共起要素のクラスタリングを用いた分布類似度計算

大平真一, 山本和英
長岡技術科学大学 電気系
E-mail:{ohira,yamamoto}@jnlp.org

1　はじめに

単語の類似度は自然言語処理において広く利用されている語彙知識であり、多義性の解消をはじめとした応用例が知られている。単語の類似度を求める手法として、人手によって構築されたシソーラスを用いるものとコーパスを用いるものが挙げられる。コーパスを用いた手法として「意味的に近い語は類似した文脈を持つ」という分布仮説に基づく分布類似度の有効性が示されている [1]。

分布類似度計算に用いられる素性として、テキスト内の共起語や係り先の語などがある [2]。従来研究ではノイズとなる素性の除去や、素性の重み付けなどが試みられてきた [3][4][5]。しかし、2 単語間で一致する素性のみを使用していたため、人名や地名など特定の単語に固有な素性が有効に活用されていなかった。

そこで本研究では、従来法において活用されていなかった素性に注目し、共起要素のクラスタリングを行うことで使われ方の近い共起要素をまとめ、その結果を用いて分布類似度計算を行う手法を提案する。

2　関連研究

分布類似度計算における単語と素性の関係に着目した研究はいくつか存在する。

相澤 [2] は単語・素性の出現頻度と類似度計算との関係から、精度向上の為に 2 つの手法を提案した。一つは語と素性の出現頻度を用いてノイズとなる素性を取り除くフィルタリング法であり、もう一つは単語ごとに異なる素性の数に上限値を設定することで偏りを低減するサンプリング法である。本研究ではクラスタリングの効果の一つとして素性の数の偏りが低減することを想定しているため、フィルタリング法のみを採用した。

柴田ら [1] は超大規模コーパスを用いた分布類似度計算を行い、コーパスの大規模化による精度の向上を示した。コーパスサイズの拡大により素性が増加し、ノイズとそうでないものが明確になったことや、類似度計算に十分な素性の数が確保できるようになった点などが要因として考えられる。本研究では、クラスタリングによって従来手法で使われていなかった素性を活用することで、類似度計算に十分な素性の数を確保できると考えた。

萩原ら [4] は分布類似度計算における計算量の削減を目的として、素性選択について多くの研究が存在する文書分類の分野における素性選択手法を適用した。精度を保ちながら多くの素性を削減しているため、我々は分布類似度計算に有効な素性は少数であると考えた。そこで、本研究では素性削減の過程でノイズとして除外される素性について注目した。

3　提案手法

本手法では以下の流れで分布類似度の計算を行う。

1. 共起ベクトルの作成
2. 共起要素のクラスタリング
3. 分布類似度計算
   (a) Weight 関数によるノイズ低減
   (b) Measure 関数による類似度計算

3.1　共起ベクトルの作成

単語の素性となる共起ベクトルの作成には柴田ら [1] の手法を用いた。単語の共起要素を係り先とし、その集合を共起ベクトルとする。単語 w と係り先の語 w'、それぞれをつなぐ格要素 r を三つ組 (w, r, w') として収集する。また、r と w' のペアを共起要素 v とした。三つ組における格要素 r には以下のものを用いた。

が, を, に, から, と, へ, まで, より, の

共起要素の収集時に行うノイズ除去と複合名詞の扱いについては朝倉ら [3] の手法を用いた。コーパス中に 1 度しか出現しない三つ組はノイズとして除外した。単語 w は複合名詞を含み、IPA 品詞体系辞書 (1) において以下の品詞が連続した単語 w を複合名詞とした。

名詞-サ変接続, 名詞-一般, 名詞-固有名詞, 名詞-接尾, 接頭詞, 未知語, 記号-アルファベット

w を「フランス」とした時の r, w' の例を示す。

「フランス」の共起ベクトル

の:大統領 (24), へ:行く (8), の:マルセイユ (6), の:コニャック (5), の:エッフェル塔 (4), ···

「の:大統領」などが共起要素、括弧内の数字は共起頻度を示している。

係り受け解析には構文解析器 CaboCha(2) を用いた。

3.2 共起要素のクラスタリング

「フランス」と「の:マルセイユ」のように、人名や地名などの共起要素は特定の単語に固有である場合が多いため、単語の特徴をよく表す共起要素であるといえる。例として国についての共起要素を挙げると、著名人や首都など国の特徴を強く表すものが多く存在している。しかし、共起要素の一致度によって類似度を計算していた従来手法では、「の:大統領」と「の:首相」など等価な立場を表すことのある共起要素でも、表現が一致しないことで類似度を低下させる要因となっていた。

また、人名や地名などの固有名詞でない共起要素についても、「に:置き換え」と「に:置換」のように同じ現象を表す共起要素が複数存在する場合、表現が分散することで有用な共起要素がノイズとして除外されてしまうことがあった。

共起要素のクラスタリングには本研究での分布類似度計算に用いる共起ベクトルの作成手法を適用し、(w, v) の形で収集した共起要素 v に対する単語 w の集合と共起頻度のデータを用いた。

例として、共起要素である「の:銀閣寺」のクラスタリングに用いる素性を示す。

京都 (5), 京都市左京区 (3), 東山 (4)

「の:銀閣寺」の素性には地名や文化に関するものが並ぶため、歴史や建物に関する共起要素と同じクラスタに所属すると考えられる。クラスタリングによって「の:銀閣寺」が関連性のある建物などとまとめられることで、建物が存在する地名に共通の共起要素が増える。その結果、従来手法と比べ類似度が上昇すべき単語に関して精度が上がるため、提案手法は有効であると考える。

