日本語話し言葉書き言葉変換による大学講義の日英翻訳の精度向上

中尾 亮太　Chenhui Chu　黒橋 禎夫
京都大学大学院情報学研究科
{nakao, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

教育の国際化により、外国人留学生が増えてきている。これに伴い、日本語を勉強中の学生をサポートする必要が増加している。この需要に応えるため、日本語講義の英訳システムを株式会社アドバンスド・メディアとの協力により開発した。本システムは講師が日本語で話す音声を、音声認識器によってテキストデータに変換し、おおよそ文単位となるように適当な位置で分割した日本語字幕と、それを機械翻訳した英語字幕を生徒のスマートフォンや講義室のスクリーンにリアルタイムで表示する。これにより講義内容の理解の補助と同時に日本語の習得を手助けできる。

本システムは講義音声を他言語字幕に翻訳する既存のシステム [1][2] と同様に、音声認識器や翻訳モデルをそれぞれ独立したコンポーネントとして構成する。音声認識器は株式会社アドバンスド・メディアが、文区切り器と日英翻訳モデルは我々が開発した。

ここで用いる機械翻訳モデルに関して、広く利用可能な日本語テキストの多くは書き言葉で記録されているため、このモデルも書き言葉のコーパスで事前訓練されている。一方、講義では話し言葉が用いられる。

そこで、本研究では日本語の話し言葉と書き言葉の違いに着目し、翻訳の前処理として書き言葉に自動変換することで翻訳精度を向上させる。そのために、大学講義の書き起こしとそれを書き言葉に変換したもの、対応する英文の３つ組からなるコーパスを作成した。次にそれを用いて話し言葉書き言葉変換モデルと日英翻訳モデルを学習させた。その結果として、話し言葉書き言葉変換が日英翻訳の精度を向上させることを示した（図 1）。

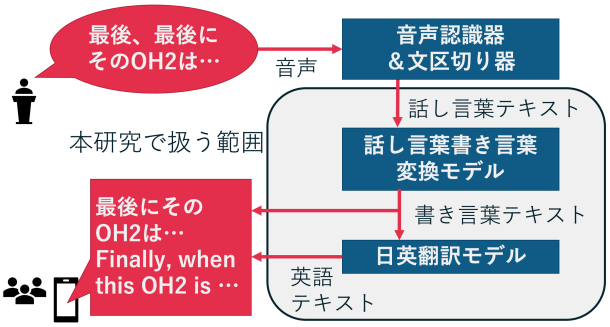

図 1　大学講義の日英翻訳システムの概要

2　関連研究

2.1　話し言葉と書き言葉の違い

話し言葉に特有の現象を、島津ら [3] は４つに大別している。

- 語彙的な現象：音韻的な縮約、口語的な助詞、敬語、感動詞
- 省略：格助詞・判定詞の省略
- 冗長な表現：言い直し、繰り返し、言い換え、言い淀み
- 文の概念がないことに関連する現象：名詞句・助詞句発話、格助詞での中止、ねじれ、倒置

他に話し言葉の性質について言及した研究として [4] があるが、作業者にとって過剰に広範かつ詳細であるため、本研究でのコーパス構築作業にあたっては [3] を参考にした。

これらのうち、どの現象を本研究の対象とするかは次章以降で議論する。

2.2　話し言葉の自動整形

日本語の話し言葉を自動で修正し読みやすくする事例として国会会議録を音声データから自動で書き起こす研究がある [5]。これを含めて音声認識分野の研究では書き起こしの忠実性や発話者の意図の尊

講義名	書き起こし発話数
有機化学基礎および演習	2,815
物理化学基礎及び演習	4,520
電気回路基礎論	4,159
化学プロセス工学基礎	2,049
基礎無機化学	4,041
合計	17,584

表 1　書き起こしの講義名と文数

重のため、言い直しや繰り返しはそのままテキストに残す場合が多く、冗長な部分を修正するとしても文末の「ですね」などに限られる。本研究では、後続の翻訳タスクでの影響を重要視し、2.1 章に示した話し言葉と書き言葉の違いを可能な限り網羅するような修正を施す。

2.3　非流暢性の除去による翻訳精度向上

[6][7] のように、フィラーや冗長な表現を除去することで翻訳精度を向上させる研究は存在する。しかし日本語では、それらだけでなく「ですんで」「こっから」のような音韻的な縮約による違いも多く見られる。本研究では非流暢性だけでなく語彙的な現象による翻訳精度への影響を調査する。また、整形されたテキストを字幕として提供することは日本語初学者が勉強するに当たり、辞書を引きやすくなる、標準的な日本語を覚えられるなどのメリットがある。

