動的時間伸縮法を用いた複数翻訳文書からの対応関係の自動抽出

中屋和樹[1] 川又泰介[1] 松田源立[1]
[1]成蹊大学 理工学部
dm226206@cc.seikei.ac.jp {kawamata,matsuda}@st.seikei.ac.jp

概要

自然言語処理のタスクを実行する上で、言い換えは非常に重要な技術である。語彙力のある人と同様に、機械も豊かな表現能力を有することが望ましい。そのためには、何らかの方法で意味的対応関係にある文を集める必要がある。本研究では複数の翻訳者によって翻訳された文書を対象として、対応関係を抽出する手法を提案する。具体的には翻訳文書を一定のルールで分割して文章群に変換し、既存の類似度指標によって文書間の距離を計算する。その後、文章を次々と組み合わせて DTW(動的時間伸縮法)[1]を適用し、その値が最小となる時のアライメントを最終的な解とする。海外の文学作品を対象とした実験の結果、組み合わせ操作を取り入れた DTW の最小化が対応関係の抽出に有効であることが示された。

1 はじめに

文章で何かを伝える時、その表現方法は一通りとは限らない。文章を構成するパーツには様々な選択肢が存在し、相手や状況に応じて最適なものを選択して使うことが重要である。例えば「このスマホは色々なことが出来ます」という表現は、携帯ショップの店員が顧客に対して説明する際には適切だが、EC サイトの商品説明欄においては「このスマホは機能面で充実しています」という表現の方が好ましいだろう。このように同一の意味内容を有し、異なる言語表現で表されたものを言い換えという。自然言語処理にはさまざまな応用タスクがあるが、言い換え技術はそれらのタスクに最も関係のある基礎的技術と言っても良い。

実際に言い換え技術を用いたタスクに取り組む際は、事前に言い換え事例を大量に集める必要がある。現在までに様々な方法が提案されているが、本研究では〈複数の翻訳者によって翻訳された文書は豊富な言い換えを含んでいる〉という考えのもと、海外文学作品を対象として対応関係の抽出を目指す。ここで「対応関係の抽出」というフレーズを使用したのは、翻訳という行為は翻訳スタイルによる個人差が生じやすく、必ずしも言い換えとは言い切れない事例が存在するためである。これまでにも複数の翻訳例を集めて言い換えを獲得する試みは行われている[2][3]。しかし、書籍規模の文書での検証を行なっており、かつ日本語を対象とした研究は確認できなかった。

[4]ではテトゥン語を対象とした機械翻訳のデータ収集の手段として、DTW を用いたアライメントを試みている。テトゥン語のような母語話者の少ない言語の機械翻訳では、モデル構築以前にデータの枯渇が問題となることがある。そこで松本らは JICA、UNIFEC によって公開されているテトゥン語と英語の文書から、テトゥンー英の対訳コーパスを構築している。文分割処理を施した後に DTW による文アライメントを実行しているが、長さの異なる系列データに一度しか適用していないため、1 対多のような対応関係から不適切な対訳が複数抽出される可能性がある。また文章の特徴量として文章長を用いることが最適であるかは定かではない。

上記の懸念は、本研究のタスクである翻訳文書からの対応関係の抽出において中心的な問題となる。原文一文を一文のまま訳す翻訳者もいれば、二つや三つに分割して訳す翻訳者もいるため、[4]の手法をそのまま適用すると句点を境界として意味が分断される文章に対して適切な対応付けができない。このような、対応関係の散乱した文章群のアライメントを人手で行うことは容易だが、コスト面からも機械での自動化が望ましい。そこで文章同士を組み合わせる操作を行い、機械でも対応付けが可能な機構を導入する。

本研究の貢献を以下に示す。

- 組み合わせ操作を取り入れた DTW の最小化が対応関係の抽出に有効であることを示した

- DTW を実行する際の文書間の距離として、文字レベルの Jaccard 係数もしくは SentenceBERT によるユークリッド距離が有効であることを示した

