テクニカルノート

国語辞典とテキストコーパスを用いた単語の類似性判別

稲子希望† 笠原 要†

テキストデータに含まれる単語を特徴ベクトルで表現して単語の類似性を判別する方法は，知的なテキスト処理の基盤技術の1つである．類似性判別の対象となる語彙の拡張を目的として，国語辞典とテキストコーパスという異なる形式のテキストデータを用いて統一的な単語のベクトル表現を行う手法を提案する．

A Method for Judging the Semantic Similarity between Words by Using a Dictionary and a Text Corpus

Nozomu Inago† and Kaname Kasahara†

Methods of judgeing the semantic similarity between words based on word vectors which were constructed from text data can be widely applied to intelligent text processing technologies. In order to increase the number of the words to be judged, we propose a method of judgement using two types of text data, a dictionary and a text corpus, together to make word vectors.

1. はじめに

テキストデータを用いて単語を多次元ベクトル（以下，ベクトル）で表現し，単語の類似性を判別する方法[1)~4)]は，文や文書等の単語からなる様々な情報をベクトル表現に変換して統一的に類似性を判別できる長所がある．実際に，情報検索[5),6)]やナレッジマネージメント支援[7)]，テキストセグメンテーション[8)]等，単語の意味を考慮した知的なテキスト処理に幅広く適用されている．

元となるテキストデータには，国語辞典あるいはテキストコーパス（以下，コーパス）が用いられ，データの種類に応じた単語のベクトル生成方法と類似性判別の方法が検討されている．国語辞典を用いた方法[3)]では，掲載されている基本的な単語に関する類似性判別を可能とするが，新語や専門語は対象とできない．一方，コーパスを用いた方法[1),2)]では，適切なコーパスを選択することにより新語や専門語の類似性判別が可能だが，コーパスに出現しない基本的な単語に関する類似性判別が行えないことが問題である．

† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation

上記の問題を解消するためには，2つのテキストデータより生成される単語のベクトルを統合的に利用することが必要である．しかし，データの種類に応じてまったく異なるベクトル空間が構成されるので，国語辞典に基づく単語のベクトルとコーパスに基づく単語のベクトルを直接比較することはできない．

そこで本稿では，国語辞典中に存在しないコーパス中の単語に関して，コーパスから作成したベクトル空間上のベクトルを，国語辞典から作成したベクトル空間上のベクトルに変換する手法を提案する．

2. 単語の類似性判別法

本提案手法の前提となる単語の類似性判別法について，国語辞典とコーパスを用いた方法をそれぞれ説明する．

2.1 国語辞典を用いた単語の類似性判別

国語辞典を用いた単語の類似性判別法[3)]では，国語辞典中の見出し語をそれぞれ多次元ベクトル（定義ベクトル）に対応させており，この定義ベクトルの集合を「概念ベース」と呼んでいる．各見出し語の定義ベクトルは，その定義文中の自立語の出現頻度をもとにした値を要素とする（**図1**）．実際には，再帰的参照，逆引き参照の線形結合や，シソーラスを用いた属性（次元）の一般化など，ベクトルの値を精錬する工夫

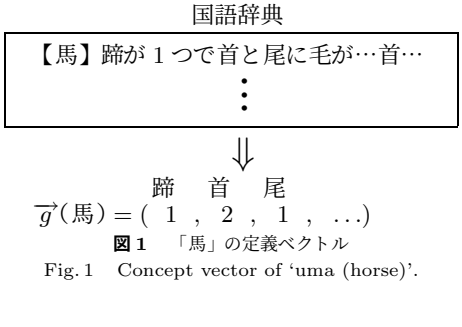

図 1 「馬」の定義ベクトル
Fig. 1 Concept vector of 'uma (horse)'.

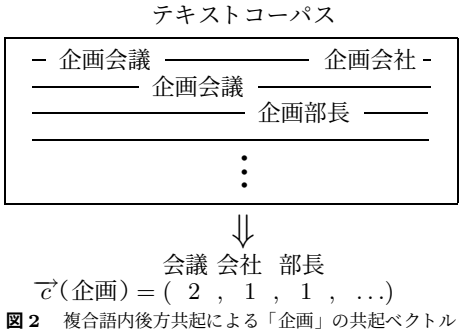

図 2 複合語内後方共起による「企画」の共起ベクトル
Fig. 2 Cooccurrence vector of 'kikaku (planning)'.

