Vol.2013-MPS-95 No.14
2013/9/26

格助詞によるクラスタリングを用いた分布類似度計算の高速化

中山 光樹[1]　山田 節夫[2]　西野 哲朗[1]

概要：本論文では，分布類似度計算の前処理として，文脈中に含まれる情報を用いて単語のクラスタリングを行うことで，分布類似度の計算を高速化する手法を提案する．格助詞と動詞を組とした文脈に対して，格助詞の種類に応じてクラスタを構築し，構築したクラスタ内で分布類似度の計算を行う．コーパスを用いて，実験と人手による評価実験を行うことで，従来手法と比べて精度が同等で高速であることを示す．

キーワード：分布類似度，類義語，クラスタリング，高速化

Efficient Distributional Similarity Calculation by Using Case Particle's Clusters

Nakayama Hiroki[1]　Setsuo Yamada[2]　Tetsuro Nishino[1]

Abstract: This paper proposes an efficient method for distributional similarity calculation by using case particle's clusters. The proposed method constructs clusters according to the case particles in the context before calculating distributional similarity. Experimental results show that the proposed method is more efficient than the existing method with almost same quality.

Keywords: distributional similarity, synonym, clustering, efficient method

1. はじめに

自然言語処理において，語の類似度計算は重要なタスクであり，語義曖昧性解消や類義語自動獲得，シソーラス自動構築などの幅広い応用がある．その際，語の類似度を測る方法として「分布類似度」が提案されている [1]．これは，語の持つ文脈の情報を利用して求める類似度であり「類似した文脈を持つ語は意味も類似している可能性が高い」という「分布仮説」に基づいた類似度である．

単語の類似度を測る方法には人手で構築されたシソーラスを利用する方法もある．しかし，人手で構築されたシソーラスは低カバレージであることや一貫性を保つことが難しいといった問題があるので，コーパスから計算される分布類似度は有用である．

これまでに様々な分布類似度の計算方法が提案され，その精度の評価が行われてきた [2][3][4][5]．Curran は使用するコーパスや Weight 関数，Measure 関数の組み合わせを変えて有効なものを示した [3]．相澤は日本語の新聞記事やウェブコーパスを利用して，文脈中の格情報の有無やいくつかの類似度尺度やフィルタリング，サンプリングの効果を示した [2]．柴田らは，超大規模コーパスを用いて分布類似度計算を行った．コーパスサイズを大きくするにつれて，精度が向上することを示した [4]．萩原らは，分布類似度計算のタスクにおける計算量の多さを問題視し，類似度計算を行うとき，本当に重要な素性は限られていることを示した [5]．しかし，分布類似度で精度を高めるためには，大量の共起データが必要となり，類似度計算の計算量が非常に多くなるという問題があった．これでは日々メンテナンスが必要な状況，たとえば新語を辞書に追加したい場合

[1] 電気通信大学
University of Electro Communications
[2] 日本電信電話株式会社
Nippon Telegraph and Telephone Corporation

に分布類似度を使いにくい．これらの計算では，語の間の分布類似度を計算する際に，類似性のない語同士で類似度を計算しているため無駄な時間がかかっている．そこで，我々はあらかじめある程度意味の近い語同士をまとめて計算すれば，効率的に分布類似度を計算できると考えた．

具体的には，文脈中に含まれる格助詞を用いたクラスタリングを行うことで，分布類似度の計算を精度を低下させずに高速化する．コーパスを用いて，実験と人手による評価を行うことで，従来手法と比べて精度が同等で高速であることを示す．

本論文の構成は以下の通りである．まず，2 節で分布類似度とその計算方法について述べる．次に 3 節では，提案手法である格助詞によるクラスタリング方法を説明する．さらに 4 節でコーパスを用いた実験結果により，提案手法が分布類似度の精度の低下を抑つつ高速化することを示し，最後に 5 節でまとめる．

2. 分布類似度

語の類似度計算の方法の 1 つとして分布類似度がある．分布類似度は，分布仮説と呼ばれる「文脈の似た語は意味も似ている可能性が高い」という考え方に基づき計算される単語間の類似度のことである [1]．すなわち，分布類似度の計算は，それぞれの語が持つ文脈がどれだけ類似しているかという指標から，語の類似度を計算する手法である．

