Triplet Lossと知識蒸留を用いたMultilingual Sentence BERTモデルの構築

Generation of Multilingual Sentence BERT Model using Triplet Loss and Knowledge Distillation

北海学園大学　大学院工学研究科教授
越前谷　博

1996年北海学園大学大学院工学研究科修士課程修了。博士（工学）。2013年～現在北海学園大学大学院工学研究科教授。機械翻訳の研究に従事。アジア太平洋機械翻訳協会（AAMT）／Japio特許翻訳研究会委員。

echi@lst.hokkai-s-u.ac.jp　TEL 011-841-1161（内線：7863）

1 はじめに

近年のディープラーニング技術の進展に伴い、多言語を一つのモデルで扱うことのできる言語モデルの研究が進んでいる。その結果、多言語文をベクトルで表現することが可能となり、様々な自然言語処理タスクにおいて多言語モデルの利用価値が高まっている。BERTモデルが単一言語を対象としたモデルであるのに対して、Multilingual BERT（mBERT）[1]は104の言語の単一言語コーパスを用いて学習された多言語モデルとなっている。このmBERTは各言語毎に単一言語のコーパスを収集し学習データとして用いており、言語間の対応は明示的に学習していないにもかかわらず、異言語の単語間で規則的な関係が成り立っており、word2vecのようにベクトルの距離が意味的な距離を表現できていることが確認されている[2]。したがって、mBERTを用いることで異言語の文間の類似度を一つのモデルで得ることが可能となる。

このような観点よりBERTやRoBERTa[3]をファインチューニングすることでより高い精度かつ短い学習時間で文間類似度を得るためのSentence-BERT（SBERT）の研究が行われている[4]。SBERTではSiameseとTriplet Network構造を用い、かつPooling層を付与して事前学習モデルのBERTをファインチューニングする。その結果、大幅な学習時間の短縮を実現しつつ精度の高い文間類似度が得られる。さらにSBERTを多言語に拡張させたモデルとしてmSBERT[5]が提案されている。mSBERTでは教師モデルに単一言語のSBERTモデル、生徒モデルに多言語のmBERTモデルを用いて知識蒸留[6]によりmSBERTを構築している。

本報告ではより高い精度で類似度を得ることが可能なmSBERTモデルを提案する。提案手法では擬似対訳コーパスに基づいてTripletデータを作成することでMargin MSE Loss[7]を用いてmBERTモデルをファインチューニングする。その後、教師モデルにSBERTモデル、生徒モデルにファインチューニングしたmBERTモデルを用いて知識蒸留によりmSBERTを構築する。性能評価実験では提案手法（mSBERT by Triplet Margin and Knowledge Distillation: mSBERT by TM and KD）、知識蒸留のみによる提案手法（mSBERT by Knowledge Distillation: mSBERT by KD）、既存のmBERTとmSBERTの4つのモデルそれぞれを用いて、The 4th Workshop on Asian Translation（WAT2017）データにある特許翻訳の英文と日本文間の文間類似度を求めた。そして、それらと人手評価との相関係数を求めることで機械翻訳に対する自動評価の観点から比較実験を行った。その結果、小規模な評価データではあるが、提案手法であるmSBERT by TM and KDが最も高い相関係数を示した。

