機械翻訳結果の自動評価値を用いて人手評価値を予測する

To predict human evaluation score of a machine translation result by automatic evaluation score

元山梨英和大学教授
江原　暉将

1967 年早稲田大学理工学部卒。同年 NHK 入局。2003 年諏訪東京理科大学教授。2009 年山梨英和大学教授。2015 年退職。アジア太平洋機械翻訳協会（AAMT）／Japio 特許翻訳研究会委員。

1　はじめに

機械翻訳の精度が向上するにつれて、翻訳結果の評価は、研究の観点からだけでなく実用の観点からも重要になってきている。翻訳は極めて人間的な作業であるため、機械で自動評価をすることが難しく、最終的には人手評価に頼らざるを得ない現状である。しかし人手評価にはコストも時間もかかるため、何とか精度の良い自動評価手法を得ようと盛んに研究が行われている [1][2]。本文では、自動評価結果に基づいて人手評価結果を推定する手法について考察する。自動評価は、まだまだ精度が低いが、現状でどの程度の誤差で人手評価結果を推定できるかを調べることが目的である。

機械翻訳の評価には、大きくシステムレベルの評価と文レベルの評価がある。システムレベルの評価とは、複数の機械翻訳システムがあった場合に、どの機械翻訳システムが精度が良いかを評価するものである。一方、文レベルの評価は、具体的な翻訳対象文 [1] をある機械翻訳システムで翻訳した場合、どの程度精度良く翻訳できたかを評価するものである。システムレベルの評価であれば、多数の試験文を翻訳させて、それらの良否を総合的に判断できるが、文レベルの評価では個々の文ごとの評価になる。従ってシステムレベルの評価より文レベルの評価のほうが難しい。

1　翻訳対象は「文」である必要はない。例えば特許のタイトルなどは「名詞句」が多い。そのため、正確には「文レベル」ではなく、「セグメントレベル」という用語を用いるが、ここでは分かりやすい「文レベル」という言い方をする。

表 1　内容の伝達レベルの評価

評点	説明
5	すべての重要情報が正確に伝達されている。（100%）
4	ほとんどの重要情報は正確に伝達されている。（80% ～）
3	半分以上の重要情報は正確に伝達されている。（50% ～）
2	いくつかの重要情報は正確に伝達されている。（20%～）
1	文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。（～20%）

特許文書の機械翻訳結果の人手評価については、文献 [3] で詳細に検討され、「内容の伝達レベルの評価」が提案されている [2]。これは、機械翻訳結果が、原文の内容をどの程度正確に伝達しているかを、【表 1】に示す 5 段階の評価基準で主観的に評価するものである。【表 1】を用いて文レベルで評価を行うものであるが、複数の文に対して内容の伝達レベルの評価を行うことで、システムレベルの評価にも用いることができる。

一方、自動評価手法としては、多くのものが提案されているが、本文では、代表的な評価基準である BLEU[4]、RIBES[5]、IMPACT[6] の三種について考察する。BLEU は自動評価基準として、最初に提案され、その後も標準的な評価手法として広く用いてられている。RIBES と IMPACT は、我が国の研究者によって考案された評価基準であり、英語と日本語あるいは中国

2　後述する WAT2016 では内容の伝達レベルの評価を JPO adequacy と呼んでいる。本文では更に簡略に adequacy と呼ぶことがある。

語と日本語など語順の大幅に異なる言語間の翻訳結果の評価に適しているといわれている。BLEU、RIBES、IMPACT ともに理想的な翻訳とみなせる参照翻訳文と機械翻訳文を比較することで評価を行う。評価値は 0 から 1 の範囲内にあり、値が大きいほど評価が高い。

