文脈情報を考慮した高速な英日文アラインメント

福田 りょう ♠, 帖佐 克己 ◊
♠ 奈良先端科学技術大学院大学　◊NTT コミュニケーション科学基礎研究所
fukuda.ryo.fo3@is.naist.jp　katsuki.chousa.bg@hco.ntt.co.jp

概要

本論文では大規模対訳コーパスの作成に向けた高精度かつ高速な文アラインメント手法を提案する．従来手法では文外文脈およびトークン単位の情報を予測時に考慮することで高精度な文アラインメントを実現していたが，予測時の計算量が大きいことから大規模なデータへの適用が難しかった．本論文では，両方の言語の文脈情報を考慮した文ベクトルを用いて対訳文書間の全てのアラインメントを同時に予測することで，小さい計算量で高精度な文アラインメントを獲得する手法を提案する．日英での新聞記事を用いた実験により，提案手法が予測速度を維持しながら高い精度でアラインメントを行えることを示した．

1　はじめに

文アラインメントは対訳となっている文書対から対訳文対を自動で抽出するタスクである．抽出された大量の対訳文対からなるデータセットは対訳コーパスといい，機械翻訳モデルを始めとした言語横断モデルの学習に用いられる．特に，高精度なニューラル機械翻訳（Neural Machine Translation; NMT）モデルの学習には高品質かつ大量の対訳文が不可欠であることが知られている [1]．そのため，Web からクロールするなどして得られる大量の対訳文書対から**高精度**な対訳文アラインメントを**高速**に抽出できる手法が必要とされている．

近年ではニューラルネットワークを用いたスコアリングに基づく文アラインメント手法が多数提案されている．Vecalign [2] は文ベクトルの類似度に基づくスコアを用いてアラインメントを行う手法である．この手法は比較的高速に文対応付けを計算できる一方で，精度が文ベクトルに用いる多言語文埋め込み表現の精度に大きく影響されるため，英語やドイツ語に比べて低資源な日本語を含む言語対では精度が大きく減少してしまう．SpanAlign [3] は言語横断スパン予測モデルとその予測時に得られるスコアを基にアラインメントを行う手法である．このスパン予測モデルは文脈情報および言語横断でのアテンションを行うことが可能で，高精度な文アラインメントを抽出することができる．一方で，モデルに入力するトークン数が増えるに従って予測にかかる計算量が非常に大きくなってしまうため，大規模データに対して適用することが難しい．

本論文では，先行研究の高速な予測速度を維持しながら高精度な文アラインメントの予測を行う手法を提案する．提案手法では，対訳文書の各文を事前学習済みの多言語モデルで文ベクトルにエンコードし，両方の言語の文脈情報を考慮した文ベクトルを求め，その文ベクトルに基づいて対訳文書間の任意のアラインメント候補に対するスコアを一度に計算する．これにより，高精度な文アラインメントが高速に自動抽出できることが期待される．日英新聞記事を用いた実験で，高速な既存手法と比べて提案手法が予測速度を維持しながら精度を F 値で 0.032 pts 向上することを確認した．

2　先行研究

文アラインメント手法は文同士の類似度を計算するスコア関数とその類似度に基づいてアラインメントを求めるアルゴリズムの 2 つのステップに大きく分解することができる．その内，スコア関数としては文脈情報を考慮しない手法 [2, 4, 5, 6, 7] および文脈情報を考慮する手法 [3] が提案されている．

近年のスコア関数はニューラルネットワークに基づく手法が主流となっている．Vecalign [2] と呼ばれる手法は多言語文埋め込み表現の LASER [8] による文ベクトルの類似度に基づいて再帰的な DP マッチングを行うことで，独仏間の文アラインメントで最高精度を達成している．また，計算量も小さいことから世界で最も大きい対訳コーパスの 1 つである ParaCrawl コーパス [9] の作成にも利用されている．一方で，文埋め込みの精度に文アラインメントの精

