E-041

Sentence-BERT による日本語文の話題分析
Topic Analysis of Japanese Sentences by Sentence-BERT

圓谷 顯信† 上原 稔† 安達 由洋‡
Kenshin Tsumuraya Minoru Uehara Yoshihiro Adachi

1. はじめに

近年、機械学習に基づく自然言語処理技術 BERT[1]やその発展形を用いて日本語テキストを話題や感情に基づいて分類する研究が盛んに行われている。しかし、教師あり学習を採用したクラス分類モデルの構築には、目的に応じたクラス集合の選択と各クラスに対応する大量の教師ラベル付きデータセットを作成する必要があり、これらの作業には多大な労力を要する。我々は Word2Vec や BERT から抽出される特徴ベクトル（分散表現）を用いて日本語テキストを教師データなしで分類・検索する研究[2, 3]を報告している。文献[4]では、コサイン類似度を用いた類似度評価に適した分散表現生成モデルとして Sentence-BERT が提案された。日本語 Sentence-BERT の構築は文献[5]で報告されている。この文献では、７つの公開されている日本語 BERT を用いてクラスタリング実験を行い、性能評価をしている。

本稿では、話題に基づく日本語文の分類・検索に適した Japanese Sentence-BERT (JSBERT)の構築法と、クラスタリング結果、クラスタのラベリング技法について報告する。

2. JSBERT

JSBERT は Pretrained Japanese BERT models [6]を事前学習モデルとして用い、JSNLI [7]を用いて訓練している。学習パラメータはバッチサイズ 128、エポック数 1、損失関数 MultipleNegativesRankingLoss である。

本研究では、文の分散表現の生成方法として、入力文を構成する全ての単語の分散表現の相加平均をとる方法(JSBERT-normal)と、入力文を構成する名詞のみの分散表現の相加平均をとる方法(JSBERT-nouns)の二つを採用する。図 1 に名詞のみを用いる分散表現の生成レイヤーを示す。

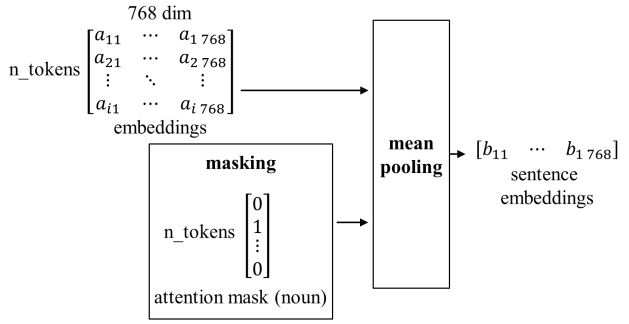

図 1　分散表現の生成レイヤー

3. 話題に基づく日本語文の分類・抽出

本節では、JSBERT-normal と JSBERT-nouns を用いたクラスタリングの性能評価結果について報告する。

3.1 クラスタリング用データセット

分類精度を評価するデータセットとして Test213 と Livedoor ニュースコーパス[8]を使用する。Test213 は我々が作成した ”大学”、”政治”、”病気” など 13 カテゴリ 213 文からなる教師ラベル付きデータセットである。表 1 に Test213 データセットの一部を示す。

表 1　Test213 データセット (一部)

sentence	label
川端康成の作品は心動かされる。	小説
いとこが九州の国公立大学に通っています。	大学
阪神・淡路大震災から今年で２５年が経ちます。	災害
高気圧の影響で、日中は暑い。	天気
ネコ科の種は、瞬発力を活かした動きで狩りを行う。	動物
診察を受けた結果、腰椎椎間板ヘルニアと認められた。	病気
麻婆豆腐をご飯にかけて食べるのが好きだ。	料理
明日はオリンピックのゴルフの予選を見に行きます。	スポーツ

