単語らしい文字 n-gram の埋め込みによる単語の分散表現

Geewook Kim[†,‡‡] 福井 一輝[††,‡‡] 羽田 哲也[‡,‡‡] 下平 英寿[††,‡‡]

† 京都大学 工学部
†† 京都大学大学院 情報学研究科
‡ 大阪大学大学院 基礎工学研究科 ‡‡ 理化学研究所 革新知能統合研究センター

{geewook, k.fukui}@sys.i.kyoto-u.ac.jp, hada@sigmath.es.osaka-u.ac.jp, shimo@i.kyoto-u.ac.jp

概要

単語らしい文字 n-gram の埋め込みによる新しい単語埋め込み手法を提案する．提案手法は単語分割せずに単語らしい文字 n-gram の分散表現を生コーパスから直接構成することで単語埋め込みを実現する．文字 n-gram の出現頻度に基づいた先行研究では，複数の単語分割に解釈できる文字列に柔軟に対応できる反面，単語とはいえない文字 n-gram まで含めて分散表現を求めてしまうため，抽出した文字 n-gram に対する有効な単語の量と，得られる分散表現の質が課題となる．この問題を解決するため，本研究では文字 n-gram の期待単語頻度に基づいて埋め込み対象を選択する．評価実験の結果から，提案手法が従来の単語埋め込み手法に比べ，量と質の面から単語の分散表現を効果的に構成できることを確認した．また，単語埋め込みを用いた名詞カテゴリ予測タスクの性能向上にも大きく貢献できることを確認した．

1 はじめに

自然言語処理において埋め込み (embedding) とは主に単語や複合語などの自然言語の構成要素に分散表現，すなわち実ベクトル表現を与えることである．近年，word2vec [1] に代表される単語埋め込み (word embedding) 手法が，教師なしで学習を行い単語の意味と関係性を内包する連続的な分散表現を取得できる方法として注目されてきた．

既存の単語埋め込み手法は一般に文字列の形式で与えられるコーパスを単語分割 (word segmentation) し，分割された文字 n-gram の共起情報を用いて単語の分散表現を構成する．つまり，単語分割された文字 n-gram (segmented character n-gram) を単語とみなして埋め込みを行う．本稿では，このような単語分割を経由する単語埋め込み手法を segmented (character) n-gram embedding といい，以下では SNE と呼ぶ．

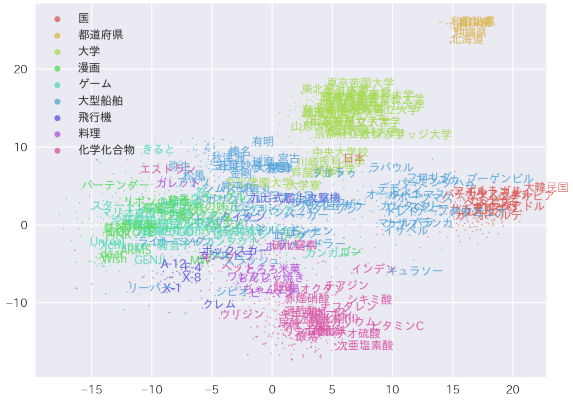

図 1: 提案手法により得られた名詞単語の分散表現の t-SNE による可視化．意味カテゴリごとに色付けした．

一方で，日本語や中国語のように単語境界が明示されない言語を扱う場合，自然言語の文章を適切な単語の列に分解することは困難であり，単語分割の過程で誤りが生じてしまう．これは「同じ文脈に現れる単語は似た意味を持つ傾向にある」という分布仮説に基づく単語埋め込みの性能を劣化させるため，SNE の問題となる．

このような問題を回避するために，単語分割を経由しない単語埋め込み手法 [2] が提案されている．この手法では単語分割は行わずに，文字列コーパスから頻出する文字 n-gram を部分文字列の重複を許して抽出し，これらを埋め込み対象とみなして分散表現を計算する．本稿ではこの手法のことを frequent (character) n-gram embedding といい，以下では FNE と呼ぶ．FNE は標準的な単語辞書では対応できない新造語や絵文字などの分散表現も辞書を使わずに獲得できる．また，単語分割が容易でない日本語や中国語などの言語を対象とした単語埋め込みに適している [3]．しかし頻出する文字 n-gram には有効な単語でないものも多いため，メモリや計算量の観点から埋め込み対象とする文字 n-gram の数を制限すると本当に学習してほしい単語が多く失われてしまう問題がある．

