非標準的に使用される単語の分散表現の補完手法

A method to complement distributed representation for non-standardly used words

魏 逸倫[1]　労 瑛瑩　韓 東力[2]

Wei Yilun[1]　Lao Yingying　Han Dongli[2]

[1]日本大学大学院 総合基礎科学研究科

[1] Graduate School of Integrated Basic Sciences, Nihon University

[2]日本大学文理学部 情報科学科

[2] Department of Information Science, College of Humanities and Sciences, Nihon University

Abstract: 自然言語の分析においては、分析の障害となりうる隠語など非標準的に使用される単語が存在する。既存研究では未知語としての隠語に関する処理方法が多く述べられているが、別単語への置き換えとしての隠語を検出･分析する技術があまり開発されてこなかった。本研究では、トピックモデルと word2vec を用いて単語の分散表現を獲得し、入力文のトピックと文中に含まれた全単語の適合性を求めることで、隠語を自動的に検出し、そして、その正しい意味を補完する手法を提案する。

1. はじめに

自然言語の処理においては、単語の意味を分散表現として学習した上で利用するのがよくある手法であるが、学習された分散表現が処理対象となる単語にうまく対応できない場合が存在する。処理対象となる単語が分散表現の辞書に登録されていない、または登録された意味と異なる意味で処理されると、解析誤りが生じる。

これには、処理対象となる単語が分散表現を学習するために利用したコーパス内に出現しなかった「未知語」である場合と、単語が「誤字」や「隠語」などのように、人為的かどうかを問わず、辞書内で予想された形ではない、非標準的に使用された場合が存在する。例えば、次の２つの文に「パンダ」という単語を含んでいるが、文中では「パンダ」が動物という一般的な意味ではなく、「パトカー」という意味で使われている。すなわち、「パンダ」という単語が「隠語」として、本来の意味と異なる非標準的な使い方をしている。

- パンダに○○県警って書いてた。
- パンダの横っ腹に○○県警って書いてた。

非標準的に使用された単語の抽出及び分析に関する研究はいくつか存在しているが、概ねに分析目標となる単語を含むデータを大量に用いる手法[1][2][3]と、ある特定の領域におけるルールを用いる手法[4]、この２つの手法のいずれに属する。

ただし、「誤字」や「隠語」は性質上、同じ形式で表すことが少なく、大量のデータを用意するのが困難である。また、非標準的に使用された単語の出現ドメインや処理ルールを予想するのに限界があり、万遍なく対応することは難しい。そのため、既存の研究は高い精度が得られたものが多い反面、汎用性が低い。

本研究では、処理対象となる非標準的に使用される目標単語が以上の処理条件を満たさない場合でも利用できる、ルールベースに頼らないかつ目標単語を含む文章だけを必要とする、汎用性の高い手法の提案を目的とする。本手法により、処理される文書から目標単語を検出・分析することで、目標単語の元の意味となる単語を特定することを期待する。

2. 目標単語の抽出

本章では、目標単語の抽出について説明する。ただし、目標単語が分散表現の辞書に存在しない場合、抽出が簡単であるため省略する。隠語として文書に出現した単語が辞書に登録済みの場合は図 1 の通り、学習データを利用して単語の分散表現を求め、そして判別対象となる文に含まれている全単語の分散表現を用いて目標単語が含まれているかどうかを判断する。含まれると判断された場合に目標単語を抽出する。以下では提案手法を詳しく説明する。

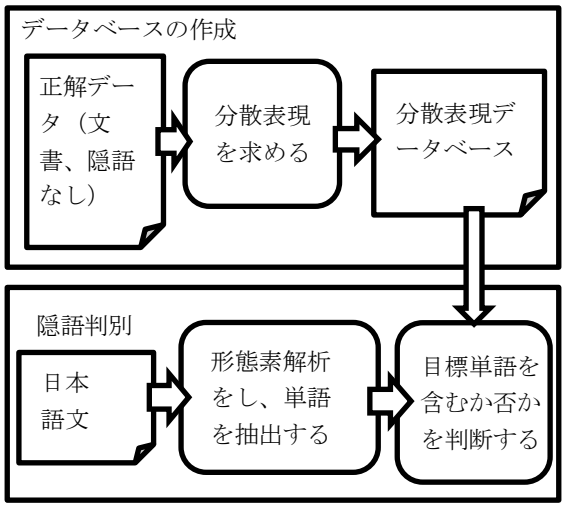

図 1 目標単語の抽出の流れ

2.1. 単語の分散表現

分散表現とは単語をベクトルに変換すること、もしくは変換したベクトルのことである。分散表現の学習方法として、word2vec とトピックモデルの一つであるＬＤＡを利用し、ベクトルの次元数およびトピック数を 200 と設定している。