を行っている．詳しくは文献 3) を参照されたい．各単語に対応する定義ベクトルどうしのなす角の余弦を類似度として用いており，本稿ではこれを「定義類似度」と呼ぶ．

2.2 テキストコーパスを用いた単語の類似性判別

コーパスを用いた単語の類似性判別の研究[1),2)]においても，概念ベースと同様に，各単語を多次元ベクトルに対応させている．コーパスには国語辞典のような見出し語に対する定義文という対応関係はないので，国語辞典を用いた類似性判別法と同様の方法を用いることはできない．しかし，コーパス中で単語ごとに同時に現れる語（共起語）を抽出することで，各単語を，共起語の出現頻度をもとにした値を要素とするベクトル（共起ベクトル）で表すことができる．コーパスから作成した共起ベクトルの集合を「共起ベース」と呼ぶことにする．

共起語としては，コーパス中において，ある単語の周辺（一定範囲内）に出現する単語[2)]や，目的語に対する述語[1)]などがあげられる．また，2 語から成る複合語に対し，前方の単語に対する後方の単語を共起語とする方法（複合語内後方共起）やその逆（複合語内前方共起），あるいは同じ文内の名詞に対する動詞を共起語とする方法（文内名詞動詞共起）も提案されている[4)]．たとえば複合語内後方共起を用いた単語 "企画" の共起ベクトル $\overrightarrow{c}$（企画）は**図 2** のようになる．

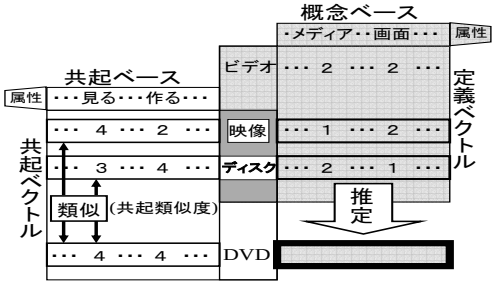

図 3 未知語の定義ベクトルの推定
Fig. 3 Estimating an unknown vector.

類似度の計算法はいくつか考えられるが，本稿では概念ベースと同様に，ベクトルどうしのなす角の余弦を用いる．本稿ではこれを「共起類似度」と呼ぶ．

3. 提案手法

本稿では，概念ベースに含まれない単語を**未知語**と呼ぶことにする．本稿の目的は，共起ベースに含まれる未知語の定義ベクトルを推定することである．未知語の定義ベクトルは，未知語に似た単語の定義ベクトルを利用することによってある程度推定できると考えられる．そのために，概念ベースと共起ベースに共通する語彙（以下，共通語彙）を利用する．まず未知語を含む共起ベースを利用して，未知語に対して共起類似度が高い語を共通語彙より検索する．そして，それらの単語の定義ベクトルを用いて，未知語の定義ベクトルを推定する．たとえば，未知語 "DVD" の定義ベクトルは，"DVD" に対して共起類似度が高い "映像" や "ディスク" の定義ベクトルを用いて推定する（**図 3**）．具体的には，共通語彙中で，未知語 k との共起類似度が大きな上位 m 語 $u_1, \ldots, u_m$（添字は共起類似度の大きさの順位）の定義ベクトルを足し合わせて，それを k の定義ベクトルと見なす．つまり，単語 u_i の定義ベクトルを $\overrightarrow{g}(u_i)$ とすると，本提案手法によって推定される未知語 k の定義ベクトル $\overrightarrow{g}'(k)$ は以下である．

$$\overrightarrow{g}'(k) = \sum_{i=1}^{m} \overrightarrow{g}(u_i) \qquad (1)$$

複数の定義ベクトルを足し合わせることによって，たとえ u_1 が k の類似語でなくても，u_2 や u_3 が類似語となっているならば，これらの定義ベクトルによりそれを補うことができると考えられる．

4. 実験

本提案手法の評価法とその実験結果を示す．

4.1 評 価 法

本提案手法は，元々存在しない定義ベクトルを推定するため，推定された定義ベクトル自体を直接評価するのは難しい．そこで，概念ベースに含まれる単語の定義ベクトルを推定し，元々の定義ベクトルと比較することによって評価する．評価の手順は以下のようになる．

