曖昧性を含む翻訳に着目した マルチモーダル機械翻訳データセットの構築方法の検討

Yihang Li　清水周一郎　Chenhui Chu　黒橋禎夫
京都大学大学院情報学研究科
{liyh, sshimizu, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

概要

翻訳には曖昧性，すなわち原言語の文に対して複数の翻訳が考えられる場合がある．既存のマルチモーダル機械翻訳データセットでは，翻訳の曖昧性を解消するために映像が役立つかどうかは明らかになっていない．そこで本研究では，曖昧性を含む翻訳に着目したマルチモーダル機械翻訳データセットの構築に取り組む．このデータセットは，(1) 原言語の文が曖昧であること (2) 映像が翻訳の曖昧性解消に役立つことの二つの条件を満たす．一つ目の条件を満たすために異なる翻訳をもつ原言語の文を集め，さらに目的言語の文の意味に基づいて選定を行った．二つ目の条件を満たすために人手による確認を行った．

1　はじめに

マルチモーダル機械翻訳 [1] は，テキスト以外のモダリティを補助的に用いる機械翻訳である．翻訳には曖昧性，すなわち原言語の文に対して複数の意味の翻訳が考えられる場合があるが，テキスト以外のモダリティを活用することによってそうした曖昧性の解消が期待できる．

これまでのマルチモーダル機械翻訳の研究では，主に画像が翻訳の曖昧性を保証するための補助的な情報として用いられてきた [2, 3]．その発展として，近年，映像を用いた機械翻訳 (video-guided machine translation; VMT) の研究が盛んになりつつある [4, 5]．映像は画像では捉えられない人・物の動きや時間経過といった情報を含むため，翻訳の曖昧性の解消により効果的であると期待される．

既存の VMT データセットは，原言語の文の曖昧性に着目しているとは言えない．VMT のデータセットの構築方法として，キャプション (映像内容の描写) にもとづく方法 [4] と，字幕 (映像内の話者の発話の書き起こし) にもとづく方法 [6] がある．現実に VMT を利用する状況，例えば映画の字幕翻訳などの場合，テキストは必ずしも映像の内容を説明している訳ではなく，映像はテキストで伝えられる内容とは別の補助的な情報として用いられる．そのため，キャプションにもとづくデータセットは実用的な観点から問題がある．また，キャプションには曖昧性が存在しないため，キャプションの翻訳は必ずしも視覚情報を必要としないという報告もある [7]．字幕にもとづくデータセットでも，曖昧性には特に注目せず，一般的な字幕を扱ったものしか存在しない．

本研究では，原言語の文が曖昧性を含む日英 VMT データセットの構築に取り組んでいる．このデータセットは，(1) 原言語の文が曖昧であること (2) 映像が翻訳の曖昧性解消に役立つことの二つの条件を満たす．

一つめの条件を満たすため，まず原言語 1 文に対して翻訳が複数存在するものを集める．ここで，原言語 1 文と複数の翻訳文の集まりを**対訳セット**と定義する．次に，収集した対訳セットから目的言語の文間の意味が違うもののみを選定する．二つめの条件を満たすためには人手による確認を行う．

選定された対訳セットの例を図 1 に示す．「放せ！」という日本語に対し，“Let me go!” と “Drop it!” の 2 通りの翻訳があり，テキストだけからはどちらが正しい翻訳か判断できない．映像を見ると，上の例では人が人を引っ張っており，下の例ではナイフを持った男が争っていることから，正しい翻訳を決定できる．

本稿では，曖昧性に着目した VMT データセットの構築手順を説明する．さらに，データセットの一部について，データの分析及び予備実験の結果を報告する．

図 1 選定された対訳セットの例．「放せ！」という日本語に対し，“Let me go!” と “Drop it!” の 2 通りの翻訳の可能性がある．映像を見ると，上の例では人が人を引っ張っており，下の例ではナイフを持った男が争っていることから，対応する翻訳を決定できる．

2 関連研究

2.1 マルチモーダル機械翻訳

マルチモーダル機械翻訳では，異なるモダリティの情報が入力データの異なる見方を与えると仮定して，複数のモダリティの情報を用いて翻訳を行う [8]．先行研究の多くは画像を用いた機械翻訳 [1, 9, 10] である．映像の有用性は特定のデータセットやタスクで議論されており [11, 7]，類似していない言語対では文法的な特徴を捉えたり翻訳の曖昧性を解消したりするのに映像が役立つ可能性がある [8]．

