NAIST Simultaneous Interpretation Corpus: Development and Analyses of Data from Interpreters of Different Levels

Kosuke Doi[1] Katsuhito Sudoh[1,2] Satoshi Nakamura[1,3]
[1]Nara Institute of Science and Technology [2]Nara Women's University
[3]The Chinese University of Hong Kong, Shenzhen
{doi.kosuke.de8, sudoh, s-nakamura}@is.naist.jp

掲載号の情報

31 巻 3 号 pp. 868-893.
doi: https://doi.org/10.5715/jnlp.31.868

概要

本論文は，我々が構築した大規模な英日・日英同時通訳コーパスである，NAIST Simultaneous Interpretation Corpus (NAIST-SIC) について述べるとともに，本コーパスを用いて行った同時通訳と翻訳字幕の比較，通訳者の経験年数に基づく分析について報告する．

近年の深層学習技術の発展によりさまざまな機械翻訳サービスが実用化されているが，話された言葉を翻訳する音声自動翻訳では，発話の終了を待って訳出を開始する．同時通訳とは，原発話の終了を待たずに訳出を開始する通訳手法であり，これを機械で実現する同時機械翻訳の研究が行われている．同時機械翻訳モデルは，人間の同時通訳者がどのように同時通訳を行なっているかを学習する必要があるが，利用可能な同時通訳コーパスの数は限られており，モデルの学習には不十分である．そのため，多くの同時機械翻訳研究では，文末を待って訳出した翻訳文データから成る，音声翻訳コーパス (例：MuST-C [1]) を学習に用いている．

そこで我々は，合計で 300 時間以上の英日・日英同時通訳データを収録し，人間の通訳者が実際に同時通訳した文の特徴の分析や，同時機械翻訳モデルの学習に利用可能な言語資源として，NAIST-SIC を構築した．本コーパスには，講演や記者会見をプロの同時通訳者が実際に同時通訳した音声とその書き起こしが収録されており，その一部には，経験年数が異なる 3 名の同時通訳者が同一の原文を訳出した「通訳比較用データ」が含まれている．収録時間の内訳は表 1 の通りである．

本論文では，表 1 でアスタリスクで示されている通訳比較用データから，14 本の講演に対する英日同時通訳データを対象とし，訳出遅延，品質，語順の観点から分析を行った．分析にあたっては，原言語 (英語) の文を基準として，原文と同時通訳の対応づけを人手で行い，さらに分析対象のうちの 3 本のデータに対しては，チャンクおよび句レベルでの対応づけも人手で行った．チャンク境界は，同時通訳者の方略 [2] に基づき人手で付与した．分析の結果，経験を積んだ通訳者は，経験が浅い通訳者よりも多くの内容語を訳出できているという点で，高い品質の訳出を行っていることが確認された．加えて，経験を積んだ同時通訳者は，翻訳字幕と比べて原言語文での語順を保持した訳出を行なっており，遅延時間と品質のバランスをよりよく保てていることが明らかになった．また，遅延時間があまりに長くなると，訳出品質に悪影響が生じていることが明らかになった．

表 1 NAIST-SIC の収録時間の内訳．アスタリスク付きの数字は，通訳比較用データの収録時間を示している．

通訳方向	収録時間
英 → 日	167 + 12*
日 → 英	114 + 12*
合計	305

参考文献

[1] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In **Proc. of NAACL**, pp. 2012–2017, 2019.

[2] 岡村ゆうき, 山田優. 「順送り訳」の規範と模範 同時通訳を模範とした教育論の試論. 石塚浩之（編）, 英日通訳翻訳における語順処理–順送り訳の歴史・理論・実践, pp. 217–250. ひつじ書房, 2023.