(1) 評価に用いる単語 k を概念ベースと共起ベースの共通語彙の中から選び，k の定義ベクトル $\overrightarrow{g}(k)$ を概念ベースから除く．

(2) 提案手法により，k の定義ベクトル $\overrightarrow{g}'(k)$ を推定する．

(3) $\overrightarrow{g}'(k)$ と $\overrightarrow{g}(k)$ の定義類似度（余弦値）を評価値（最大値 1）とする．

$\overrightarrow{g}(k)$ は正解の定義ベクトルであるので，この評価値が大きいほど，より正解に近い定義ベクトルが推定されたことになる．実際には，評価用の単語として 100 語を共通語彙中からランダムに選択し，その評価値の平均を最終的な評価値とする．この評価法は被験者による評価が不要なため，大規模な自動評価実験が可能である．そこで，足し合わせる定義ベクトルの数 m を変化させて，定量的な評価を行った．

実験では，学研国語大辞典[9]から構築した概念ベースを用いた．約 9 万語に対する定義ベクトルが含まれ，ベクトルの次元数は約 3 千である．また，コーパスとしては CD–毎日新聞 95 版の経済欄を用い，複合語内後方共起，複合語内前方共起，文内名詞動詞共起[4]によって構築した 3 種類の共起ベースを利用した．共起ベースに含まれる単語数はそれぞれ 8 千語，6 千語，2 万 4 千語であり，概念ベースとの共通語彙数はそれぞれ 5 千語，4 千語，1 万 5 千語である．さらに，これら 3 種類の共起による共起類似度計算法を併用する方法も用いた．併用の方法としては，文献 4) にならい，それぞれの共起による共起類似度の平均を最終的な共起類似度とした．

4.2 実 験 結 果

実験結果を**図 4** に示す．横軸は未知語の定義ベクトルを推定するときに足し合わせる定義ベクトルの数 m であり，縦軸は評価値（定義類似度）の平均値である．

図 4 中の点線は，概念ベース中の全定義ベクトルの平均ベクトルを未知語の定義ベクトル $\overrightarrow{g}'(k)$ と見なしたときの評価値（0.21）である．これは，概念ベースのみから未知語の定義ベクトルを推定する手法の評価値と見なすことができ，共起ベースを用いて定義ベクトルを推定する本提案手法に対するベースラインとなるものである．いずれの共起類似度計算法においても，足し合わせる定義ベクトル数を増やすことで，その評価値がこのベースラインを越えており，本提案手法の有効性を示している．ただし，評価値の最大値 1 に対し，今回の評価値は 0.3 程度であり，今後，方式のさらなる改良が必要である．

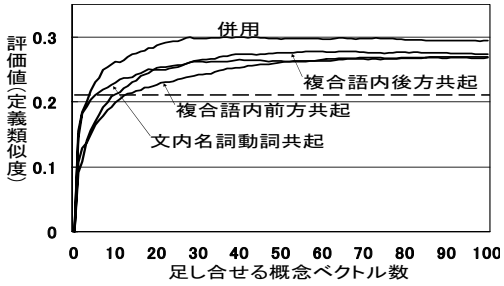

図 4 実験結果
Fig. 4 Result.

文献 4) では，本稿と同じ 4 種類の共起類似度計算法を用いた単語の類似性判別実験が行われている．これによると，3 つの共起を併用した共起類似度計算法（以下，併用法）が最も高い評価となっている．本稿の実験結果においても，併用法を利用した場合に最も高い評価となっており，共起を用いた単語の類似性判別の精度が本提案手法の精度に反映されていると考えられる．

また，いずれの共起類似度計算法においても，足し合わせる定義ベクトルの数が 1 のときは評価値が非常に低く，既存の定義ベクトル 1 つで未知語を表現するのは難しいことが分かる．足し合わせる数を増やしていくと評価値が上がっていき，あるところでピークを迎えて，その後序々に下がっていく傾向にある．たとえば併用法では，足し合わせる数が 30 あたりになるまで評価値が上がり，その後はゆるやかに下がっている．共起による類似性判別では，単語の定義ではなく，その周辺の語（共起語）を比較するために，共起類似度が大きい単語どうしは必ずしも類似語ではないが，関連が大きい単語である場合が多い．このような場合，足し合わせる数を増やすことによってベクトルの要素を補完し合い，適切な定義ベクトルに近づけていると考えられる．さらに足し合わせる数を増やすと，今度は関係のない単語が増えていくために，評価値が落ちると考えられる．

併用法を用い，足し合わせる数を 30 個として “委員長” の定義ベクトルを推定し，概念ベース中から類似度が高い単語 10 語を検索した．結果を**表 1** に示す．“委員長” に類似した単語が上位に検索されていることが分かる．

表 1 「委員長」に対する類似語検索結果
Table 1 Similar words retrieval for "Iincho".