本研究では単語らしい文字 n-gram を埋め込むことで単語埋め込みを行う手法 word-like (character) n-gram embedding を提案する．これを WNE と呼ぶ．有効な単語をより多く選択するために，単語境界確率を用いてそれぞれの文字 n-gram がコーパス中で単語として出現する期待頻度を計算し，その値に基づいて埋め込み対象を選択する．そして生コーパスからそれらの共起情報を集計し分散表現を構成する．提案手法 WNE は単語分割の誤りによる問題を回避しつつ単語らしい文字 n-gram を優先的に埋め込むことで FNE を改善する．

日本語版 Wikipedia コーパスを用いた評価実験からは提案手法が既存手法より多くの単語を効果的に埋め込めることを確認した．また，単語埋め込みを用いる下流タスクの性能向上にも貢献することを確認した．

2 関連研究

2.1 Segmented n-gram embedding

従来手法の SNE では，文字列コーパスを単語分割してから単語埋め込みを行う．単語分割とは入力文の中に単語境界がどこに存在するかを決定することである．例えば，長さ (文字数) N のコーパス $\mathbf{x}_1^N = x_1 x_2 \cdots x_N$ に対する単語分割は，変数 Z_i を

$$Z_i = \begin{cases} 1 & x_i, x_{i+1}\text{の間に単語境界が存在する} \\ 0 & x_i, x_{i+1}\text{の間に単語境界が存在しない} \end{cases}$$

とすると，単語境界の存在を表す変数の列 $\mathbf{Z}_0^N = Z_0 Z_1 \cdots Z_N$ を出力として返す．ここで最初の文字の前と最後の文字の後ろには単語境界が存在するとみなせるので Z_0 と Z_N は 1 である．英語など分かち書きされた言語では空白からほぼ自明に分割が定まる．また，日本語など分かち書きされない言語では一般的に形態素解析を用いて分割を定めることができる．SNE では単語分割の結果から与えられた単語境界に基づいて，隣接する単語境界の間に存在する文字 n-gram を単語としてみなして埋め込みの対象とする．

コーパスの単語分割が定まると，SNE では一般に窓幅 s を決めて，注目している単語とその単語の前後の窓幅以内の文脈に存在する単語ペアの頻度を共起情報として集計する．これらの共起情報を用いて埋め込みの対象として選択した文字 n-gram の分散表現を構成する．具体手法としては，共起行列の行列分解に基づいて分散表現を構成する LSA や文脈に出現する単語の分散表現から注目している単語の分散表現を予測する (またはその逆を予測する) タスクを通して分散表現を学習する word2vec などが提案されている．

2.2 Frequent n-gram embedding

最近提案された FNE では，まず文字列コーパス中に出現する文字 n-gram の出現頻度を集計する [2, 3]．そして集計された文字 n-gram の中から出現頻度順で上位 k 個の文字 n-gram を埋め込み対象として選択する．

次に，選択した埋め込み対象の共起情報をコーパスから集計する．FNE では単語分割を経由しないので，単語境界が明示されていない生コーパスから共起情報を集計する．よって，SNE とは共起情報の集計方法が異なる．例えば，先行研究 [2] では，埋め込み対象として選択された文字 n-gram が互いに隣接して出現する頻度を共起情報として用いた．つまり，コーパス中に「…素晴らしい学術論文を…」という文字列があるとして,「素晴らしい」,「学術」,「論文」,「学術論文」が埋め込み対象として選択されたとすると，この文字列から (素晴らしい, 学術)，(素晴らしい, 学術論文)，(学術, 論文) が語彙とコンテキストのペアとして集計される．

このように埋め込み対象を選択してそれらの共起情報が得られれば，従来の単語埋め込み手法と同様に分散表現を構成する．

3 提案手法

提案手法の word-like n-gram embedding (WNE) では，単語らしい文字 n-gram を埋め込みの対象として選択する．文字 n-gram の単語らしさは期待単語頻度を指標とする．文字 n-gram の期待単語頻度とは，文脈情報を考慮してその文字 n-gram がコーパス中で単語として出現する期待頻度である．これは文献 [4] で確率的単語分割コーパスにおける単語 1-gram 頻度として提案されているものである．期待単語頻度は単なる文字 n-gram の出現頻度とは異なる概念である．例えば,「美容院でカラーリングした」という入力文が与えられた時，この文には「リング」という文字 3-gram が出現しているが，この文脈において「リング」は単語として出現しているとはいえない．

WNE では期待単語頻度を計算するために，文字列コーパス内の全ての隣接する文字間に対して単語境界が存在する確率を求める．2.1 節で述べたように従来の SNE では単語分割から単語境界を表す変数 Z_i を定めたが，WNE では Z_i を確率変数として，$P(Z_i = 1) = P_i$,