2.1 分布類似度計算

分布類似度の計算は，単語の出現頻度を重みを表すベクトルに変換する Weight 関数と 2 つのベクトル間の類似度を計算する Measure 関数により計算される [3]．本論文では，文献 [4] の検証の結果，最も精度の高かった Weight 関数と Measure 関数の組み合わせである，相互情報量と Simpson-Jaccard 係数の組み合わせを用いた．以下では，2.1.1 節で相互情報量，2.1.2 節で Simpson-Jaccard 係数について説明する．

2.1.1 相互情報量

相互情報量は，ある 2 語の関連度を測る指標である．語がよく共起しているほど関連度が高くなり，相互情報量の値も大きくなる．相互情報量 $MI(w,v)$ の計算式は (1) 式のように表される [4]．

$$MI(w,v) = \log \frac{P(w,v)}{P(w)P(v)} \tag{1}$$

ここで $P(w)$，$P(v)$ はそれぞれ語 w，v のコーパス中での出現確率，$P(w,v)$ は語 w と v がともに出現する確率を示す．MI がしきい値 β 以上の共起語のみを計算に使用し，それ以外は削除する．

2.1.2 Simpson-Jaccard 係数

2 つの単語 w_1, w_2 に対する共起語の集合をそれぞれ V_1, V_2 とすると，Simpson-Jaccard 係数は (2) 式のように表される [4].

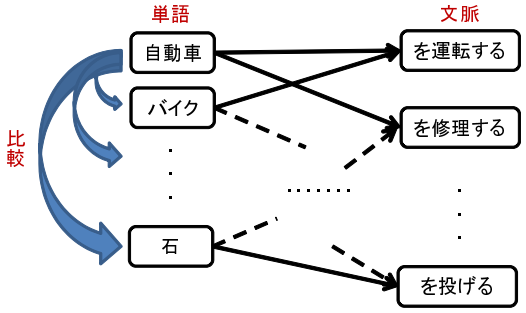

図 1 計算方法
Fig. 1 Computation method.

$$sim = \frac{Jaccard(w_1,w_2) + Simpson(w_1,w_2)}{2} \tag{2}$$

ここで，$Jaccard(w_1,w_2)$ と $Simpson(w_1,w_2)$ はそれぞれ (3) 式と (4) 式とする．

$$Jaccard(w_1,w_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} \tag{3}$$

$$Simpson(w_1,w_2) = \frac{|V_1 \cap V_2|}{\min(|V_1|,|V_2|)} \tag{4}$$

3. 格助詞によるクラスタリング

本論文では，ある程度意味の近い語同士でクラスタを構築し，そのクラスタの中だけで類似度を計算する．ここで，共起語としてはある単語に対する係り受け関係を利用する．また，以下では係り受け関係で修飾される方の語を文脈と呼ぶことにする．

3.1 従来手法の問題点

従来手法では大量の共起データを用いたときに，分布類似度計算の計算量が非常に多くなるという問題があった．原因としては，従来手法では n 個の単語があった場合，分布類似度の計算時に $\frac{n(n-1)}{2}$ 回の比較を行わなければならなかったことが挙げられる．しかし，この比較の中には類似していない単語同士の無駄な比較が多く含まれている．この無駄な比較を減らせば計算量を減らすことができると考えられる．たとえば**図 1** のような単語と文脈があるときに，「自動車」と「石」や「バイク」と「石」は類義語ではないので，類似度計算時にこれらの単語の比較をしないようにすることで計算量を減らす．

3.2 格助詞によるクラスタリング

本論文では，単語をあらかじめ意味の近い語にクラスタリングし，意味の近い語が集まったクラスタの中で類似度計算を行うことで，計算量を減らして分布類似度計算を高速化する手法を提案する．このクラスタを構築するために，文脈中に含まれる格助詞を用いることで単語のクラスタリングを行う．ここで，文脈は「格助詞+動詞」とした．

■『自動車』の文脈例
「を運転する」「を修理する」「に乗る」「が跳ね飛ばす」「で訪問する」
■『自転車』の文脈例
「を運転する」「を修理する」「に乗る」「がカーブする」「で行ける」
■『石』の文脈例
「を投げる」「を蹴る」「で遊ぶ」「を並べる」「に触れる」
■『文法』の文脈例
「を暗記する」「に詳しい」「と格闘する」「からやり直す」「から学ぶ」

図 2 単語と文脈の例
Fig. 2 Example of words and contexts.

