Japio

機械翻訳精度の各種自動評価の比較

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 2009 年度特許版産業日本語委員会委員。

北海学園大学工学部電子情報工学科助教
越前谷　博

PROFILE
1991 年北海学園大学工学部卒。1998 年北海道大学大学院工学研究科博士後期過程退学。博士（工学）。同年北海学園大学工学部電子情報工学科助手。2008 年より現職。AAMT ／ Japio 特許翻訳研究会委員。

沖電気工業株式会社
サービスプラットフォームセンタ
下畑　さより

PROFILE
1988 年同志社大学文学部卒。同年沖電気工業株式会社入社。機械翻訳を中心とする自然言語処理技術の研究開発に従事。AAMT ／ Japio 特許翻訳研究会委員。

東京工業大学大学院情報理工学研究科准教授
藤井　敦

PROFILE
1998 年東京工業大学大学院博士課程修了。博士（工学）。筑波大学大学院准教授等を経て、2009 年より現職。自然言語処理、情報検索、音声言語処理の研究に従事。
2009 年度特許版産業日本語委員会委員。

独立行政法人情報通信研究機構　マスタープロジェクト
言語翻訳グループ　主任研究員
内山　将夫

PROFILE
1992 年筑波大学卒業。1997 年同大学院工学研究科修了。博士 (工学)。2001 年情報通信研究機構研究員。現在、同機構主任研究員。

筑波大学システム情報工学研究科
コンピュータサイエンス専攻教授
山本　幹雄

PROFILE
1986 年豊橋技術科学大学大学院修士課程了。(株) 沖テクノシステムズラボラトリ研究開発員等を経て、1995 年筑波大学電子・情報工学系講師。1998 年同助教授。2008 年筑波大学システム情報工学研究科教授。博士 (工学)。

筑波大学大学院システム情報工学研究科
知能機能システム専攻准教授
宇津呂　武仁

PROFILE
1994 年京都大学大学院工学研究科 博士課程電気工学第二専攻 修了。京都大学博士 (工学)。京都大学 情報学研究科 知能情報学専攻 講師等を経て、2006 年より筑波大学 大学院システム情報工学研究科 知能機能システム専攻 助教授。2007 年より同准教授。

国立情報学研究所情報社会相関研究系教授
神門　典子

PROFILE
1994 年慶應義塾大学大学院博士課程 修了。博士 (図書館・情報学)。学術情報センター助教授、国立情報学研究所 助教授を経て、2004 年より同教授。

1 はじめに

機械翻訳技術は日々進歩している。新しい翻訳システムを適用したときに、古いシステムと比較して精度が良くなったかどうかを調べたい。システムを新しくしなくとも、辞書を更新したときの効果を測定したい場合もある。このようなときには、試験文を翻訳させてみて、人間が評価する方法が、まず考えられる。しかし、評価を公平なものにするためには、かなり大量の試験文を用

いなければならず、人手評価にはコストがかかる。何とか自動評価ができないであろうか。

こうした要望から生まれたのが自動評価基準BLEU*1である。BLEUでは正解とみなせる基準翻訳文(reference)を用意して、機械翻訳結果と基準翻訳文の近さをある式に基づいて測り、近い機械翻訳結果は精度が高く、遠い場合は低いと評価するものである。翻訳では基準翻訳文が一通りとは言えないので、複数の基準翻訳文を用意して測定するのが一般的である。このようにBLEUは機械翻訳の自動評価方法を初めて提案したという意味で画期的であるが、問題も多く、その後、多く種類の自動評価基準が提案されている*2。

本文では、いくつかの自動評価基準の良否をメタ評価した結果について概説する。詳細は文献*3を参照されたい。

2 評価方法

自動評価基準の良否をメタ評価する評価方法として、人手評価の結果である正確さ(adequacy)と流暢さ(fluency)と自動評価結果との相関係数を比較することを行った。人手評価との相関が大きいほど良い自動評価基準である。相関係数としては、Pearsonの相関係数とSpearmanの順位相関係数を用いた。試験文としては、人手評価結果のあるNTCIR-7 (NII Test Collection for Information Retrieval systemsの第7回目のworkshop*4)のデータを用いた。原文は日本語で100文である。これらはすべて、公開特許公報から選ばれた文である。英語の基準翻訳文の数は各原文に対して4文ずつを用意した。そのうち、1文はUSPTOのPatent grant dataの中で、patent familyとして日本文に対応する英文を用いた。他の3文は、新たに人手で翻訳して作成した。機械翻訳システムとしてはNTCIR-7に参加した日英システムの中で、14システムを対象とした。これらのシステムのうち、2システムは規則方式、11システムは統計方式、1システムは用例方式の機械翻訳システムである[1]。よって機械翻訳文の数は1,400文である。人手評価は3人の評価者によって行い、相関係数の計算には3人の中央値を用いた。メタ評価の対象となる自動評価基準としては、筆者らが提案した基準であるIMPACTとNMG-WNを含め、以下の8種類のものを用いた。

- IMPACT (Recursive acquisition of Intuitive comMon PArts ConTinuum) *5
- ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) *6
- BLEU[2] (Bilingual Evaluation Understudy) *1
- NIST (National Institute of Standards and Technology) *7
- WN (Word Number)
- NMG-WN (Normalized Mean Grams, Word Number) *8
- METEOR (Metric for Evaluation of Translation with Explicit ORdering) *9
- WER (Word Error Rate) *10