図 2: 単語境界確率列 $\mathbf{P}_1^N$ の例

$P(Z_i = 0) = 1 - P_i$ となる確率 P_i を文脈情報から与える．これをコーパス全体に適用して単語境界確率列 $\mathbf{P}_1^N = P_0 P_1 \cdots P_N$ を出力する（図 2）．ここでも最初の文字の前と最後の文字の後ろには単語境界が存在するとみなせるので P_0 と P_N は 1 である．文献 [4] では自動単語分割システムにより単語境界と判定された点では $P_i = 0.987$，そうでない点では $P_i = 1 - 0.987$ としていた．本稿では 4.4 節で述べるように文脈情報を説明変数とするロジスティック回帰によって単語境界確率を与える．

単語境界確率列を用いて，以下のモデルで期待単語頻度を求める [4]．まず i 番目の文字から長さ $n = j - i + 1$ の文字 n-gram $\mathbf{x}_i^j = x_i x_{i+1} \cdots x_j$ が出現していた時，それが有効な単語である確率を

$$P(i,j) = p_i p_{j+1} \prod_{k=i+1}^{j} (1 - p_k) \qquad (1)$$

と与える．これは現実を単純化して，確率変数 $Z_1, \ldots, Z_{N-1}$ が独立であることを仮定したモデルである．コーパスにおける文字 n-gram w の全ての出現を $I(w) = \{(i,j) \mid \mathbf{x}_i^j = w\}$ とすると，w が有効な単語として出現していた期待頻度は (1) の合計

$$D(w) = \sum_{(i,j) \in I(w)} P(i,j) \qquad (2)$$

であり，これを期待単語頻度と呼ぶ．WNE では，以上の計算で求められる期待単語頻度が大きいものを埋め込み対象として選択し，FNE と同様に共起情報を集計して分散表現を構成する．

4 実験

4.1 実験 1: 名詞カテゴリ予測

学習した分散表現の品質を評価するために，単語ベクトルに基づいて名詞単語を意味カテゴリに分類するタスクを行った．コーパスは日本語版 Wikipedia[1]を用い，単語ベクトルはすべて 200 次元とした．評価用の名詞単語は wikidata[2]から事前に用意した意味カテゴリ[3]に属する名詞単語を抽出して用いた．また，単語ベクトルから名詞カテゴリ予測を行う分類器として C-SVM を用い，埋め込み対象の中で wikidata の名詞単語に含まれる単語ベクトルの 6 割を訓練データ，残りの 4 割をテストデータとした．

表 1: 名詞カテゴリ予測 (実験 1)

Method	Recall	micro F-score
SGNS-SNE	0.011	0.735
SGNS-FNE	0.008	0.706
SGNS-WNE	**0.024**	**0.844**

4.2 実験 2: 埋め込み対象の選択基準の評価

コーパス内の文字 n-gram をその出現頻度と期待単語頻度それぞれを用いてソートし，上位 k 個の文字 n-gram を埋め込み対象とする．別途用意した単語辞書の単語がどのくらい含まれているかを Precision-Recall Curve によって調べた．標準的な単語辞書の mecab-ipadic と新造語なども多く収録されている mecab-ipadic-neologd[4]の両方を用いた．

4.3 単語埋め込み手法

実験 1 では以下の SNE, FNE, WNE を比較した．実験 2 では FNE と WNE を比較した．

SGNS-SNE: 単語分割を経由する単語埋め込み (SNE) を行った．単語埋め込みには Skip-gram model with Negative Sampling (SGNS)[5][1] を用いた．形態素解析器 MeCab と mecab-ipadic 辞書[6] を用いて分割されたユニークな単語を全て埋め込み対象とした．

SGNS-FNE: SGNS を FNE で拡張した手法[7][3] を用いた．また，埋め込み対象は上位 $k = 2 \times 10^6$ 個の頻出 n-gram である．

SGNS-WNE (提案手法): SGNS を WNE で拡張したものを用いた．期待単語頻度の意味で上位 $k = 2 \times 10^6$ 個の文字 n-gram を埋め込みの対象とする．共起情報を SGNS-FNE と同様に集計し，この共起情報に基づいて SGNS の目的関数で文字 n-gram の分散表現を学習する．

[1] https://dumps.wikimedia.org/jawiki/20171201/jawiki-20171201-pages-articles1.xml-p1p168815.bz2
[2] https://dumps.wikimedia.org/wikidatawiki/entities/20171120/wikidata-20171120-all.json.bz2
[3] { 人, 企業, 映画, 鉄道駅, 化合物, 都市, 日本の漫画, テレビゲーム, 本, 大学, 食品, 大型船舶, 都道府県, 国, 歌, 飛行機 }
[4] https://github.com/neologd/mecab-ipadic-neologd
[5] https://code.google.com/archive/p/word2vec/
[6] http://taku910.github.io/mecab/
[7] https://github.com/oshikiri/w2v-sembei

