の相関係数を求めることで行う。自動評価法のスコアと人手評価値が近いほど相関係数は高くなり、高い評価性能を有する自動評価法と位置付けられる。

また、WMT では自動評価法のメタ評価は複数文を一括で評価した場合のシステムレベルのメタ評価と一文ごとの評価によるセグメントレベルのメタ評価の２種類が行われている。システムレベルの相関係数については多くの自動評価法が 0.9 を超えており、評価精度は高い。それに対してセグメントレベルの相関係数は十分に信頼できる評価精度を示すまでには至っていない。したがって、本稿ではセグメントレベルのメタ評価結果についてのみ言及する。表１に多言語から英語方向の訳文を用いたセグメントレベルのメタ評価結果を示す。また、表２には英語から多言語方向の訳文を用いたセグメントレベルのメタ評価結果を示す。

表１と表２の en は英語、cs はチェコ語、de はドイツ語、et はエストニア語、fi はフィンランド語、ru はロシア語、tr はトルコ語、そして、zh は中国語を表している。相関係数は自動評価法による２つのスコアと人手評価による２つの値の大小関係が等しいか異なるかをカウントした相対評価に基づく Kendall's τ を用いて得る。そのため表中の“Human Evaluation”の“DARR”は相対評価であることを意味する。また、n は比較したスコアのペア数を示している。各言語ペアと “Avg.”（全言語ペアの相関係数の平均）に存在する太字の数値は全自動評価法における最大値である。

2.2 使用した自動評価法

ここで表１と表２に登場する自動評価法について述べる。本稿では著者らが提案している自動評価法 IMPACT（Intuitive comMon PArts ConTinuum）[3] および WE_

表1　多言語から英語方向の訳文を用いたセグメントレベルのメタ評価結果

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	Avg.
Human Evluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR	
n	5,110	77,811	56,721	15,648	10,404	8,525	33,357	
Correlation	τ	τ	τ	τ	τ	τ	τ	
WE_WPI_fastText	0.298	0.462	0.329	0.210	0.270	0.230	0.205	0.286
WE_WPI_BERT_LAY1	0.286	0.448	0.323	0.207	0.254	0.213	0.204	0.276
WE_WPI_fastText_BERT	0.305	0.471	0.343	0.224	0.280	0.228	**0.225**	0.297
IMPACT	0.258	0.443	0.302	0.161	0.248	0.166	0.194	0.253
BEER	0.295	0.481	0.341	0.232	0.288	0.229	0.214	0.297
BLEND	0.322	0.492	0.354	0.226	0.290	0.232	0.217	0.305
CHARACTER	0.256	0.450	0.286	0.185	0.244	0.172	0.202	0.256
CHRF	0.288	0.479	0.328	0.229	0.269	0.210	0.208	0.287
CHRF++	0.288	0.479	0.332	0.234	0.279	0.218	0.207	0.291
ITER	0.198	0.396	0.235	0.128	0.139	-0.029	0.144	0.173
METEOR ++	0.270	0.457	0.329	0.207	0.253	0.204	0.179	0.271
RUSE	**0.347**	**0.498**	**0.368**	**0.273**	**0.311**	**0.259**	0.218	**0.325**
SENTBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178	0.234
UHH_TSKM	0.274	0.436	0.300	0.168	0.235	0.154	0.151	0.245
YISI-0	0.301	0.474	0.330	0.225	0.294	0.215	0.205	0.292
YISI-1	0.319	0.488	0.351	0.231	0.300	0.234	0.211	0.305
YISI-1_SRL	0.317	0.483	0.345	0.237	0.306	0.233	0.209	0.304
newstest2018								

