The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

日本語大規模 SNS+Web コーパスによる単語分散表現のモデル構築

Constructing of the word embedding model by Japanese large scale SNS + Web corpus

松野 省吾 Shogo Matsuno　水木 栄 Sakae Mizuki　榊 剛史 Takeshi Sakaki

株式会社ホットリンク
Hottolink, Inc.

Abstract: In this paper, we present the word embedding model constructed by Japanese text existing on SNS including Twitter. This model is created from a Japanese large-scale corpus using multiple categories such as SNS data, Wikipedia, and Web pages as media. Perorming the evaluation by the word similarity calculation task with Speaman's rank correlation coefficient as the evaluation index for the created word embedding model resulted in a performance of about 7 points better than the model created by only Wikipedia as the learning corpus was obtained. The presented word embedding model in this paper is planned to be released through the website, and we hope that by utilizing this model, natural language processing research for SNS data will become more active.

1. はじめに

自然言語処理の研究開発を進める上で，単語分散表現モデルを用いる機会が多くなっている．単語分散表現は単語の周辺文脈を含めて単語の意味をベクトルとして表現する．これを用いることで，文書分類，評判分析，機械翻訳など，様々な自然言語処理タスクの性能が向上することが示されてきた．そのため，コーパスから作成した単語分散表現のモデルが活発に研究・公開されているが，大規模なコーパスはその殆どが英語を対象とするものであり，公開されている日本語を対象としたコーパスとその単語分散表現は限られている．主要な日本語大規模コーパスとしては，現代日本語書き言葉均衡コーパス(BCCWJ)[前川 2008]，国語研日本語ウェブコーパス(NWJC)[Asahara 2014]，京都大学テキストコーパス[河原 2002]，Kyoto University and NTT Blog コーパス(KNB)[橋本 2011]などが知られている．一方で，Twitter のような SNS に着目したコーパスはプライバシー保護の観点などから公開が難しく，単語分散表現モデルとしての公開も GloVe[Pennington 2015] や Twitter word2vec model[Godin 2015]といった，英語を対象としたモデルに限られる．そこで，筆者らの研究開発上で利用するリソースとして作成した単語分散表現モデルを一般に利用可能な形で配布することにした．筆者らの分散表現モデルは日本語 Wikipedia に加え，自動収集した Web ページ，ブログや Twitter 等の SNS データを中心としてコーパスを構築しており，砕けた日本語表現やネットスラングなどの新語・未知語を多く含むという特徴を持つ．

本稿では，配布する分散表現モデルの作成方法と，その学習コーパスである日本語大規模 SNS+Web コーパスについて紹介する．

2. コーパス構築方法

本稿で紹介する日本語大規模 SNS+Web コーパスは SNS データ，日本語 wikipedia，自動収集 Web ページの 3 種類の媒体から作成した日本語大規模コーパスである．複数の媒体から語彙を収集したため，語彙情報が豊富であり，応用向きであることを特徴とする．各媒体の規模は SNS（ブログ・Twitter）データ＞自動収集 Web ページ＞日本語 wikipedia の順である．

連絡先：松野省吾，株式会社ホットリンク，東京都千代田区富士見 1-3-11 富士見デュープレックスビズ 5F，shogo.matsuno@hottolink.co.jp

各媒体は以下の 4 種類の処理を順に実施した．

1. 平文コーパスの収集・構築
2. 前処理
3. 分かち書きコーパスの構築
4. 全媒体を統合，単語分散表現(Word2Vec)の学習

以下，各処理について説明する．

2.1 平文コーパスの収集・構築

平文コーパスを構築するためにデータを収集した．データ収集の期間は以下の通りである．

- ブログ：2015 年 1 月～2016 年 6 月
- Twitter：2016 年に公式アプリからの投稿の一部
- 日本語 Wikipedia：2015 年 11 月 23 日付 dump file
- 自動収集 Web ページ：2009 年 9 月～2016 年 6 月

なお，Twitter データはリツイートによる投稿を除外している．

2.2 前処理

収集した平文コーパスから本文を抽出する．各媒体から本文抽出時に加えた処理を以下に示す．

- ブログ：HTML タグの除去
- Twitter：ReTweet ヘッダ，URL，メンションタグ，ハッシュタグの除去
- 日本語 Wikipedia：タグ，テーブル等のメタ情報の除去（WikiExtractor を使用）
- 自動収集 Web ページ：HTML タグの除去