表 1 文脈中の格助詞の種類
Table 1 Type of case particles in the context.

単語	文脈中に含まれる格助詞
自動車	「を，に，が，で」
自転車	「を，に，が，で」
石	「を，に，で」
文法	「を，に，と，から」

「格助詞+動詞」を文脈とした理由は，文献 [2] によると格の情報を用いる場合と用いない場合では用いた場合の方が性能が高くなるからである．また格助詞によるクラスタリングを行うのは，「文脈中に含まれる格助詞の種類が異なる単語同士は類義語ではない可能性が高い」という仮説に基づく．

格助詞によるクラスタリングについて例を挙げて説明する．**図 2** に単語と文脈の例を挙げる．ここで単語は『自動車』『自転車』『石』『文法』であり，それ以外は文脈である．図 2 の文脈に含まれる格助詞に注目すると，**表 1** のようになる．ここで表 1 は単語とその文脈中に含まれる格助詞を示している．たとえば『自動車』という単語の場合，その文脈中には「を，に，が，で」の 4 つの格助詞が含まれていたということである．表 1 より，『自動車』と『自転車』の文脈中には同一の格助詞が含まれていることが確認できる．

表 1 のような場合，**表 2** のようなクラスタに分けることができる．ここで表 2 は，クラスタとそのクラスタに含まれる単語について示している．たとえば「を，に，が，で」という格助詞のクラスタの場合は，そのクラスタの中には『自動車』と『自転車』という単語が含まれるということである．このクラスタ内で類似度を計算すれば，「自動車」と「石」等の類似していない単語同士の類似度を計算することがなくなるため，計算量を減らすことができる．

3.3 処理の流れ

図 3 で提案手法における分布類似度計算の処理の流れを説明する．まず，係り受け関係をコーパスから抽出する．抽出した係り受け関係は出現頻度で共起ベクトルにする．共起ベクトルに対して，相互情報量で重み付けを行いしきい値 β 以下の文脈を削除する．これにより，関連度の低い文脈が削除される．従来手法ではこの後に Simpson-Jaccard 係数で類似度計算を行うが，提案手法では図 3 で示すように，類似度計算の前にクラスタリングを行う．クラスタリングをした後，各クラスタ内で Simpson-Jaccard 係数による類似度計算を行う．次に，類似度の値に基づいて単語のランク付けを行う．そしてランクの高い順に単語対を抽出する．

表 2 クラスタリングの例
Table 2 Example of clustering.

クラスタ	クラスタに含まれる単語
「を，に，が，で」	自動車，自転車
「を，に，で」	石
「を，に，と，から」	文法

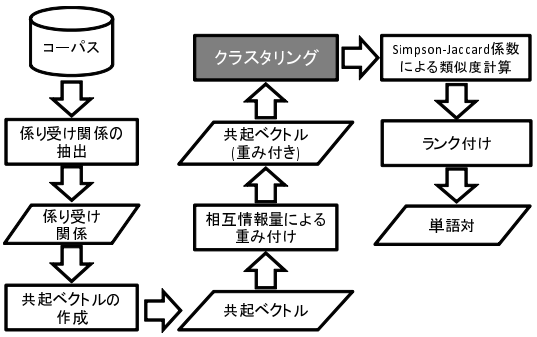

図 3 提案手法の処理の流れ
Fig. 3 Flow of the proposed method.

4. 実験

4.1 実験環境

実験に用いたコーパスは，日本語係り受けコーパス（JDC）の NCV 形式 [6] である．このコーパスは，日本語ウェブコーパス 2010[7] より係り受け解析器 CaboCha[8] を用いて，格助詞を介した語と語の係り受け関係を抽出したものである．NCV の N は名詞 (Noun)，C は格助詞 (Case particle)，V は動詞 (Verb) を表している．**図 4** に NCV 形式の例を挙げる．NCV 形式では，1 行につき「名詞　格助詞　動詞　出現頻度」の情報が書かれている．たとえば，図 4 の 1 行目「間隔　で　ブロードキャストする　6」は名詞が「間隔」，格助詞が「で」，動詞が「ブロードキャストする」，出現頻度が 6 であることを示す．以降では，日本語係り受けコーパスの 1 行を 1 ペアと呼ぶ．