2.2 VMT データセット

VMT のデータセットとして VaTeX [4] と How2 [6] がある．VaTeX は既存の映像データセットに対し英語と中国語でキャプションを付けたデータセットである．VaTeX はキャプションのデータセットであり，VMT を用いる現実的なシナリオを考えにくい．How2 は YouTube から集められた教育ビデオに，その英語字幕と対応するポルトガルとの翻訳，及び映像の英語の要約が付けられたデータセットである．いずれのコーパスもテキストに曖昧性がないため，翻訳に必ずしも視覚情報を必要としない点が本研究と異なる．

3 データセット構築手順

原言語に曖昧性を含む日英の VMT データセットを構築する手順は以下の通りである．まず，既存の字幕データセットである OpenSubtitles [12] の日英の字幕から対訳セットを集める．さらに，Sentence-BERT によって対訳セット内の複数の英語訳の意味が異なるかどうかを判定し，対訳セットの選定を行う．次に，字幕に対応する映画やテレビの映像を切り取る．最後に，人手で字幕と映像をチェックし，字幕に曖昧性があり映像によりそれが解消できることを確認する．

3.1 対訳セットの収集

OpenSubtitles は映画やテレビのデータから収集された大規模な字幕のデータセットで，計 60 言語にわたる 26 億文の字幕がある．本研究ではこのうち 1,213,468 文からなる日英の対訳データを用いる．日本語を原言語として選んだのは，日本語がプロドロップ言語であり，曖昧性を原言語に含むデータセットを作成する上で適当であると考えたためである．また，OpenSubtitles では各映像に対し IMDb[1] の ID が付けられており，映像の情報を得ることができる．字幕もこの ID に基づいて分類されている．翻訳の曖昧性を確保するために，原言語 1 文と複数の翻訳文の集まりである対訳セットごとに字幕をまとめる．

3.2 目的言語の文間の類似度を利用した対訳セットの選定

複数の目的言語の翻訳が存在するだけでは，原言語の文が曖昧性を含むことにはならない．それらが単に表現として違っているだけで，意味的には同じである可能性があるからである．そこで，各対訳セット内の目的言語の文同士の類似度を Sentence-BERT [13] を用いて計算し，類似度の閾値を設定して，閾値よりも類似度が小さい (すなわち意味の異なりが大きい) 対訳セットのみを残すことで曖昧な文を含む対訳セットの選定を行う．ここで，1 対訳セット内に 3 文以上の目的言語の文が含まれる場合には，全ての文の組み合わせについて類似度を計算し，最も類似度の小さいものをその対訳

1) https://www.imdb.com/

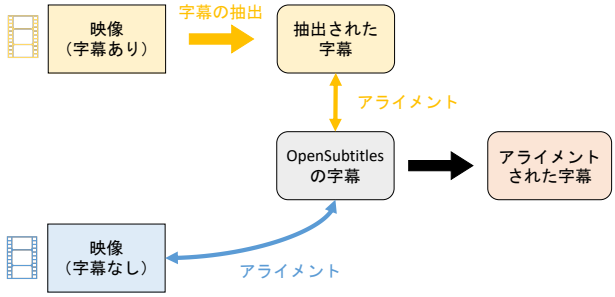

図 2　字幕のアライメントの手順．映像に字幕がついている場合 (上側) とついていない場合 (下側) で手順を分ける．

セットの目的言語文間の類似度として用いる．

3.3　字幕と映像のアライメント

OpenSubtitles の字幕にはその表示時刻の始まりと終わりを示すタイムスタンプがついているが，それらは必ずしも映像のタイムスタンプと一致しているわけではない．OpenSubtitles で用いられた映像と本研究で用いた映像でフレームレートに差があったり，定数時間のずれが生じていたりする場合があるからである．そこで，OpenSubtitles の字幕と使用する映像の間でタイムスタンプのアライメントをとる必要がある．なお，OpenSubtitles の字幕では原言語と目的言語それぞれにタイムスタンプがついているが，それらは概ね一致しているため，例外は手作業で取り除いて英語のタイムスタンプを OpenSubtitles のタイムスタンプとする．