3　コーパスの構築

ここで示すデータは京都大学で 2019 年度に行われた 5 つの講義の書き起こしをもとに構築した。表 1 に講義名と分割後の文の数を示す。以後、この書き起こしのデータを $J_{話}$ と表記する。

3.1　話し言葉書き言葉変換コーパスの構築

2.1 章で述べた話し言葉と書き言葉の違いを具体例とともに作業者に提示して変換を依頼した。ただし、それらの話し言葉と書き言葉違いのうち、「名詞句・助詞句発言」「ねじれ」はコンテキストへの依存性の高さや専門的な内容に言及している場合の判断の難しさから、作業者には変換を指示しなかった。また、講義で使われる敬語の多くは丁寧語であり、書き言葉でもよく見られる表現であるため、「敬語」に関する違いも変換対象としなかった。変換後のデータを以後 $J_{書}$ と表記する。$J_{話}$-$J_{書}$ の訓練・開発・テストセットの文数を表 2 に示す。

分割	$J_{話}$-$J_{書}$	$J_{話}$-$J_{書}$-E
訓練	11,072	3,353
開発	1,384	275
テスト	1,384	367
合計	13,840	3,995

表 2　構築したデータセットの文数

3.2　対訳コーパスの構築と統合

講義の書き起こし文から「こんにちは。」などの改めて翻訳する価値の低い文や内容の重複する文を除いて翻訳業者に依頼し、日英対訳コーパスを作成した。この英訳を以後 E と表記する。これと話し言葉書き言葉変換コーパスを統合し、$J_{話}$-$J_{書}$-E データセットを構築した。統合されたデータの文数は表 2 に示す。

4　話し言葉書き言葉変換

4.1　手法

LaserTagger[8] は、Sequence to Sequence の問題を Sequence Labeling として解くことが出来るモデルで、小さいデータセットでも良い性能を出せることと推論が高速であることが特徴である。話し言葉と書き言葉変換のタスクは、2.1 章で示したように、多くは語彙の変換や削除で対応できるため、LaserTagger が適用できる。ただし、倒置の修正は多くのトークンにまたがる変換であるため LaserTagger では解くことが難しいというデメリットがある。一方 Sequence to Sequence の手法はそのような制限はないため解ける可能性がある。

LaserTagger に必要となる事前訓練された BERT[9] モデルは、日本語 Wikipedia で訓練された $BERT_{large}$[10] を用いる。また、LaserTagger の Decoder には Transformer Decoder を使用した。

比較対象として、Sequence to Sequence のタスクで標準的なモデルである Transformer[11] を単言語コーパスで事前訓練できる手法である MASS[12] を用いて日本語話し言葉コーパス [13] で事前訓練したモデルを用いる。パラメータは MASS 著者による参考実装と同じものを用いた。

$J_{話}$-$J_{書}$ データに対して上記の 2 モデルを fine tuning し、評価には SARI[14] という指標を用いる。文の単純化のタスクで提案された指標で、1-gram から

手法	維持	追加	削除	SARI
LaserTagger	**91.9**	**65.6**	**87.7**	**81.7**
Transformer + MASS	69.8	34.6	70.1	58.1

表 3　話し言葉書き言葉変換の評価

4-gram ごとの追加、維持の F1 スコアと削除の適合率の平均をスコアとする。

4.2　結果と考察

結果を表 3 に示す。Transformer+MASS に比べ、LaserTagger が高い精度を持っていることがわかる。これは話し言葉と書き言葉でほとんどのトークンが同一であるというタスクの特性によるところが大きい。

5　話し言葉書き言葉変換の日英翻訳への影響

5.1　手法

日本語話し言葉書き言葉変換モデルには、4 章で述べた LaserTagger を用いる。

日英翻訳モデルには Transformer を用いる。科学技術振興機構 (JST) が ASPEC[15] と同様の手法で、科学技術論文の概要の和文と英文をアラインメントして構築した日英対訳コーパス（訓練データ数：約 13.8M ペア）を用いて事前訓練し、$J_{話}$-$J_{書}$-E コーパスの $J_{話}$-E または $J_{書}$-E で fine tuning を行う。本実験では参考実装である tensor2tensor[16] に実装された transformer_big モデルを利用する。2,000 ステップごとに開発データの翻訳を行って BLEU スコアを算出し、10 回連続で精度向上が見られなかったときに学習を中断する early stopping を行う。ただし fine tuning 時は 200 ステップごとに算出する。

