第9回アジア翻訳ワークショップ(WAT2022)報告

Report of the 9th Workshop on Asian Translation (WAT2022)

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@nlab.ci.i.u-tokyo.ac.jp　TEL 03-5841-8986

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、2022年の第9回（WAT2022 [1]）は韓国の慶州市（キョンジュ市）で10月12日から17日に開催されるCOLING2022の併設ワークショップとして開催予定である。本稿執筆時点ではまだ開催前であるため、本稿ではWAT2022の概要を述べる。なお招待講演はニューヨーク大学のDuygu Ataman氏により行われる予定である。Duygu Ataman氏は低資源言語の言語処理に精通しており、WATの対象とする言語のいくつかも低資源言語であるため、有用な講演になると期待される。

2 研究論文

WATでは機械翻訳に関する研究論文の募集も行っている。WAT2022では9件の研究論文の投稿があり、そのうち4件を採択した。以下に採択した論文のタイトルと簡単な概要を挙げる。

— Does partial pre-translation can improve low resourced-languages pairs?

本論文は日本語・フランス語間の翻訳において、人手ルールにより入力文を事前編集することで翻訳精度の向上を目指している。

— Multimodal Neural Machine Translation with Search Engine Based Image Retrieval

本論文は言語だけのコーパスだけでマルチモーダル機械翻訳を可能とするものである。入力文に含まれるキーワードと画像検索エンジンを用いることで、入力文に関連した画像をインターネットから収集し、これを用いてマルチモーダル機械翻訳を実現するものである。

— Comparing BERT-based Reward Functions for Deep Reinforcement Learning in Machine Translation

本論文は強化学習によりNMTの最適化の目的関数を様々に変えた実験をおこなっている。BLEUのような単語の表層に基づく目的関数を用いるよりも、BERTに基づく指標を用いる方が翻訳精度が向上することを示している。

— Improving Jejueo-Korean Translation With Cross-Lingual Pretraining Using Japanese and Korean

本論文はチェジュ語と韓国語間の翻訳において、日本語と韓国語のモノリンガルコーパスを利用してmBARTの事前学習を行うことで、翻訳精度を向上させられることを報告するものである。

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

3 Shared Tasks

WAT2022 では以下の Shared Task を設定した。なお各タスクの後ろに括弧付きで示されている数字は、各タスクの参加チーム数である。数字の示されていないタスクは、参加チームがなかったことを表している。

・文書単位翻訳タスク：
- ASPEC+ParaNatCom：英→日　科学技術論文
- BSD Corpus：英⇔日　ビジネスシーン対話
- JIJI Corpus：英⇔日　ニュース
- NICT-SAP：ヒンディー／タイ／インドネシア／マレー（1）／ベトナム⇔英　非構造的文書
- 日中韓⇔英　構造的文書（1）

・マルチモーダル翻訳タスク：
- Hindi Visual Genome：英語→ヒンディー語（3）
- Malayalam Visual Genome：英語→マラヤーラム語（2）
- Bengali Visual Genome：英語→ベンガル語（3）
- Ambiguous MS COCO：英⇔日
- Video Guided Ambiguous Subtitling：日→英
- Khmer Speech Translation：クメール→英／仏

・インド諸語翻訳タスク（2）：
- アッサム／ベンガル／グジャラート／ヒンディー／カンナダ／マラヤーラム／マラティ／ネパール／オリヤー／パンジャーブ／シンド／シンハラ／タミル／テルグ／ウルドゥ⇔英

・特許翻訳タスク：
- JPC3：英中韓⇔日

・制限翻訳タスク：
- ASPEC：英（3）／中⇔日

・対訳コーパスフィルタリングタスク：
- JParaCrawl：日⇔英（1）

昨年度は 26 チームの参加があったのだが、残念ながら今年度は 8 チームの参加にとどまった。表 1 に参加者のリストを示す。

表 1　WAT2022 Shared Task 参加者リスト

Team ID	Organization	Country
TMU	Tokyo Metropolitan University	Japan
NICT-5	NICT	Japan
Sakura	Rakuten Institute of Technology Singapore, Rakuten Asia.	Singapore
CNLP-NITS-PP	NIT Silchar	India
NITR	NIT Rourkela	India
HwTscSU	Huawei Translation Services Center, 2012 Lab, Huawei co. LTD; School of Computer Science and Technology, Soochow University	China
SILO_NLP	Silo AI	Finland
nlp_novices	SCTR's Pune Institute of Computer Technology	India

制限翻訳タスクは WAT2021 で新たに追加されたタスクであるが、今年度は新たに日中の言語対が追加された。しかしながら日中のタスクへの参加者は今回はなかった。

対訳コーパスフィルタリングタスクは今年度から新たに追加されたタスクである。本タスクは、オーガナイザーから提供されるノイズを含む大規模対訳コーパスから、機械翻訳の訓練に有害となりそうな対訳文を除外することが目標である。参加者はノイズ除去済みの対訳コーパスを提出し、全く同じ設定で NMT の訓練を行い、テストデータでの翻訳精度を競う。世界最大の機械翻訳ワークショップである WMT[2] では 2018 年より同様のタスクを実施しているが、日英の言語対で行われるのは今回が初である。

今回は NTT が一般公開している JParaCrawl v3.0[2][3] を対象としてタスクが実施され、科学技術論文コーパスである ASPEC[3][4] をテストデータとして評価が行われた。

4 まとめ

本稿では WAT2022 の概要について報告した。冒頭にも述べた通り、本稿執筆時点ではまだ WAT2022 は開催されていたないため、開催後の報告についてはまた別の機会に行いたい。

2　https://www.statmt.org/wmt22/

3　https://www.rd.ntt/cs/team_project/icl/lirg/jparacrawl/

4　https://jipsti.jst.go.jp/aspec/

今年度は残念ながら参加チームが少なかったため、来年度以降、より多くの参加者が集まるよう、工夫していきたいと思う。

参考文献

[1] Nakazawa, T., Mino, H., Goto, I., Dabre, R., Higashiyama, S., Parida, S., Kunchukuttan, A., Morishita, M., Bojar, O., Chu, C., Eriguchi, A., Abe, K., Oda, Y., Kurohashi, S., 2022. Overview of the 9th Workshop on Asian Translation, in Proceedings of the 9th Workshop on Asian Translation (WAT2022). Association for Computational Linguistics, Gyeongju, Republic of Korea.

[2] Morishita, Makoto, Suzuki, Jun, Nagata, Masaaki, 2020. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus, in Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, pp. 3603-3609.

[3] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, Hitoshi Isahara, 2016. ASPEC: Asian Scientific Paper Excerpt Corpus, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC2016). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 2204-2208.