第10回アジア翻訳ワークショップ(WAT2023)報告

Report of the 10th Workshop on Asian Translation (WAT2023)

東京大学　大学院情報理工学系研究科　特任研究員

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特任研究員。機械翻訳の研究に従事。

nakazawa@nlab.ci.i.u-tokyo.ac.jp

1 はじめに

アジア翻訳ワークショップ（Workshop on Asian Translation, WAT）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、2023年の第10回（WAT2023[1]）は中国のマカオで9月4日から8日に開催されたMT Summit 2023の併設ワークショップとして開催された。参加者は現地参加とオンラインと合わせて約50名程度で、現地参加が2/3ほどであった。

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。WAT2023では1件の研究論文を採択した。“Mitigating Domain Mismatch in Machine Translation via Paraphrasing”である。

本論文は入力文の言い換えと翻訳結果のリランキングにより、訓練データとテストデータでのドメインミスマッチの影響を低減する方法を提案している。言い換えは単語レベルの言い換えと文レベルの言い換えを用いている。提案法は機械翻訳モデルのファインチューニングが不要であり、オンラインの翻訳サービスなどのブラックボックスな機械翻訳に対しても適応可能であるという特徴がある。

実験により提案手法の有効性が示され、特に言い換えの手法に関しては文単位での言い換えよりも単語単位の言い換えの方が精度が高くなる傾向にあることがわかった。これに関して会場にいた参加者から「単語単位の言い換えの方が精度がいいという結果はLSP (Language Service Provider) にとっては朗報で、他のドメインでの効果も検証してほしい」とのコメントが出た。これはおそらく、単語単位での言い換えはLSPが独自に持っている辞書などを用いることで容易に行うことができるため、提案手法が実用的にも期待できるからではないかと思われる。

3 招待講演

招待講演はWikimedia FoundationのPrincipal Software Engineerである Santhosh Thottingal 氏より“Machine Translation at Wikipedia”というタイトルで行われた。Wikipediaは320以上の言語で提供されており、記事の翻訳に機械翻訳を使用している。さまざまな機械翻訳サービスを利用することができる統合ツールにより最初の翻訳を提供し、編集者による修正を経てから翻訳された記事が公開される。この方法で現在までに約160万件の記事が翻訳されている。

講演ではテキストのみの機械翻訳と人手による情報のキュレーションを組み合わせることで高品質な立地テキスト翻訳を提供することに焦点を当てて、ヒューマン・イン・ザ・ループの製品設計が紹介された。また翻訳品

質と翻訳者に関する洞察と分析が共有された。フリーでオープンソースのシステムから、セルフホスト型のサービス、外部の有料 API まで、Wikipedia で採用している機械翻訳エンジンが紹介され、これらを用いることで 198 の言語を翻訳していると述べられた。

Thottingal 氏の講演で用いられたスライドは WAT2023 のホームページ でダウンロードすることができるので、興味のある方は是非ご覧いただきたい。

4 翻訳タスク

WAT2023 では以下の Shared Task を設定した。

・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
- NICT-SAP: ヒンディー / タイ / インドネシア / マレー / ベトナム → 英　非構造的文書
- 日中韓 → 英　構造的文書

・マルチモーダル翻訳タスク：

- Hindi Visual Genome: 英語 → ヒンディー語
- Malayalam Visual Genome: 英語 → マラヤーラム語
- Bengali Visual Genome: 英語 → ベンガル語
- 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：

- 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 / サンタル / カシミール / マイティリー / サンスクリット / ウルドゥ ó 英

・特許翻訳タスク：

- JPC3: 英中韓 → 日

・制限翻訳タスク：

- ASPEC: 英 / 中 → 日

・対訳コーパスフィルタリングタスク：

- JParaCrawl: 日 → 英

・非反復翻訳タスク：

- JIJI Corpus: 日 → 英

残念ながら今年度はインドの言語のマルチモーダル翻訳タスクに参加した 2 チームの参加にとどまった。

5 まとめ

本稿では WAT2023 全体の概要を報告した。残念ながら WAT2023 では論文の投稿数や翻訳タスクの参加者は非常に少なかったが、ワークショップ当日の参加者はそれなりにあった。また招待講演、研究発表、翻訳タスクの結果それぞれで議論も活発に行われたことから、ワークショップとしては成功だったと言える。

WAT2024 は 11 月 12 日から 16 日にアメリカ、フロリダ州のマイアミで開催される EMNLP2024 の併設ワークショップとして開催される予定である。なお同じ EMNLP2024 で、世界最大の機械翻訳のワークショップである WMT も開催されるため、WAT2024 の翻訳タスクは特別に WMT2024 の元で開催することになっている。なお研究論文の募集や招待講演などは独立に行う予定である。

WAT は今後も継続して開催予定であり、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。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] Toshiaki Nakazawa,Kazutaka Kinugawa,Hideya Mino, Isao Goto, Raj Dabre,Shohei Higashiyama,Shantipriya Parida,Makoto Morishita, Ondˇrej Bojar,Akiko Eriguchi,Yusuke Oda, Chenhui Chu and Sadao Kurohashi. 2023. Overview of the 10th workshop on Asian translation. In Proceedings of the 10th Workshop on Asian Translation (WAT2023), Macau SAR, China. MT-Summit 2023.