次に，意表記を吸収するための正規化処理を実施する．正規化処理を以下に示す．

- 全角英数字を半角に置換
- 半角カタカナを全角に置換
- ハイフンの類似文字をハイフンマイナス(U+002D)に置換
- 長音記号の類似文字を全角長音記号(U+30FC)に置換
- 1 回以上連続する長音記号，スペースを 1 回に置換
- チルダに類似する文字を削除
- 全角記号，全角スペースは半角に置換
- ただし，句読点，中黒，等号，カギ括弧は全角記号に置換
- テキスト先頭・末尾・文字間に含まれるスペースを削除

なお，正規化処理の実装は Python::neologdn package を用いた．

2.3 分かち書きコーパスの構築

前処理済の平文コーパスの分かち書きを実施する. 分かち書き器として MeCab を用い, システム辞書に mecab-ipadic-NEologd を用いた. ただし, Twitter 媒体のみ分かち書き器として Juman を用いている. Juman を用いる理由としては以下で述べる. また, 最小形態素数は日本語 wikipedia のみ 5 とし, それ以外の媒体では 10 とした.

Twitter はくだけた文体が多いことから, 通常の形態素解析器では分かち書きの性能が十分でないことが知られている[北川 201, 森 2016]. そのため, Twitter の分かち書き精度向上を目的としたコーパスの構築も試みられている[大崎 2016]. また, Juman は未知語モデルを搭載しているため, くだけた文体に対しても比較的ロバストな処理が可能であるという報告もある[笹野 2014]. そこで, 実際に Twitter の本文の分かち書きを行い, いずれの方法が良いかを調査した.

調査対象となる文は Twitter から 129 文をランダムで抽出し, 以下に示す形態素解析器と辞書・モデルの組み合わせを用いて分かち書きを行った. 文ごとの分かち書き結果を確認し, 特に誤りが少ないように思えるモデルを選び(複数選択可), 選ばれた回数を各組み合わせで比較した. 表 1 に分かち書き器と辞書・モデルの組み合わせと選択回数の結果を示す.

表 5 から, 今回の調査では Juman による分かち書きが最も誤りが少ないように思えるという結果となった. そこで, 日本語大規模 SNS コーパスでは Twitter の分かち書き器として Juman を用いる.

表 1: 分かち書き器と辞書・モデルの組み合わせと選択回数

分かち書き器	辞書・モデル	選択回数
Juman	default	47
MeCab	mecab-ipadic-NEologd	41
MeCab	ipadic	37
KyTea	default (BCCWJ + UniDic)	16
KyTea	Twitter Corpus (re-training)	14

2.4 単語分散表現の学習

作成した分かち書きコーパスを用い, 単語分散表現(Word2Vec [Mikolov 2013])を学習する. ここで, 実装は Python::genism package を使用した. 表 2 に Word2Vec の学習パラメータを示す.

表 2: Word2Vec の学習パラメータ

パラメータ	値
アルゴリズム	Word2Vec [CBOW,Skip-Gram]
次元数	200
最低単語頻度	10
Context window size	5
負例サンプリング	25
初期学習率 α	0.025
Context Distribution Smoothing	0.75
Down-sampling ratio	1e-5
Iteration	20
単語表現	w

3. 評価

3.1 コーパスの概要

表 3, 4 に構築したコーパスの規模を示す. また, 表 5 に分かち書きコーパスの統計量を示す.

表 3: 平文コーパスの規模

媒体	行数 [Mil]	ファイルサイズ [GB]
SNS データ	288	36.2
Wikipedia	7	2.2
Web ページ	126	25

表 4: 分かち書きコーパスの規模

媒体	行数 [Mil]	ファイルサイズ [GB]
SNS データ	180	37.1
Wikipedia	7	2.5
Web ページ	95	28

表 5: 分かち書きコーパスの統計量

指標	値
行数 [Mil]	282
トークン数[Giga]	12
頻度 10 回以上のユニークな形態素数	2,067,629

3.2 分散表現の性能評価

日本語大規模 SNS+Web コーパスによる分散表現と Wikipedia による分散表現の性能を比較する. 比較評価対象として, 本コーパスによる分散表現と同様の方法で作成した Wikipedia による分散表現, および東北大により公開されている日本語 Wikipedia エンティティベクトル[Suzuki 2016]を用いる. 加えて, 単語分散表現の性能評価用に単語ペアを厳選したデータセットである, 日本語単語類似度・関連度データセット JWSAN-1400[猪原 2018]を用い, Speaman の順位相関係数を評価指標とすることで, モデル性能評価を実施した.

