第8回アジア翻訳ワークショップ（WAT2021）開催報告

Report of the 8th Workshop on Asian Translation (WAT2021)

東京大学大学院情報理工学系研究科客員研究員
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）。現在は東京大学大学院情報理工学系研究科客員研究員。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との併設ワークショップとして開催しており、2021年の第8回（WAT2021 [1]）はタイのバンコクで開催されたACL-IJCNLP 2021の併設ワークショップとして、2021年8月6日にオンラインで開催された。

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、WAT2021では5件の研究論文と、ACL-IJCNLP2021から2件のfindings論文を採択した。また2件の招待講演も行われた。1件目はFacebook AIのFrancisco Guzmán氏およびAngela Fan氏により“Massively Multilingual Translation and Evaluation”というタイトルで行われ、2件目はカーネギーメロン大学のGraham Neubig氏により“Understanding and

Team ID	Organization	Country
TMU	Tokyo Metropolitan University	Japan
NTT	NTT Corporation	Japan
NICT-2	NICT	Japan
NICT-5	NICT	Japan
NLPHut	Idiap Research Institute Switzerland, IIT BHU, BITS Pilani India, KIIT University India, Silicon Techlab pvt. Ltd India, University of Chicago	Switzerland, India, USA
TMEKU	Tokyo Metropolitan University, Ehime University, Kyoto University	Japan
*goodjob	Dalian University of Technology	China
YCC-MT1	University of Technology (Yatanarpon Cyber City)	Myanmar
YCC-MT2	University of Technology (Yatanarpon Cyber City)	Myanmar
NECTEC	National Electronics and Computer Technology Center (NECTEC)	Thailand
mcairt	CAIR	India
nictrb	NICT	Japan
sakura	Rakuten Institute of Technology Singapore, Rakuten Asia.	Singapore
IIT-H	International Institue of Information Technology	India
*gauvar	Amazon	Singapore
*JBJBJB	Indivisual participant	Korea
SRPOL	Samsung R&D Poland	Poland
NHK	NHK	Japan
CFILT	Computing for Indian Language Technology	India
iitp	IIT Patna	India
Volta	International Institute of Information Technology Hyderabad	India
coastal	University of Copenhagen	Denmark
CFILT-IITB	Indian Institute of Technology Bombay	India
CNLP-NITS-PP	NIT Silchar	India
Bering Lab	Bering Lab	South Korea
tpt_wat	Transperfect Translations	USA

図1　WAT2021 shared task参加者リスト

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

Improving Context Usage in Context-aware Translation”というタイトルで行われた。

WAT2021では日英、日中、日韓の特許文翻訳タスクなどを含む、18の言語を対象とした14のshared taskが行われ、世界中から26のチームが参加した。図1に参加者のリストを示す。

WAT2021で新たに追加されたタスクは以下の通りである：

・英語からマラヤーラム語（インドで話されている言語）へのマルチモーダル翻訳
・曖昧性のある動詞を対象とした日英のマルチモーダル翻訳
・10言語のインド諸語と英語間のマルチリンガル翻訳
・使用するフレーズが指定された日英の制限翻訳タスク

本項では特許翻訳タスクの結果の報告と、今回新しく追加された日英・英日制限翻訳タスクのタスク概要と結果を報告する。

2 特許翻訳タスク

特許翻訳タスクは特許庁より提供された特許対訳コーパスJPC[2]を用いて行われ、日英・日中・日韓の双方向の翻訳タスクで構成されている。翻訳評価は自動評価と人手評価を行った。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。人手評価は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」[3]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。

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%)

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

3 https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html

日英・日韓には3チーム（TMU, Bering Lab, tpt_wat）が、日中には2チーム（Bering Lab, tpt_wat）が自動評価サーバーに翻訳結果を提出したが、人手評価用の翻訳結果を提出したのは1チーム（TMU）のみであった。また予算の都合上、実際に人手評価を行ったのは日英、英日タスク1チーム分のみである。図2に自動評価結果を、図3および図4に人手評価結果を示す。