度が大きく影響されてしまうため，日英間では独仏間のような言語対と比べて精度が大幅に減少してしまうことが確認されている．SpanAlign [3] はある言語の文書中からもう一方の言語の文に対応する文のスパンの開始/終了位置を予測し，その予測から得られるスコアに基づいて ILP を行うことで文アラインメントを実現する．この言語横断スパン予測モデルは，先行研究によって予測された文アラインメントの結果を学習データとして事前学習済み多言語言語モデルから作成することができ，文脈情報の利用および言語横断でのアテンションを行うことで，日英間の文アラインメントにおいて最高精度を達成している．一方で，モデルに同時に入力することができるトークン数の上限から，文書に含まれる文数が多くなると sliding window という方法で文書の一部に対する予測を組み合わせて文書全体に対する予測を行う必要があるため，予測の際の計算量が非常に大きくなってしまう．

また，文アラインメントと似たタスクとして単語アラインメントタスクがある．このタスクでは対訳文間の単語の対応関係を抽出する．このタスクでも言語横断スパン予測モデルに基づく手法 [10] が複数の言語対で最高精度を達成している．さらに，Procopio ら [11] は，対訳文間のすべての単語ベクトル同士の直積から得られる類似度行列を用いてアラインメントを抽出する手法を提案している．ここで用いられる単語ベクトルはスパン予測モデルと同様の方法で文脈情報および言語横断のアテンションに基づいて計算される．これにより，スパン予測に基づく手法と同等の精度をより高速に実現している．

3 提案手法

本論文では，先行研究の高速な予測速度を維持しながら高精度な文アラインメントの予測を行うために，高速かつ高精度にスコア関数を計算する手法を提案する．図 1 に示す提案手法は，まず対訳文書の各文を事前学習済みの多言語モデルで同じ空間のベクトル表現へと変換し（3.1.1 節），次に言語内および他言語の文脈情報を考慮した文 n-gram ベクトルを計算し（3.1.2 節），最後に対訳文書間の任意の文アラインメントの候補に対するスコアを同時に計算する（3.1.3 節）．提案手法は，文ベクトルへダウンサンプリング，およびスコアの同時計算により高速な予測が可能である．更に文脈情報を考慮したベクトルによる高精度な文アライメントが期待できる．

3.1 モデル

3.1.1 多言語モデルによる文ベクトル

異なる言語の文を同じ空間の埋め込み表現へと変換するために，本手法では LaBSE [12] を使用する．LaBSE は 109 言語以上に対応している多言語文埋め込みモデルで，対訳文検索（bitext retrieval）では非常に高い精度を実現することが報告されている．

LaBSE を用いて文書 E の文ベクトル行列 $H^e \in \mathbb{R}^{|E|\times d}$ を以下のように求める．

$$H^e = [\mathbf{h}^e_1, \mathbf{h}^e_2, ..., \mathbf{h}^e_{|E|}] \quad (1)$$

$$\mathbf{h}^e_i = \texttt{LaBSE}(e_i), 1 \le i \le |E| \quad (2)$$

$\mathbf{h}^e_i$ は e_i に対応する文ベクトルであり，d は文ベクトルの次元数である．文書 F の文ベクトル行列 $H^f \in \mathbb{R}^{|F|\times d}$ も同様に求める．

3.1.2 文脈横断の文 n-gram ベクトル

次に，言語内およびもう一方の言語の文脈情報を考慮した文ベクトルを，それぞれの文書の文ベクトルを結合して self attention [13] を計算することで求める．

$$C^e_1, \_, C^f_1 = \texttt{self\_attn}([H^e; \mathbf{s}; H^f]) \quad (3)$$

ここで，self_attn は l 層の Transformer Encoder，$\mathbf{s}$ は 2 つの文書の区切りを表す学習可能パラメータ，$C^e_1 \in \mathbb{R}^{|E|\times d}, C^f_1 \in \mathbb{R}^{|F|\times d}$ はそれぞれ H^e, H^f に対応する Transformer Encoder の出力を表す．続いて，多対多の文対応を抽出するために，文ベクトルを文 n-gram ベクトルに拡張する．

$$C^e_{1:N} = [C^e_1; ...; C^e_N] \quad (4)$$

$$C^e_n = \texttt{max\_pooling}_n(C^e), 1 \le n \le N \quad (5)$$

ここで，N は文 n-gram の最大文数を表すハイパーパラメータを，$\texttt{max\_pooling}_n$ はカーネルサイズ $(n, 1)$，ストライド 1 の Max Pooling 層を表す．$C^e_{1:N} \in \mathbb{R}^{|E_{1:N}|\times d}$ は C^e の文 n-gram への拡張である（ただし，$|E_{1:N}| = \frac{N(2|E|-N+1)}{2}$）．$C^f$ の拡張 $C^f_{1:N}$ も同様に計算する．