2 提案手法におけるmSBERTモデルの構築

本報告で提案するmSBERT by TM and KDのアーキテクチャを図1に示す。

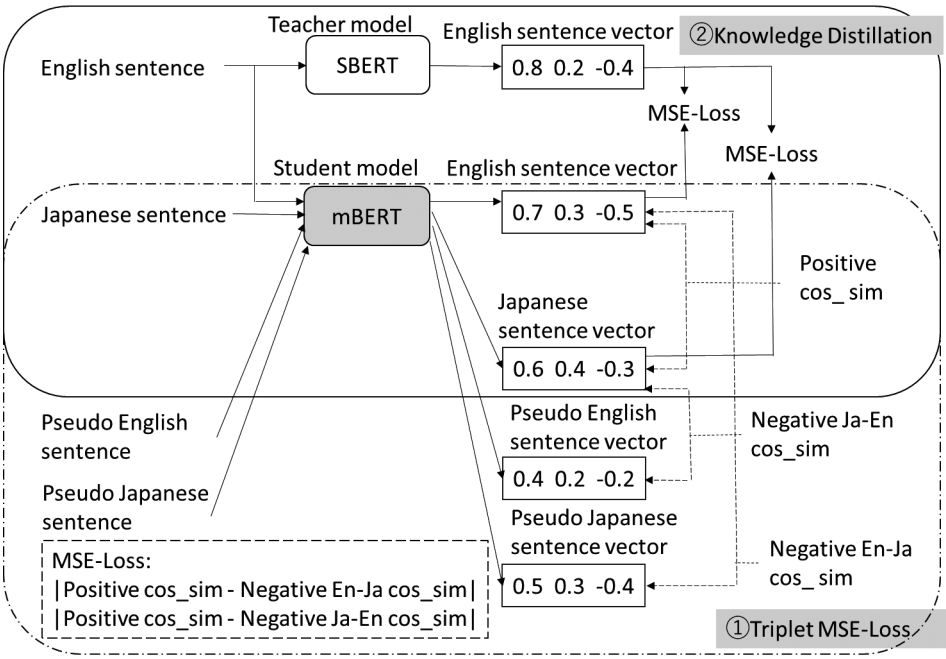

図 1　提案手法のアーキテクチャ

提案手法ではまず Triplet データを用いて Margin MSE Loss に基づくファインチューニングを行い、その後、知識蒸留により mSBERT by TM and KD モデルを生成する。Triplet MSE Loss に基づくファインチューニングでは対訳文と擬似対訳文を用いて mBERT をファインチューニングする。その際、対訳文の英文と日本文の間のコサイン類似度を Positive cos_sim と位置付ける。また、対訳文の英文をクエリとした場合、英文と擬似日本文の間のコサイン類似度を Negative En-Ja cos_sim とする。そして、Gold Similarity である｜Positive cos_sim–Negative En-Ja cos_sim｜の値を用いて学習を行う。さらに日本文をクエリとした場合には、日本文と擬似英文の間のコサイン類似度を Negative Ja-En cos_sim とする。そして、Gold Similarity である｜Positive cos_sim–Negative Ja-En cos_sim｜の値を用いて学習を行う。

次いで、類似度に特化した文ベクトルを生成するために知識蒸留により mBERT をファインチューニングする。その際に教師モデルには単一言語の SBERT、生徒モデルには Triplet MSE Loss に基づきファインチューニングされた mBERT を用いる。このように mBERT を事前に Triplet データによりファインチューニングし、そのうえで知識蒸留によりファインチューニングすることでより高い精度で類似度を得るための言語モデルの生成を図る。

3　擬似対訳コーパスの作成とコサイン類似度の付与

擬似対訳コーパスは英日対訳コーパスの英文と日本文をそれぞれ Google 翻訳により翻訳した文を用いる。すなわち、英文を Google 翻訳により日本文に翻訳することで擬似日本文、日本文を Google 翻訳により英文に翻訳することで擬似英文を作成する。

また、これらの擬似対訳コーパスに付与する Gold Similarity は英語の SBERT モデルを用いてコサイン類似度を算出することで得る。日本文をクエリとした Negative Ja-En cos_sim は英文と擬似英文との間のコサイン類似度を SBERT より算出することで得る。また、英文をクエリとした Negative En-Ja cos_sim は擬似日本文をさらに Google 翻訳により擬似英文を作成し、その擬似英文と対訳コーパスの英文との間のコサイン類似度を SBERT により算出することで得る。なお、Positive cos_sim については対応関係にある対訳コーパスを用いることから、コサイン類似度の最大値である 1.0 を固定値として用いる。

4　性能評価実験

4.1　実験方法

評価実験は提案手法によるモデル（mSBERT by TM and KD モデル）、知識蒸留のみを用いたモデル

(mSBERT by KD モデル)、そして、既存の mBERT モデルである xlm-roberta-base[1] と mSBERT モデルである paraphrase-multilingual-mpnet-base-v2[2] を用いて比較実験を行った。paraphrase-multilingual-mpnet-base-v2 は教師モデルに SBERT（paraphrase-mpnet-base-v2[3] モデル）、生徒モデルに mBERT（xlm-roberta-base モデル）を用いて知識蒸留により生成された mSBERT モデルである。mSBERT by TM and KD モデルと mSBERT by KD モデルについては学習データとして OPUS にある TED2020[4] の英日対訳コーパスの 10 万の対訳文を使用した。また、mSBERT by TM and KD モデルの生成において用いた Triplet データは英文をクエリとしたデータが約 32,000、日本文をクエリとしたデータが約 65,000 である。評価の際には 4 つのモデルそれぞれより WAT2017 の英日データと日英データを用いて類似度を求め、それらを自動評価スコアとした。そして、自動評価スコアと人手評価によるスコアとの間で相関係数を求めた。また、mSBERT by TM and KD モデルと mSBERT by KD モデルでは、コサイン類似度は英語版 SBERT モデルである paraphrase-MiniLM-L12-v2[5] を用いて得た。一方、図 1 の SBERT モデルには paraphrase-mpnet-base-v2、mBERT モデルには xlm-roberta-base を用いた。