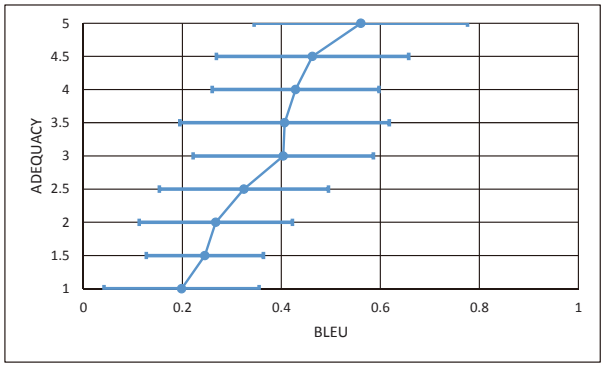
2 自動評価値と人手評価値の関係

自動評価値から人手評価値を推定するにあたって、まず同一の機械翻訳文に対する人手評価値と自動評価値の関係を調査する。このようなデータとして WAT2016（Workshop on Asian Translation, 2016）での評価結果データを利用する [7]。WAT2016 では、特許文書の機械翻訳タスクとして英日、中日、韓日とその逆方向の計 6 タスクが行われ、人手評価として、pairwise evaluation と JPO adequacy（内容の伝達レベルの評価）が用いられた。参加者が提出した機械翻訳結果のうち pairwise evaluation での上位 3 位の結果に対して、JPO adequacy での評価が行われた。JPO adequacy の評価者は翻訳の専門家であり 2 名である。人手評価の文数は各タスクごとに 200 文である。本文では人手評価結果として JPO adequacy を用いる。

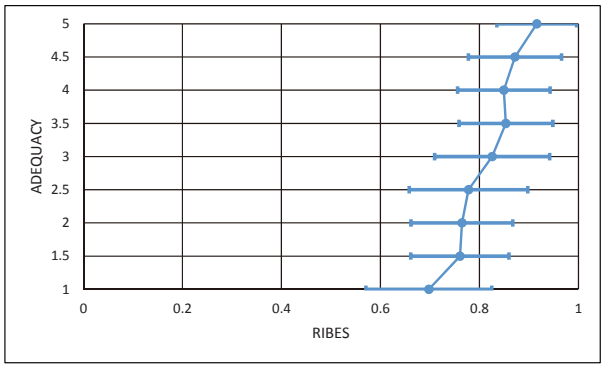
【図 1】に中日の翻訳結果に対する自動評価結果と人手評価結果の関係を示す。人手評価は 1 から 5 の 5 段階評価であるので 2 名の平均値は 1，1.5，2，2.5，3，3.5，4，4.5，5 の 9 個の値をとりうる。人手評価の各値に対して、その値をとるデータに対する自動評価値の平均値の変化を【図 1】に示す。【図 1】には標準偏差の範囲も示している。

人手評価値の 5 と 1 の間で自動評価値の平均値の差は BLEU、IMPACT、RIBES の順に小さくなるが、標準偏差の値も、その順に小さくなっている。そこで人手評価値が 5 の平均値から 1 の平均値を引いた値（DA）と各人手評価値に対する標準偏差の平均値（SD）および SD を DA で割った値（SD/DA）を【表 2】に示す。

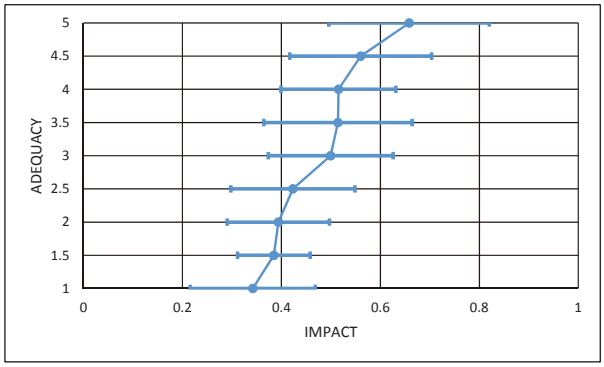
【表 2】から SD/DA は BLEU、RIBES、IMPACT の順に小さくなっており、IMPACT が人手評価値を推定するのに誤差が最も小さいことがわかる。