3.1.3 スコア計算

最後に，対訳文書間の任意のアラインメント候補に対するスコア行列 $\hat{A} \in \mathbb{R}^{|E_{1:N}|\times|F_{1:N}|}$ を，文 n-gram ベクトル同士の直積を MLP 層 mlp に入力すること

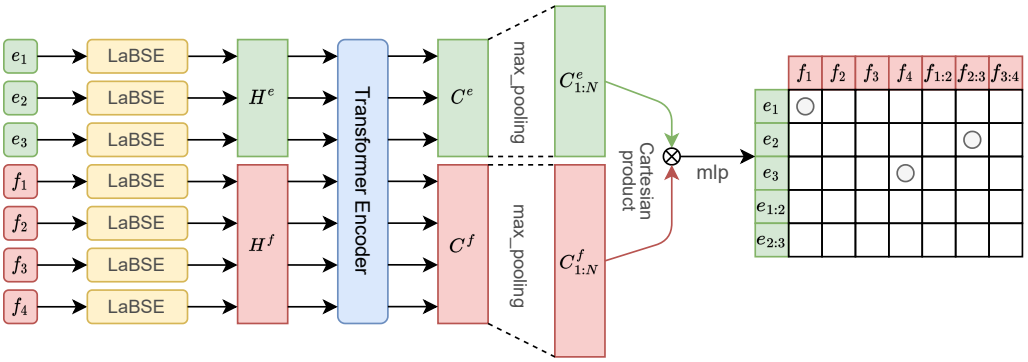

図 1　提案手法の概要

で求める．

$$\hat{A} = \mathtt{mlp}(C^e_{1:N} \otimes C^f_{1:N}) \tag{6}$$

(6) 式は一度で全てのアライメント候補のスコアを計算できるため，スパン予測モデル [3] のスコア計算と比べて非常に高速である．

3.2　損失関数

文アラインメントのためのスコア関数計算を，正解の文 n-gram 対応を表す行列 $A \in \mathbb{R}^{|E_{1:N}|\times|F_{1:N}|}$ を予測する問題としてモデルを学習する．ただし $A_{ij} \in \{0,1\}$ であり，$A_{ij} = 1$ は i に対応する文書 E の文 n-gram と j に対応する文書 F の文 n-gram が対応関係にあることを示す．モデルの学習時，損失関数として Binary Cross Entropy（3.2.1 節）または Softmax Cross Entropy（3.2.2 節）を用いた．

3.2.1　Binary Cross Entropy (BCE) loss

文 n-gram の対応を，二値分類タスクとして学習する．学習時の損失 $\mathcal{L}_{bce}$ は予測されたスコア行列 $\hat{A}$ と，正解の文 n-gram 対応 A 間の交差エントロピーとして (7) 式で定義される．

$$\mathcal{L}_{bce} = -\frac{1}{|E_{1:N}||F_{1:N}|}\sum_{i=1}^{|E_{1:N}|}\sum_{j=1}^{|F_{1:N}|}\Big\{a_{ij}\log\sigma(\hat{a}_{ij}) + (1-a_{ij})\log(1-\sigma(\hat{a}_{ij}))\Big\} \tag{7}$$

ここで，σ はシグモイド関数であり，また $\sigma(\hat{a}_{ij})(0 \le \sigma(\hat{a}_{ij}) \le 1)$ と $a_{ij} \in \{0,1\}$ はそれぞれ $\hat{A}$ と A の (i,j) 成分に対応する．$\mathcal{L}_{bce}$ によって，モデルは多対多の対応付けを直接的に学習できる．また A の各行及び各列に，値が 1 の要素は高々 1 つしか含まれないため，DP や ILP といったアライメントアルゴリズムの適用が容易である．一方で，N が大きいほど行列がスパースになり，学習が難しくなる欠点がある．

