アジア言語を中心とした機械翻訳評価ワークショップ(WAT)の取り組み

Introduction of the Workshop on Asian Translation (WAT)

東京大学大学院情報理工学系研究科
中澤　敏明

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）。2018 年 4 月より東京大学大学院情報理工学系研究科特任講師。機械翻訳の研究に従事。

nakazawa@logos.t.u-tokyo.ac.jp　　TEL 03-5841-6650

1 はじめに

アジア翻訳ワークショップ（Workshop on Asian Translation, WAT）[1] はアジア言語を中心とした評価型機械翻訳ワークショップであり、2014 年に第 1 回（WAT2014）を開催して以降、毎年開催している。本稿の著者は初回からオーガナイザーの一人としてワークショップの運営を行っている。2016 年の第 3 回（WAT2016）以降は自然言語処理の国際会議の併設ワークショップとして開催しており、2019 年の第 6 回（WAT2019[1]）は EMNLP-IJCNLP 2019 の併設ワークショップとして、2019 年 11 月 4 日に香港で開催された。2020 年の第 7 回（WAT2020）は AACL-IJCNLP2020 の併設ワークショップとして 12 月に開催予定であるが、今回はオンラインでの開催が予定されている。

ワークショップは様々な機械翻訳の評価タスクの実施に加えて機械翻訳に関する研究論文の募集も行っており、2019 年度は 6 件の論文を採択した。また毎年招待講演も行っており、2019 年度はコペンハーゲン大学の Desmond Elliott 氏により “Multitask Learning from Multilingual Mutimodal Data” というタイトルで行われた。

WAT で行われた機械翻訳タスクの変遷を図 1 に示

1　http://lotus.kuee.kyoto-u.ac.jp/WAT/

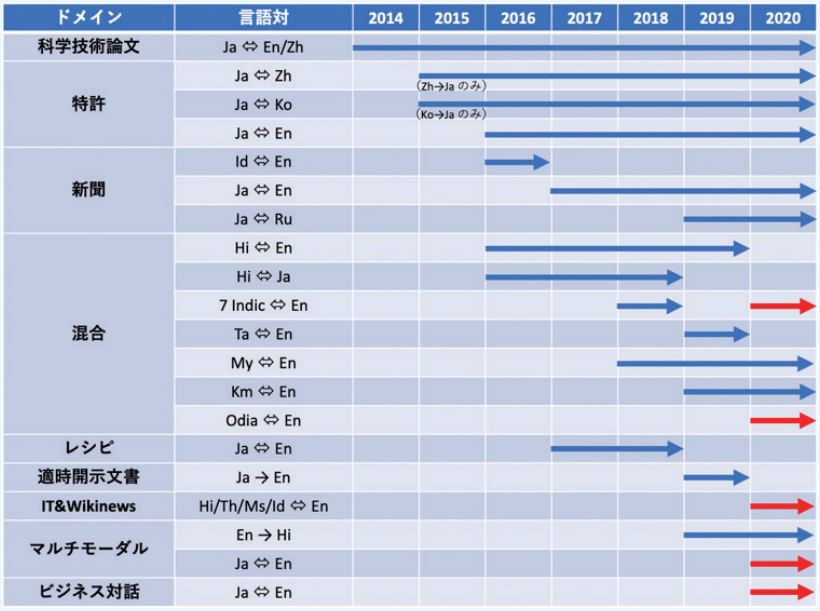

図 1　WAT での翻訳タスクの変遷

す。開始当初は JST および NICT より提供された科学技術論文の対訳コーパス ASPEC[2] を用いた日英・日中の翻訳タスクのみであったが、翌年に特許庁より提供された特許対訳コーパス JPC[3] を用いた日中・日韓の翻訳タスクが追加された。その後も毎年新たなドメイン・新たな言語が追加され、ワークショップは拡大し続けている。

赤色で示されているのは 2020 年度に新たに追加された翻訳タスクである。また 2020 年度は新たな試みとして、科学技術論文[4]およびビジネス対話[5]の日英翻訳において、文書単位での翻訳タスクを設定した。近年の翻訳技術の進歩により文単位での翻訳の精度はかなり向上したためであり、今後は文をまたぐ情報を活用した、より文脈に適した翻訳が生成できるような技術が必要であると思われるためである。

図２に各年の翻訳タスクの参加チーム数を示す。WAT2019 ではこれまでで最多の 25 チームが参加した。海外からは FacebookAI、Microsoft、SYSTRAN など機械翻訳のビッグプレーヤーの参加もあり、WAT に対する世界的な注目度が高まったのではないかと言える。