3 評価結果

評価結果を図1から図3に示す。図1はadequacyとの相関係数で、棒グラフの左側はPearsonの相関係数、右側はSpearmanの順位相関係数である。図2はfluencyとの相関係数であり、棒グラフの意味は図1と同じである。図3は統計方式(SMT)と規則方式(RBMT)の相関係数の比較である。棒グラフの左側が統計方式、右側は規則方式である。ここでは、adequacyに対するPearsonの相関係数を示した。

これらの図から以下のことが言える。

自動評価基準の中で、IMPACT、ROUGE-L、NMG-WN、WERが比較的に性能が良い。

Pearsonの相関係数とSpearmanの順位相関係数では大きな差がない。

adequacyとfluencyを比較すると、adequacyの

[1] 機械翻訳の方式は、大きく、規則方式(RBMT)、用例方式(EBMT)、統計方式(SMT)の3種類に分けられる。
[2] 元来文書単位の評価基準であるBLEUを文単位の評価に用いるため、若干変更している。

方が全般に相関が大きい。つまり、正確さ (adequacy) の評価の方が流暢さ (fluency) の評価より、自動評価の性能が高い。

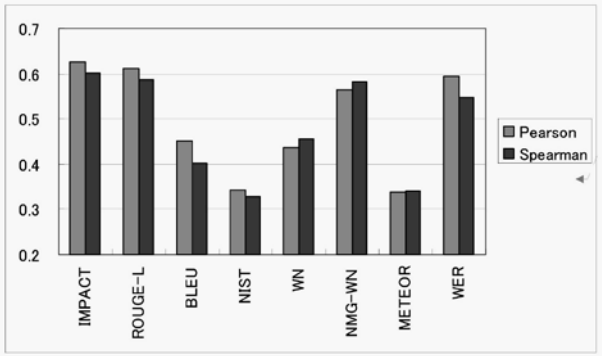

図 1　Adequacy との相関係数
Pearson：左、Spearman：右

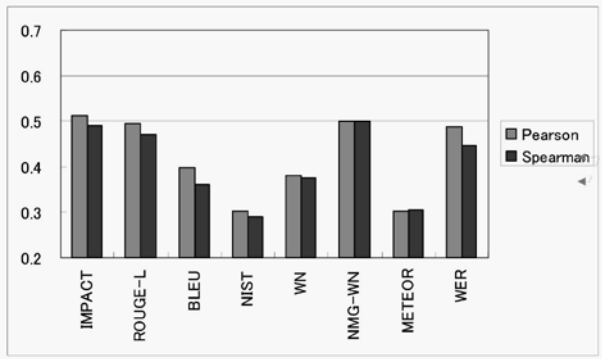

図 2　Fluency との相関係数
Pearson：左、Spearman：右

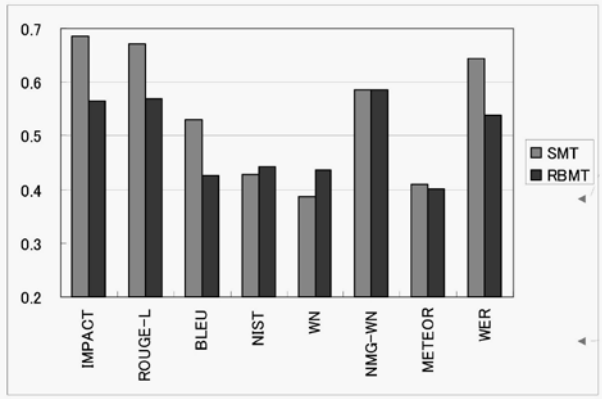

図 3　Adequacy との Pearson の相関係数
統計方式機械翻訳：左、規則方式機械翻訳：右

図 3 より、IMPACT は統計方式で高性能を発揮し、NMG-WN は統計方式・規則方式によらず、一定の性能を持っていることが分かる。

4 おわりに

8 種類の自動評価基準について、人手評価との相関をとることで、基準としての良否をメタ評価した。その結果、筆者らが提案する IMPACT や NMG-WN が比較的良い基準であることが分かった。しかしながら、正確さ (adequacy) の評価より流暢さ (fluency) の評価で、相関が低いことが分かり、課題が残った。今後は評価基準に構文情報を考慮するなどして、一層の性能向上を図りたい。

参考文献

*1 Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu : BLEU: a Method for Automatic Evaluation of Machine Translation, Proc. of ACL2002, 2002.

*2 安田圭志、隅田英一郎：機械翻訳の研究・開発における翻訳自動評価技術とその応用、人工知能学会誌、Vol.23、No.1、pp.2-9、2008.

*3 Hiroshi Echizen-ya, Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro and Noriko Kando : Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7, Proc. of MT Summit XII Workshop on Patent Translation, pp.9-16, 2009.

*4 Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro : Overview of the Patent Translation Task at the NTCIR-7 Workshop, Proc. of 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information

Access, pp.389-400, 2008.

*5 Hiroshi Echizen-ya and Kenji Araki : Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, Proc. of MT Summit XI, pp.151-158, 2007.

*6 Chin-Yew Lin and Franz Josef Och : Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, Proc. of ACL2004, pp. 606-613, 2004.

*7 NIST : Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf, 2002.

*8 江原暉将：新しい機械翻訳自動評価基準 NMG の提 案、Japio 2007 Year Book、pp.238-241、2007。

*9 Satanjeev Banerjee and Alon Lavie : METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Proc. of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp.65-72, 2005.

*10 Gregor Leusch, Nicola Ueffing and Hermann Ney : A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation, Proc. of MT Summit IX, pp.240-247, 2003.