モデルの予測時，(6) 式のスコアは $\hat{A}$ の各要素をシグモイド関数で正規化する．

3.2.2　Softmax Cross Entropy (SCE) loss

文 n-gram の対応を，多クラス分類タスクとして学習する．学習時の損失 $\mathcal{L}_{sce}$ は [14] に従い，$E \to F$ 方向の対応付け損失 $\mathcal{L}_{E\to F}$ と $F \to E$ 方向の対応付け損失 $\mathcal{L}_{F\to E}$ の平均で定義する．

$$\mathcal{L}_{sce} = (\mathcal{L}_{E\to F} + \mathcal{L}_{F\to E})/2 \tag{8}$$

$$\mathcal{L}_{E\to F} = -\frac{1}{|E_{1:N}|}\sum_{i=1}^{|E_{1:N}|}\sum_{j=1}^{|F_{1:N}|}\left\{a_{i,j}\log\frac{e^{(\hat{a}_{i,j}-m)}}{e^{(\hat{a}_{i,j}-m)} + \sum_{c=1,c\neq j}^{|F_{1:N}|} e^{\hat{a}_{i,c}}}\right\} \tag{9}$$

ここで，m は正例と負例のスコアの分離を促進するためのマージンである．$F \to E$ 方向の対応付け損失 $\mathcal{L}_{F\to E}$ も (9) 式と同様に計算する．

モデルの予測時，(6) 式のスコアは $\hat{A}$ に対して行方向と列方向にソフトマックス関数をかけて正規化した値の平均を用いる．

絶対的な対応付けスコア予測を学習する $\mathcal{L}_{bce}$ とは対照的に，$\mathcal{L}_{sce}$ では相対的な対応付けスコアの予測を学習するため，学習が容易になることが期待できる．

4　実験

提案手法の有効性を確認するため，日本語と英語の実際の新聞記事の対訳データを用いて文アラインメントの精度と速度の評価を行った．

学習および開発，評価のデータには読売新聞の記事とその翻訳となっている The Japan News の記事を使用した[1)]．表 1 に各データの平均文数を示す．学

1)　https://database.yomiuri.co.jp/about/glossary/

表 1 データセットの平均文数

	学習	開発	評価
英語	19.7	17.9	23.8
日本語	23.9	18.9	26.2

習データには，内山らの手法 [5] を用いて自動抽出した対訳文書 2,989 本とその文書内の文アラインメントを擬似正解データとして用いた．これらのデータは，2012 年に発行された日本語記事 317,491 本と英語記事 3,878 本から抽出した．開発および評価データには，2013 年に発行された記事から人手でアラインメントを行ったものをそれぞれ 15 本ずつ用いた．評価には，文アラインメントの評価尺度として一般的に用いられている，文アラインメント単位での F_1 スコアを用いた[2]．

ベースラインには，Vecalign [2] と SpanAlign [3] を用いた．Vecalign での文アラインメントには著者実装[3] を用いた．アラインメントに含まれる最大文数は 8，文埋め込みベクトルを作成する際に結合する最大文数は 10 とした．また，多言語文埋め込みには LASER [8] を用いた．SpanAlign での文アラインメントについても著者実装[4] を用いた．各種設定に関しては論文の設定に準拠した．

提案手法に含まれる LaBSE は，事前学習済みモデル[5] をパラメータを固定して用いた．Transformer Encoder は 1 層の encoder layer で構成し，multi-head attention の head 数を 8，隠れ層の次元を 768，Feedforward 層の次元を 2048 とした．文 n-gram の最大文数 $N = 2$ とした．アラインメントアルゴリズムとして整数計画法を適用し，ソルバには IBM ILOG CPLEX 12.8.0.0 を用いた．

表 2 に，各手法における英日文アラインメントの精度と評価データの推論にかかった実行時間を示す．提案手法の中で，SCE loss を用いたモデルは BCE loss を用いたモデルより高い F1 スコアを得た．相対的なスコアの学習により，スパース性による学習の難しさが緩和されたことが示唆される．また Vecalign との比較では，提案手法による 0.32pt の F1 スコア向上が見られた．文ベクトル計算時の文脈情報の利用や，ILP による非単調な文対応の抽出が精度向上に寄与したと考えられる．一方で，