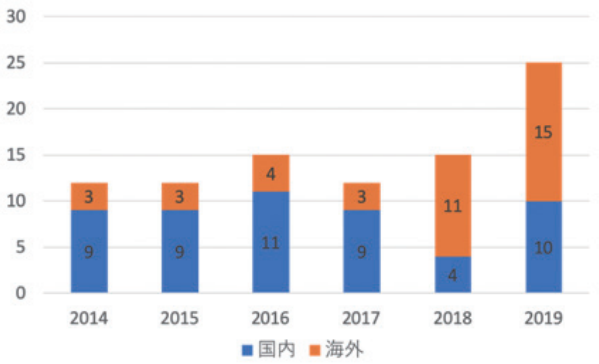

図２　翻訳タスクの参加チーム数

2 翻訳評価

翻訳評価は自動評価と人手評価を行い、人手評価はさらに２つの方法で行った。

2.1 自動評価

WAT では自動評価サーバーを用意し、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ている。自動評価システムはワークショップ終了後も利用可能であり、引き続き翻訳結果の提出・修正を受け付けている。

自動評価尺度としては BLEU[2]、RIBES[3] 及び AM-FM[4] を用いている。AM-FM は正確さと流暢さの両方を考慮したような評価手法である。日本語と中国語に関しては単語分割基準の違いによりスコアが変化するため、いくつかの単語分割ツールを使って翻訳結果を単語に分割し、それぞれの基準でのスコアを算出する。

2.2 人手評価

人手評価は一対比較評価（Pairwise Evaluation）と特許庁基準での専門家による正確性評価（JPO Adequacy Evaluation）の二段階で行った。ただし、一対比較評価については参加チーム数の多かった一部のタスクに対してのみ実施している。

一対比較評価は各翻訳システムの結果を文ごとにベースライン（SMT または NMT）と一対比較し、その勝敗数をスコア化することで行う。WAT2019 では各文ペアにつき５名の評価者で評価を行い、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た値が＋２以上となれば最終判断を勝ち、－２以下ならば負け、それ以外ならば同程度とした。一対比較評価はテストセットの中からランダムに選択された 400 文を対象としている。

次に、タスクごとに一対比較評価上位３チーム、一対比較評価を実施しなかったタスクについては全てのチームの翻訳結果に対して、専門家による評価を行った。専門家による評価の基準として、特許庁が公開している「特許文献機械翻訳の品質評価手順」[6] の中の「内容の伝達レベルの評価」に従って行った。これは機械翻訳結果が原文の実質的な内容をどの程度正確に伝達しているかを、人手翻訳の内容に照らして、下記５段階の評価基準で主観的に評価するものである。内容の伝達レベルの評価で

2 http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
3 http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/
4 http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/aspec_doc.html
5 http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/bsd.html
6 https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html

表1　JPO の機械翻訳評価基準

評価値	評価基準
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%)

はそのうちのさらに 200 文に対象を絞って行った。また内容の伝達レベルの評価においては、2名の評価者が独立に評価を行なっている。

なお英語→ヒンディー語のマルチモーダル翻訳に対しては、Direct Assessment[5] という指標で評価を行った。これは WMT（Conference on Machine Translation）での公式スコアリング指標として用いられているもので、評価者は各翻訳に対して0から 100 の間での翻訳の質を整数値で付与することで評価を行うものである。

次節では特許翻訳タスクの結果などについて述べる。

3 特許翻訳タスク

昨年度の結果の詳細は WAT2019 の overview paper[1] をご覧いただくことにして、ここではこれまでの各年度の正確性評価におけるトップチームの成績を図3から図7に示す（日→韓翻訳タスクについては参加者が少ないため省略する）。

WAT 開始当時はまだニューラル機械翻訳（NMT）が注目を浴びる前であり統計翻訳（SMT）や用例ベース翻訳（EBMT）が使われていたが、その後 Recurrent Neural Network（RNN）方式の NMT があっという間に統計翻訳を追い越した。近年ではさらに Transformer 方式の NMT を採用したチームが大多数を占めている。翻訳精度に関しては SMT/EBMT < RNN < Transformer となっており、最近ではどの言語対でも正確性評価で平均 4.5 以上を達成している。