2 関連研究

2.1 DTW(動的時間伸縮法)[1]

DTW は二つの時系列データ同士の類似度を動的計画法によって求めるアルゴリズムである。具体的には、二つの時系列データが与えられた時、DTW は各系列の点同士の距離を総当たりで計算し、波形の距離が最短となるような経路を求める。二つの時系列データを$X = \{x_1, x_2, , , x_m\}$ $Y = \{y_1, y_2, , , y_n\}$とすると以下のように定式化される。

$$DTW(X, Y) = \min_{path}(table(X, Y, dis, path))$$

ここで$table()$はデータ列X, Y及び距離関数$dis()$が与えられた元で生成される DP テーブルであり、$path$は波形間の距離が最短となるような経路を表す。

2.2 文の類似度指標・埋め込み生成モデル

本研究では文章の類似度を評価する指標として、Jaccard 係数、文ベクトル間のユークリッド距離を用いる。

Jaccard 係数は集合同士の類似度を測る指標であり、文字や単語を集合の要素とみなして文同士の評価にも使用することができる。以下に式を示す。

$$Jaccard(Sent1, Sent2) = \frac{|\Omega(Sent1) \cap \Omega(Sent2)|}{|\Omega(Sent1) \cup \Omega(Sent2)|}$$

ここで$\Omega(Sent)$はある文$Sent$に対する文字レベルの集合である。

文章をベクトルに変換するモデルを総称して文章埋め込み生成モデルと呼ぶ。現在までに様々な手法が提案されているが、それらは主にカウントベースの方法、単語ベクトルの加重平均を取るなどして生成する方法、深層学習で直接ベクトルを生成する方法、のいずれかに分類されることが多い。近年は、距離学習によって文章をベクトル空間上に適切にマッピングする手法(上記の 3 番目)が主流となっている。本研究では文章埋め込み生成モデルとして、fasttext[5]、SentenceBERT[6]を用いる。

fasttext は 2016 年に Facebook が公開した単語埋め込み生成モデルである。基本的には Word2vec[7]の技術を踏襲した形となっており、Word2vec と比較して高速にベクトル生成可能な点が特徴である。本研究では fasttext の日本語学習モデルを利用し、得られたベクトルの平均をとって文ベクトルとする。

SentenceBERT は 2019 年に公開された文章埋め込み生成モデルである。この手法が提案される以前は文章の類似度を測るために BERT の出力を文ベクトルとして用いることがあったが、文ベクトルとしての精度は低いことが知られていた。そこで著者らは BERT に対して距離学習を適用することで、高品質な文ベクトルの獲得を可能にした。

3 提案手法

提案手法ではまず作品を単純なルールに従って分割し、文章群へと変換する。その後、文類似度指標を用いて文書間の距離を計算し、文章対を key、それらの距離を value とした辞書を構築する。最後に、文章を次々と組み合わせて DTW を計算していき、その値が最小となるアライメントを対応関係にあるとみなす。

3.1 分割処理

アライメントを計算する前の事前処理として、それぞれの作品を文単位に分割する。本研究では以下の分割ルールを採用した。

- 句点、疑問符、感嘆符、鉤括弧、丸括弧を基本単位として分割する

読点を分割単位とすると、文の粒度が低くなってしまい計算時間の増加につながるため、採用していない。なお、「（）」のような入れ子構造が存在した場合、flag 変数を用いることで不適切な分割が行われないように処理している。また、作品によっては鉤括弧や丸括弧が半角や全角になっていたり、ニューラルモデルの Tokenizer で UnKnown トークンと認識されてしまう文字も存在する。そこで NFKC 正規化処理を行い、文字の表記ゆれを一括で修正した。

3.2 DTW による対応関係の抽出

以下にアライメントの手順を示す。詳細は図 1 を参照されたい。

1. 3.1 のルールに従って 2 つの作品を文書群に変換する。(List1、List2 とする)

