

講演動画の言語横断字幕生成のための 英日マルチモーダル対訳コーパスの構築

寺面 杏優¹ 近藤 里咲¹ 梶原 智之² 二宮 崇²

¹ 愛媛大学工学部 ² 愛媛大学大学院理工学研究科

{teramen@ai., kondo@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp

概要

本研究では、英語の講演動画とその日本語字幕からなるマルチモーダル対訳コーパスを構築し、公開する。動画からの言語横断字幕生成では、音声認識のエラーが伝播して翻訳品質が劣化してしまう。この課題に対処するために、音声認識文に加えて音声や画像を参照するマルチモーダル機械翻訳が有望である。我々は、このようなマルチモーダル機械翻訳の研究に取り組むために、既存の英日対訳コーパスに対して画像・音声・音声認識文を付与した約10万文対のマルチモーダル対訳コーパスを構築した。英日翻訳における実験の結果、音声認識文の誤り訂正によって、翻訳品質の改善を確認できた。

1 はじめに

音声翻訳 [1] とは、原言語の音声から目的言語のテキストや音声への機械翻訳の技術である。音声アシスタントやオンライン会議などのサービスの普及により、人と人や人と機械の間のグローバルなコミュニケーションを円滑に進めるために、音声翻訳の技術が期待されている。

伝統的な音声翻訳は、音声認識 [2] と機械翻訳 [3] のパイプラインモデルとして実現され、前段の音声認識における誤りが後段の機械翻訳の品質を劣化させてしまう課題がある。この課題に対して、機械翻訳の際に音声認識テキストだけでなく音声や画像を参照するマルチモーダル機械翻訳 [4] のアプローチが有望である。マルチモーダル機械翻訳の先行研究では、音声翻訳 [5-7] や画像を用いる機械翻訳 [8-10] など、対訳コーパスに対して他の1種類のモダリティのデータのみを追加する事例が多い。しかし、例えば講演の発表資料などにアクセスできる状況では、講演の音声と発表資料の画像の両方から有益な情報を得ることが期待できるため、言語・

音声・画像の3種類のモダリティのデータを組み合わせることで翻訳品質のさらなる改善が見込める。3種類のモダリティを組み合わせる先行研究には、How-2 [11] や QED [12] などの動画を用いる機械翻訳がある。ただし、How-2 は英語からポルトガル語の言語対のみを扱い、QED は教育ドメインのみを扱うため、適用範囲が限定されている。

本研究では、言語・音声・画像の3つのモダリティを扱うコーパスの拡充のために、講演動画の英日翻訳を対象とするマルチモーダル機械翻訳コーパスを構築し、英日の言語横断字幕生成に取り組む。既存の IWSLT2017 の英日対訳コーパス [13] をもとに、新たに音声と画像を付与し、さらに音声から音声認識文を生成し、画像・英語音声・音声認識英語文・書き起こし英語文・日本語参照訳の5つ組からなる約10万件のデータセット TAIL¹⁾を作成する。評価実験の結果、音声認識英語文をそのまま機械翻訳するよりも、音声認識文と書き起こし英語文を用いて訓練した音声認識の誤り訂正を間に挟むことで、翻訳品質を改善できることが明らかになった。

2 TAIL コーパス

本研究では、英語から日本語への言語横断字幕生成のために、TED²⁾の講演動画を対象に、画像・英語音声・音声認識英語文・書き起こし英語文・日本語参照訳の5つ組コーパス TAIL を構築する。IWSLT2017 [13] において TED 講演動画の書き起こし英語文と日本語参照訳からなる英日対訳コーパスが公開されているため、本研究ではこれに画像・英語音声・音声認識英語文を新たに付与して5つ組を得る。図 1 にコーパス構築の流れを示し、図 2 には本研究で構築する TAIL コーパスの例を示す。ここで、画像が3枚あるのは、音声の開始・中間・終了の