表2　英語から多言語方向の訳文を用いたセグメントレベルのメタ評価結果

	en-cs	en-de	en-et	en-fi	en-ru	en-tr	en-zh	Avg.
Human Evluation	DARR	DARR	DARR	DARR	DARR	DARR	DARR	
n	5,413	19,711	32,202	9,809	22,181	1,358	28,602	
Correlation	τ	τ	τ	τ	τ	τ	τ	
WE_WPI_fastText	0.472	0.644	0.531	0.488	0.382	0.387	0.331	0.462
WE_WPI_BERT_LAY1	0.468	0.633	0.471	0.452	0.349	0.330	0.347	0.436
WE_WPI_fastText_BERT	0.488	0.662	0.539	0.515	0.389	0.380	**0.351**	0.475
IMPACT	0.416	0.643	0.445	0.404	0.342	0.302	0.322	0.411
BEER	**0.518**	0.686	0.558	0.511	0.403	0.374	0.302	0.479
BLEND	-	-	-	-	0.394	-	-	-
CHARACTER	0.414	0.604	0.464	0.403	0.352	0.404	0.313	0.422
CHRF	0.516	0.677	0.572	0.520	0.383	0.409	0.328	0.486
CHRF++	0.513	0.680	**0.573**	**0.525**	0.392	0.405	0.328	**0.488**
ITER	0.333	0.610	0.392	0.311	0.291	0.236	-	-
SENTBLEU	0.389	0.620	0.414	0.355	0.330	0.261	0.311	0.383
YISI-0	0.471	0.661	0.531	0.464	0.394	0.376	0.318	0.459
YISI-1	0.496	0.691	0.546	0.504	**0.407**	**0.418**	0.323	0.484
YISI-1_SRL	-	**0.696**	-	-	-	-	0.310	-
newstest2018								

WPI（**W**ord **E**mbedding-based automatic MT evaluation using **W**ord **P**osition **I**nformation）[4] について紹介する。

2.2.1　表層情報に基づく自動評価法：IMPACT

著者らが 2007 年に提案した自動評価法 IMPACT はトレーニングおよび言語知識を必要としない表層情報に基づく自動評価法である。ルールベースの機械翻訳や統計的機械翻訳の評価を目的としており、自動評価法としては初期に提案されたものの一つである。自動評価法全般における位置付けとしては、表層情報のみに基づいており、静的な言語情報を必要としないため特定の言語に依存することなく、どのような言語であっても容易に

評価可能であることが特徴である。すなわち、単語分割された訳文と参照訳のみがあればスコアを容易に算出することができる。

技術的には IMPACT は単語レベルの最長共通部分列（LCS）に基づいてスコアを算出する。LCS を用いることで訳文と参照訳との間で表層的に一致している単語列を決定することができる。しかし、LCS では得られる単語列は同じ並びで出現するものに限定される。つまり、一致単語列が左右交互に出現する場合には左右いずれかの一致単語列は無視される。そのため LCS は語順の違いに過度に厳しい。そこで IMPACT では語順の違いに柔軟に対応可能な仕組みを取り入れている。具体的には左右交互に位置する一致単語列についてはどちらかを完全に無視するのではなく、一致単語列の出現位置の情報をパラメータとすることで、どの程度出現位置の異なる一致単語列をスコアに反映させるかを制御する。その結果、語順の違いに柔軟な自動評価法を実現している。

表 1 と表 2 における IMPACT の相関係数は他の自動評価法に比べて低い。“Avg.”のみを比較した場合、順位は下位に位置している。IMPACT は特定の言語に依存しないことから容易にスコアを得ることができる反面、単語の意味情報を使用していないため評価精度は十分とはいえない。しかし、同様のアプローチである SENTBLEU に比べると高い相関係数を示しており、迅速に評価を行いたい場合には有効な自動評価法である。

