単語の分散表現を用いた同位語の抽出

水越俊希†， 杉本徹‡
†芝浦工業大学大学院　理工学研究科，‡芝浦工業大学　工学部
{ma15082, sugimoto} @shibaura-it.ac.jp

1. はじめに

単語間の関係は、単語の意味を表す上で重要な要素である。単語間の関係の 1 つに同位語がある。同位語とは、共通の上位概念をもつ単語の集合のことをいう。

同位語抽出は情報検索や話題の転換などに応用できる。例えば、情報検索で SNS の情報を知りたいが、「twitter」という単語しか知らないとき、「twitter」で検索すると「twitter」の情報ばかりが出力される。一方で、「twitter」の同位語である「Facebook」や「google+」などを抽出して、それらの OR 検索をすることで、SNS についての情報を広く出力することができる。

同位語の抽出には大きく分けて 2 種類の方法が用いられる。1 つはシソーラス上の特定の概念の下位に属する概念を同位語として抽出する方法である。この方法には、新しい語が生まれるたびに人手でシソーラスの更新をする必要があるという問題がある。もう 1 つは大規模なコーパスを解析し、同位語が満たす特徴を持つ単語を同位語として抽出する方法である。この方法には、大規模なコーパスを解析する必要があるために大きな計算コストがかかるという問題がある。

本研究では単語の分散表現[1]を用いて同位語を抽出する方法を提案する。単語の分散表現を用いることで、機械的に低コストで同位語を抽出することを目指す。

2. 関連研究

同位語抽出に関する研究はいくつか存在している。Ghahramani ら[2]はベイズ推定の枠組みを用いて、与えられた単語を含む同位語集合 D が存在すると仮定したときに、単語の相互情報量を用いて D に含まれる確率が高い単語を順に抽出する方法を提案している。

新里ら[3]は HTML 文書中の箇条書きや表で同列に扱われる単語は、意味的にも同列の単語である場合が多い点に着目し、大量の HTML 文書から箇条書きや表で同時に出現する単語集合を抽出することで、同じクラスに属する単語集合を抽出する方法を提案している。

山口ら[4]は、同位語の関係にある単語は検索されるときに似た単語と一緒に検索されやすいという点に着目し、大量の検索クエリログから、同じ単語と検索されている単語を同位語候補として抽出する方法を提案している。

3. 提案手法

提案手法の概要を図 1 に示す。
まず、出力したい単語と同位語の関係にある単語w_iを 2 個以上、任意の数だけ入力する。複数の単語を入力する理由は、1 つの単語では適切な同位語候補を特定できないためである。例えば、入力単語「犬」に対して「鳥」が同位語かを考えると、「動物」や「生物」を上位概念とみなすと同位語であるが、「哺乳類」や「犬科の動物」を上位概念とみなすと同位語ではない。一

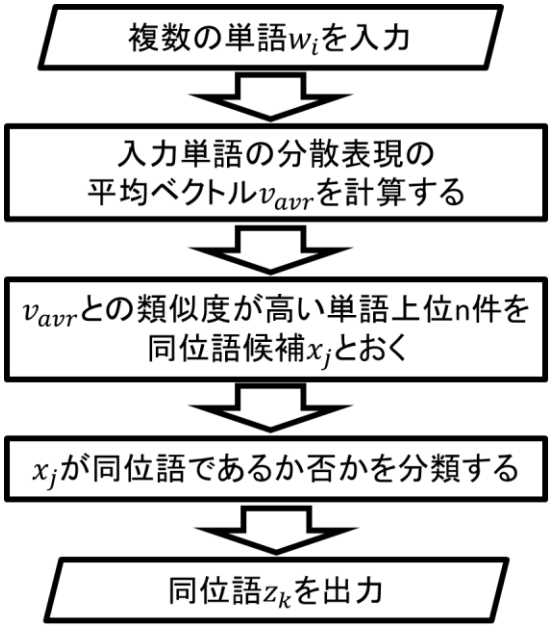

図 1　提案手法の概要

方で、入力単語「犬」と「猫」に対して「鳥」が同位語かを考えると、「犬」と「猫」の共通の上位概念は「哺乳類」と特定できるため、「鳥」は同位語ではないと判断できる。

次に、入力単語に対応する分散表現ベクトルv_iを求め、その算術平均ベクトル$\mathbf{v}_{avr}$を求める。そして、v_{avr}とすべての名詞の分散表現ベクトルとのコサイン類似度を求め、類似度が高いベクトル上位 n 件を求める。求めた単語ベクトルに対応する単語を同位語候補x_jとおく。

表 1 は、関東にある大学である「早大」「明治大」と関西にある大学である「神戸大」「関西大学」をそれぞれ入力したときの同位語候補、上位 20 件である。このように、入力単語の平均ベクトルと類似度が高い単語を求めることで、入力単語と関連する単語が抽出されやすくなる。

分散表現は単語の意味を 1 つのベクトルで表現しているため、類似度が高い単語の中には同位語以外の関係（上位-下位関係、全体-部分関係、反義語など）にある単語も含まれている。そのため、最後に同位語候補x_jが同位語であるか否かを分類し、同位語であると分類したものを同位語z_kとして出力する。同位語であるか否かの