4.2 クラスタリングに用いる格助詞

文献 [9] によると，格助詞は「を，に，が，で，と，から，の，へ，より，や」の 10 種類である．ここで，日本語

間隔	で ブロードキャストする	6
悪口	を 言い放つ	81
休み	と 書く	657
階段	が 構える	12
潔白	と 言える	37
代わり	に 開け放つ	16
書込み	を みつける	7
動物	に 譲る	10
...		

図 4 日本語係り受けコーパスの例
Fig. 4 Example of the JDC.

表 3 文脈中の格助詞の割合
Table 3 Percentage of case particles.

格助詞	割合 [%]
を	26.9
に	22.0
が	20.6
で	16.4
と	5.38
から	5.27
の	1.74
へ	0.911
より	0.835

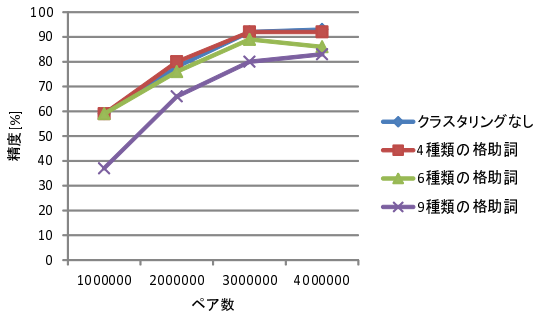

図 5 精度とペア数
Fig. 5 Precision and the number of the pairs.

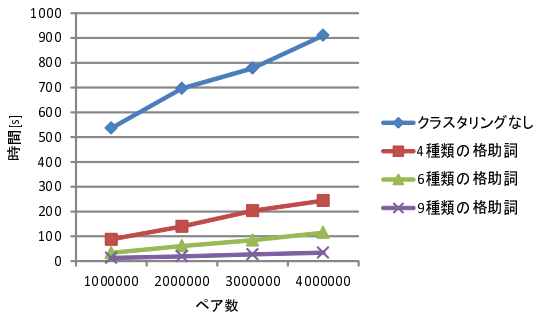

図 6 実行時間とペア数
Fig. 6 Execution time and the number of the pairs.

係り受けコーパスの NCV 形式を用いて格助詞の割合を調査した．結果は**表 3** の通りである．表 3 は，格助詞とその出現割合を示している．格助詞「や」が出現しないのは，「や」が並列の格助詞だからだと考えられる．つまり，「や」は名詞と名詞を結びつける格助詞なので，「名詞+格助詞+動詞」の NCV 形式では出現しなかった．

本論文では，表 3 の調査結果より，出現割合が 10%以上の 4 種類の格助詞「を，に，が，で」と，5%以上の 6 種類の格助詞「を，に，が，で，と，から」と，これに「の，へ，より」を加えた 9 種類の格助詞でクラスタリングを行った．

4.3 実験内容

ペア数はそれぞれ 1,000,000，2,000,000，3,000,000，4,000,000 で実験を行った．また，相互情報量のしきい値は β=0.0〜3.0 まで 0.5 刻みで設定した結果，最も性能が良かった β=2.5 を使用した．実験内容としては，以下の (1)-(4) の 4 種類のクラスタリングを設定し，「格助詞+動詞」を文脈として，類似度が上位 100 位までの単語対の抽出をした．

(1) クラスタリングなし（従来手法）
(2) 4 種類の格助詞「を，に，が，で」でクラスタリング
(3) 6 種類の格助詞「を，に，が，で，と，から」でクラスタリング
(4) 9 種類の格助詞「を，に，が，で，と，から，の，へ，より」でクラスタリング

4.4 評価方法

抽出した単語対は，精度と実行時間で評価する．精度は，以下のように計算する．

精度=抽出した類義語数/抽出した単語対の数

また文献 [10] にならい，人手により以下の 3 項目に該当した場合に類義語と判定した．

(1) ほとんど同じ意味
(2) 同じカテゴリに属する
(3) 上位下位の関係

(1) の例としては『道端』と『道ばた』や『急激』と『急速』などが挙げられる．(2) の例としては『七月』と『八月』や『特急』と『快速』などが挙げられる．(3) の例としては『弦楽器』と『ヴァイオリン』や『製法』と『蒸留』などが挙げられる．Weblio 類語辞典 [11] に載っている単語も (1) の中に入れて類義語とした．