2.2.2　単語の意味に基づく自動評価法：WE_WPI

著者らが 2019 年に提案した自動評価法 WE_WPI は EMD（Earth Mover’s Distance）[5] に基づいている。EMD は輸送問題の最適解を求めるタスクとして 2 つの分布間の距離を計算する。それぞれの分布は複数の特徴量で構成されており、この特徴量を訳文と参照訳の単語と捉えることでスコア計算を行う。また、特徴量は輸送問題の観点より重み（荷物）を有しており、それを運ぶための作業量は輸送量と分布間の距離との積で表される。そのため、EMD を自動評価法に適用するにあたっては特徴量、重み、そして、距離を定義する必要がある。WE_WPI では特徴量として単語の意味に相当する単語分散表現、重みに文レベルの tf・idf、そして、距離計算にはコサイン距離を用いている。重みに文レベルの tf・idf を用いる理由は機能語と内容語を差別化するためである。また、WE_WPI では語順をスコアに反映させるためにコサイン距離に対して単語の出現位置の相対的なずれを負の重みとして用いている。このように単語の出現位置の情報を EMD に適用することは非常に有効であり、WE_WPI の評価精度を大幅に向上させた。

表 1 と表 2 において WE_WPI に基づく自動評価としては単語分散表現モデルの利用の観点より、WE_WPI_fastText、WE_WPI_BERT_LAY1、そして、WE_WPI_fastText_BERT の 3 つを用いた。WE_WPI_fastText は単語の分散表現を得る際に事前学習された fastText のモデルを使用している。WE_WPI_BERT_LAY1 は事前学習された BERT のモデルを使用している。その際には最終層から抽出した分散表現を用いている。WE_WPI_fastText_BERT は単語間の距離を求める際に fastText から抽出した単語分散表現より得られるコサイン距離と BERT から抽出した単語分散表現より得られるコサイン距離の積を最終的な距離として用いている。

WE_WPI に基づく自動評価法（WE_WPI_fastText、WE_WPI_BERT_LAY1、WE_WPI_fastText_BERT）は単語分散表現モデルを用いているが、それらは全て事前学習されたモデルであり、サイト上で無償提供されているため各自で構築する必要はない。また、文レベルの tf・idf も訳文と参照訳のみから求めることができる。したがって、比較的容易にスコアの算出を行える点が提案手法の大きな利点である。

表 1 と表 2 より、WE_WPI に基づく自動評価法においては WE_WPI_fastText_BERT が最も高い相関係数を示した。また、他の自動評価法との比較においては、表 1 より多言語から英語方向への訳文の評価においては“Avg.”は 17 の自動評価法において 5 番目であり、上位に位置している。表 2 より英語から多言語方向においては“Avg.”は 11 の自動評価法において 5 番目であり、中間に位置している。しかし、最も高かった CHRF++ との“Avg.”の差は 0.013 と小さいことから評価精度に大きな差はない。また、言語ペア別にみると WE_WPI_fastText_BERT は表 1 と表 2 共に英語と中国語の言語ペア（zh-en、en-zh）において他の自動評価法に比べて最も高い相関係数を示した。

3 WAT2017の特許データにおける自動評価法の性能

特許データにおける自動評価法の性能について考察する。本稿では、WAT2017におけるJPC-EJ（English-Japanese JPO Patent Corpus）の英文をベースラインシステム（Phrase-based SMT）とニューラル機械翻訳システムに翻訳させ、その訳文（日本文）を表層情報に基づく自動評価法IMPACTと単語の意味に基づく自動評価法WE_WPI_fastText_BERTに評価させた。訳文は400文であり、それに対応する参照訳もまた400文である。評価精度の計算はベースラインとニューラル機械翻訳システムの訳文に対する2つのセグメントレベルのスコアをペアとしたペアワイズ評価に基づいて行った。その結果、IMPACTとWE_WPI_fastText_BERTの評価精度はそれぞれ15.7%と24.1%となり、IMPACTよりもWE_WPI_fastText_BERTの方が高い評価精度を示すことが確認された。図1にIMPACTとWE_WPI_fastText_BERTの評価結果の具体例を示す。