1) TAIL: English-to-Japanese Translation Corpus with Audio and Images from Lecture Video <https://github.com/EhimeNLP/TAIL>

2) <https://www.ted.com>

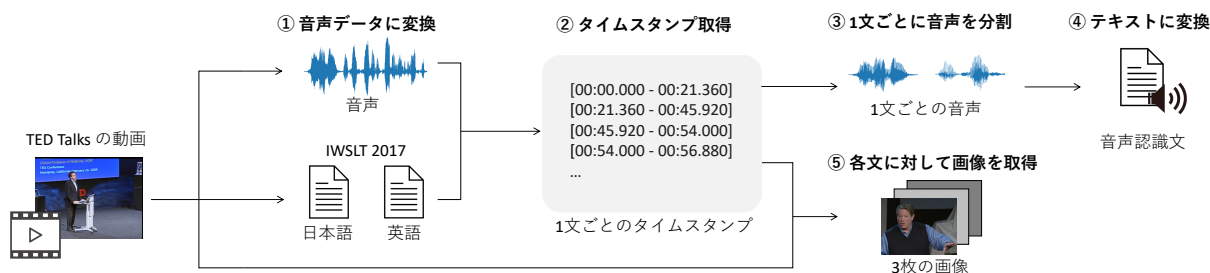


図1 コーパス構築の全体像

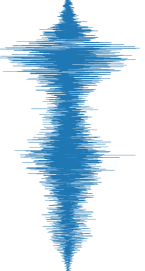



		ASR	just as the biosphere is being severely eroded so too is the ethnosphere and if anything at a far greater rate
		IWSLT - en	And just as the biosphere has been severely eroded, so too is the ethnosphere -- and, if anything, at a far greater rate.
		IWSLT - ja	生物圏が激しく侵されてしまったのと同様に 民族文化圏もーしかももっと急速に 侵食されています (Japanese) Seibutsuken ga hageshiku oka sarete shimatta noto douyou ni minzoku bunkaken mo - shikamo motto kyusoku ni shinshoku sarete imasu

図2 TAIL コーパスの例 (画像・英語音声・音声認識英語文・書き起こし英語文・日本語参照訳の5つ組) ただし、画像は音声の開始・中間・終了の各時刻に対応する3枚を収集。

各時刻に対応する画像を収集しているためである。

2.1 音声アノテーション

本節では、IWSLT2017 英日対訳コーパスに対して、英語音声および音声認識英語文を付与する。

音声データの取得 IWSLT2017 英日対訳コーパスのメタデータに含まれる URL から、MP4 形式の動画データを取得できる。この動画を ffmpeg³⁾ を用いて FLAC 形式の音声データに変換する。これは、図1の手順1に対応する。

音声とテキストの対応付け IWSLT2017 英日対訳コーパスの書き起こし英語文および手順1で取得した音声を講演単位で入力し、aeneas⁴⁾ を用いて音声とテキストを対応付ける。その際に、1文ごとに開始時刻および終了時刻のタイムスタンプを取得するとともに、ffmpeg を用いて音声を分割する。これは、図1の手順2 (タイムスタンプの取得) および手順3 (音声の分割) に対応する。

音声認識 手順3で文ごとに分割した音声を、Google Speech Recognition⁵⁾ を用いてテキストに変換する。これは、図1の手順4に対応する。

2.2 画像アノテーション

本節では、これまで構築してきたコーパスに対して、さらに画像を付与し、5つ組を得る。ここで、TED の講演動画には、講演者のみが写るなど必ずしも講演内容を十分に表現しないシーンも多い。そこで本研究では、各文に対して複数枚ずつの画像を収集する。具体的には、各文のタイムスタンプの開始・中間・終了の各時刻に対応する3枚の画像を用いる。動画および手順2で取得したタイムスタンプを講演単位で入力し、OpenCV⁶⁾ を用いて1文あたり3枚ずつの画像を取得した。

