アジア言語を中心とした機械翻訳の評価
－第１回アジア翻訳ワークショップ概要－

Evaluation of Machine Translation Focusing on Asian Languages
– Overview of the 1st Workshop on Asian Translation –

国立研究開発法人科学技術振興機構　情報企画部研究員　**中澤　敏明**

PROFILE　2010 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。機械翻訳の研究に従事。

国立研究開発法人情報通信研究機構　先進的音声翻訳研究開発推進センター先進的翻訳技術研究室専門研究員　**美野　秀弥**

PROFILE　2004 年東京工業大学情報理工学研究科計算工学専攻修士課程修了後、NHK に入局。2013 年から NICT に出向。機械翻訳の研究に従事。

日本放送協会　放送技術研究所ヒューマンインターフェース研究部専任研究員　**後藤　功雄**

PROFILE　2014 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。1997 年 NHK 入局。自然言語処理の研究に従事。

1　はじめに

アジア翻訳ワークショップ（The Workshop on Asian Translation, WAT）はアジア言語を対象とした、新しい評価型機械翻訳ワークショップである。本ワークショップを通じて得られた知見を共有することで、機械翻訳研究において今必要なことが明らかとなり、アジア各国間の機械翻訳が実用的なものになることが期待される。WAT のキーワードとして「オープンイノベーションプラットホーム」がある。テストデータを含む全てのデータがあらかじめ公開されており、定められたテストデータでの翻訳評価を繰り返し行うことで、１システムでの翻訳精度の経年変化を見ることや、翻訳システムごとの翻訳精度の違いを見ることを可能にする。

第１回目のワークショップ（WAT2014）[1] では科学技術論文を対象として、日英（JE）・英日（EJ）、日中（JC）・中日（CJ）翻訳の評価を行った。評価には 12 チームが参加した。報告会は 2014 年 10 月４日に行われた。

本稿では WAT2014 の概要や結果などをまとめて報告する。なお評価結果の詳細や各参加チームの翻訳システムの説明などは、全て WAT2014 のウェブサイトで確認することができる（http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2014/）。

2　データセット

データは JST より研究利用目的で一般に公開されている「アジア学術論文抜粋コーパス（ASPEC）」を利用した。ASPEC は、約 300 万対訳文からなる日英論文抄録コーパス（ASPEC-JE）と約 68 万対訳文からなる日中論文抜粋コーパス（ASPEC-JC）とで構成される世界初の大規模な論文対訳コーパスである。本コーパスは機械翻訳での利用を想定し、「訓練データ」「開発データ」「開発試験データ」「試験データ」の４つの部分に分けられている。

2.1 ASPEC-JE

ASPEC-JE は、JST 所有の約 200 万件の学術論文日英抄録対から、内山・井佐原の方法 [2] により、情報通信研究機構（NICT）が作成したものである。抄録対から文単位の対応を自動抽出することで作成されたコーパスであるため、各対訳文は必ずしも完全な対訳になっているとは限らない。各対訳文には内山・井佐原の方法により計算された類似度が付与されており、類似度の高い順に並べられている。つまり最初の方は対訳文としての質は高いが、後ろの方に行くにつれ質が低下するため、訓練データの使用には注意が必要である。

開発データ・開発試験データ・試験データは、JST 所有の学術論文日英抄録のうち、訓練データに含まれない抄録から対訳文を抽出したものであり、それぞれ 400 抄録（約 1,800 文）ずつからなる。これらのデータに関しては、文対応を自動で付け、全ての文が 1 対 1 で対応づいたもののみを利用しており、訓練データとは異なり、元の抄録全体を復元可能である。

各対訳文には、アルファベット 1 文字からなる分野記号が付与されている。これは元の抄録がどの学術分野のものかを表すものであり、分類の詳細は JST 分類コード（http://opac.jst.go.jp/bunrui/）に記載されている。