4.5 実験結果

図 5 に精度の結果，**図 6** に実行時間の結果を示す．図 5 では，横軸がペア数，縦軸が抽出した単語対中の類義語の精度を表している．図 6 では，横軸がペア数，縦軸が実行時間を表している．図 5 より，精度の結果は，従来手法と比べてクラスタリングに用いる格助詞の数が増えると低下

する傾向にあるが，クラスタリングに用いる格助詞の数が少なければ低下を抑えられていることがわかる．図 6 より，実行時間の結果は従来手法と比べて提案手法では，4 種類の格助詞でクラスタリングを行ったときに約 4 倍，6 種類の格助詞で約 8 倍，9 種類の格助詞で約 27 倍ほど計算を高速化することができている．以上より，精度の低下を抑えつつ計算を高速化できることがわかった．

4.6 考察

図 6 より，文脈に含まれる格助詞を用いてクラスタリングすることで，実行時間を短くできることがわかった．1 つのクラスタに多くの語が集中せず，複数のクラスタへの語の分散がある程度成功していると言える．

図 5 より，クラスタリングに用いる格助詞の種類を増やすと，精度は低下する傾向にあることがわかった．これは，クラスタリングしたことにより，本来は類義語となるべき語が別々のクラスタに分類されてしまい，類似度の計算が行われなかったためだと考えられる．

クラスタリングなしでは「月曜日」と「火曜日」,「六月」と「十一月」といった単語が抽出されたが，4 種類の格助詞でクラスタリングしたときにはこれらの単語の抽出ができなかった．理由としてはこれらの単語が別々のクラスタに分けられてしまったからである．別々のクラスタに分けられた原因としては，そもそも用いたコーパスに誤った係り受け関係が含まれるので，類義語でも同じクラスタにならないということが考えられる．これに対しては，係り受け解析の精度が低下するウェブコーパスを用いるのではなく，新聞記事のコーパスを用いるという対策が考えられる．

5. おわりに

従来手法では，大量の共起データを使うと類似度計算の計算量が多くなるという問題があった．この問題を解決するため，本論文では格助詞を用いたクラスタリングによる分布類似度計算を提案した．コーパスを用いて，従来手法とクラスタリングを行う提案手法で類似度上位 100 までの単語の抽出をしたところ，提案手法では精度の低下を抑えながら高速に類義語の抽出ができることを確認した．以上より，分布類似度計算を行う際に提案手法が有効であることを示せた．今回の実験では出現頻度の多い順に 4，6，9 種類の格助詞をクラスタリングに用いたが，今後の課題としては，さらに精度向上を目指してクラスタリングに用いる格助詞の種類を変えたときの精度の検証が挙げられる．

参考文献

[1] J.R.Firth: Studies in Linguistic Analysis, *ChapterA synopsis of linguistic theory* (1957).
[2] 相澤彰子：大規模テキストコーパスを用いた語の類似度計算に関する考察，情報処理学会論文誌， Vol. 49, No. 3, pp. 1426–1436 (2008).
[3] Curran, J.: From Distributional to Semantic Similarity, PhD Thesis, University of Edinburgh. College of Science (2004).
[4] 柴田知秀，黒橋禎夫：超大規模ウェブコーパスを用いた分布類似度計算，言語処理学会年次大会, D4-7, pp. 705–708 (2009).
[5] 萩原正人，小川泰弘，外山勝彦：分布類似度のための文脈素性選択，言語処理学会 NLP 若手の会第 2 回シンポジウム, 発表， Vol. 11 (2007).
[6] 林部祐太：日本語係り受けコーパス，（オンライン），入手先 ⟨http://hayashibe.jp/jdc/⟩ （参照 2012-11-3）.
[7] 矢田晋：日本語ウェブコーパス 2010，（オンライン），入手先 ⟨http://s-yata.jp/corpus/nwc2010/⟩ （参照 2013-8-24）.
[8] 工藤拓，松本裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌， Vol. 43, No. 6, pp. 1834–1842 (2002).
[9] 田近洵一：くわしい国文法，文英堂，1nd edition (2012).
[10] 中渡瀬秀一：2 部グラフ構造を用いた類義語の抽出，情報知識学会研究報告会講演論文集， No. 10, pp. 29–34 (2002).
[11] weblio: Weblio 類語辞典，（オンライン），入手先 ⟨http://thesaurus.weblio.jp/⟩ （参照 2013-8-20）.