ＬＤＡとは、文書中の単語がどのようなトピックを表しているかを確率的に求める言語モデルの一つである。この手法を利用すれば、文書に出現した全単語がそれぞれのトピックに属する確率がベクトルとして得られる。word2vec も単語のベクトルを得る手法であり、単語の周辺語を利用し、単語の意味をベクトルの各次元で解釈する。本研究では word2vec とＬＤＡを比較した上で、word2vec で得られた単語ベクトルの各次元をトピックと見なす。

日本語としての正しさを考慮し､Wikipedia 日本語版のテキストを学習データとして用いて、単語をベクトル化する。このようにして得られた 870374 文書、合計 593,670 単語をベクトル化し、単語のベクトルの各次元の値を単語が各トピックとの関連度と考える。ただし、word2vec の結果でベクトルの次元の値がマイナスとなるとき、単語はトピックと反対する方向に意味を持つと考えられるので、トピックとの関連度を値の絶対値をとる。

2.2. 目標単語の判別

任意の１文に対して、少なくとも１つのトピックが含まれていると考えられる。また、文に含まれている全単語が文のトピックに寄与する。

文があるトピックに属すると思われるとき、トピックが文内の全単語と何らかの関連をもつと考えられる。すなわち、文に含まれるすべての単語がトピックとの関連度が高いほど、文全体がそのトピックに属していると判断するのが適切である。したがって､文があるトピックとの関連度 *AP* は各単語とトピックの関連度を用いて式(1)の通り相乗平均､あるいは式(2)の通りエントロピーを用いて評価する。

$$AP_{t,c} = (\sum_{w \in t} Rw,c)/n \quad (1)$$

$$AP_{t,c} = -(\sum_{w \in t} R_{w,c} log\left(R_{w,c}\right)) \quad (2)$$

t: 文
c:トピック
$AP_{t,c}$：文 t がトピック c に属する関連度
w: 単語
$R_{w,c}$：単語 w とトピック c の関連度
n ：文に含まれる単語数

文は全トピックに対して最大となる関連度をもつトピックと最も関連しており、文中の全単語も同一トピックに属すと考える。さらに、文中の全単語が同じトピックに属す可能性を式(3)のように計算し、その結果を一貫性と定義する。一貫性は小さいほど、全単語が同一の文に含まれる可能性が小さくなり、すなわち文に存在すべきではない目標単語が存在する可能性が大きくなる。

$$\boldsymbol{S_t = \max\left(AP_{t,c}\right)}(\mathrm{c} \in \boldsymbol{C}) \quad (3)$$

t: 文
c:トピック
C:200 個のトピックの集合
S_t：文 t の一貫性

図 1 の通り、分析対象となる各文に対し、形態素解析した上で、各単語の分散表現を用いて、一貫性を計算し、目標単語の存在を判別する。

2.3. 一貫性の有用性実験

2.3.1. 実験データ

Apple Store [5]の商品レビューから人手で隠語を含まない 200 文を収集し、ランダムに 100 文ずつに分け、データセット 1 とデータセット 2 とする。データセット 1 の各文に対し、ランダムに選定した１つの単語を同じ発音を持つ別の単語に置き換えることで作成された目標単語を含むデータセット 3 を作成する。ここでは、選定された単語に対し、Mecab の辞書を用いて読み仮名が同一となる単語候補の中からランダムに１つ選択する。最後に３つのデータセットに対し一貫性を求める。

表 1 抽出の結果実験

分散表現の計算方法	文の関連度の計算方法	一貫性の平均			目標単語の認識度
		データセット1	データセット2	データセット3	
Word2vec	相乗平均	3.9706	3.9195	3.8501	79%
	エントロピー	3.6971	3.6930	3.6862	66%
	相乗平均×エントロピー	13.5855	13.4777	13.1148	82%
LDA モデル	相乗平均	0.0209	0.0208	0.0188	66%
	エントロピー	3.3363	3.3382	3.0817	61%
	相乗平均×エントロピー	0.0581	0.0582	0.0535	60%