2.2 ASPEC-JC

ASPEC-JC は、文献データベース JDream Ⅱ搭載の和文抄録と、電子ジャーナルサイト J-STAGE（科学技術情報発信・流通総合システム）搭載の情報処理学会、言語処理学会、人工知能学会論文誌の和文論文を各学協会から許諾を得て中国語に翻訳することで構築した対訳コーパスである。翻訳対象は抄録、もしくは本文の段落単位である。

開発、開発試験、試験データは、ASPEC-JC 全体で 1 段落しか含まれていない論文からランダムに抽出したものであり、それぞれ 400 段落（論文）（約 2,100 文）ずつからなる。つまり訓練データや、他の開発、開発試験、試験データには、これらのデータと同じ論文に属する文は含まれていない。

3 ベースラインシステム

人手評価は特定のベースラインシステムとの比較に基づいて行った。この比較基準となる特定のベースラインシステムとして、フレーズベース統計的機械翻訳システムを選択した。

フレーズベース統計的機械翻訳システムに加えて、3 種類の他の統計的機械翻訳システム、5 つの商用ルールベース機械翻訳システム、2 つのオンライン機械翻訳システムもベースラインシステムとして利用した。ベースラインシステムの統計的機械翻訳システムは、公開されているソフトウェアで構成し、システムの構築方法と翻訳方法の手順は WAT 2014 のウェブサイトで公開している。ベースラインシステムの統計的機械翻訳システムには Moses を利用し、英語と中国語の構文解析器には Berkeley parser を利用した。ベースラインシステムと適用したサブタスクを表 1 に示す。表 1 では商用システムおよびオンラインシステムのシステム ID は匿名にしている。

表 1　ベースラインシステム

システム ID	システム	種類	JE	EJ	JC	CJ
SMT Phrase	Moses フレーズベース統計的機械翻訳	統計ベース	✓	✓	✓	✓
SMT Hiero	Moses 階層フレーズベース統計的機械翻訳	統計ベース	✓	✓	✓	✓
SMT S2T	Moses String-to-Tree 統計的機械翻訳および Berkeley parser	統計ベース	✓		✓	
SMT T2S	Moses Tree-to-String 統計的機械翻訳および Berkeley parser	統計ベース		✓		✓
RBMT X	The 翻訳 V15（商用システム）	ルールベース	✓	✓		
RBMT X	ATLAS V14（商用システム）	ルールベース	✓	✓		
RBMT X	PAT-Transer 2009（商用システム）	ルールベース	✓	✓		
RBMT X	J 北京 7（商用システム）	ルールベース			✓	✓
RBMT X	蓬莱 2011（商用システム）	ルールベース			✓	✓
Online X	Google translate (July, 2014)	（統計ベース）	✓	✓	✓	✓
Online X	Bing translator (July, 2014)	（統計ベース）	✓	✓	✓	✓

4 自動評価

4.1 自動評価スコアの計算手法

機械翻訳の自動評価は、機械翻訳結果と参照訳（翻訳の正解となる訳）との類似度を計算することで、翻訳結果の品質を数値化する。WAT2014 では、2 種類の異なる自動評価尺度 BLEU[3], RIBES[4] を用いた。自動評価の詳細な手順は、WAT2014 のウェブサイトにて公開している。

4.2 自動評価システム

WAT2014 では、自動評価システムを用意し、参加チームが機械翻訳結果の自動評価結果をいつでも確認できるようにした。翻訳結果は WAT2014 のウェブサイトからいつでも提出することができる。提出された翻訳結果は即時に自動評価サーバーによって自動評価が行われ、スコアが出力される。翻訳結果の提出の際には下記の情報を入力してもらい、提出時にスコアの公開を許可した（下記の iv の項目を可とした）場合は、自動評価後に WAT2014 のウェブサイト上にて提出ファイルの自動評価スコアがランキング形式で公開される。

i) タスク：日本語⇔英語，日本語⇔中国語
ii) 手法：統計的機械翻訳，ルールベース翻訳，統計的機械翻訳とルールベースの両方を用いた手法，その他の手法
iii) ASPEC 以外のデータ（対訳データや単言語データなど）の利用の有無
iv) 自動評価スコアのウェブサイト上での公開の可否

