自律型日本語学習のための単語分散表現モデルの構築

豊田哲也（東邦大学）

Construction of a Word Embedding Model for Autonomous Japanese Language Learning

Tetsuya TOYOTA, Toho University

要旨：筆者は e-learning 上での自律的な日本語学習を実現するための手段として、学習資源の不足に着目し、これまでに問題文とその解答を自動生成する手法を提案してきた。中でも類語選択の問題においては、妥当な類語となる正解の選択肢と、それ以外の誤選択肢を複数用意する必要があり、この選択肢を単語分散表現から得ることを試みた。本研究では、単語の分散表現モデルによって得られる類語抽出の精度を向上させるために、コーパスの一部を絞り込むことによって、日本語学習教材の作成に特化した分散表現モデルを提案する。

キーワード：単語分散表現、Word2Vec、問題の自動生成、e-learning、Wikipedia

1. はじめに

新型コロナウィルス感染症の拡大によって、e-learning やオンライン学習といった遠隔教育の需要が高まっている。これまでオンライン学習などの適用範囲とはみなされていなかった分野においても対策の必要性に迫られており、日本語教育もその分野の 1 つと言ってよい。その中で、e-learning システムの利用によるメリットについての再評価がなされており、時間や場所を選ばずに学習が可能な点において、今後も e-learning を用いた新たな学習活動が継続されるであろうと考えられる。教師と学習者が対面で学ぶ形式の教育・学習とは異なり、e-learning を含むオンライン学習においては、教師が学習コンテンツを e-learning 上に実装する必要があるため、学習者に提供可能な学習リソースには限界がある。この問題の解決策として、学習者に提供する問題の自動生成に関する研究が行われており、筆者も日本語教育に関する問題を Web 上のテキストから収集して、これを問題文に変換する研究を行っている(豊田ら 2019)。これに加えて、近年自然言語処理で注目を集めている、単語分散表現を利用し、類語に関する問題生成に利用している（豊田 2019)。単語分散表現とは、これまでの one-hot ベクトルによる文章のベクトル表現とは異なり、単語そのものを多次元の固定長ベクトルで表現したモデルのことで、得られた単語のベクトル間の類似度計算によって単語の類義語を抽出することが可能となっている。これまでの研究では、類似単語の穴埋めや選択の問題を作成する「類義語抽出タスク」において、回答選択肢中の正解選択肢が他の選択肢に比べて、対象単語と単語間とのベクトル類似度が高くなる傾向にあることを確認している。ただし、現在提案されているいくつかのモデ

ルを用いて類義語抽出タスクを適用すると、必ずしも大規模なコーパスによって構築されたモデルで良い結果が得られるとは限らず、日本語 Wikipedia をコーパスとしたモデルが最も性能が高い結果となっている。ただ、日本語 Wikipedia のコーパスにおいても、類語辞典に掲載されていない単語の類似度が高くなるという結果が得られており、問題自動生成には適していない単語も多くみられる。そのため、日本語学習における問題自動生成に適した単語分散表現モデルを獲得することが必要となる。

そこで本研究では、日本語学習の類語問題を自動生成することに特化した、より精度の高い類語抽出を可能にする単語分散表現モデルを構築する。これは、単語分散表現モデルを構築するうえで必要なコーパスを Wikipedia の一部のカテゴリに属する記事に限定することでの解決を試みる。具体的には、Wikipedia の本文の中から、任意のカテゴリに属する記事群を選別して、モデル構築に必要なコーパスを調整可能にする。これにより、モデル構築に不要なテキストデータを削減し、類語抽出の精度向上に必要な分野を絞り込むことが可能となり、より精度の高い類語抽出が可能になると考えられる。得られたモデルは、日本語教育における類義語抽出タスクとして、市販教材の問題集を用いて検証を行う。

2. 提案手法

2.1 単語の分散表現モデル

本節では、単語の分散表現モデルについていくつか説明する。数学的なモデル説明は参考文献に譲り、ここでは分散表現モデルの仕組みと、コーパスとの関係性に焦点を当てて説明する。