2.3.2. 実験結果および考察

表 1 は実験結果のまとめである。文に目標単語を入れると、目標単語が文のトピックに属ないため、適切性が下がる。どの手法においても、目標単語を含むデータセット 3 がそのもととなるデータセット 1 より一貫性の平均値が小さくなっていることを確認できた。目標単語の認識度は、同じ文に対し、目標単語を含む方の一貫性が小さくなる割合を表しており、最大 82%で最小でも 60%といずれの手法を利用しても一定の有用性が確認できた。

また、異なる文を利用したデータセット 2 とデータセット 3 を比べても、目標単語を含むデータセット 3 の一貫性が小さく、一貫性は目標単語を含むかどうかを判断する一つの基準として利用できる可能性が示唆された。

さらに具体的な精度を求めるために、人手で選んだ目標単語を含む文 100 文と含まない 50 文を抽出しテストした。結果としてエントロピーを利用して文関連度を計算する場合に、精度が一番高く、f=0.9 となる。ただし目標単語を判別する閾値として、表 1 に示された、データセット 2 とデータセット 3 の一貫性の中間値を利用した。

2.4. 目標単語の抽出

目標単語が含まれていると判定された文に対し、文内に含まれた全単語の平均ベクトルとの cos 類似度が小さい単語を目標単語として抽出する。目標単語が 1 個だけ含まれると仮定し、類似度が最小となる単語だけを目標単語として抽出した場合、2.3.2 のデータに対して、0.69 の精度で目標単語を抽出できた。

3. 意味の補完

手法に汎用性を持たせるために、すべての単語がもっている、「意味」、「発音」「形」の３つの要素を利用することで目標単語の意味を分析する。また、計算速度を上げるために、文脈から意味の候補を抽出してから目標単語の分析を行う。流れは図 2 の通りになる。

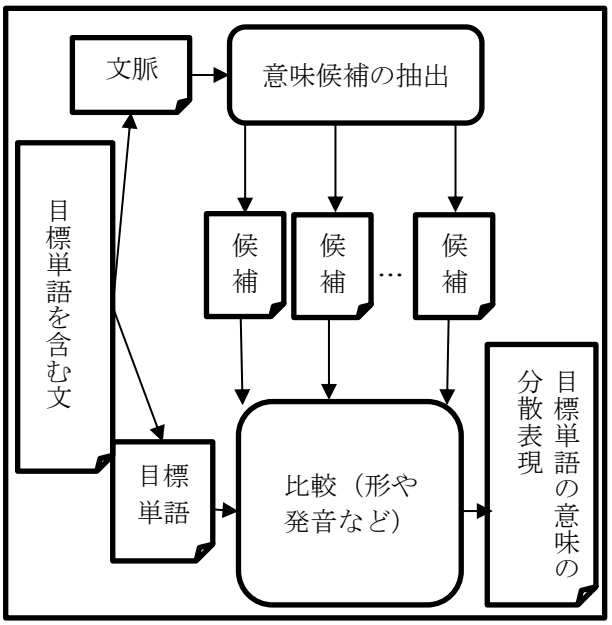

図 2 補完手法の流れ図

3.1. 文脈からの候補抽出

すべての日本語単語と比較して正解を探し出すことは現実的ではないため、文脈を利用することで、文章に関連する単語だけを取り出す。目標単語と関連を持つ可能性が大きい手がかりとして以下の４つを利用する。

- 目標単語を含む文書のタイトル
- 目標単語を含む文書全体
- 目標単語を含む文
- 目標単語の周辺語単語

文章に関連する単語は４つの手がかりから圧縮して得られたベクトルとの cos 類似度が高い単語と定

義する。圧縮方法は図 3 の通り、ニューラルネットワークを利用して学習したものとなる。入力データは手がかりごとに含む全単語の平均ベクトルである。また本研究では分散表現ベクトルの長さを 200 と設定している。