4.4 単語境界確率の推定

本稿では SGNS-WNE の前処理で必要となる単語境界推定にロジスティック回帰を用いた．説明変数とする文脈情報は，注目している単語境界から窓幅 s 以内で隣接している文字 n-gram のペア a, b の Association

$$A(a,b) = \log\left(\frac{\frac{f(ab)}{N}}{\frac{f(a)}{N}\frac{f(b)}{N}}\right) \tag{3}$$

を用いた [5]．ここで N はコーパスの長さ（文字数），$f(a)$ は文字列 a の出現頻度である．Association は一種の PMI (poitwise mutual information) である．例えば「私は研究者です」という入力文に対して「研」と「究」の間の単語境界確率を推定する時，窓幅 $s = 2$ であれば，A(研, 究), A(は研, 究), A(研, 究者), A(は研, 究者) を説明変数として用いる．

本実験ではロジスティック回帰のパラメータ推定は $s = 8$ として，MeCab (辞書は mecab-ipadic) を用いて生成した正解データを利用した．図 2 はこのようにして得られたロジスティック回帰の動作例である．

4.5 実験結果

実験 1 の結果を表 1 に示す．本来はコーパス中の単語を正解とすべきだがこれは得られないので，wikidata の名詞単語を「正解」とみなして埋め込み対象として選ばれた単語の割合 (Recall) を計算している．提案手法 (WNE) が既存手法 (SNE, FNE) よりも多くの名詞単語の分散表現を学習しており，なおかつ分類器の予測精度 (micro F-score) も高い．また，提案手法によって学習した分散表現の t-SNE によるマッピングを図 1 に示す．単語分割を経由せずに多くの名詞単語の分散表現が構成でき，意味カテゴリごとにクラスタが形成されていることが確認できる．以上の結果から，提案手法は既存手法より高いカバレッジを有しながら，得られた分散表現の質も高いことが確認できる．

次に，実験 2 の結果を図 3 に示す．これは埋め込み対象の n-gram 数を $k = 1$ から 2×10^6 まで変化させたときの Recall と Precision をプロットした．FNE と WNE の AUC 比は 4.06 であり，WNE がより多くの有効な単語を選択している．この結果から，期待単語頻度が文字 n-gram の単語らしさを表す指標として優れていると考えられる．

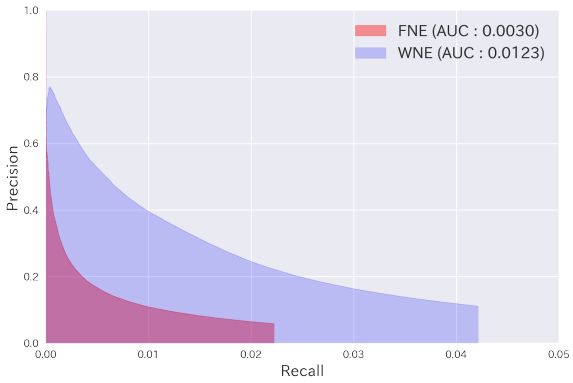

図 3: Precision-Recall Curve (実験 2)

5 おわりに

本稿では，単語らしい文字 n-gram の埋め込みによる単語埋め込み手法を提案した．評価実験により，提案手法が既存手法と比べ，単語として有効な n-gram を多く埋め込みながら質の高い分散表現を獲得できることが確認できた．そして単語境界確率から推定する期待単語頻度が単語らしさを表す指標として適切であることも確認できた．本稿では分かち書きされない言語，特に日本語における単語埋め込みを扱ったが，中国語や韓国語などの他の言語に対しても有効に機能することが期待できる．

参考文献

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[2] 押切孝将, 下平英寿. 単語分割を経由しない単語埋め込み. 言語処理学会第 23 回年次大会論文集, pp. 258–261, 筑波大学, 3 2017. 言語処理学会.

[3] Takamasa Oshikiri. Segmentation-free word embedding for unsegmented languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 767–772, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[4] 信介森, 大介宅間, 岳人倉田. 確率的単語分割コーパスからの単語 n-gram 確率の計算. 情報処理学会論文誌, Vol. 48, No. 2, pp. 892–899, feb 2007.

[5] R. Sproat and C. Shih. *A Statistical Method for Finding Word Boundaries in Chinese Text*, Vol. 4. Computer Processing of Chinese and Oriental Languages, 1990.