2) 次のスクリプトの *strict* スコアを使用した．https://github.com/thompsonb/vecalign/blob/master/score.py
3) https://github.com/thompsonb/vecalign
4) https://github.com/nttcslab-nlp/spanalign
5) https://huggingface.co/sentence-transformers/LaBSE

表 2 英日文アラインメントの精度と評価データの推論にかかった実行時間（秒），および文ベクトルの計算に文脈情報を用いない場合（w/o `self_attn`）の精度

	Precision	Recall	F1	時間
Vecalign [2]	.591	.658	.623	39
SpanAlign [3]	.734	.675	.703	120
提案手法				
BCE loss	.634	.667	.650	29
SCE loss	.598	.724	.655	29
w/o `self_attn`				
BCE loss	.475	.482	.478	-
SCE loss	.390	.355	.372	-

SpanAlign との比較では，提案手法が 0.48pt 下回った．SpanAlign はトークン単位の文脈情報を考慮することに対し，提案手法は文単位に圧縮された文脈情報を利用する．この文脈情報の粒度の違いが精度の差に繋がったと考えられる．評価データの推論にかかる実行時間は，提案手法が SpanAlign の 1/4 程度であった[6]．以上の比較から，提案手法は実用レベルの精度を維持しつつ，既存手法と比べて高速にアライメントを行えると結論づける．

文脈情報の有効性を検証するため，提案手法から `self_attn` を除いたモデル（w/o `self_attn`）を作成した．アライメントの精度を表 2 の下部に示す．`self_attn` を除き，文ベクトルの計算に文脈情報を用いないことで，大きく精度が低下した．このことから，両方の言語の文脈情報が精度向上に作用したことが示唆された．

5 まとめ

本研究では，大規模対訳コーパスの構築に向けた文対応付けの高速化を検討した．提案手法は，文単位に圧縮された文脈情報を活用することで，高精度な既存手法である SpanAlign よりも高速な文対応付けが行える．また実験では，提案手法が高速な既存手法である Vecalign と比べて高精度に文アライメントを抽出できることを示した．今後の課題として，スパース性の問題を解消して対応付け可能な最大文数を増やすことや，非連続な対応関係の抽出を可能にすることなどが挙げられる．

6) モデルやデータの読み込み処理時間を含む．

参考文献

[1] Huda Khayrallah and Philipp Koehn. On the impact of various types of noise on neural machine translation. In **Proceedings of the 2nd Workshop on Neural Machine Translation and Generation**, pp. 74–83, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] Brian Thompson and Philipp Koehn. Vecalign: Improved sentence alignment in linear time and space. In **Proceedings of EMNLP-2019**, pp. 1342–1348, 2019.

[3] Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. SpanAlign: Sentence alignment method based on cross-language span prediction and ILP. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 4750–4761, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[4] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. **Computational Linguistics**, Vol. 19, No. 1, pp. 75–102, 1993.

[5] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning japanese-english news articles and sentences. In **Proceedings of the ACL-2003**, pp. 72–79, 2003.

[6] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In **Proceedings of the RANLP-2005**, pp. 590–596, 2005.

[7] Rico Sennrich and Martin Volk. Iterative, MT-based sentence alignment of parallel texts. In **Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)**, pp. 175–182, Riga, Latvia, May 2011. Northern European Association for Language Technology (NEALT).

[8] Mikel Artetxe and Holger Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings. In **Proceedings of the ACL-2019**, pp. 3197–3203, 2019.

[9] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics.

[10] Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. A supervised word alignment method based on cross-language span prediction using multilingual bert. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)**, Online, November 2020. Association for Computational Linguistics.

[11] Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. Multimirror: Neural cross-lingual word alignment for multilingual word sense disambiguation. In Zhi-Hua Zhou, editor, **Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21**, pp. 3915–3921. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[12] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the NIPS 2017**, pp. 5998–6008, 2017.

[14] Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. **arXiv preprint arXiv:1902.08564**, 2019.