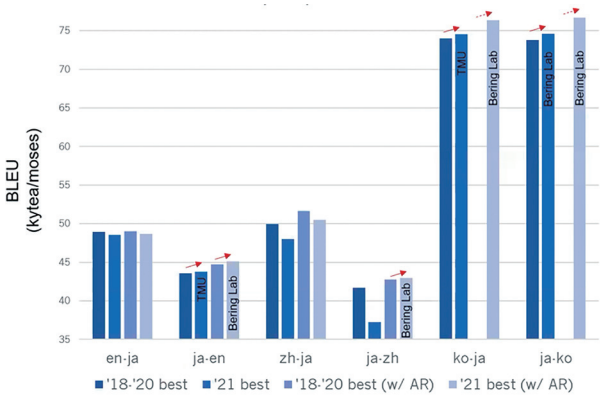

図2 特許翻訳タスク自動評価結果（BLEU）

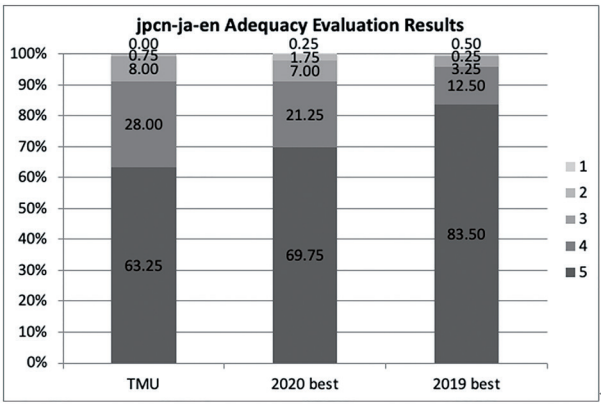

図3 特許翻訳タスク人手評価結果（日英）

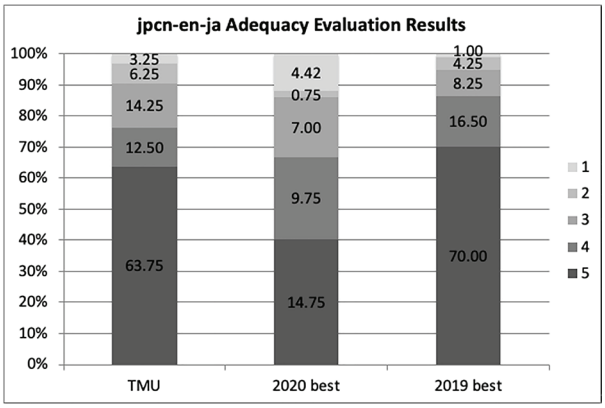

図4 特許翻訳タスク人手評価結果（英日）

Bering Lab は JPC コーパスに加えて、自前で用意した 1300 万文からなる特許対訳コーパスも合わせて用いており、これのおかげで高い BLEU スコアを達成している。TMU は fine-tuning された日本語の BART[5] モデルを利用しており、韓日翻訳において最も良い AM-FM スコアを達成した。日英・英日の人手評価結果を見ると、残念ながら過去の WAT でのベストなシステムと比べるとやや低い精度という結果であった。しかしながら、BART という新たな NMT の枠組みを利用しても、過去のシステムに引けを取らない精度であることが示された。

3 日英・英日制限翻訳タスク

日英・英日制限翻訳タスクは WAT2021 で新たに追加されたタスクである。NMT は訳語統一が不得意であることはよく知られているが、専門用語や固有名詞を特定の用語に常に正しく翻訳することが求められるような文書の種類も多い。この問題に対して、現在の精度を知ることや解決方法を模索するために提案されたのがこのタスクである。

入力文とともに出力文で必ず使わなければならない訳語のリストが与えられるので、システムはこれらの訳語を必ず含むように翻訳を生成しなければならない。なお与えられるのは訳語のリストのみであり、入力文のどの語に対応する訳語なのかは与えられない。今回は ASPEC[4] の日英データ（dev/devtest/test）を対象とし、10 人のバイリンガルに依頼して専門用語（制限用語）の抽出を行なった。表 2 に文単位の制限用語の平均数を示す。