3.2 JSBERT によるクラスタリング

本実験ではまず、Test213 に対して JSBERT-normal と JSBERT-nouns を用いたクラスタリング精度評価を行った。

実験手順を下記に示す：

手順 A1：JSBERT-normal と JSBERT-nouns のそれぞれに Test213 データセットを入力し各文の分散表現を出力する。

手順 A2：分散表現を Agglomerative Clustering (ward 法)でクラスタリングする。

手順 A3：クラスタの中で最も多い教師ラベルをそのクラスタのラベルに設定する。

手順 A4：クラスタのラベルをそのクラスタの属する文の予測ラベルとし、精度評価を行う。

表 2 に各モデルの精度評価結果を示す。

表 2　Test213 データセットによる分類精度

Topic	JSBERT-normal			JSBERT-nouns		
	Prec	Rec	F1	Prec	Rec	F1
大学	0.000	0.000	0.000	1.000	0.833	0.909
政治	0.750	1.000	0.857	0.923	1.000	0.96
病気	0.696	0.941	0.800	0.761	0.941	0.842
服装	0.800	0.800	0.800	1.000	1.000	1.000
料理	0.958	0.920	0.939	0.961	1.000	0.980
スポーツ	1.000	0.778	0.875	0.933	0.777	0.848
災害	0.640	1.000	0.780	0.800	1.000	0.888
天気	1.000	0.750	0.857	1.000	0.812	0.896
ゲーム	0.941	0.762	0.842	1.000	1.000	1.000
動物	1.000	0.944	0.971	1.000	1.000	1.000
小説	1.000	0.800	0.889	1.000	0.933	0.965
勉強	0.577	0.938	0.714	0.875	0.875	0.875
IT	1.000	1.000	1.000	1.000	0.857	0.923
average	0.864	0.886	0.860	0.942	0.925	0.929

† 東洋大学 総合情報学研究科　Graduate School of Information Sciences and Arts, Toyo University
‡ 東洋大学 工業技術研究所　Research Institute of Industrial Technology, Toyo University

表 2 より、JSBERT-nouns で生成した分散表現を用いたクラスタリングは、JSBERT-normal で生成した分散表現を用いたクラスタリングよりも F1 Score の平均が約 7%高い。なお、JSBERT-normal で生成した分散表現は ”楽しい” や ”怖い” といった感情語の情報も埋め込んでいるため、感情語に基づいたクラスタが形成されることがある。

次に Livedoor ニュースコーパスを用いて、JSBERT-normal と JSBERT-nouns の精度評価を行う。Livedoor コーパスでは、文単位分類ではなく記事単位で分類を行い、各記事の分散表現は記事中の文の分散表現の相加平均で求めている。表 3 に Livedoor コーパスに対する評価結果を示す。

表 3　Livedoor ニュースコーパスに対する分類精度

Category	JSBERT-normal			JSBERT-nouns		
	Prec	Rec	F1	Prec	Rec	F1
独女通信	0.560	0.748	0.640	0.562	0.708	0.626
IT ライフハック	0.477	0.303	0.371	0.000	0.000	0.000
家電チャンネル	0.000	0.000	0.000	0.305	0.300	0.302
Livedoor HOMME	0.211	0.360	0.266	0.000	0.000	0.000
MOVIE ENTER	0.768	0.890	0.824	0.741	0.913	0.818
Peachy	0.583	0.461	0.515	0.477	0.620	0.539
エスマックス	0.439	0.680	0.534	0.423	0.717	0.532
Sports Watch	0.876	0.807	0.840	0.877	0.856	0.871
トピックニュース	0.698	0.839	0.762	0.682	0.806	0.739
average	0.576	0.636	0.594	0.582	0.703	0.633

3.3 ラベル語評価値に基づくラベリング

クラスタリングで得た各クラスタに、自動的に適切なラベルを付加できると各クラスタの内容を把握しやすくなり、クラスタの個数の決定に有益な情報となる。

各クラスタに自動的にラベルを付加する手順を示す：

手順 B1：分析対象全文を構成する名詞の cluster-based TF-IDF（通常の TF-IDF の document を cluster に置き換えたもの）を求める。

手順 B2：各名詞の分散表現とその名詞が属するクラスタのセントロイドとの cosine 類似度を求める。

手順 B3：ラベル語評価値 e を次式で求める。

$$e = \alpha \times TF\text{-}IDF + (1-\alpha) \times cosim$$

ここで、$TF\text{-}IDF$はその名詞の cluster-based TF-IDF であり、$cosim$はその名詞とセントロイドとの cosine 類似度である。