順位	単語	定義類似度
1	代表	0.6143
2	教育長	0.5432
3	総督	0.5358
4	総裁	0.5152
5	団長	0.5144
6	局長	0.5113
7	総長	0.5045
8	長	0.4924
9	県庁	0.4919
10	国務長官	0.4914

もともと概念ベースに含まれる定義ベクトルは，値が 0 でない要素の数（非零要素数）が比較的少ない．たとえば，今回使用した概念ベースにおいて，非零要素数のベクトルごとの平均は約 50 であった．本提案手法は，未知語の定義ベクトルを推定する際に複数の定義ベクトルを足し合わせるため，その非零要素数は比較的大きくなる．したがって，本提案手法によって推定した定義ベクトルを用いた類似性判別では，実際には未知語と関連性が小さな単語に対しても類似すると判別されてしまう可能性がある．文献 3) では，定義ベクトルにおいて値が小さな要素を除去することによって類似性判別の精度が向上している．本提案手法においても同様の処理を施すことによって，精度が向上することが考えられる．

5. おわりに

単語の類似性判別法として，国語辞典中の定義文，テキストコーパス中の単語共起をそれぞれ用いる 2 種類の方法に対し，これらを統合する手法を検討した．評価実験において，提案手法の有効性を確認した．本提案手法により，ある分野のコーパスにのみ含まれるような専門用語を知らないユーザでも，自分が知っている国語辞典中の単語を使うことで，その分野に対する情報検索等の処理が可能となることが期待できる．

参考文献

1) Hindle, D.: Noun Classification from Predicate-Argument Structures, *Proc. ACL*, pp.268–275 (1990).
2) Schütze, H.: Dimensions of Meaning, *Proc. Supercomputing 92*, pp.787–796 (1992).
3) 笠原 要，松澤和光，石川 勉：国語辞書を利用した日常語の類似性判別，情報処理学会論文誌，Vol.38, No.7, pp.1272–1284 (1997).
4) 稲子希望，笠原 要，松澤和光：複合語内単語共起による名詞の類似性判別，情報処理学会論文誌，Vol.41, No.8, pp.2291–2298 (2000).
5) Schütze, H. and Pedersen, J.: Information retrieval based on word senses, *4th Annual Symp. on Document Analysis and Information Retrieval*, pp.161–175 (1995).
6) 熊本 睦，島田茂夫，加藤恒昭：概念ベースの情報検索への適用—概念ベースを用いた検索の特性評価，情報処理学会研究報告，SIG-ICS 115, pp.9–16 (1999).
7) 加藤恒昭，笠原 要，北 寿郎：概念検索に基づく技術内容からのエキスパートの推定，電子情報通信学会技術研究報告，NLC2000-8, pp.55–62 (2000).
8) 別所克人：単語の概念ベクトルを用いたテキストセグメンテーション，情報処理学会論文誌，Vol.42, No.11, pp.2650–2662 (2001).
9) 金田一春彦，池田弥三朗：学研国語大辞典第二版，学習研究社 (1988).

(平成 14 年 4 月 5 日受付)
(平成 14 年 9 月 5 日採録)

稲子 希望（正会員）

昭和 48 年生．平成 10 年九州大学大学院システム情報科学研究科情報理学専攻修士課程修了．同年日本電信電話（株）入社．大規模知識ベースの研究に従事．現在，NTT コミュニケーション科学基礎研究所所員．

笠原 要（正会員）

昭和 39 年生．平成 3 年東京工業大学総合理工学研究科電子化学専攻修士課程修了．同年日本電信電話（株）入社．知識処理技術，特に大規模知識ベースの研究に従事．現在，コミュニケーション科学基礎研究所研究主任．平成 10 年 11 月より平成 11 年 11 月までスタンフォード大学 CSLI 滞在．平成 10 年人工知能学会奨励賞受賞．