アライメントの手順を図 2 に示す．アライメントには Alass[2] (Automatic Language-Agnostic Subtitle Synchronization) というツールを用いる．映像に付けられた字幕には，容易に字幕を抽出できる場合と，字幕がついていない (但し OpenSubtitles には対応する字幕がある) 場合がある．なお，映像に字幕が埋め込まれているような映像は使用しない．字幕が抽出できる場合には，FFmpeg[3] を用いて映像から字幕を抽出し， Alass を用いて OpenSubtitles の字幕とのアライメントを取る．このアライメントはかなり正確に行うことができる (Alass の論文内の実験では 100% 正確であったと主張されている)．字幕がついていない場合には，Alass の音声区間検出をベースにした手法で映像と OpenSubtitles の字幕のアライメントを取ることで字幕を得る．この場合，抽出された字幕のタイムスタンプは必ずしも正確でないため，アライメントが正しくできているかどうかを人手で確認する．

2）https://github.com/kaegi/alass
3）https://ffmpeg.org/

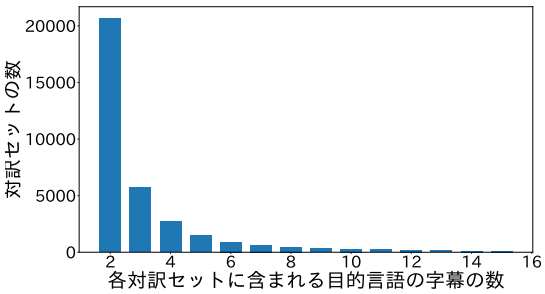

図 3　対訳セットを含まれる目的言語の文数にもとづき分類した結果．

3.4　映像の分割

字幕と映像のアライメントに基づき，映像を分割する．各映像は字幕のタイムスタンプの区間の中央値の前後 5 秒間をとって 10 秒間とし，25 fps の mp4 ファイルとする．なお，集められた映像についた音声の多くは目的言語の英語であり，音声がついていると翻訳文が音声認識で得られることになってしまうため，映像から音声は取り除く．

4　データの分析及び予備実験

データセット構築手順において，対訳セットの選定はデータセットの質に大きく関わる．本章では，収集・選定した対訳セットのデータの統計と予備実験結果を示す．

4.1　対訳セットの収集

対訳セットを，含まれる目的言語の数にもとづき分類した結果を図 3 に示す．全部で 35,023 の対訳セットが得られた．2 文のみ目的言語の文を含む対訳セットが最も多く，対訳セット数で見ると全体の 59%，文数では全体の 30% を占めていた．3 文以上の目的言語の文を含む対訳セットについては，2 文のみのものに比べて曖昧性を含む対訳が含まれる割合は高かったが，文の長さが短いものが多く含まれていた．

4.2　対訳セットの選定のための人手評価データセットの作成

対訳セットの選定のための目的言語の文間の類似度の閾値を設定するため，まず人手で 100 個の対訳セットを選んで曖昧さの正解を与えた．対訳セットの正解を与えるための手順は以下の通りである．(1)

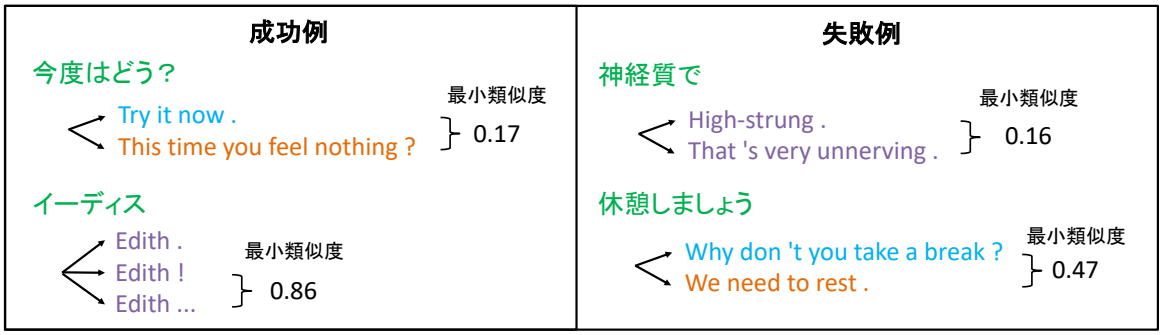

図 4　Sentence-BERT のスコアによる対訳セットの選定の成功例と失敗例.