図1よりIMPACTのスコアはベースラインの訳文に対しては0.578、ニューラル機械翻訳の訳文に対しては0.570であり、ベースラインの訳文に対する評価がニューラル機械翻訳の訳文よりも若干高い。しかし、人手評価ではニューラル機械翻訳の訳文の方が良いという結果であり、IMPACTの評価結果とは異なっている。それに対して、WE_WPI_fastText_BERTではベースラインとニューラル機械翻訳のスコアはそれぞれ0.660と0.725であり、ニューラル機械翻訳の方がスコアは高く、人手評価と一致している。IMPACTは語順を考慮した自動評価法ではあるが、ベースラインの訳文と参照訳との間では構成単語の多くが一致するためベースラインの訳文のスコアが若干高くなったと考えられる。また、ニューラル機械翻訳の訳文中の“を”、“す”は参照訳中の“は”、“される”とそれぞれ表層レベルでは一致しないためニューラル機械翻訳の訳文のスコアを低下させる要因になったと考えられる。一方、WE_WPI_fastText_BERTは表層レベルで一致していない単語間であっても意味レベルの類似性がスコアに反映されるため人手評価の結果と一致したと考えられる。

4 おわりに

本稿では自動評価法の現状について、WMT2018の評価タスクによるメタ評価とWAT2017の特許データを用いた性能評価実験により述べた。著者らが提案する、単語の意味に基づく自動評価法WE_WPI_fastText_BERTを通して意味に相当する分散表現の利用が自動評価法において有効となることを示した。また、特許データにおいても表層情報に基づく自動評価法IMPACTよりも単語の意味に基づく自動評価法WE_WPI_fastText_BERTの方が高い評価精度を有していることを示した。現在の自動評価法はセグメントレベルにおいては人手評価との相関係数は高いとはいえず課題が残されているが、着実に進歩していることも事実である。また、近年では参照訳を用いない手法[6]も提案されており、新たなアプローチから自動評価の研究が行われている。その結果、評価精度や操作性などユーザーが何を優先するかによって使い分けることが可能となる。

今後は評価精度の向上のための研究を進めると共に、機械翻訳システムの開発に向けてより有効利用できるように、スコアの出力だけでなくスコアに対して根拠となる説明も付与可能な自動評価法の研究を行う予定である。

参考文献

[1] Qingsong Ma, Ondřej Bojar and Yvette

入力文	The photoresist formulations for the lithographic evaluation are shown in Table 3 , below .	IMPACTのスコア	WE_WPI_fastText_BERTのスコア	人手評価の比較結果
参照訳	リソグラフィ評価のためのフォトレジスト配合物は以下の表3に示される。			
ベースラインの訳文	フォトレジストは、リソグラフィのための配合物の評価は以下の表3に示す。	0.578	0.660	*loss*
NMTの訳文	リソグラフィ評価のためのフォトレジスト製剤を以下の表3に示す。	0.570	0.725	*win*

図1　IMPACTとWE_WPI_fastText_BERTの評価結果の具体例

Graham : Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance, Proceedings of the Third Conference on Machine Translation (WMT) , Volume 2: Shared Task Papers, October - November 2018, pages 682–701.

[2] Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig and Sadao Kurohashi : Overview of the 4th Workshop on Asian Translation, Proceedings of the 4th Workshop on Asian Translation, November 2017, pages 1–54.

[3] Hiroshi Echizen-ya and Kenji Araki : Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, Proceedings of the Eleventh Machine Translation Summit (MT SUMMIT XI), September 2007, pages 151-158.

[4] Hiroshi Echizen' ya, Kenji Araki and Eduard Hovy : Word Embedding-Based Automatic MT Evaluation Metric using Word Position Information, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT 2019), Volume 1, June 2019, pages 1874-1883.

[5] Yossi Rubner, Carlo Tomasi and Leonidas J. Guibas : A Metric for Distributions with Applications to Image Databases, Proceedings of the Sixth International Conference on Computer Vision (ICCV), January 1998, pages 59–66.

[6] 嶋中宏希，梶原智之，小町守：事前学習された多言語の文符号化器を用いた機械翻訳の品質推定，言語処理学会第 26 回年次大会発表論文集，2020 年 3 月，pages 913-916.