機械翻訳結果の自動評価と人手評価の比較

山梨英和大学人間文化学部人間文化学科教授 **江原　暉将**

PROFILE

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 年より現職。アジア太平洋機械翻訳協会 (AAMT) ／ Japio 特許翻訳研究会副委員長

独立行政法人情報通信研究機構　ユニバーサルコミュニケーション研究所多言語翻訳研究室専門研究員 **後藤　功雄**

PROFILE

1997 年早稲田大学大学院修士課程了。同年 NHK 入局。2008 年より情報通信研究機構に出向し現職。自然言語処理の研究に従事。

1 はじめに

筆者らは規則方式機械翻訳 (Rule Based Machine Translation、RBMT と記す) に統計的後編集 (Statistical Post Editing、SPE と記す) を組み合わせた翻訳方式を研究している [1,2]。この方式は、人手によるきめ細かな翻訳規則を備えた規則方式の利点と大量の対訳コーパスから翻訳規則を自動獲得する統計方式の利点を兼ね備えて一層の精度向上を図ろうとするものである。我々のシステムは RBMT 部として市販の日英機械翻訳システムを用い、SPE 部としては句レベルの統計的機械翻訳ソフトである Moses[3] を用いている。

この方式を用いて、ワークショップ型共同研究である NTCIR-8[1] に参加し、特許文書に関する日英機械翻訳のタスクで好成績を得た [4]。そのときの評価基準は自動評価基準 BLEU[2] である。BLEU は人手で正確に翻訳した訳文 (基準翻訳文という) と機械翻訳文を自動比較して、基準翻訳文により近い機械翻訳文の評価値を高くするというものである。BLEU は 0 から 1 の間の値をとり 1 に近いほど評価が高い。

一方、BLEU のような自動評価とは別に人手による評価が行われる場合もあり、人手評価と自動評価の間にずれが見られるという報告もある。そこで本文では、NTCIR-8 の試験文から 200 文を無作為に選んで機械翻訳した結果に対して人手評価を行い自動評価の結果と比較する。

2 評価の方法

本稿での人手評価は、試験文の原文である日本語文と機械翻訳結果の英文を評価者に提示し、翻訳の良し悪しを評価してもらうものである。評価者は英語を母語とする人で、日本語も理解できる。評価者数は 3 名である。評価基準は NTCIR-9 の日英特許翻訳タスクで用いている基準 (acceptability) であり AA、A、B、C、F の 5 段階評価である [6]。原文と RBMT で翻訳した文、RBMT の出力に SPE を加えて翻訳した文 (以後、単に SPE と書く) を評価者に提示して評価してもらった。基準翻訳文を評価者に提示することは、しなかった。

一方、自動評価には BLEU および IMPACT を用いた。IMPACT は人手評価との相関が比較的高い自動評価基準である [5]。

1) National Institute of Informatics Test Collection for Information Retrieval Systems

2) BLEU および後述する IMPACT については文献 [5] の参考文献を参照されたい。

3 評価の結果

人手評価と自動評価の結果を表 1 に示す。人手評価結果の値 (HE と書く) は以下のようにして求めた。各文 (200 文) に対する各評価者 (3 名) の評価結果 (600 データ) の各データに対して、RBMT と SPE に対する評価結果を比較し、RBMT のほうが評価が高ければ HE(RBMT)=1.0、HE(SPE)=0.0 とし、SPE のほうが高ければ HE(RBMT)=0.0、HE(SPE)=1.0 とする。両者で評価値が同じ場合は HE(RBMT)=HE(SPE)=0.5 とする。このようにして求めた HE の値をデータ数 (600) で割った平均値を人手評価値とした。

RBMT と SPE を比較すると人手評価では RBMT のほうが評価値が高く、自動評価では BLEU も IMPACT もともに SPE のほうが評価値が高い。このように人手評価と自動評価では逆の結果となった。

評価値の信頼性を調べるために、人手評価について符号検定 [7] を実施した。HE(RBMT) > HE(SPE) となるデータ数は 167、HE(RBMT) < HE(SPE) となるデータ数は 97 であり、有意水準 1% で統計的に有意に RBMT のほうが SPE より人手評価が高かった。また、自動評価と共通の方法として、ブートストラップリサンプリング法 [7] も実施した。人手評価値とそれに対する 95% 信頼区間を図 1 に示す。こちらでも有意水準 5% で有意に RBMT のほうが SPE より人手評価値が高かった。一方、自動評価値については、逆に有意水準 5% で有意に SPE のほうが RBMT より自動評価値が高かった。