単語の分散表現とは、単語を多次元固定長の実数値ベクトルで表現したものであり、単語をベクトルで表現することで単語の類似度を計算することや、ある単語を別の単語の演算によって求めることが可能となる。分散表現以前のモデルには、その文書中に含まれる単語の出現頻度や、単語の有無といった one-hot ベクトルによる文章のベクトル表現があるが、これに対して単語の分散表現モデルは、単語自身と単語間の関係性をベクトルの類似度によって得ることが可能となる。例えば、「大阪」という単語と「福岡」という単語のベクトルは類似した値を取り、「東京」－「日本」＋「イギリス」は「ロンドン」に近い値を取るようになる。このように、単語をベクトルで表現することで、様々な応用に用いられている。この単語分散表現を獲得するための代表的なモデルが、 Mikolov らによって提案された Word2Vec である（Mikolov et.al 2013）。Word2Vec は、単語の意味が文脈上の周囲の単語によって決まる「分布仮説」に従ったモデルであり、コーパスの統計情報からニューラルネットワークを使って学習される。単語の分散表現を得るためには、モデル構築のための「単語の学習」が必要となり、一般的にはコーパスを学習モデルに適用させることによってこれらを得る。日本語 Wikipedia をコーパスとして Word2Vec に適用して得られた単語分散表現（鈴木他 2016）は、日本語 Wikipedia の本文データをコーパスとして作成されたモデルであり、他にも日本語 Wikipedia をコーパスとして利用する研究が

散見される。他にも、国立国語研究所の「国語研日本語ウェブコーパス」から作成された NWJC2vec は、258 億語規模のウェブテキストコーパスから構築された単語分散表現である（Asahara 2018）。モデル構築に単語の学習が必要であるということは、コーパス中に任意の単語が存在しなければその単語のベクトルを得ることはできない。単語を表現する固定長ベクトルの次元は、コーパス中の単語数の数に応じて決定する必要があり、モデル構築のパラメータの 1 つとなっている。提案されているモデルの多くは次元数を 100 以上に設定していることが多い。

大規模なコーパスを利用することは、日本語という言語の中の 1 単語について、普遍的な意味合いと他の単語との距離関係を表現するためのものと考えることができる。そのため、学習する必要のない単語を除外すれば、学習される単語間の類似度が上昇する可能性があり、なおかつ、目的に応じて局所的な学習用コーパスを設定できれば、より類似する単語の距離を短くすることができるのではないかと考えられる。

2.2 Wikipedia のカテゴリデータ

分散表現モデルを構築するために利用されるリソースとしてオンライン事典の Wikipedia がある。Wikipedia は言語リソースとして十分な量と、これらのデータがデータベースの型式等で提供されており、誰でもこの言語リソースを利用できることが特徴である。Wikipedia の 1 つの記事には、複数のカテゴリと呼ばれる見出しがつけられており、検索を容易にすることを目的として用意されているが、カテゴリは上位/下位の階層構造を有しており[1]、カテゴリ階層的な構造からオントロジー構築に関する試みもなされている。本研究ではこのカテゴリデータに着目し、カテゴリの特徴を利用してモデル構築に必要なコーパスを調整する方法を提案する。

まず、Wikipedia 上の任意のカテゴリ$C_i \in C$(C は全カテゴリ集合)に対して、C_iの下位にあるサブカテゴリを求める関数$sub(C_i)$を定義する。$sub(C_i)$は、Wikipedia のデータベースデータ「categorylinks」から求めることができる。$sub(C_i)$により、得られたC_iのサブカテゴリ集合$SC_i \subset C$を基に、サブカテゴリがなくなるまで$sub(SC_i)$を再帰的に繰り返し、C_iをルートとしたC_iより下位に位置するサブカテゴリ集合$ASC_i \subset C$を得る。このとき、$SC_i \subset ASC_i$であるが、$sub(SC_i) = \emptyset$のときに限り$SC_i = ASC_i$である。一方で、Wikipedia 上の任意の記事$P_j \in P$(Pは全記事集合)は、必ず 1 つ以上のカテゴリに属しており[2]、記事P_jが属するカテゴリ集合$C_{P_j} \subset C$を定義する。ここで、任意のカテゴリC_iのすべての下位カテゴリASC_iにC_{P_j}が含まれている場合、記事P_jの本文テキストをモデル構築のコーパスに加える。以上から、任意のカテゴリC_iを決定すると、その下位のカテゴリ集合ASC_iをカテゴリとして持

[1] 厳密には木構造のような完全な階層構造ではなく、閉路を含まない有向非巡回グラフであるとされている（）が、変更不可の上位カテゴリが設定されていることに加えて、下位のカテゴリをサブカテゴリとして構造化しているため、階層構造と見なす.