表 1 大学名を入力したときの同位語候補

	A.「早大」「明治大」を入力したとき	B.「神戸大」「関西大学」を入力したとき
1	慶大*	神戸大**
2	東大*	関西学院大**
3	早稲田大学*	大阪大**
4	同大	同大**
5	学習院大*	一橋大学
6	慶応大*	同志社大学**
7	明大*	京大**
8	慶応大学*	大阪大学**
9	中央大学*	立命館大**
10	東京大学*	京都大**
11	京大	大阪市立大学**
12	一橋大*	明治大学
13	法政大学*	東京大学
14	日大*	龍谷大**
15	一橋大学*	京都府立医科大**
16	中大*	中央大学
17	東海大*	琉球大学
18	法政大*	法政大学
19	法大*	東京経済大学
20	上智大学*	慶応大

*はAの出力のうち関東にある大学を表す
**はBの出力のうち関西にある大学を表す

分類には SVM を用いる方法または検索サジェストを用いる方法のどちらかを用いる。

3.1. SVM を用いる方法

同位語候補x_jと入力単語w_iの差のベクトルを SVM に入力し、同位語に分類されたものを同位語とみなす。SVM の学習は以下の手順で行う。

EDR 電子化辞書[5]を用いて指定した上位概念の下位概念にあたる単語集合L_1を取得する。その中から、分散表現ベクトルの生成に用いるコーパスの中で出現頻度上位 20,000 件に含まれる単語集合L_2を求める。L_2から無作為に抽出した 2 語w_1, w_2を提案手法に入力して、同位語候補x_jを 20 件求める。以上の方法で得た結果から、

$$(\text{入力}) = (w_1 \text{または} w_2 \text{と} x_j \text{の差ベクトル})$$

$$(\text{出力}) = \begin{cases} 1 & (x_j \text{が} L_1 \text{に含まれる}) \\ -1 & (x_j \text{が} L_1 \text{に含まれない}) \end{cases}$$

から成る学習データを 15,000 件作成し、SVM に学習させる。

3.2. 検索サジェストを用いる方法

山口ら[4]の検索クエリを用いた同位語抽出の手法を参考に、以下の方法で同位語分類を行う。

Google Suggest API を用いて、同位語候補x_jおよび入力単語w_iと一緒に検索に使われる単語の集合 X,W を取得する。X と W のコサイン類似度を求め、その値が閾値以上であれば同位語であるとみなす。

4. 評価実験

単語 2 語を入力したとき、v_{avr}との類似度が高い単語上位 20 件を同位語候補とするときの正答率を求めた。また、SVM および検索サジェストを用いる方法で同位語を分類したときの F 値を求めた。

分散表現ベクトルは、word2vec[1]を用いて 200 次元のベクトルを生成した。コーパスには日本語 Wikipedia 全文と毎日新聞の記事 2 年分の本文において単語の活用形を基本形に直したものを用いた。

評価データには、EDR 電子化辞書[5]から表 2 の 12 種類の概念の下位概念にあたる単語集合それぞれに対して、分散表現ベクトルの生成に用いたコーパスの中で出現頻度上位 20,000 件に含まれる単語集合からランダムに抽出した 2 語を入力単語とする問題 100 問を用いた。

表 2 使用した概念の種類

	概念の種類	名詞の種類	単語の例
area_proper	地域	固有名詞	インドネシア、地中海、小笠原
area_generic		普通名詞	州、南方、大都市
building_proper	建物	固有名詞	ホワイトハウス、伊勢神宮
building_generic		普通名詞	病院、武道館、議事堂
place_proper	場所	固有名詞	秋田、英国、太平洋
place_generic		普通名詞	空港、スーパー、河川
org_proper	組織	固有名詞	松下電器、オックスフォード
org_generic		普通名詞	学校、寺院、百貨店
machine_generic	機械	普通名詞	自家用車、家電、自転車
color_generic	色	普通名詞	桃、白、すみれ色
mammal_generic	哺乳類	普通名詞	猫、犬、キツネ
plant_generic	植物	普通名詞	柏、椿、栗

SVM を用いる方法では、それぞれの評価データに対して、表 2 の 12 種類の概念から作成した学習データを用いて学習した場合の精度を調べた。SVM のカーネル関数は RBF カーネルを、パラメータは gamma=2^{-4},cost=2^{13}を用いた。

検索サジェストを用いる方法では、閾値を 0 から 0.5 までの間で10^{-4}ずつ変化させたときの分類結果を求め F 値が最も高いものを用いた。

5. 実験結果と考察

SVM を用いる方法では、評価データと学習データの上位概念の種類が一致するときに精度が最も高かった。同位語候補 20 件を抽出したとき、評価データと同じ種類の学習データを用いて学習した SVM で分類したとき、検索サジェストで分類したときの 3 種類において出力した同位語候補数とその正答率を表 3 にまとめる。