評価対象の MT 訳には WAT2017 データにある JPO の英文 200 文に対する Japio による日本文の MT 訳、また、JPO の日本文 200 文に対する Japio による英文の MT 訳を用いた。人手評価には英日と日英共に adequacy の観点より行った 2 名の評価者による 5 段階評価の平均を用いた。

mSBERT by TM and KD、mSBERT by KD、xlm-roberta-base、そして、paraphrase-multilingual-base-v2 の 4 つのモデルの評価精度はそれぞれのモデルが原文とその MT 訳に対して出力した文ベクトルに対してコサイン類似度を求め、そのコサイン類似度と人手評価スコアとの間の相関係数を求めることで行った。

4.2 実験結果

mSBERT by TM and KD、mSBERT by KD、そして、xlm-roberta-base、paraphrase-multilingual-base-v2 の 4 つのモデルの相関係数は、英日ではそれぞれ 0.1689、0.0851、-0.1204、0.0656、日英ではそれぞれ 0.0652、0.1049、0.0296、0.1149 となり、英日と日英の平均値はそれぞれ 0.1171、0.0950、-0.0908、0.0903 となり提案手法の mSBERT by TM and KD の相関係数が最も高かった。表 1 に英日翻訳における原文（英文）、その MT 訳（日本文）、参照訳（日本文）、人手評価スコア、そして、4 つのモデルにより得られたコサイン類似度の具体例を示す。

4.3 考察

提案手法の mSBERT by TM and KD モデルは英日の MT 訳においては人手評価との間で最も高い相関係数を示した。しかし、日英の MT 訳においては mSBERT by KD モデルと paraphrase-multilingual-base-v2 モデルよりも低い相関係数となった。これは Triplet Margin Loss による学習において英日の Triplet データに比べ、日英の Triplet データでは適切なコサイン類似度が付与されなかったため適切な学習ができなかったことが原因と考えられる。日本文をクエリとした日英の Triplet データでは Negative Ja-En cos_sim の値、すなわち、対訳コーパスの英文と擬似英文の間においてコサイン類似度がそれほど高くないものも数多く存在した。擬似英文は対訳コーパスの日本文を Google 翻訳により英文に翻訳することで得られるが、その場合、日本文が意訳されていると擬似英文と対訳コーパスの英文との間で意味的なずれが生じてコサイン類似度が低下してしまう。その結果、適切ではないコサイン類似度が付与された英文が生成され、日英の Triplet データでは十分な学習ができなかったと考えられる。一方、英文をクエリとした英日の Triplet データでは英文から日本文、

[1] https://huggingface.co/xlm-roberta-base

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[3] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