[2] Wikipedia の運営方針であるため、カテゴリがない記事には警告が出る.

つすべての記事群を単語学習用のコーパスとして加えることができる。本研究では、任意のカテゴリを決定することで、下位カテゴリを限定し、その下位カテゴリをカテゴリとして設定されている記事の本文データをコーパスとして利用する。

3. 評価実験

3.1 実験環境と利用データ、評価指標

評価実験で利用するデータについて説明する。モデル構築には、Python において Word2Vec を利用可能な gensim のライブラリを利用する。モデル構築の際の各種パラメータについては、ベクトルの次元数を 100 次元で固定し、それ以外のパラメータは min count = 20, window size = 10, epocks = 5 である[3]。なお、コーパス中の単語分割のための形態素解析器には MeCab を利用し、NEologd を辞書として使用する。この辞書は、新語や固有表現の抽出精度が高いことから採用している。モデル構築のためのコーパスとして、日本語 Wikipedia の dump データ「jawiki-latest-page-article.xml」を用いる。データは 2021 年 1 月時点での最新版である。また、Wikipedia 中の各記事がどのカテゴリに属しているかを調べるために、「jawiki-latest-page」および「jawiki-latest-categorylinks」の 2 つのデータベースを用いる。類語抽出の精度検証のために、市販されている日本語能力試験の問題集（日本語能力試験問題研究会 2010，森本ら 2018）から必要な要素を選別し、各レベルの問題の対象となる語と正解/不正解の 4 つの選択肢、計 5 単語を抽出する。ここで選別とは、単語分散表現でベクトル化された単語を対象とすることを指し、ベクトル化されていない単語や語句、文章などの類似度計算ができない要素は除外するという意味である。また、本実験でのカテゴリの絞り込みは、Wikipedia のカテゴリの最上位項目である「主要カテゴリ」のサブカテゴリである、「主題別分類」の中から選択する。そのほかの候補である「学科別分類」は学問基準であることから、1 カテゴリに様々な分野の記事が含まれており、「指標別分類」は様々な指標を基にしたカテゴリ構造のため一貫性がない。その点、「主題別分類」は記事の主題別に分類されており、記事内容が類似するものがまとめられていると考えられることから、主題別分類を選択した。主題別分類の下位カテゴリは 10 項目[4]存在し、10 カテゴリから下位のカテゴリを抽出して、コーパスに利用する記事の絞り込みに利用する。ただし、下位カテゴリには Wikipedia の運営上のカテゴリや、重複関係にあるなどの不要なカテゴリが数多く含まれているため、不要なカテゴリを削除したうえで利用する。

次に、評価指標について述べる。コーパスごとのモデル評価は、対象語のベクトルと正解語のベクトルの類似度より求める。ベクトル間類似度は 2 つのベクトルの内積より求める。全単語集合 W のうち、ある単語 $w_i \in W$ の分散表現は、n 次元のベクトル v_{w_i} で

[3] min count は単語の出現数以下を対象としないことを指し、window size は文脈中の学習単語の前後幅を示す．epochs は学習の回数である．

[4] 科学、技術、自然、社会、総記、地理、哲学、人間、文化、歴史の 10 項目

表 1：各モデルの概要と実験結果

モデル		全記事	75%	50%	25%	10%
記事数		2,050,871	1,537,826	1,025,832	512,901	204,811
学習単語数		403,094	337,885	263,785	169,609	92,195
正解選択肢類似度	平均	0.4582	0.4918	**0.4978**	0.4821	0.4838
	最大	0.8883	0.8902	0.9023	**0.9071**	0.9011
	最小	0.0265	0.0749	0.0557	0.0393	**0.0810**

表現され、これは任意の単語 $w \in W$ に対して同様である。今、対象語w_tのベクトルv_{w_t}と正解語w_aのベクトルv_{w_a}の類似度は、

$$\cos(v_{w_t}, v_{w_a}) = \frac{v_{w_t} \cdot v_{w_a}}{|v_{w_t}||v_{w_a}|} \quad \cdots\cdots\cdots \quad (1)$$