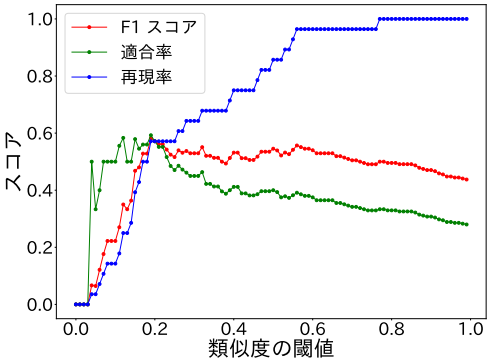

図 5　類似度の閾値を変化させたときの適合率，再現率，及び F1 スコア.

表 1　対訳セット選定の各段階における対訳セット数.

確認した対訳セット	100
原言語の文が曖昧なもの	28
テキストのみから曖昧性を解消可能と推測されたもの	22
映像を見て曖昧性が解消できると確認されたもの	12

対訳セット内の全ての対訳を見て，原言語の文の意味に曖昧性があるかどうかを判断する．(2) 原言語の文に曖昧性があると判断した場合，対訳セット内の全ての対訳を見て，目的言語の複数の文が複数の意味を持ち，映像があればその曖昧性が解消できるかどうかを映像なしで予測する．(3) 曖昧性が解消できると予測した対訳セットについて，実際に映像を見て，翻訳を一意に決定できるかどうかを判断する．

各段階の対訳セット数を表 1 に示す．最終的に 100 個のうち 12 個の対訳セットが，映像に基づいて曖昧性を解消できる対訳を含んでいた．

4.3　目的言語の文間の類似度を利用した対訳セットの選定

Sentence-BERT で計算した文間の類似度のスコアの閾値を 0 から 1 まで 0.01 刻みで変化させ，曖昧性を人手で判断した 28 個を正解としてどの程度正しく対訳セットを分類できるかどうか計算した．閾値を変化させたときの適合率，再現率，及び F1 スコアを図 5 に示す．図 5 を見ると，類似度が 0.2 の付近で F1 スコアが最も高くなっていることが分かる．

閾値を 0.2 に設定したときの，対訳セットの選定の成功例と失敗例を図 4 に示す．類似度が閾値より小さく目的言語の文の意味が異なる場合，または閾値より大きく意味が同じ場合が成功であり，それ以外が失敗である．

今後，大規模な対訳セットの集まりに対し，この閾値を利用して対訳セットの選定を行った上で，映像の曖昧性解消可能性の確認のためにクラウドソーシングを行って，曖昧性を含む翻訳に着目した大規模な VMT データセットを構築する予定である．

5　おわりに

本研究では，VMT における翻訳の曖昧性の問題を指摘し，対訳セットの収集・選定によるデータセットの構築方法を示した．既存の字幕データセットをもとに，原言語の文に対して複数の翻訳が存在する対訳セットを集め，文の類似度を利用して対訳セットを選定することで，原言語の文の曖昧性を確保した．さらに，人手の確認により映像が翻訳の曖昧性解消に役立つことを確認した．本稿では，映像が翻訳の曖昧性解消に役立つかどうかについての評価は小規模に人手で行ったが，今後これをクラウドソーシングで行い，曖昧性を含む翻訳に着目した大規模な VMT データセットを構築する予定である．

謝辞　本研究は科研費#19K20343 の助成を受けたものである．

参考文献

[1] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In **Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers**, pp. 543–553, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In **Proceedings of the 5th Workshop on Vision and Language**, pp. 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. Double attention-based multimodal neural machine translation with semantic image regions. In **Proceedings of the 22nd Annual Conference of the European Association for Machine Translation**, pp. 105–114, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[4] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 4581–4591, 2019.

[5] Tosho Hirasawa, Zhishen Yang, Mamoru Komachi, and Naoaki Okazaki. Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020. In **The First Workshop on Advances in Language and Vision Research: Video-guided Machine Translation (VMT) Challenge**, July 2020.

[6] 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 **NeurIPS**, 2018.

[7] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4159–4170, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] 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, No. 2, pp. 97–147, 2020.

[9] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In **Proceedings of the Second Conference on Machine Translation**, pp. 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[10] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**, pp. 304–323, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[11] Desmond Elliott. Adversarial evaluation of multimodal machine translation. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2974–2978, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[12] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[13] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.