付　　録

ペア数 4,000,000 としたときのクラスタリングなしの従来手法と 4 種類の格助詞でクラスタリングした結果得られた単語対上位 50 位を**表 A·1** と**表 A·2** に示す．単語 a と単語 b が得られた単語対である．評価の欄の数字はそれぞれ，1 がほとんど同じ意味，2 が同じカテゴリに属する，3 が上位下位の関係，4 が単語 a と単語 b に関係がないことを示している．

表 A·1 クラスタリングなし
Table A·1 Result of extraction by non clusters.

順位	単語 a	単語 b	評価
1	急激	急速	1
2	挙句	挙げ句	1
3	道端	道ばた	1
4	太もも	太股	1
5	目途	目処	1
6	日差し	陽射し	1
7	七月	八月	2
8	車椅子	車いす	1
9	出だし	初っ端	1
10	咆哮	雄叫び	1
11	月曜日	火曜日	2
12	世紀	四半世紀	2
13	精子	精液	1
14	中隊	師団	2
15	六月	十一月	2
16	見た目	見かけ	1
17	攻勢	猛攻	2
18	精子	卵子	2
19	緑色	白色	2
20	七月	三月	2
21	特急	快速	2
22	中指	親指	2
23	夕方	昼過ぎ	2
24	面持ち	顔つき	1
25	小刀	包丁	1
26	航空	汽船	2
27	戦艦	巡洋艦	3
28	悪行	悪事	1
29	中隊	戦隊	1
30	一月	五月	2
31	一月	十月	2
32	演習	実習	2
33	兵士	兵隊	1
34	十月	五月	2
35	長期	永年	1
36	絶叫	雄叫び	1
37	太もも	太腿	1
38	先住民	原住民	1
39	面持ち	無表情	3
40	長期	長期間	1
41	強風	暴風	2
42	高温	低温	2
43	師団	戦隊	2
44	長年	多年	1
45	戦艦	戦隊	2
46	四月	十月	2
47	製法	蒸留	3
48	大使	公使	1
49	漁船	小船	2
50	緑色	黄色	2

表 A·2 4 種類の格助詞でクラスタリング
Table A·2 Result of extraction by four case particles.

順位	単語 a	単語 b	評価
1	急激	急速	1
2	挙句	挙げ句	1
3	道端	道ばた	1
4	太もも	太股	1
5	目途	目処	1
6	日差し	陽射し	1
7	七月	八月	2
8	車椅子	車いす	1
9	咆哮	雄叫び	1
10	精子	精液	1
11	中隊	師団	2
12	見た目	見かけ	1
13	精子	卵子	2
14	緑色	白色	2
15	特急	快速	2
16	中指	親指	2
17	夕方	昼過ぎ	2
18	小刀	包丁	1
19	航空	汽船	2
20	戦艦	巡洋艦	3
21	中隊	戦隊	1
22	兵士	兵隊	1
23	絶叫	雄叫び	1
24	太もも	太腿	1
25	先住民	原住民	1
26	強風	暴風	2
27	師団	戦隊	2
28	戦艦	戦隊	2
29	漁船	小船	2
30	緑色	黄色	2
31	咆哮	絶叫	1
32	見地	観点	1
33	天ぷら	刺身	2
34	爆音	絶叫	4
35	南方	北方	2
36	電子	原子	2
37	底面	下面	1
38	艦隊	大隊	2
39	囚人	捕虜	1
40	神社	稲荷	3
41	議会	国会	1
42	禅師	上人	2
43	足元	足下	1
44	八月	十二月	2
45	中期	末期	2
46	砂漠	荒野	2
47	連隊	小隊	2
48	日の出	雪解け	4
49	突き	一撃	2
50	粘液	精液	2