2.3 対訳コーパスフィルタリング

タイムスタンプや対応付けの誤りに起因する対訳コーパスのノイズの影響を軽減するために、自動的な対訳コーパスフィルタリングを実施する。本研究では、音声認識英語文 (ASR) および書き起こし英語文 (REF) の文対を用いて、文長の比および単語誤り率による対訳コーパスフィルタリングを行う。まず、文長差が大きいほどノイズの多い文対であると考え、文長の比による対訳コーパスフィルタリングでは $0.8 \leq \text{len}(\text{ASR})/\text{len}(\text{REF}) \leq 1.2$ の事例のみを残す。ここで、 $\text{len}(\cdot)$ は文の単語数を表す。また、単語誤り率 (WER) が大きいほどノイズの多い文対で

3) <https://ffmpeg.org>

4) <https://github.com/readbeyond/aeneas>

5) https://github.com/Uberi/speech_recognition

6) <https://opencv.org>

表1 音声認識文とその誤り訂正の誤りの分布および訂正の良し悪しの内訳

		誤りを含む文の割合 (%)		訂正の内訳 (%)		
		音声認識文	誤り訂正文	改善	悪化	その他
文体	大文字	27	4	22	0	6
	小文字	85	8	83	1	7
記号	追加	2	31	1	28	5
	削除	100	31	99	3	27
単語	追加	11	19	4	15	7
	削除	58	41	24	9	37
置換	同音異義語	64	50	13	5	45
	表記揺れ	12	6	7	1	5
	活用	20	20	7	8	14
	その他	19	39	2	8	34

表2 TAIL コーパスの統計

	訓練用	評価用
IWSLT2017 英日対訳コーパス	223,108	9,340
5つ組を取得できた文対数	211,604	8,840
対訳コーパスフィルタリング後	101,953	5,040

あると考え、 $WER(ASR, REF) \leq 0.5$ の事例のみを残す。ここで、WER の計算においては、前処理としてテキストを小文字化し、記号を除去した。

表2に、TAIL コーパスの文対数を示す。一部の動画にアクセスできなかったため、IWSLT2017 英日対訳コーパスのうち約5%の文対は本研究では使用していない。対訳コーパスフィルタリングを経て、最終的に約10万文対のマルチモーダル対訳コーパスを構築した。

2.4 音声認識の誤り分析

TAIL コーパスから無作為に100文対を抽出し、音声認識の誤りについて分析した結果を表1に示す。本コーパスの音声認識誤りは、大きく「大文字/小文字」「記号の追加/削除」「語彙の置換」の4種類に分類できる。

大文字/小文字 書き起こし英語文においては、大文字であるが音声認識英語文では小文字として認識されている誤りが、85%の文において見られた。特に、音声認識英語文の文頭は多くの場合に小文字として出力されてしまっていた。反対に、小文字であるべき文字を誤って大文字としてしまう例も27%の文において見られた。

記号の追加/削除 書き起こし英語文に含まれるピリオドや疑問符などの記号は、全ての音声認識英語文において1つ以上が削除されていた。これらの記号の削除は機械翻訳の性能に影響を与える可能性が高いため、機械翻訳の前に適切に復元することが望ましいと考えられる。

語句の追加/削除 書き起こし英語文に含まれない語句の追加に比べて、書き起こし英語文に含まれる語句が音声認識英語文において削除されている事例が多かった。特に、音声とテキストの対応付けのエラーの影響で、文頭の語句が削除されている事例が多かった。

語句の置換 同音異義語・表記揺れ・活用の違い・その他の4種類の語句の置換を調査した。いずれかの語句の置換を伴う音声認識誤りが78%の文に見られ、特に同音異義語や活用の誤りが多かった。

3 評価実験

TAIL コーパスを用いて、音声認識の誤り訂正および音声認識英語文から日本語への機械翻訳の評価実験を行う。誤り訂正および機械翻訳の性能は、SacreBLEU [14] を用いて BLEU [15] を評価する。