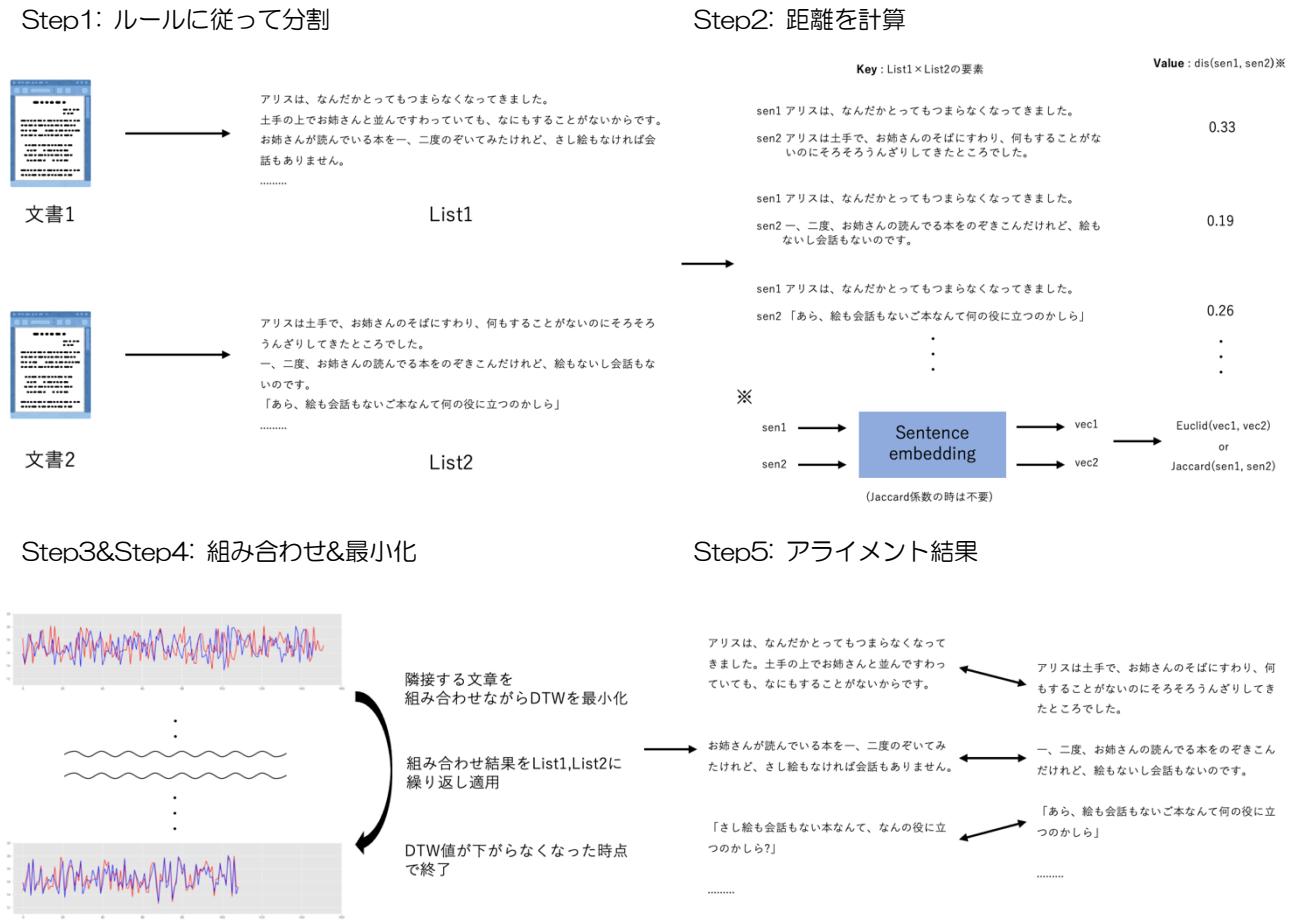

図 1　DTW によるアライメント(例文は[8][9]より引用)

2. List1 と List2 の直積を求め、その要素を key、要素を構成する文対の距離を value とした辞書を構築する。(距離を計算)
3. List1、List2 の要素全てに対し組み合わせ候補を順番に 2 つ選択して結合する。結合した状態での List1 と List2 の DTW 値を計算し保存するとともに、組み合わせた文対とそれらの距離を辞書に追加する。
4. DTW 値が最小となる組み合わせ候補を求め、その候補を適用する形で List1 もしくは List2 を更新する。
5. DTW 値が下がらなくなるまで 3、4 の操作を繰り返す。操作が終了した時の DTW によるアライメントを求める解とする。