（a）BLEU



（b）RIBES



（c）IMPACT

図 1　自動評価値 (横軸) と人手評価値 (縦軸) の関係

表 2　DA、SD、SD/DA の値

	DA	SD	SD/DA
BLEU	0.362	0.175	0.483
RIBES	0.218	0.103	0.473
IMPACT	0.316	0.125	0.396

3 自動評価値から人手評価値を推定する

【図 1】と【表 2】を参考にして自動評価値（IMPACT）から人手評価値（JPO adequacy）を推定する関数（グラフ）を【図 2】のように定める。

【図 2】の関数を用いて、IMPACT 値から JPO adequacy を推定した推定値を実測値から引いた残差

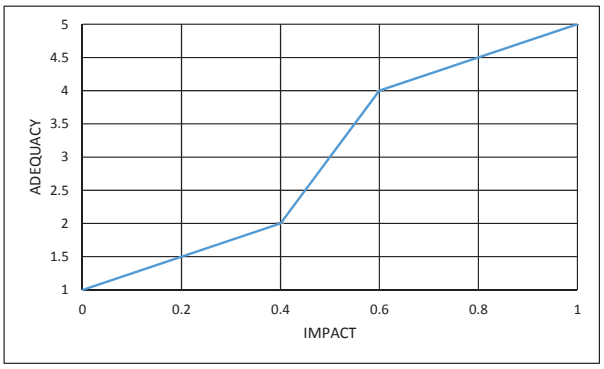
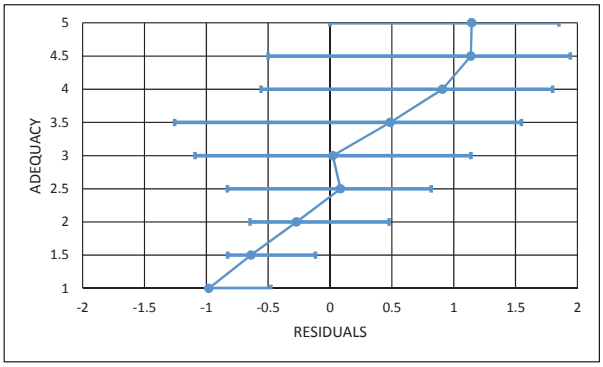
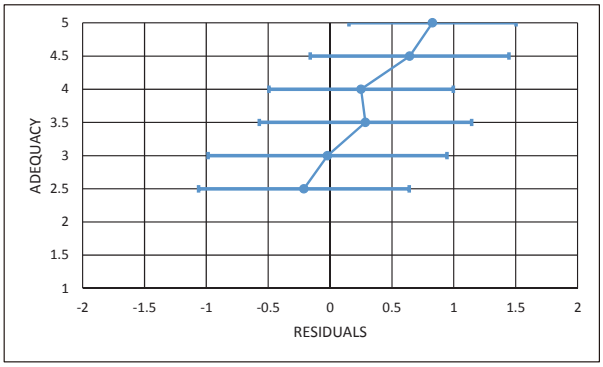


図２　自動評価値（横軸）から人手評価値（縦軸）を推定する関数

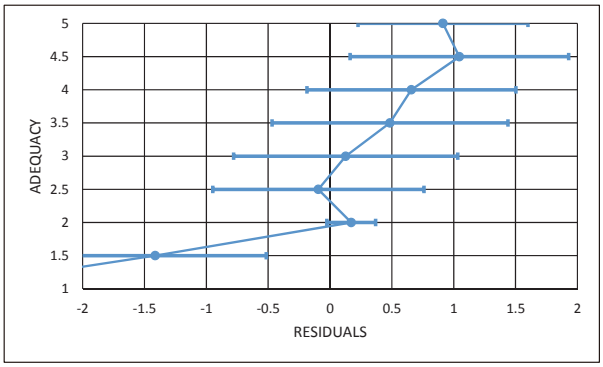
の値を【図 3(a)】に示す。【図 3】には、実測値の各値に対する残差の平均値と標準偏差を示している。また、【図 3】には、中日の結果に加えて韓日と英日の結果も示してある。韓日と英日における IMPACT 値から JPO adequacy を推定する関数は中日と同一で【図 2】に示すものである。



(a) 中日



(b) 韓日



(c) 英日

図３　人手評価値の実測値（縦軸）と残差（横軸）の関係

4 推定結果の考察

【図 3】を見ると、英日の実測値 1 と 1.5 を除き、残差は、ほぼ±1.5 の範囲にある。実測値が小さい場合は、残差が負（実測値が推定値より小さい）となる傾向が高く、実測値が大きい場合は、残差が正（実測値が推定値より大きい）となる傾向が高い。なお、韓日の実測値 1，1.5，2 はデータが存在しなかった。また、英日の実測値 1 のデータ数は 1 であり、1.5 のデータ数は 12 であった。

英日の実測値 1 と 1.5 の場合の推定値と実測値の齟齬について考察する。実測値 1 の原文・参照翻訳文・機械翻訳文は以下の通りである。

Another embodiment provides a green tea polyphenol esterified with at least two fatty acids.