表 6 に評価結果を示す. 本コーパスによる分散表現は Wikipedia のみで学習した 2 つのモデルと比較し, 良い性能を獲得することができた. Wikipedia のみで学習した 2 つのモデルはほぼ同程度の性能を示していることから, 分散表現の学習方法に関しても妥当であると考えられる.

表 6: 分かち書きコーパスの統計量

モデル	相関係数
日本語大規模 SNS+Web コーパス	0.548
Wikipedia(ホットリンク)	0.478
Wikipedia(東北大)	0.472

4. まとめ

本稿では筆者らの作成した日本語大規模 SNS+Web コーパスとその単語分散表現について, その構築方法と統計量の概要を紹介した. 本コーパスは SNS データを含む大規模な日本語コーパスである. 日本語を対象とした文・単語の分散表現モデルで公開されているものは限られ, SNS や Web 上の文書を学習コーパスとした分散表現のモデルには価値があると筆者らは考えている. 評価実験の結果から, 本コーパスから作成した単語分散表現モデルは Wikipedia から作成したモデルと比較して, 単語類似度算出タスクにおいて 7 ポイント程度良い性能

を獲得できた．本稿で紹介した単語分散表現モデルは Web サイトを通じて公開する予定である．今回のモデルが活用されることで，特に SNS データを対象とした自然言語処理研究が一層盛んになることを期待したい．

今後の予定としては，日本語の文分散表現モデルの作成や，最近の新語，流行語などを取り入れた単語分散表現モデルの更新について取り組みたい．

参考文献

[前川 2008] 前川喜久雄：KOTONOHA『現代日本語書き言葉均衡コーパス』の開発，日本語の研究，4(1), pp. 82-95, 2008.

[Asahara 2014] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, Hikari Konishi: Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan', Alexandria, Vol 26, No.1-2, pp.129-148. 2014.

[河原 2002] 河原大輔，黒橋禎夫，橋田浩一：「関係」タグ付きコーパスの作成．言語処理学会 第 8 回年次大会予稿集，pp.495‐498，2002.

[橋本 2011] 橋本力，黒橋禎夫，河原大輔，新里圭司，永田昌明：構文・照応・評判情報つきブログコーパスの構築．自然言語処理 Volume 18, Number 2, pp.175-201. 2011.

[Pennington 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning: GloVe: Global Vectors for Word Representation, Empirical Methods in Natural Language Processing (EMNLP), pp. 1532- 1543, 2014.

[Godin 2015] Godin, F., Vandersmissen, B., De Neve, W., & Van de Walle, R.: Named entity recognition for Twitter microposts using distributed word representations. Workshop on Noisy User-generated Text, ACL 2015.

[北川 2016] 北川善彬，小町守：深層ニューラルネットワークを利用した日本語分割，言語処理学会第 22 回年次大会，pp.933-936, 2016.

[森 2016] 森信介：多様なテキストの言語処理，第 112 回音声言語情報処理研究会，招待講演，2016

[大崎 2016] 大崎彩葉，唐口翔平，大迫拓矢，佐々木俊哉，北川善彬，堺澤勇也，小町守：Twitter 日本語形態素解析のためのコーパス構築，言語処理学会第 22 回年次大会，pp.16-19, 2016.

[笹野 2014] 笹野遼平，黒橋禎夫，奥村学：日本語形態素解析における未知語処理の一手法-既知語から派生した表記と未知オノマトペの処理-，自然言語処理，22(6), pp. 1183-1205, 2014.

[Mikolov 2013] T. Mikolov, I. Sutskever, K. Chen, G. S. Correado, and J. Dean: Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, 26, pp. 3111-3119, 2013.

[鈴木 2016] 鈴木正敏，松田耕史，関根聡，岡崎直観，乾健太郎：Wikipedia 記事に対する拡張固有表現ラベルの多重付与，言語処理学会第 22 回年次大会，2016. (http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector)

[猪原 2018] 猪原敬介，内海彰：日本語類似度・関連度データセットの作成，言語処理学会第 24 回年次大会，pp. 1011-1014, 2018. (http://www.utm.inf.uec.ac.jp/JWSAN/)