「文を組み合わせて DTW を計算する」という流れを 1 ループとすると、1 ループ前後での List1 と List2 はそのほとんどが変化しない。よって、DTW の過程で逐一文章同士の距離計算を行うことは効率の面で問題がある。よって、最初の段階で辞書を構築することで、List1×List2 の距離は辞書にアクセスして取得することが可能となる。組み合わせた文書に関しては key が存在しないので、随時距離を計算して辞書に追加すればよい。

4 実験設定

4.1 データ

不思議の国のアリス[8][9]、白鯨[10][11]、ハックルベリー・フィンの冒険[12][13]、罪と罰[14][15]、自負と偏見[16][17]、風と共に去りぬ[18][19]を用いた。今回は多くの作品での適応可能性の調査を行うため、使用する文は前書き等を除いた冒頭付近の範囲に限定した。尚、以降では以下のように作品名を省略して表記する。

不思議の国のアリス → 不思議の国
ハックルベリー・フィンの冒険 → ハックルベリー

表 1 実験結果(結果は 正解対/対の総数 の形式となっている)

作品名	分割時の要素数(L1:L2)	Sent_len	Jaccard 係数	fasttext	SentenceBERT
不思議の国	152 : 147	63/153	108/120	60/70	103/109
白鯨	137 : 122	90/116	106/110	88/88	106/108
ハックルベリー	111 : 94	39/92	66/70	24/28	60/68
罪と罰	116 : 107	82/106	91/96	65/65	87/89
自負と偏見	104 : 97	41/98	81/81	30/34	64/64
風と共に去りぬ	77 : 68	34/69	62/62	33/33	60/60

4.2 評価方法

適切なアライメントが取れているか自動で判定することは困難であるため、目視でのチェックを行った。この時、得られた結果に本来組み合わせる必要の無い文章が存在する場合があり、例えば、「」「」のような会話が複数回連続した文章はそれぞれ個別にアライメントされることが望ましい。しかし本研究の目的はあくまでも対応関係の自動抽出であるため、少なくとも意味関係が等しいことが明らかであれば正しいアライメントとみなした。

5 実験

表 1 に 6 作品で実験を行った際のアライメント結果を示す。結果の形式について、(正解対/対の総数)の形式での記載とした。意味関係を適切に考慮してアライメントしつつも、なるべく最小限の組み合わせ回数で終了することが理想の流れである。よって、最終的に得られる対の総数は多い方が良い。

先行研究で提案されていた文章長(Sent_len)による DTW は全ての指標・特徴量において最も抽出精度が悪く、不適切な箇所へのアライメントが顕著に見られた。これは最終的な List1 と List2 の要素数が異なっており、正確な一対一の対応関係を抽出できなかったことが原因である。fasttext はおおむね適切な対応関係が取れているものの、過度な組み合わせによる長文でのアライメントが散見された。結果的に他の指標・特徴量に比べて、得られたアライメントの数が少なくなっている。単語ベクトルは足し合わせるほど原点に近づく傾向があり、文章を組み合わせる操作を行なった際に座標が大幅に変化する。それに伴って波形が大きく変化し、最適化に多くの組み合わせ操作を要した結果、非常に長い文のアライメントが発生したと考えられる。Jaccard 係数と SentenceBERT は両方とも適切に対応関係を抽出できており、その数に差はあまり見られない。しかし、SentenceBERT の方がやや圧縮率が高く、不必要な組み合わせが行われている傾向が見られた。今回採用した海外の文学作品は、翻訳者によって翻訳スタイルは違うものの、表面的には似た箇所が多い。Jaccard 係数を単語レベルではなく文字レベルの集合で求めることで、共通して登場する単語に加えてある程度類似した箇所を考慮したアライメントが可能であったと解釈できる。SentenceBERT の出力ベクトルを用いて二文の類似度を比較する方法は、その精度の高さから近年では頻繁に用いられている。しかし、比較する一方の文に余計な要素を追加すると類似度が上昇する、など不安定な挙動が確認されることも多い。今回の実験では、組み合わせ操作を取り入れたことによって、その挙動の不安定さが結果に現れた形となった。