図の中でアスタリスク（＊）がついている年の結果は、WAT が提供している公式の訓練データ（各言語対 100 万文ずつ）以外に、独自で持っている訓練データ（論文によるとおよそ 1000 万文対程度）も併せて使用した結果となっている。当然ながら、これらのシステムの結果は公式データのみを使ったシステムよりも良いものとなっているのだが、興味深いのは 2019 年の公式データのみを使ったシステムの精度が、大量の訓練データを使ったシステムと遜色がないという点である。2019 年のシステムはベースとなる NMT モデルとして Transformer を採用し、他にも様々なテクニックを駆使してこの精度を達成している。わずか数年のうちに、機械翻訳の技術が大幅に進歩しているということがわかる。使われたテクニックの一部を一言ずつで説明すると以下のような感じである（詳細は各論文を参照いただきたい）。

・Relative positional encoding

Transformer の位置エンコーディングに相対位置を用いる[6]。

・Data augmentation

訓練データを仮想的に増やす方法。Backtranslation[7]（単言語コーパスを機械翻訳して対訳コーパスとして用いる）や Multi-source translation[8]（目的言語文が同じである複数言語の対訳文を同時に用いる）などがある。

・Right-to-left reranking

順方向（文の先頭から末尾）で生成された翻訳文を逆方向の翻訳モデルも用いてリランキング[9]

これらの結果を見ると、文単位での翻訳精度は飽和状態となってきていると言えるため、WAT2020 では一部のタスクについて文書単位での翻訳タスクを実施している（特許翻訳については残念ながら引き続き文単位の翻訳のみである）。文書単位での翻訳タスクを実施する上での最大の問題は、文書単位のデータの準備である。評価用のデータ（数千文程度）のみ文書単位にするという方法もあるが、これでは翻訳エンジンの訓練において