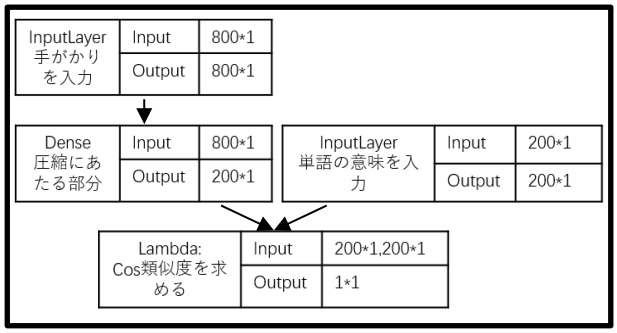

図 3 圧縮を学習するために利用した NN 構造

3.2. 候補の順番付け

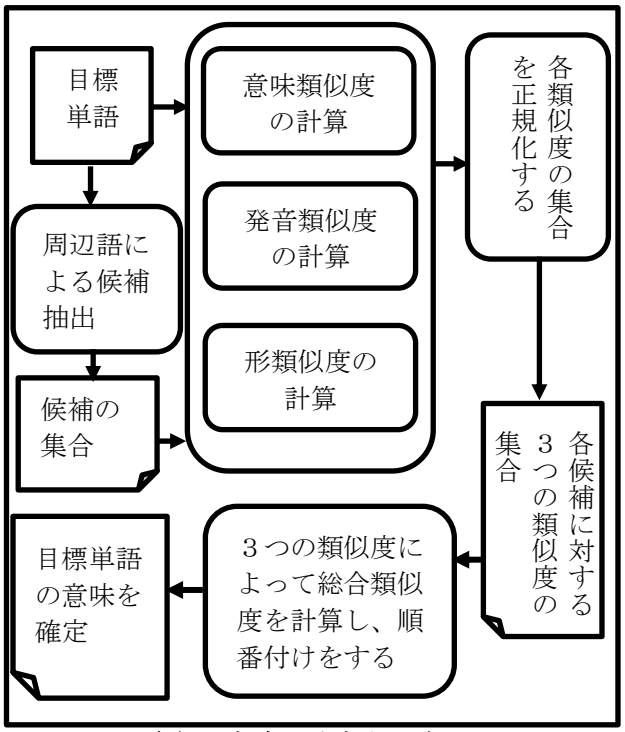

図 4 意味を確定する流れ

図 4 では候補を目標単語と比較することで、元の意味を確定する流れを示す。すべての候補に対し、目標単語と「意味」「発音」「形」で比較する。比較は以下のように類似度を求めることにより行う。

- 意味の類似度：分散表現ベクトルの cos 類似度
- 発音の類似度：単語ローマ字の編集距離
- 形の類似度：学習したＣＮＮモデル

また、総合類似度は式(4)の通りに計算する。ただし、正規化により、類似度の値が 0 から 1 までの数字に変換する。

総合類似度 =
MAX(意味類似度,発音類似度,形類似度)　　(4)

総合類似度が高いほど、候補と元の意味が一致する可能性が高い。

3.3. 事例による手法有用性の考察

ネット掲示板から抽出した目標単語を含む文を利用して、意味補完の効果を確かめた。元の意味は著者が文脈から判断し、決めたものとなる。

3.3.1. 事例 1

- 入力文:' 一年間乙
 来年は分かりやすい相場だといいな～

'
- タイトル:'○○○○年○○会を語る'
- 目標単語:乙
- 分散表現の類似単語:'丙', '甲', '丑', '卯', '支路', '癸', '巳', '辰丸', '支文', '酉'
- 元の意味:お疲れ様
- 分析結果:

表 2 事例 1 の分析結果

順位	候補	文脈との意味類似度	発音類似度	形類似度	意味類似度
1	オウガ	0.9332	0.987	1	0.3932
2	オワタ	0.8169	1	1	1
3	イジリ	0.71065	0.3859	1	0.4007
4	イワオ	0.68545	0.987	1	0.2244
5	イチカ	0.64475	0.7048	1	0.2621
6	ルウム	0.41155	0.3859	1	0.6071
7	メラク	0.2024	0.7048	1	0.27045
8	綴る	0.7197	0.3859	0.2178	0.99995
9	負わさ	0.7032	0.99995	0.707	0.58795
10	緒形	0.5909	0.99995	0.1491	0.4882
11	わた	0.5143	0.99995	0.46555	0.68945
12	終わら	0.3541	0.99995	0.4621	0.6122
…	…	…	…	…	…