によって求める。このとき、(1)式が 1 に近いほど類似していることとなる。評価用データセットの（対象語、正解語）のペアに対して類似度を式(1)により計算し、各コーパスにおける類似度の平均値を「正解選択肢類似度」とし、コーパスを絞り込んだことによる影響を分析する。次に、誤選択肢の評価指標について説明する。誤選択肢w_eは、各問題に 3 つ用意されている。この 3 つの単語と対象語との類似度を計算し、それぞれ最も類似度の高い誤選択肢と、中間、最も類似度の低い誤選択肢ごとにまとめて平均や標準偏差等により分析を行う。この評価指標を「誤選択肢類似度」とする。

3.2 実験結果

ページ数の都合上、各カテゴリで絞り込んだモデルの分析結果は本発表において公表することとし、ここでは提案手法のコンセプトが妥当であるか否かについて評価する。すなわち、コーパス中の本文データを絞り込むことの有効性を評価するため、全記事で構築したモデルと、ランダムに抽出した記事で構築したモデルを比較することで検証する。各モデルの実験結果を表 1 に示す。なお、表 1 の列先頭に表記されている%は、乱数による記事抽出確率である。正解選択肢類似度は、評価用データセットの（対象語、正解語）の類似度における各統計値である。表 1 の結果から、ランダムに記事を絞り込んだ場合で学習したモデルは、全記事で学習したモデルよりも高い類似度を得ることができている。平均値は 50%絞り込みのモデルが最も高く、最大値は 25%絞り込み、最小値で最も値の高かったモデルは 10%絞り込み、という結果となった。ただし、10%のモデルでは、学習単語数が全記事でのモデルと比べて 2 割程度まで減少しており、学習済み単語の中に正解データセットの単語が存在しない場合があったことから、大幅な絞り込みは単語の学習に悪い影響を及ぼしていることがわかった。これを踏まえると、カテゴリによる記事の絞り込みは、カテゴリによっては記事数が少ない場合が存在すること、さらには、記事本文の記述量に

よって、モデル構築に大きな影響を与える可能性がある。本発表では、カテゴリごとのモデルの評価指標の結果、および具体的な対象語と選択肢の関係について説明する。

4　おわりに

本研究では、日本語学習の類語問題を自動生成することに特化した、より精度の高い類語抽出を可能にする単語分散表現モデルを構築する手法を提案した。これは、単語分散表現モデルを構築するためのコーパスを日本語 Wikipedia の一部のカテゴリに属する記事に限定することで、任意のカテゴリに属する記事群を選別して、モデル構築に必要なコーパスを調整可能にした。これにより、単語分散表現モデルを構築する際に不要なデータを削減し、類語抽出の精度向上に必要な分野を絞り込むことが可能となった。評価実験によって、データの絞り込みが類語抽出の精度を向上されていることを確認したが、記事数を抑えすぎると、単語の学習に悪影響を与えていることを確認した。今後は、カテゴリの組み合わせなどを考慮したモデル構築手法の改良を検討する予定である。

参考文献

Asahara, M. (2018) *NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus'*, Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, Vol. 24, No. 2. pp.7-25.

Mikolov, T. et.al (2013) *Distributed representations of words and phrases and their compositionality*, In advances in Neural Information Processing Systems, Vol. 26, pp.3111-3119.

鈴木正敏、松田耕史、関根聡、岡崎直観、乾健太郎 (2016)「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(NLP2016)』

豊田哲也、島田めぐみ、保坂敏子 (2019) 「Web テキストを用いた日本語学習問題自動生成システムの構築」『東アジア日本語教育・日本文化研究』、 Vol. 22、 pp.127-138

豊田哲也 (2019)「Web ブラウザを介した学習者の語彙情報の獲得と語彙力測定の試み」『8th international Conference on Computer Assisted Systems For Teaching & Learning Japanese』

日本語能力試験問題研究会 編 (2010)「日本語能力試験直前対策 N1/N2/N3 文字・語彙・文法」『国書刊行会』

森本智子、高橋尚子、黒岩しづ可 (2018)「日本語能力試験 N1/N2/N3 直前対策ドリル＆模試 文字・語彙・文法」『J リサーチ出版』

コーパス

日本語 Wikipedia dump データ　https://dumps.wikimedia.org/jawiki/latest