以上の内容を踏まえると、日本語の文書からの DTW による対応関係の抽出において、文字レベルの Jaccard 係数を用いる、もしくは SentenceBERT によるユークリッド距離を用いる、のいずれかが有効であると考えられる。

6 まとめ

本研究では複数の翻訳者によって翻訳された文書を対象として、対応関係を抽出する手法を提案した。海外の文学作品を対象とした実験の結果、組み合わせ操作を取り入れた DTW の最小化が対応関係の抽出に有効であり、DTW を実行する際の文書間の距離として、文字レベルの Jaccard 係数もしくは SentenceBERT によるユークリッド距離が有効であることを示すことができた。本研究では日本語の文学作品を対象として実験を行ったが、今後は DTW の高速化なども視野に入れつつ、他言語への適応可能性も調査したい。

謝辞

本研究はJSPS科研費JP21K12036の助成を受けたものである。

参考文献

[1] Sakoe, H. and Chiba, S.: Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transaction on Acoustics, Speech, and Signal Processing, Vol. ASSP-26, No. 1, pp. 43-49 1978.
[2] Regina Barzilay and Kathleen R. McKeown. Extracting Paraphrases from a Parallel Corpus. In Proc. of ACL 2001, pp. 50–57, 2001.
[3] Kiyonori Ohtake and Kazuhide Yamamoto. 2003. In Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation, pages 380–391, Sentosa, Singapore. COLIPS PUBLICATIONS.
[4] 松元 航太郎 速水 悟 田村 哲嗣 “テトゥン語を対象としたニューラル機械翻訳の研究” 言語処理学会 第 26 回年次大会 発表論文集 pp.1065-1068 2020.
[5] https://fasttext.cc (2023 年 1 月 12 日アクセス)
[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional trans- formers for language understanding. 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 Pa- pers), pp. 4171–4186. Association for Computational Lin- guistics, June 2019.
[7] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, In Proceedings of ICLR Workshops Track (2013).
[8] Lewis Carroll (1865) Alice's Adventures in Wonderland(ルイス・キャロル 河合祥一郎(訳) 不思議の国のアリス+鏡の国のアリス 2 冊合本版(2015) 角川文庫 P12)
[9] Lewis Carroll (1865) Alice's Adventures in Wonderland (ルイス・キャロル 多田幸蔵(訳) 不思議の国のアリス (2012) グーテンベルク２１ P2)
[10] Herman Melville (1851) Moby-Dick (ハーマン・メルヴィル 富田彬(訳) 白鯨(上) (2015) 角川文庫)
[11] Herman Melville (1851) Moby-Dick (ハーマン・メルヴィル 高村勝治(訳) 白鯨(上) (2011) グーテンベルク２1)
[12] Mark Twain (1885) Adventures of Huckleberry Finn (マーク・トウェイン 山本長一(訳) マーク・トウェインコレクション ハックルベリィ・フィンの冒険 (2014) 彩流社)
[13] Mark Twain (1885) Adventures of Huckleberry Finn (マーク・トウェイン 土屋京子(訳) ハックルベリー・フィンの冒険 〈上〉 (2014) 光文社古典新訳文庫)
[14] Dostoevskii (1866) Crime and Punishment (ドストエフスキー工藤精一郎(訳) 罪と罰（上下）合本版 (2021) 新潮文庫)
[15] Dostoevskii (1866) Crime and Punishment (ドストエフスキー 米川正夫(訳) 罪と罰 上 (2022) 角川文庫)
[16] Jane Austen (1813) Pride and Prejudice(ジェイン・オースティン 小山太一(訳) 自負と偏見 (2016) 新潮文庫)
[17] Jane Austen (1813) Pride and Prejudice (ジェイン・オースティン 小尾芙佐(訳) 高慢と偏見〈上〉 (2016) 光文社古典新訳文庫)
[18] Margaret Munnerlyn Mitchell (1936) Gone With the Wind (マーガレット・ミッチェル 鴻巣友季子(訳) (2015) 風と共に去りぬ 第 1 巻 新潮文庫)
[19] Margaret Munnerlyn Mitchell (1936) Gone With the Wind (マーガレット・ミッチェル 大久保康雄,竹内道之助(訳) 風と共に去りぬ（一）(2016) グーテンベルク２１)