元の意味が 2079 番目にあり、全日本語の範囲内で考えると、上位であると言えるが、実用できる水準に至らなかった。

実際に上位となる単語は、総合類似度が高いが、「お疲れ様です」と比べると、文脈との関係性が低

いので、文脈を考慮する総合類似度の求め方の改善が課題となる。

3.3.2. 事例 2

- 入力文: 'オワタ<&喜びを表す顔文字>

'
- タイトル:'○○○○年○○会を語る'
- 目標単語: オワタ
- 分散表現の類似単語: '綴る', 'アドヴァタイジングスローガン', 'スラング', '悲しい', '失恋', 'キャプテン・ファルコン'
- 元の意味:おわった
- 分析結果:

表 3 事例 2 の分析結果

順位	候補	文脈との意味類似度	発音類似度	形類似度	意味類似度
1	載っ	0.8877	1	0.6741	0.7750
2	ボツ	0.7162	1	0.9765	0.7643
3	コツ	0.7099	1	0.5332	0.8115
4	ほっ	0.4915	1	0.8836	0.6521
5	モツ	0.4854	1	0.9437	0.9741
6	卒	0.3633	1	0.9481	0.9698
7	ボッ	0.3568	1	0.9943	0.6034
8	乙種	0.314	0.9968	0.6764	1
9	コフィー	0.2921	0.8283	1	0.3653
10	すっ	0.5217	0.9992	0.9999	0.9094
11	蝦	0.3408	0.4283	0.5934	0.9999
…	…	…	…	…	…
2079	お疲れさま	0.8887	0.9646	0.3356	0.5134
…	…	…	…	…	…

元の意味と近いと著者が判断した単語「終わら」が 12 番目に出現した。他の上位単語と比べると、周辺語との関係性が低く、自動的に判断するには難しいため、意味の補完は、図 5 のように、すべての上位単語をユーザーに見せ、人手で最終意味を決定するという半自動的な形式で行う予定となっておる。

分析対象となる文章に出現した目標単語は、他の文書でも同じ意味で目標単語として使われる可能性が大きく、また同じ文書内で使われた他の目標単語と関連を持つ可能性が大きいため、既知の目標単語を新しい目標単語の元の意味を分析する手がかりとして応用す手法も検討している。

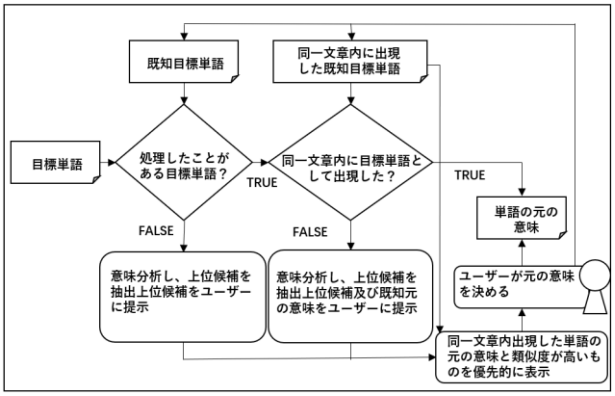

図 5 半自動的で意味を決める方式

4. まとめ

本研究では、自然言語処理の精度を上げるために、前処理として非標準的に使用される単語を検出し、その分散表現を補完する手法を提案した。

単語の分散表現を利用して、文のトピックと全単語の適合性を求めることで、隠語を自動的に検出する手法を提案した。評価実験では、約 60%の精度で隠語が存在するかどうかを判断できることが判明し、手法の有用性が示された。

また、目標単語の意味分析において、目標単語の文脈と単語自身の特徴（形や発音など）を利用することで、意味を特定する手法を開発しているが、精度や文脈の利用などいくつかの課題が存在している。

今後の予定としては客観的な評価実験を実施し、本提案手法の有用性を確かめる。

参考文献

[1] 大西 洋,田島 敬史:語の出現の偏りに基づく新たな目標単語の発見, 日本データベース学会論文誌, Vol. 12, No. 1,pp. 49-54, June 2013

[2] 日田 仁:www サイト内の不正コンデンツ検出支援システムの構築,分散システム/インターネット応用技術,24-6 2001.11.30

[3] 乾 亮,山村 毅:情報科学部視覚的「読み」を用いた分割表記文字の処理,言語処理学会第 25 回年次大会発表論文集, P1-33,2019

[4] 屋 誠司:常識的判断システムにおける未知語処理方式,人工知能学会論文誌,17, 667-675, 2002

[5] https://www.apple.com/app-store/