表 1 人手評価と自動評価の比較

	RBMT	SPE
人手評価	0.558	0.442
自動評価 (BLEU)	0.217	0.338
自動評価 (IMPACT)	0.425	0.521

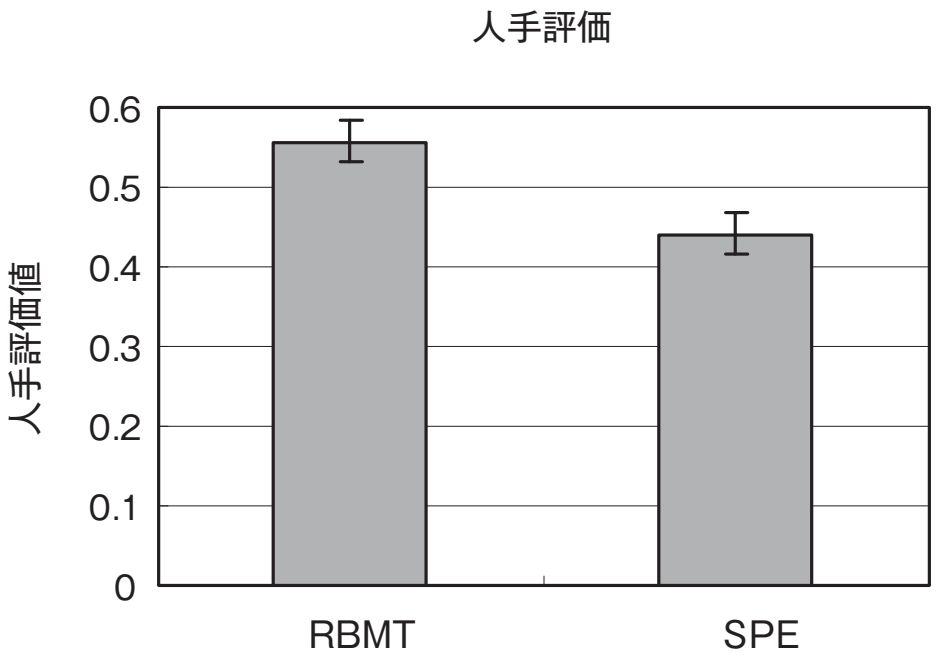

図 1 RBMT と SPE の人手評価値と 95% 信頼区間

4 評価結果の分析

SPE のほうが RBMT より人手評価値が低く、自動評価値では逆に SPE のほうが高かった試験文を目検で調べたところ SPE の結果に訳抜けがあり人手評価値が下がったと思われる例が目立った。例えば表 2 のような例がある。

RBMT では文末に "is arranged" が存在するが、SPE では抜けている。なお表 2 中、SRC は原文 (Source)、REF は基準翻訳文 (Reference) の意味である。この例について、訓練データを調べた結果、以下のことがわかった。まず訓練データ 152,072 文対の中で翻訳元部分 (我々の場合 RBMT の出力にあたるので RBMT と呼ぶ) が "is arranged ." で終わっている文は 191 文あった。それらの文に対する翻訳先部分 (我々の場合基準翻訳文であるので REF と呼ぶ) を調べると文末の "." は REF でも文末の "." に対応するデータが大部分 (190 データ) であった。一方、RBMT の単語 "arranged" に対応する REF の単語と文末の "." との距離 (単語数) が 6 以上のデータが 141 文あった。例えば表 3 のような例がある。RBMT の "arranged" は REF の "disposed" に対応しており、"." との距離は 13 である。このようになった理由は RBMT と REF で構文が全く異なるように訳されているためである。句レベルの統計的後編集では、距離が離れた対応の確率が低くなるので、RBMT の "is arranged ." が REF の "." のみに対応してしまい、その結果、訳抜けに繋がったものと考えられる。

5 おわりに

規則方式機械翻訳の結果 (RBMT) と、それに統計的後編集を加えた結果 (SPE) について、人手評価と自動評価を実施した。その結果、自動評価では SPE のほうが評価値が高かった一方、人手評価では RBMT のほうが評価値が高く、両者で矛盾した結果となった。そこで自動評価と人手評価が逆転する事例を調べたところ SPE を施すことで「訳抜け」が生じることが原因の一つであることがわかった。さらに訳抜けの生ずる原因として、訓練データの RBMT 部分と基準翻訳部分 (REF) で構文が大きく異なることが一因であることがわかった (表 3)。

表 2　SPE で訳抜けが生ずる例

[SRC] この実施例では、バッテリ電流検出器18−1が配されている場合について説明する。
[REF] In this embodiment-1, the case where the battery current detector 18-1 is provided is explained.
[RBMT] This example explains the case where battery current detector 18-1 is arranged.
[SPE] This embodiment explains a case in which the battery current detector 18-1.

表 3　SPE の訓練データの例

[SRC] この第９水槽49に隣接かつ密着して第10水槽50が配置されている。
[RBMT] It adjoins and sticks to this 9th tank 49 , and the 10th tank 50 is arranged.
[REF] The tenth water tank 50 is disposed adjacent to the ninth water tank 49 and inabutting relation thereto .

今後の課題としてRBMTとREFの間の構文の違いを吸収できる後編集方式とする必要があり、例えば構文レベルの統計的後編集の利用が考えられる。

参考文献

[1] 江原暉将：句レベルを用いた統計的後編集の精度向上、Japio 2008 YEAR BOOK、pp.262-265、2008。

[2] 江原暉将：規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳、Japio 2010 YEAR BOOK、pp.280-283、2010。

[3] 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, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

[4] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya and Sayori Shimohata : Overview of the Patent Translation Task at the NTCIR-8 Workshop, Proceedings of NTCIR-8 Workshop Meeting, June, 2010.

[5] 江原暉将ほか：機械翻訳精度の各種自動評価の比較、Japio 2009 YEAR BOOK、pp.272-275、2009。

[6] http://ntcir.nii.ac.jp/PatentMT/

[7] Philipp Koehn : Statistical Machine Translation, Cambridge University Press, pp.234-237, 2010.