手順 B4：各クラスタ内の名詞（ラベル候補単語）から評価値 e の上位 10 個をそのクラスタのラベルに設定する。

表 4 に“病気”クラスタに属する文の一部と教師ラベルを、表 5 に自動的に付加されたラベルを示す。

表 4　“病気”クラスタに属する文 (一部)

sentence	label
診察を受けた結果、腰椎椎間板ヘルニアと認められた。	病気
熱が下がらずぐったりしており、肺炎の疑いがある。	病気
定期検査で、脳に腫瘍が認められた。	病気
椎間板ヘルニアが収まって嬉しいです。	病気
両側の唾液腺に腫脹が見られ，流行性耳下腺炎の疑いを認める。	病気
気管支喘息および慢性呼吸不全が疑われるため、テオフィリン血中濃度の計測を行う。	病気
頸椎損傷した患者の容態が経過良好なので、投薬３０日分で様子をみる。	病気

表 5　“病気”クラスタのラベル

$TF\text{-}IDF$	$cosim$	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$
疑い	呼吸	呼吸	呼吸	呼吸
呼吸	息	疑い	疑い	疑い
とき	腹	運動	運動	運動
ヘル	椎	息	気管	気管
neer	運動	腹	不全	とき
運動	疑い	気管	患者	不全
気管	不全	不全	慢性	患者
慢性	患者	患者	病気	慢性
不全	慢性	椎	久々	病気
ふく	腰	慢性	息	久々

表 5 は、α が 0（$cosim$のみ）、0.25、0.50、0.75、および 1（$TF\text{-}IDF$のみ）の場合のラベルをラベル評価値の降順で表示している。この ”病気”クラスタでは、$\alpha = 0.5$のとき $TF\text{-}IDF$のみや$cosim$のみでは現れない ”病気” という適切なラベルが付加されている。Test213 データセットに対する 13 クラスタへのクラスタリングでは、$\alpha = 0.5$ としたときに他のα値のときよりも大多数のクラスタに適切なラベルが付加された。

4. まとめ

話題に基づく日本語文の分類・検索に適した分散表現を生成する JSBERT の構築法を提案し、JSBERT により生成した分散表現を用いて日本語文クラスタリングの精度評価を行った。そして、文の分散表現を名詞の単語のみから生成することで、話題に基づく文のクラスタリング精度が上がることを検証した。また、形成されたクラスタに対し cluster-based TF-IDF と cosine 類似度を用いて適切なラベルを自動的に付加する技法を考案した。

今後の課題として、JSNLI にその他の学習データも増やして Sentence-BERT を訓練し、話題に基づいたクラスタリングの精度を向上させることがある。また、JSBERT で生成した分散表現を用いて、問合せ文に対して類似の話題を含む文を検索する類似文検索技法の開発も課題である。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv:1810.04805v2 (2018).

[2] Adachi, Y. & Negishi, T. “Development and evaluation of a real-time analysis method for free-description questionnaire responses”, In Proceedings of the 15th IEEE International Conference on Computer Science and Education (2020).

[3] 圓谷顯信, 高橋宏和, 安達由洋, ”BERT による日本語文の感情分析と話題分析”, 情報処理学会第 84 回全国大会 (2022)

[4] Reimers, N. & Gurevych, I. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019).

[5] Shibayama, N. & Shinnou, H. (2021). Construction and Evaluation of Japanese Sentence-BERT Models, In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp.731–738.

[6] Tohoku NLP Group. (2022/06/01, accessed). Pretrained Japanese BERT models, https://github.com/cl-tohoku/bert-japanese.

[7] Yoshikoshi, T., Kawahara, D. & Kurohashi, S. “Multilingualization of a Natural Language Inference Dataset Using Machine Translation”, SIG Technical Reports, Vol.2020-NL-244 No.6 (2020).

[8] RONDHUIT. (2022/06/01, accessed). Livedoor news corpus, https://www.rondhuit.com/download.html#ldcc.