別の実施形態は、少なくとも 2 つの脂肪酸でエステル化された緑茶ポリフェノールを提供する。

別の実施形態は、少なくとも 2 つの脂肪酸を有するグリーン緑茶 esterified を提供する。

機械翻訳文は原文に含まれる技術用語 “polyphenol” が訳されておらず、“esterified” も原文のままである。従って、人手評価値が 1 となったと思われる。一方、訳文の前半は参照翻訳文と機械翻訳文が一致しており IMPACT の値が大きく（0.671）なったと推定される。実測値が 1.5 の場合で、IMPACT 値が 0.6 以上の原文・参照翻訳文・機械翻訳文のデータは以下の 4 データである。

FIG. 15 is a representation of one unit of a carboxymethyl cellulose molecule.

図 15 は、カルボキシメチルセルロース分子の 1 つの単位の模式図である。

図 15 は、したセルロース分子の 1 つのユニットの表現である。

The coatings comprising a SOF are resistant to surface wear or damage.

SOF を含むコーティングは、表面摩耗または損傷に対して耐性である。
＼を含むコーティングは、表面の摩耗または損傷に抵抗される。

This process and the reaction performing the transformation is known as substitution.
このようなプロセスおよび変換を行う反応は、置換として知られている。
変換を行うこの処理及び反応は、代替として知られている。

The resin composition 66 is a composition for forming the resin layer 63 by hardening.
樹脂組成物 66 は、硬化により樹脂層 63 を形成する組成物である。
66 は、硬化樹脂組成物によって樹脂層 63 を形成するための組成物である。

第 1 と第 2 の例は、技術用語が誤訳となっている。第 3 の例は and の並列範囲を間違えており、第 4 の例は、“resin composition” の訳出し位置が誤っている。いずれの例でも、誤りは一部の重要な用語の誤訳または訳出し位置の誤りであり、文全体としては参照翻訳文と機械翻訳文は類似している。

このように文の一部ではあるが重要な部分に誤りがある場合、人手評価値は低くなる。一方、文全体の類似性を計測して評価する自動評価手法では評価値が高くなってしまう傾向がある。

5 まとめ

機械翻訳の文レベルの評価において、自動評価値（IMPACT）から人手評価値（内容の伝達レベル）を推定する手法を試みた。その結果、5 段階評価の中で、多くの場合±1.5 の範囲で推定することができた。

人手評価の実測値は小さいが、自動評価値からの推定値が大きく、両者の齟齬が大きい例について考察したところ、文の一部ではあるが重要な部分に誤りがある例が多かった。今後、文の構成要素の中での重要性に着目した自動評価手法の開発が望まれる。特許文献の場合は、特に技術用語の訳語の正確性を考慮した自動評価手法が必要である。

今回行った手法は、非常に単純なものである。用いている自動評価規準も IMPACT のみであり、推定するための関数も単純な区分線形関数である。今後、多数の自動評価規準を組み合わせることや、より複雑な推定手法を用いることなどが考えられる。

参考文献

[1] 越前谷博：自動評価手法がもたらした歓喜と失望、そして、希望、特許文書の機械翻訳結果評価方法検討会発表資料、2012年9月7日。
http://aamtjapio.com/kenkyu/files/discussion01/AAMT_Japio_discus(20120907)-01.pdf

[2] 磯崎秀樹：翻訳自動評価法―翻訳の質を推定する技術の進化―、第4回特許情報シンポジウム発表資料、2016年11月25日、pages 31-38。
http://aamtjapio.com/kenkyu/files/symposium2016/AAMT_symposium_20161125.pdf

[3] 特許庁：特許文献機械翻訳の品質評価手法に関する調査報告書、平成26年2月。
https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/kikai_honyaku/h25_01.pdf

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu：BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), July 2002, pages 311-318.

[5] 平尾努、磯崎秀樹、Kevin Duh、須藤克仁、塚田元、永田昌明：RIBES：順位相関に基づく翻訳の自動評価法、言語処理学会第17回年次大会発表論文集、2011年3月、pages 1115-1118.

[6] Hiroshi Echizen-ya, 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), Sept. 2007, Page.151-158.

[7] Toshiaki Nakazawa, Chenchen Ding, Hideya Mino, Isao Goto, Graham Neubig, and Sadao Kurohashi：Overview of the 3rd Workshop on Asian Translation, Proceedings of the 3rd Workshop on Asian Translation (WAT2016), pages 1-46, Dec. 2016.