3.1 誤り訂正

本実験では、音声認識の誤り訂正を行う。TAIL コーパスの音声認識英語文と書き起こし英語文の対を用いて、事前訓練済みの BART [16] をファインチューニングして誤り訂正器を構築する。

表3 音声認識の誤り訂正の実験結果 (BLEU)

	音声認識文	誤り訂正文
$0.05 \leq \text{WER} \leq 0.15$	60.96	80.09
$0.20 \leq \text{WER} \leq 0.50$	40.56	57.66

モデル 誤り訂正器として、fairseq⁷⁾ [17] を用いて Base 設定の BART⁸⁾ をファインチューニングした。最適化には Adam [18] を使用し、バッチサイズを 2,048 トークンとして、4 バッチごとにパラメータを更新した。検証用データにおけるクロスエントロピー損失が 10 回改善されなければ訓練を停止した。

データ 表 2 に示した 101,953 文対の中から、誤りを含まない文対 ($\text{WER} \leq 0.05$) を除去した 80,859 文対の音声認識英語文および書き起こし英語文を用いて、誤り訂正を訓練した。評価用の 5,040 文対からは、 $0.05 \leq \text{WER} \leq 0.15$ の誤りが少ない評価用データと $0.20 \leq \text{WER} \leq 0.50$ の誤りが多い評価用データを無作為に 1,200 文対ずつ抽出し、その他を検証用データとして使用した。前処理には、BART のトークナイザを用いてサブワード分割をした。

実験結果 表 3 に、BLEU による評価結果を示す。上段の誤りが少ない評価用データと下段の誤りが多い評価用データのそれぞれにおいて、19 ポイントおよび 17 ポイントの大きな BLEU の改善が見られた。この結果から、本コーパスが音声認識の誤り訂正のために有用であると言える。なお、本コーパスに含まれる音声や画像などの他のモダリティのデータを考慮することで、誤り訂正の性能をさらに改善できると期待できるが、これは今後の課題である。

表 1 に訂正結果の詳細な分析を示す。大文字/小文字の誤りや記号の漏れ、文頭における語句の抜けが大きく改善できた。一方で、記号を過剰に追加したり文中の語句を誤削除する事例も見られた。

3.2 機械翻訳

本実験では、IWSLT2017 英日対訳コーパス [13] を用いて訓練した Transformer [3] および JParaCrawl v3.0 [19, 20] の事前訓練済みモデル⁹⁾ を IWSLT2017 英日対訳コーパス上でファインチューニングした Transformer の 2 つの英日翻訳器を用いて、音声認識英語文から日本語への機械翻訳を行う。そして、3.1 節の技術で音声認識の誤りを訂正するこ

表4 英日翻訳の実験結果 (BLEU)

事前訓練なし	音声認識	誤り訂正	書き起こし
$0.05 \leq \text{WER} \leq 0.15$	7.90	8.60	8.73
$0.20 \leq \text{WER} \leq 0.50$	6.52	7.17	8.66
事前訓練あり	音声認識	誤り訂正	書き起こし
$0.05 \leq \text{WER} \leq 0.15$	11.56	11.80	12.54
$0.20 \leq \text{WER} \leq 0.50$	8.87	9.50	11.87

とにより、翻訳品質が改善できることを確認する。

モデル 翻訳器として、fairseq を用いて Base 設定の Transformer [3] を訓練した。最適化には Adam を使用し、バッチサイズを 4,096 トークンとして、8 バッチごとにパラメータを更新した。検証用データにおけるクロスエントロピー損失が 10 回改善されなければ訓練を停止した。

データ 表 2 に示した 211,604 文対の書き起こし英語文および日本語参照訳を用いて、機械翻訳を訓練した。検証用と評価用データは、3.1 節と同じものを用いた。前処理として、英語には Moses Tokenizer¹⁰⁾ [21]、日本語には MeCab (IPADIC)¹¹⁾ [22] を用いて単語分割した。そして、SentencePiece¹²⁾ [23] を用いて Byte Pair Encoding [24] による語彙サイズ 32,000 のサブワード分割を行った。