この実験では表 2 の $J_{話}$-$J_{書}$ のデータ全体は使わず、$J_{話}$-$J_{書}$-E の対応がついている部分のみを用いる。LaserTagger には $J_{話}$-$J_{書}$ の変換を学習させ、Transformer は $J_{書}$-E のペアで fine tuning を行う。これによって話し言葉を自動で書き言葉に変換した後に日英翻訳を行うことができる。比較対象として、$J_{話}$-E を学習させた Transformer に $J_{話}$ のテストデータを入力する場合と、$J_{書}$-E を学習させた Transformer に $J_{書}$ のテストデータを入力する場合を考える。前者は話し言葉書き言葉変換による精度の変化を定量化するのに使い、後者は話し言葉書き言葉変換による精度向上の上限の指標とする。

最終的な評価指標には sacrebleu[17] による BLEU スコアを使用する。

手法	BLEU
$J_{話}$-E	20.7
$J_{書}$-E	**21.4**
LT-E	**21.4**

表 4　話し言葉書き言葉変換による翻訳精度への影響

5.2　結果と考察

以後、LaserTagger による出力を LT と表記し、*-E はそれぞれ対応する日英翻訳モデルからの出力を表すこととする。

結果を表 4 に示す。話し言葉書き言葉変換により BLEU スコアが向上しており、$J_{書}$-E と同等のスコアを示した。

実際の変換結果を表 5 に示す。

表 5（上）は書き起こし $J_{話}$ に繰り返しを含んでいる例である。LaserTagger は繰り返しを適切に削除することで簡潔な英訳の生成に成功している。

表 5（中上）では音韻的な縮約と話し言葉に特有の終助詞「ね」を含む例で、LaserTagger は修正に成功している。また、修正後のテキストの英訳は「できる」の意味が訳出されておりより適切なものとなっている。しかし、このような語彙的な現象の修正による英訳の改善は珍しいことがわかった。例えば音韻的な縮約「んです」を含む 48 例の英訳を調査したところ、過半数は意味のある変化が見られず、改善した例と悪化した例の量の差も大きくなかった。

表 5（中下）は倒置を含んでおり、LaserTagger は修正することができずに単に「、ちょうど」を削除している。また、「なんです」という音韻的な縮約を「なのです」に修正しているが、この例では $J_{話}$-E に問題は見られない。

表 5（下）は LaserTagger が変換に失敗した影響で英訳にも問題が伝播している例である。「次の次」は繰り返しにはなっているが削除すべき部分ではなく、これを削除してしまったために英訳の情報が欠けている。また、英訳には影響は見られないものの「であれば」が不自然に挿入されていたり「は」だけが残っていたりと文法的におかしくなっている。

6　おわりに

本研究では日本語の話し言葉と書き言葉の違いに着目し、両者の違いを表すデータセットを構築した

データ	具体例
$J_{話}$	で次、B ということで、原子の分率座標、分率座標と投影図。
$J_{書}$	次、B ということで、原子の分率座標と投影図。
LT	次、B ということで、原子の分率座標と投影図です。
$J_{話}$-E	Next, B. The atomic fraction coordinates, the fraction coordinates, and the projection drawing.
$J_{書}$-E	Next, B, the atomic fraction coordinates and the projection drawing.
LT-E	Next, B is the atomic fraction coordinate and the projection drawing.
E	Next, regarding B, the fractional coordinates of the atom and the projection drawing.
$J_{話}$	だから圧力が変わってるときに、エントロピーがどう変わるかっていうのは 実際に計算できるんですね。
$J_{書}$	だから圧力が変わっているときにエントロピーがどう変わるかというのは 実際に計算できるのです。
LT	だから圧力が変わっているときに、エントロピーがどう変わるかというのは 実際に計算できるのです。
$J_{話}$-E	It calculates actually how the entropy changes when the pressure changes.
$J_{書}$-E	So, how entropy changes when pressure changes can be actually calculated.
LT-E	So, it is actually possible to calculate how the entropy changes when the pressure changes.
E	Therefore, when the pressure changes, we can calculate how the entropy changes.
$J_{話}$	この微分が 0 になるところがここなんです、ちょうど。
$J_{書}$	この微分が 0 になるところがちょうどここなのです。
LT	この微分が 0 になるところがここなのです。
$J_{話}$-E	The point where this differential becomes zero is here.
$J_{書}$-E	There is a differential of zero.
LT-E	This is where the differential becomes zero.
E	This is the point where the differential becomes zero.
$J_{話}$	でその次だったり次の次は○○先生が担当されて、…
$J_{書}$	その次と次の次は○○先生が担当され、…
LT	その次であればは○○先生が担当されて、…
$J_{話}$-E	The next one is Dr. Xxx, …
$J_{書}$-E	Next and next, Dr. Xxx is in charge of giving a lecture …
LT-E	In the next class, Dr. Xxx will give you his lectures …
E	Professor Xxx will give you the next class, and the next after the next, …