5 人手評価

機械翻訳の人手評価には、1. 非常に多くの時間とお金がかかる、2. 様々な基準が存在する、3. 評価者間の一致度が低いなど、様々な解決すべき問題が存在する。WAT2014 ではクラウドソーシングを利用することで問題 1 を解決した。クラウドソーシングを利用した翻訳の評価は、他のワークショップにおいても採用されている（IWSLT2011, 2012 や WMT2012, 2013 など）。今回は様々存在するクラウドソーシングサービスの中からランサーズを利用した。ランサーズを利用した理由は二つあり、一つは依頼する作業のカテゴリーを指定できる点、もう一つは、「本人確認済」の作業者を指定できる点である。これらの機能を使うことで、より適切な作業者が作業を行うことが期待できる。

問題 2 については、ベースラインとなる機械翻訳結果を用意しておき、これと各システムの翻訳結果を 1 文ずつ比較し、その勝敗数をスコア化（*HUMAN* スコア）することで各システムを評価するという方法を採用した。評価者には入力文とベースラインおよび評価対象システムの翻訳の 3 つが提示され、どちらの翻訳がより良いか、または同程度かの 3 択で評価を行う。ベースラインに対する勝ち数を W、負け数を L、引き分け数を T とすると、*HUMAN* スコアは以下の式で計算できる：

$$HUMAN = 100 \times \frac{W - L}{W + L + T}$$

HUMAN スコアは -100 から 100 の値を取り、正の値は全体としてベースラインより良い翻訳結果であり、負の値は逆に悪い翻訳結果であるという傾向を示す。

クラウドソーシングの性質上、各文ペアの評価は異なる評価者が行うことになる。ここで問題 3 の影響を軽減するために、各文ペアの評価を複数の異なる評価者に行ってもらい、意見を集約することで評価を安定させた。評価対象システムの翻訳がベースラインよりも良いという判断を +1、悪いという判断を -1、同程度を 0 としたとき、全ての評価者の判断を足し合わせて正の値となれば最終判断を勝ち、負の値ならば負け、0 ならば同程度とした。

WAT2014 では人手評価対象文として、テストデータからランダムに 400 文を選択した。また各文の勝敗は異なる 3 人の評価者の判断を集約することで決定した。なお評価者による 1 つの文ペアの評価費用は 5 円と設定した。1 システムの評価には異なる 3 人ずつに 400 文を評価してもらう必要があるため、1 システムの 1 つの翻訳結果の評価にかかる費用は 3 人× 400 文× 5 円で 6,000 円となる。

6 評価結果

紙面の都合上、評価結果の要点のみ報告する。詳細な報告 [1] および各チームの報告は WAT2014 のサイトからオンラインで入手可能である。図 1 に自動評価結果、図 2 に人手評価結果を示す[1]。横軸はシステム、縦軸はスコアを表す。人手評価の結果から、次のことが確認された。

- 最高性能の統計的機械翻訳システムはルールベースシステムより良い評価であった。
- ベースラインシステム間の比較による訳質の順は次のようであった。フレーズベース統計的機械翻訳＜階層フレーズベース統計的機械翻訳＜ Tree-to-String/String-to-Tree 統計的機械翻訳
- Forest-to-String 統計的機械翻訳システム [5] が