表 3 同位語抽出の精度

	20件の同位語候補の正答率	SVMで分類後の正答率	suggestで分類後の正答率
area_proper	15.87/20 =0.794	14.90/16.41 =0.908	15.12/18.07 =0.837
area_generic	4.02/20 =0.201	3.02/3.51 =0.860	3.51/15.89 =0.221
building_proper	12.31/20 =0.616	12.22/12.28 =0.995	9.89/11.28 =0.877
building_generic	4.72/20 =0.236	4.33/4.78 =0.906	3.30/11.96 =0.276
place_proper	14.02/20 =0.701	12.68/14.76 =0.859	13.52/17.89 =0.756
place_generic	5.30/20 =0.265	3.26/5.06 =0.644	4.63/16.06 =0.288
org_proper	14.43/20 =0.722	12.51/14.66 =0.853	13.57/17.80 =0.762
org_generic	5.61/20 =0.281	4.47/5.25 =0.851	5.52/19.37 =0.285
machine_generic	5.07/20 =0.254	4.60/4.81 =0.956	3.26/8.72 =0.374
color_generic	7.49/20 =0.375	7.39/7.39 =1.000	5.70/12.18 =0.468
mammal_generic	5.94/20 =0.297	5.81/5.83 =0.997	4.50/8.10 =0.556
plant_generic	4.07/20 =0.204	3.38/3.69 =0.916	2.67/7.26 =0.368
平均	8.24/20 =0.412	7.38/8.20 =0.900	7.10/13.72 =0.518

SVM の分類結果の F 値を求め、学習データを変えたときに F 値が高いもの上位 4 件（SVM1～SVM4）と検索サジェストの分類結果の F 値を比較した結果を表 4 にまとめる。

表 4　検索サジェストと SVM の学習データを変えたときの分類精度の比較

	suggest	SVM1	SVM2	SVM3	SVM4
area_proper	0.891	0.923 area_proper	0.901 place_proper	0.885 org_proper	0.136 building_proper
area_generic	0.353	0.802 area_generic	0.562 place_generic	0.248 org_proper	0.215 area_proper
building_proper	0.838	0.994 building_proper	0.811 org_proper	0.811 place_proper	0.554 area_proper
building_generic	0.396	0.912 building_generic	0.540 place_generic	0.283 org_proper	0.269 area_proper
place_proper	0.847	0.881 place_proper	0.861 area_proper	0.847 org_proper	0.132 building_proper
place_generic	0.434	0.629 place_generic	0.297 area_proper	0.290 org_proper	0.277 area_generic
org_proper	0.842	0.860 org_proper	0.825 area_proper	0.810 place_proper	0.139 building_proper
org_generic	0.442	0.823 org_generic	0.320 area_proper	0.309 org_proper	0.282 place_proper
machine_generic	0.473	0.931 machine_generic	0.352 area_proper	0.299 place_proper	0.283 org_proper
color_generic	0.580	0.993 color_generic	0.368 area_proper	0.278 org_proper	0.228 place_proper
mammal_generic	0.641	0.987 mammal_generic	0.384 place_proper	0.380 org_proper	0.346 area_proper
plant_generic	0.471	0.871 plant_generic	0.285 area_proper	0.270 place_proper	0.240 org_proper

表 3 から SVM や検索サジェストを用いることで、同位語抽出の精度は向上することがわかった。特に、SVM を用いた場合は 9 割の正答率が得られた。

一方で表 4 をみると SVM を用いる方法では学習データと評価データの種類が同じ、または似ている場合には分類精度が高くなるが、そうではない場合は極端に精度が低下することがわかる。そのため、応用する際には事前におおまかなカテゴリごとに SVM の学習をしておき、入力単語に応じて分類に使用する SVM を切り替えることが必要であると考えられる。また、検索サジェストを用いる方法を、SVM で 3 位以下の学習データで学習したときの精度と比べると、どの評価データにおいても分類精度が高いことがわかる。そのため、事前に用意できないカテゴリがある場合には検索サジェストを用いることが有効であると考えられる。

6. おわりに

同位語抽出を機械的に低コストで行う方法として、単語の分散表現を用いる方法を提案した。評価実験の結果、正答率の平均は 90.0%であった。しかし、SVM を用いる際に評価データと学習データに含まれる同位語が属する上位概念の種類が異なると精度が低下する。

今後の課題として、入力単語の意味カテゴリを推定することで、その意味カテゴリに合うデータで学習された SVM を同位語抽出に利用する手法の検討が挙げられる。

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, Advances in Neural Information Processing Systems, pp.3111–3119, 2013.

[2] Zoubin Ghahramani and Katherine A. Heller: Bayesian Sets, Advances in Neural Information Processing Systems 18 (NIPS 2005), 2005.

[3] 新里圭司,鳥澤健太郎: HTML 文書からの単語意味クラスの単純な自動獲得手法, 情報処理学会論文誌 48(6), pp.2140-2152, 2007.

[4] 山口雅史,大島裕明,小山聡,田中克己: サーチエンジンのクエリログを利用した同位語の発見, 日本データベース学会 letters 5(2), pp.17-20, 2006.

[5] 日本電子化辞書研究所: EDR 電子化辞書 2.0 版 仕様説明書, 2001.