表 5 $J_{話}$-$J_{書}$-E コーパスとモデルによる変換・翻訳の具体例（「…」は編集により省略した箇所。赤字は話し言葉書き言葉変換による修正を期待する箇所、緑字は逆に修正すべきでない箇所、青字は修正ミスを表す。）

上で、話し言葉から書き言葉に自動で変換する手法を用いて日英翻訳の精度が向上することを示した。

今後の課題として、本研究では取り扱わなかったが、音声認識に誤りがある場合はそれがそのまま英訳されてしまう問題がある。今回のように翻訳の前処理としてそれらを修正したり、翻訳モデルを誤りに対してロバストにすることで対処するか、音声から直接翻訳テキストを生成する手法によりこれを解決できる可能性がある。

また、データ量の不足から関西弁などの方言を標準語に修正することはできていない。翻訳コーパスとして方言を含んだ日本語テキストは非常に少なく翻訳モデルの精度に影響があると考えられ、また外国人留学生をサポートするという目的においても標準語での字幕を提供することには価値がある。

謝辞 本研究は、科研費#19K20343 の助成を受けたものである。また、京都大学における講義翻訳プロジェクトを立ち上げられた大嶋 正裕 工学部長、協力いただいた株式会社アドバンスド・メディアに感謝致します。

参考文献

[1] Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 82–86, San Diego, California, June 2016. Association for Computational Linguistics.

[2] 須藤克仁, 林輝昭, 西村優汰, 中村哲. 授業アーカイブの翻訳字幕自動作成システムの試作. 研究報告自然言語処理 (NL), Vol. 2019, No. 15, pp. 1–4, 2019.

[3] 島津明, 中野幹生, 堂坂浩二, 川森雅仁. 話し言葉対話の計算モデル. コロナ社, 第 1 版, 2 2014.

[4] 杉戸清樹, 前川喜久雄, 小磯花絵, 西川賢哉, 間淵洋子, 小椋秀樹, 山口昌也, 丸山岳彦, 高梨克也, 内元清貴, 藤本雅子, 菊池英明, 五十嵐陽介, 塚原渉. 日本語話し言葉コーパスの構築法, 3 2006. https://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/CSJ_rep.pdf.

[5] 秋田祐哉, 三村正人, 河原達也. 会議録作成支援のための国会審議の音声認識システム. 電子情報通信学会論文誌, Vol. 93-D, No. 9, pp. 1736–1744, 2010.

[6] Wen Wang, Gokhan Tur, Jing Zheng, and Necip Fazil Ayan. Automatic disfluency removal for improving spoken language translation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5214–5217, 2010.

[7] Elizabeth Salesky, Susanne Burger, Jan Niehues, and Alex Waibel. Towards fluent translations from disfluent speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 921–926, 2018.

[8] Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. Encode, tag, realize: High-precision text editing. In *EMNLP-IJCNLP*, 2019.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, Vol. abs/1810.04805, , 2018.

[10] 柴田知秀, 河原大輔, 黒橋禎夫. Bert による日本語構文解析の精度向上. 言語処理学会 第 25 回年次大会, pp. 205–208, 3 2019.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, Vol. abs/1706.03762, , 2017.

[12] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pp. 5926–5936, 2019.

[13] 前川喜久雄. 『日本語話し言葉コーパス』の概要. 日本語科学, Vol. 15, pp. 111–133, apr 2004.

[14] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 401–415, 2016.

[15] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2204–2208, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).

[16] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. Tensor2tensor for neural machine translation. *CoRR*, Vol. abs/1803.07416, , 2018.

[17] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Belgium, Brussels, oct 2018. Association for Computational Linguistics.