1 評価に参加した 12 チームのうち、人手評価を希望した 11 チームの翻訳結果に対して人手評価も実施した。

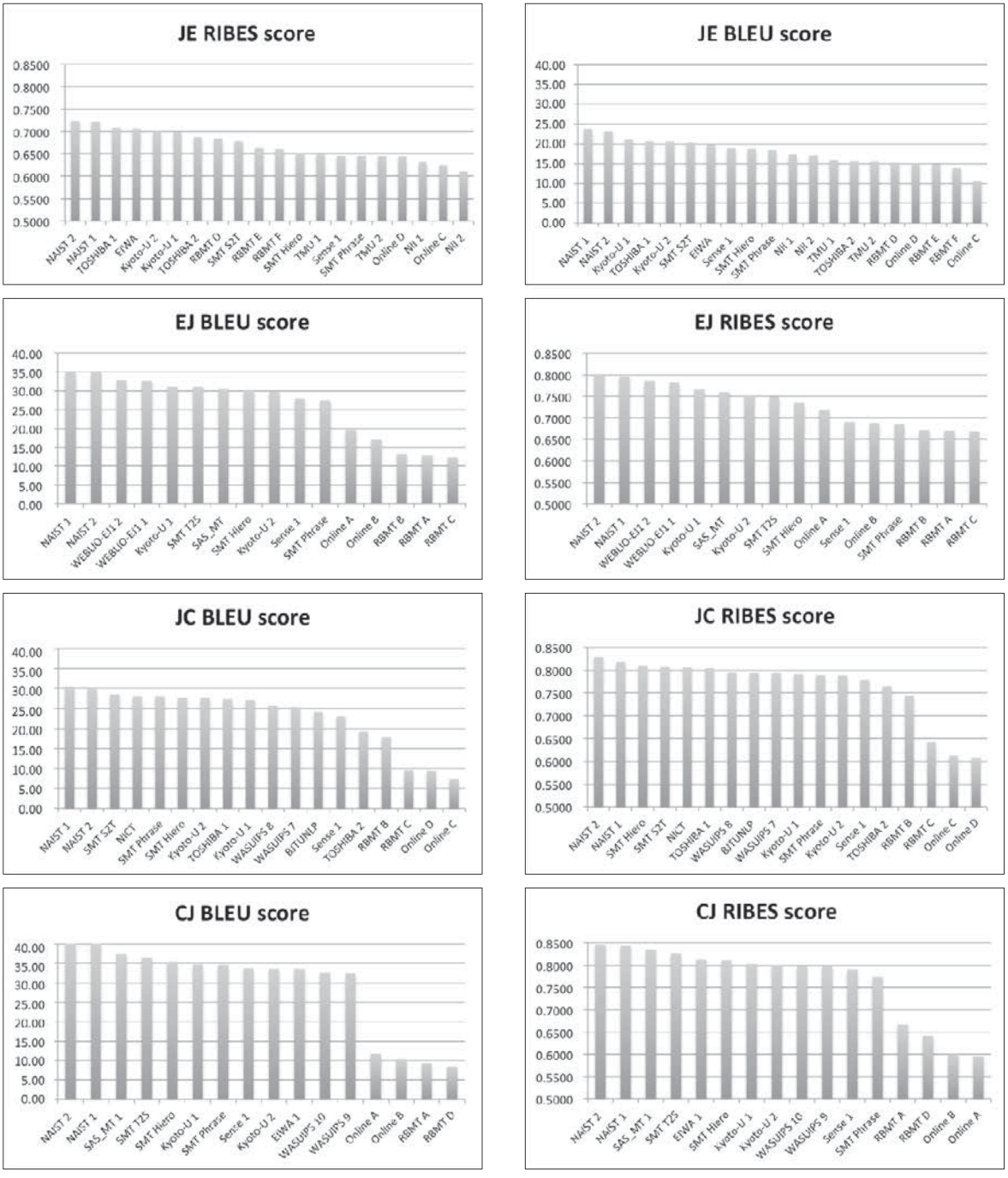

図 1　自動評価結果

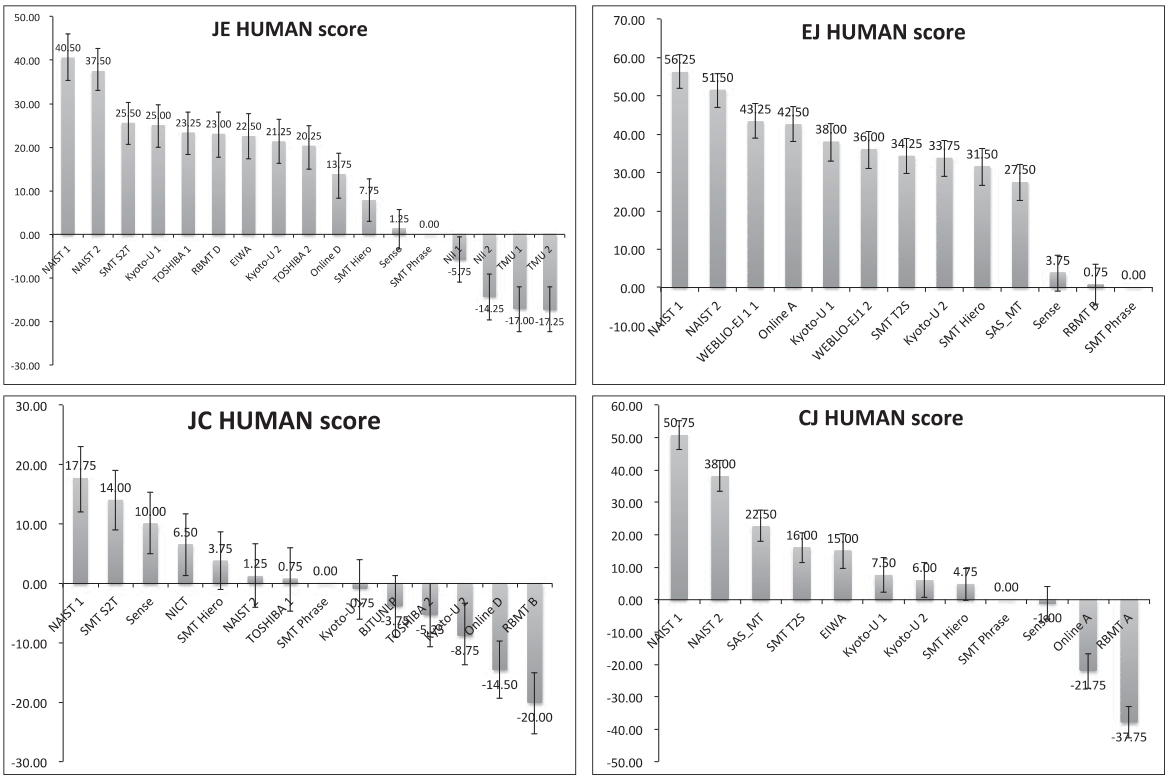

図2　人手評価結果

全ての翻訳方向で最高評価を達成した。

7 まとめと今後の展望

本稿では WAT2014 の概要や結果などについて概説した。初の試みであったが国内外から 12 チームが参加し、様々な手法での翻訳結果が集まり、これらを分析することで様々な知見が得られた。

WAT は今後も開催する予定である。現在実施中の WAT2015 では、JPO より提供された中日、韓日の特許文書からなるデータセットを利用した評価も行っている。WAT2015 の結果は 2015 年 10 月 16 日の報告会にて発表される予定である。

参考文献

[1] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the 1st Workshop on Asian Translation. In Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 1-19.

[2] Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English Patent Parallel Corpus. In Proceedings of MT summit XI. Pages 475-482.

[3] Kishore Papineni, Salim Roukos, ToddWard, andWeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of ACL, pages 311-318.

[4] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 944-

952.

[5] Graham Neubig. 2014. Forest-to-String SMT for Asian Language Translation: NAIST at WAT 2014. In Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 20-25.