実験結果 表 4 に、BLEU による自動評価の結果を示す。音声認識の自動的な誤り訂正によって、誤りが多い評価用データと誤りが少ない評価用データの両方において翻訳品質が改善できた。JParaCrawl を用いた事前訓練によって全体に翻訳品質が改善するものの、事前訓練済み翻訳器に対しても音声認識誤り訂正は有効であることがわかった。

4 おわりに

本研究では、英日の言語横断字幕生成のために、講演動画を対象に約 10 万文対の画像・英語音声・音声認識英語文・書き起こし英語文・日本語参照訳の 5 つ組コーパスを構築した。そして、音声認識文と書き起こし文を用いた音声認識の誤り訂正によって、音声翻訳の性能を改善した。今後の課題として、音声や画像を使ったマルチモーダル誤り訂正およびマルチモーダル機械翻訳を行い、言語横断字幕生成の性能をさらに改善したい。

7) <https://github.com/facebookresearch/fairseq>8) <https://huggingface.co/facebook/bart-base>9) <https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl>10) <https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl>11) <https://taku910.github.io/mecab/>12) <https://github.com/google/sentencepiece>

謝辞

本研究成果は、国立研究開発法人情報通信研究機構（NICT）の委託研究（課題番号：22501）により得られたものです。

参考文献

- [1] Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. Recent Advances in Direct Speech-to-text Translation. In **Proc. of IJCAI**, pp. 6796–6804, 2023.
- [2] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. In **Proc. of ICASSP**, pp. 4960–4964, 2016.
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Proc. of NIPS**, pp. 5998–6008, 2017.
- [4] Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. Multimodal Machine Translation through Visuals and Speech. **Machine Translation**, Vol. 34, pp. 97–147, 2020.
- [5] 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.
- [6] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The Multilingual TEDx Corpus for Speech Recognition and Translation. In **Proc. of INTERSPEECH**, pp. 3655–3659, 2021.
- [7] Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus. In **Proc. of LREC**, pp. 4197–4203, 2020.
- [8] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German Image Descriptions. In **Proc. of VL**, pp. 70–74, 2016.
- [9] Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In **Proc. of EMNLP**, pp. 715–729, 2022.
- [10] Shantipriya Parida, Ondrej Bojar, and Satya Ranjan Dash. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. **Computación y Sistemas**, Vol. 23, No. 4, pp. 1499–1505, 2019.
- [11] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A Large-scale Dataset for Multimodal Language Understanding. In **Proc. of NeurIPS**, 2018.
- [12] Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In **Proc. of LREC**, pp. 1856–1862, 2014.
- [13] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. Overview of the IWSLT 2017 Evaluation Campaign. In **Proc. of IWSLT**, pp. 2–14, 2017.
- [14] Matt Post. A Call for Clarity in Reporting BLEU Scores. In **Proc. of WMT**, pp. 186–191, 2018.
- [15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In **Proc. of ACL**, pp. 311–318, 2002.
- [16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In **Proc. of ACL**, pp. 7871–7880, 2020.
- [17] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. Fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In **Proc. of NAACL**, pp. 48–53, 2019.
- [18] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In **Proc. of ICLR**, 2015.
- [19] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In **Proc. of LREC**, pp. 3603–3609, 2020.
- [20] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In **Proc. of LREC**, pp. 6704–6710, 2022.
- [21] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In **Proc. of ACL**, pp. 177–180, 2007.
- [22] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In **Proc. of EMNLP**, pp. 230–237, 2004.
- [23] Taku Kudo and John Richardson. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In **Proc. of EMNLP**, pp. 66–71, 2018.
- [24] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In **Proc. of ACL**, pp. 1715–1725, 2016.