表 2　文単位の制限用語の平均数

	英→日 （フレーズ数、文字数）	日→英 （フレーズ数、単語数）
dev	(2.8, 16.4)	(2.8, 6.6)
devtest	(3.2, 18.2)	(3.2, 7.3)
test	(3.3, 18.1)	(3.2, 7.4)

自動評価は BLEU スコアと一貫性スコアにより行い、最終的なシステムのランキングはこれらを組み合わせたスコアにより行なった（表 3 中の final）。一貫性スコアは全ての制限用語が出力できた文の割合であり、最終ランキングは全ての制限用語が出力できた文のみで計算した BLEU スコアにより決定した。人手評価は Direct Assessment および Contrastive Assessment [6] により行なった。

英日翻訳には 3 システムが参加し、日英翻訳には 4 システムが参加した。表 3 に自動評価と人手評価の結果を示す。

表 3　制限翻訳タスクの評価結果

En-Ja		Human Eval.	
Team	final	src-based DA	src-based CA
NTT	**57.2**	**77.5**	**79.7**
NHK	33.9	74.1	77.2
NICTRB	28.8	73.6	77.1
(human ref.)	—	73.4	76.4
Ja-En		**Human Eval.**	
Team	final	src-based DA	src-based CA
NTT	**44.1**	**75.6**	**74.4**
NHK	37.5	73.9	73.5
NICTRB	31.8	72.1	71.8
TMU	22.6	50.2	48.3
(human ref.)	—	74.1	72.9

全てのシステムが入力文に制限用語を付加して入力し、翻訳時には制限用語を可能な限り出力するように翻訳の探索を行うという方針をとっており、これらの手法は実際に指定された訳語を適切に出力するのに有効であることが示された。

タスクの仕様として訓練データには制限用語が付加されていないため、訓練時にはなんらかの方法で制限用語を準備する必要がある。各システムの違いを見ると、ここでの工夫の仕方が最終的な精度に影響を与えているようである。NHK は固有表現抽出技術により抽出された固有表現と、ベースラインとなる NMT において誤訳となった表現を制限用語として用いるという手法を提案している。一方で NTT は LeCA [7] と呼ばれる手法を用いている。結果を見ると NTT が提案した手法の方が精度良く制限用語を出力し、翻訳精度も高いということが示された。

興味深い点として、元の対訳文の精度（表 3 中の human ref）はそれほど高くないということが示された。今後対訳コーパス自体のクリーニングといったことが必要になる可能性がある。またほとんどのシステムが human ref よりも高精度を達成しており、現時点でも

4　https://jipsti.jst.go.jp/aspec/

人間の翻訳に匹敵するような精度であることがわかる。

4 まとめ

本稿では WAT2020 における特許翻訳タスクと制限翻訳タスクの結果を報告した。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。

NMT の翻訳精度は年々向上しているが、訳抜け、過剰訳、訳語の一貫性、長文の対応など未解決の問題はまだまだ多く残されている。WAT2021 で行われた制限翻訳タスクはこのうちの一貫性の問題に注目したものであり、実用的にも重要なタスクとなっている。今後も特定の現象に絞ったようなタスク設計が重要と考えており、引き続きデータの収集や評価指標の定義などを行なっていく予定である。

WAT は今後も継続して開催予定であるが、来年度の開催予定は未定である。また 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。

[1] Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa Pa, W., Kunchukuttan, A., Parida, S., Bojar, O., Chu, C., Eriguchi, A., Abe, K., Oda, Y., Kurohashi, S., 2021. Overview of the 8th Workshop on Asian Translation, in: Proceedings of the 8th Workshop on Asian Translation (WAT2021). Association for Computational Linguistics, Online, pp. 1–45.

[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] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs, stat] .

[6] Federmann, C., 2018. Appraise Evaluation Framework for Machine Translation, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Santa Fe, New Mexico, pp. 86–88.

[7] Chen, G., Chen, Y., Wang, Y., Li, V.O.K., 2020. Lexical-Constraint-Aware Neural Machine Translation via Data Augmentation, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Presented at the Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}, International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, pp. 3587–3593. https://doi.org/10.24963/ijcai.2020/496