文書情報を利用することができないため不十分である。文書単位での対訳データを持っている組織は、これらのデータの公開を是非ともご検討いただきたい。

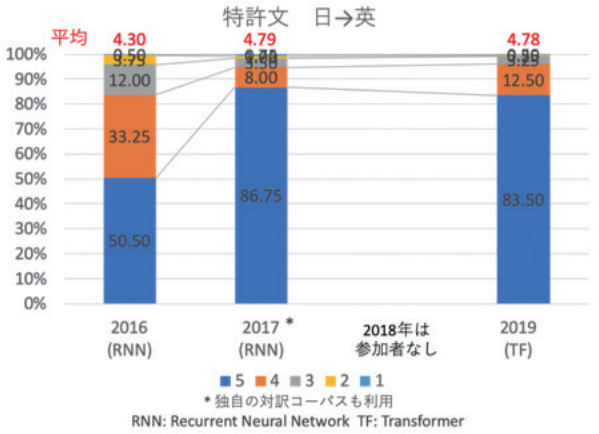

図３　日→英翻訳の評価結果の変遷

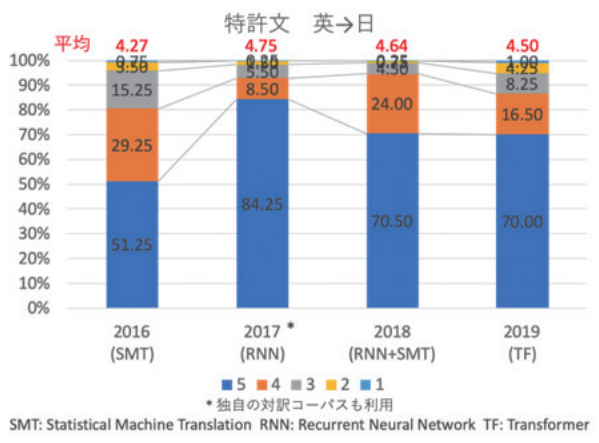

図４　英→日翻訳の評価結果の変遷

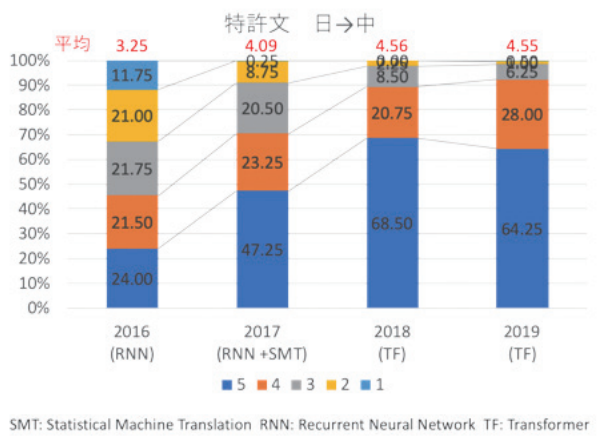

図５　日→中翻訳の評価結果の変遷

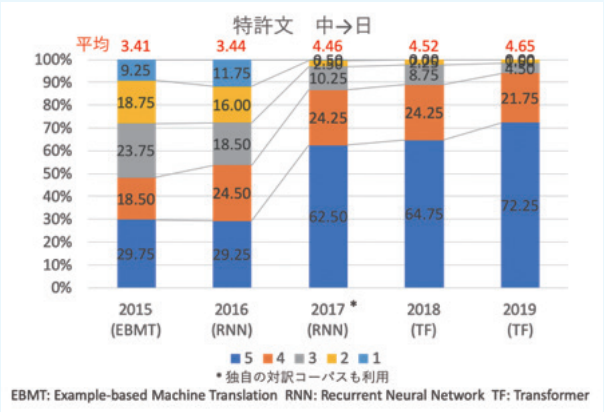

図６　中→日翻訳の評価結果の変遷

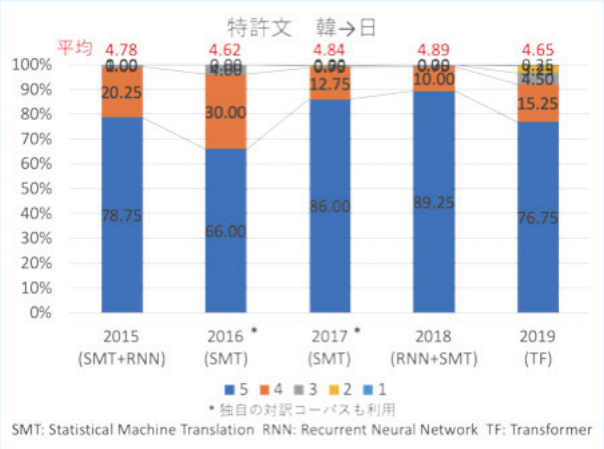

図７　韓→日翻訳の評価結果の変遷

4 まとめ

本稿では WAT の取り組みの紹介と特許翻訳タスクの結果について報告した。アジアの翻訳研究の活性化、データ整備等を目的として 2014 年に始めた WAT は、ドメイン数や言語数の増加、参加者数の増加など一定程度の成果を得ており、WAT を通じてアジア地域の機械翻訳研究コミュニティーの連携等が行えると良いと考えている。

今後は実用上問題となる訳抜け、専門用語や低頻度語の翻訳、訳語の統一、文脈レベルの翻訳などを評価できるタスク設計が重要と考えており、これにはデータの収集や評価指標の定義などが必要となる。また長文への対応も実用上非常に重要である。WAT での翻訳タスクを含め、多くの機械翻訳評価のデータセットは、長文（例えば 100 文字以上など）を取り除いていることがほとんどである。このため現状の機械翻訳システムの多くは、特許文で頻繁に現れる非常に長い文を高精度に翻訳する

ことはできない。これは NMT モデルの制約として長文を扱えない場合や、計算機の能力上扱えない場合などいくつかの要因がある。いずれにせよ、長文であっても精度の低下が起きないような手法の開発も必要である。

WAT は今後も継続して開催予定である。WAT では翻訳評価にかかる費用等のためのスポンサーを募集しているため、興味のある方はご連絡いただければ幸いである。

参考文献

[1] Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Sadao Kurohashi. 2019. Overview of the 6th Workshop on Asian Translation. In Proceedings of the 6th Workshop on Asian Translation (WAT2019), pages 1-35.

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

[3] 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.

[4] Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating mt in the continuous space model framework. IEEE/ACM Trans. Audio, Speech and Lang. Proc., 23 (3):472–482, March.

[5] Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Lamia Tounsi, Teresa Lynn. 2016. Is all that glitters in MT quality estimation really gold standard? In Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan.

[6] Peter Shaw, Jakob Uszkoreit, Ashish Vaswani, 2018, Self-Attention with Relative Position Representations, In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 464–468.

[7] Rico Sennrich, Barry Haddow, Alexandra Birch, 2016, Improving Neural Machine Translation Models with Monolingual Data, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96.

[8] Barret Zoph, Kevin Knight, 2016, Multi-Source Neural Translation, In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 30–34.

[9] Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on targetbidirectional neural machine translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 411–416.