[4] https://opus.nlpl.eu/TED2020.php

[5] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L12-v2

表１　英日の比較実験におけるコサイン類似度の具体例

例１				
原文	The turbomachine 2 includes an electrical motor 4 connected to a compressor 6 .			
MT 訳	ターボ機械２は、圧縮機６に接続された電気モータ４を含む。			
参照訳	ターボ機械２は圧縮機６に接続された電気モータ４を含む。			
コサイン類似度				人手評価
mSBERT by TM and KD：0.9581	mSBERT by KD：0.9560	xlm-roberta-base：0.9957	paraphrase-multilingual -mpnet-base-v2：0.8377	5
例２				
原文	Typical fibrillated carboxymethyl cellulose is shown in FIGS. 1-6 .			
MT 訳	典型的な解繊カルボキシメチルセルロースセルロースを図１～６に示す。			
参照訳	図１～６に典型的なフィブリル化カルボキシメチルセルロースを示す。			
コサイン類似度				人手評価
mSBERT by TM and KD：0.8652	mSBERT by KD：0.8865	xlm-roberta-base：0.9957	paraphrase-multilingual -mpnet-base-v2：0.6437	4
例３				
原文	Water as the heat medium flows through the inside of the first heat medium passage 232 .			
MT 訳	熱媒体は、第１熱媒体通路 232 の内部を通って流れる。			
参照訳	第１熱媒体経路 232 はその内部に熱媒体としての水を通流させる。			
コサイン類似度				人手評価
mSBERT by TM and KD：0.8317	mSBERT by KD：0.9340	xlm-roberta-base：0.9955	paraphrase-multilingual -mpnet-base-v2：0.9044	3
例４				
原文	As understood from FIG. 27 , data value 0 is added for expansion into two-dimensions.			
MT 訳	図 27 から理解されるように、拡張のためにデータ値０を２次元に拡張する。			
参照訳	図 27 から分かるように、２次元拡張のためにデータ値０が追加される。			
コサイン類似度				人手評価
mSBERT by TM and KD：0.8870	mSBERT by KD：0.9083	xlm-roberta-base：0.9969	paraphrase-multilingual -mpnet-base-v2：0.8314	2.5

そして、その日本文から英文に翻訳したものと対訳コーパスの英文との間でコサイン類似度を求めるため、意訳の悪影響を受けずに適切なコサイン類似度が付与されたTriplet データを用いて学習が行われたと考えられる。したがって、擬似対訳コーパスを生成する際には良質な対訳コーパスを用いることが重要である。

また、表 1 の例 1 から例 3 の具体例より英日の場合、比較的人手評価が高い MT 訳文については mSBERT by TM andKD モデルは人手評価に追随できているが、例 4 のように人手評価が低い場合、過度に高いコサイン類似度を与えていることがわかる。そのような傾向は mSBERT by KD モデルと xlm-roberta-base モデルにおいてより強く見られた。ただし、人手評価が 1 から 5 の中央値である 3 を下回るデータは例 4 の原文と MT 訳のみであったためデータ量は不十分であり、一定量のデータに基づく検証が必要である。paraphrase-multilingual-mpnet-base-v2 モデルについては MT 訳にカタカナ表記が多くなるに伴い、今回の評価実験では類似度が低くなる傾向が見られた。カタカナ表記を対象としたファインチューニングを強化することでこの点は改善される可能がある。提案手法の mSBERT by TM and KD モデルについては平均値の相対的比較においては最も高い相関係数を得ることができたが、絶対値としては低く、また日英の評価精度は相対的にも不十分であるため改善が必要である。

5 まとめ

本報告では異言語の文間類似度を求めるための多言語対応の新たな言語モデルを提案した。提案手法では英日対訳コーパスと英日擬似対訳コーパスを用いて Triplet Margin により多言語モデル mBERT をファインチューニングした。その後、単言語の SBERT モデルを教師モデル、ファインチューニングされた mBERT を生徒モデルとした知識蒸留により mBERT をファインチューニングすることで mSBERT モデルを生成した。性能評価実験では WAT2017 の特許翻訳文の原文と MT 訳間の類似度を求めることでそれらを評価スコアとし、人手評価との相関係数を求めた。その結果、提案手法の相関係数が従来の言語モデルに比べて高い値を示し、提案手法の有効性を確認した。

今後は英日の対訳コーパスだけではなく、他の言語の対訳コーパスについても擬似対訳コーパスを生成し、それらを用いたファインチューニングを行う予定である。また、多言語の原文と MT 訳間の類似度を求めることで大規模な性能評価実験を行う。そして、WMT データをファインチューニングのための学習データに用いることでより自動評価に適した言語モデルを構築する予定である。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of NAACL-HLT 2019, pp. 4171–4186 (2019)

[2] Telmo Pires, Eva Schlinger and Dan Garrette, “How multilingual is Multilingual BERT?,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4996–5001 (2019)

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv preprint arXiv:1907.11692 (2019)

[4] Nils Reimers and Iryna Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3982–3992 (2019)

[5] Nils Reimers and Iryna Gurevych, “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation,” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 4512–4525 (2022)

[6] Geoffrey Hinton, Oriol Vinyals and Jeff Dean, “Distilling the Knowledge in a Neural Network,” arXiv preprint arXiv:1503.02531 (2015)
[7] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, Allan Hanbury, “Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation,” arXiv preprint arXiv:2010.02666 (2022)