共起要素のクラスタリングにはデータクラスタリングツール bayon[3] を用いた。bayon は使用する際にクラスタリング手法の選択を行うことができるが、計算の速度と精度に優れていることから Repeated Bisection 法を採用した。

3.3 分布類似度計算

本研究では、分布類似度計算の関数として柴田らの行った比較実験で最も高い精度を示した Weight 関数と Measure 関数の組み合わせを用いた。Weight 関数は頻度から重みを計算する関数であり、Measure 関数はベクトル間の類似度を計算する関数である。

3.3.1 Weight 関数

Weight 関数には単語 w と共起要素 v の相互情報量 MI が閾値 β より大きい場合に 1、β 以下の場合 0 となる関数 P_β を用いた。P_β の式を以下に示す。

$$P_\beta = 1 \text{ if } MI > \beta : \text{ otherwise } 0$$

MI の式を以下に示す。ここで、$freq(w)$ を w の出現頻度、$freq(w, v)$ を w と v の共起頻度とする。

$$MI(w, v) = \log \frac{freq(w, v)}{freq(w), freq(v)}$$

3.3.2 Measure 関数

Measure 関数には単語 w_1 と w_2 の共起要素集合の重なりを見る $Jaccard$ 係数と $Simpson$ 係数を相加平均した $sim(w_1, w_2)$ を用いた。$sim(w_1, w_2)$ の式を以下に示す。

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

$Jaccard$ 係数と $Simpson$ 係数を以下に示す。ここで、単語 w_1 の共起要素の集合を V_1、単語 w_2 の共起要素の集合を V_2 とする。

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

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

4 実験

4.1 実験の条件

本実験ではコーパスに「日本経済新聞全記事データーベース 1990-2004 年度版」[4] を使用した。コーパスから (w, v) のペアを 3,500,150 件収集し、145,057 個の単語 w と 158,057 個の共起要素 v を獲得した。

共起要素のクラスタリングの際に設定するクラスタ数と分布類似度計算の結果との関係を見るため、クラスタ数は 12,500、25,000、50,000、75,000、100,000、125,000 の場合について実験を行った。クラスタリングに用いる共起要素や単語のデータには出現頻度に大きな偏りが存在する。そのため、bayon の idf オプションを用いて共起要素のクラスタリングに用いる素性を共起要素 v に対する単語 w の集合と w の idf とした。

4.2 評価セット

朝倉ら [3] の評価セットを用いて評価を行った。この評価セットでは、類義語と非類義語だけでなく、分類語彙表 [5] の階層に応じて類似度に「強」や「中」などの段階を設定しているため、類義語と非類義語のみの評価セットと比べ難易度が高く、詳細の分析が可能となる。

以下にそれぞれの類義語ペアの例を示す。

強類義語ペア ⇒ フランス:ドイツ
中類義語ペア ⇒ フランス:欧州
弱類義語ペア ⇒ フランス:日本人
非類義語ペア ⇒ フランス:建物

本実験では 4 種類の類義語ペアの数の偏りを解消するため、ランダムに 800 ペアずつ抽出したものを使用した。類似度の異なる 2 つの類義語セットを組み合わせた、「強＋中」セット、「中＋弱」セット、「弱＋非」セットの 3 パターンについて 2 値分類を行い、上位 800 ペアを類似度の高いセット、下位 800 ペアを類似度の低いセットとして 2 値分類の誤り数を求め評価を行った。

4.3 Weight 関数の閾値

ノイズ低減に用いる Weight 関数には、相互情報量 MI が閾値 β 以下の共起要素をノイズとして除外する関数を用いる。閾値 β の値は 2 値分類を行うセットの類似度と関係する。それぞれの評価セットに用いる β の値を決定するため、クラスタリングを行っていない共起ベクトルを用いて 2 値分類の誤り数と閾値の関係についての実験を行った。

各評価セットにおいて、最も誤り数の少なくなる場合の β の値は、「強＋中」セットと「中＋弱」セットで 1.8、「弱＋非」セットで 0.1 であった。

評価セットごとに最適な β の値が異なっているが、クラスタリングを行ったことによる最適な β の変化は 0.1 程度であった。

4.4 実験結果

提案手法との比較に用いるベースラインは柴田らの手法とする。Weight 関数によるノイズ低減の後、Measure 関数を用いて類似度計算を行った。提案手法では収集した共起要素をクラスタリングし、その結果を用いて柴田らの手法と同様にノイズ低減と類似度計算を行った。

ベースラインでの誤り数、提案手法において誤り数が最小となった場合の結果およびそのときのクラスタ数を表 1 に示す。

表 1: 各手法における判定誤り数

評価セット	柴田らの手法	提案手法	クラスタ数
「強＋中」	582	**554**	**50,000**
「中＋弱」	440	**418**	**50,000**
「弱＋非」	192	**190**	**75,000**

実験の結果、各評価セットにおける提案手法での誤り数は「強＋中」セットで 554(-28)、「中＋弱」セットで 418(-22)、「弱＋非」セットで 190(-2) となった。括弧内の数値はベースラインから変化した数である。全てのセットにおいて判定誤りの減少を確認した。

提案手法を用いた場合、評価セットごとに最適なクラスタ数が異なっていた。「強＋中」セットと「中＋弱」セットの場合では 50,000 クラスタ、「弱＋非」セットの場合は 75,000 クラスタのときに最も誤り数が減少した。

図 1、図 2 および図 3 に各評価セットでのクラスタ数と誤り数に関する実験結果を示す。それぞれの評価セットで傾向が異なることがわかる。

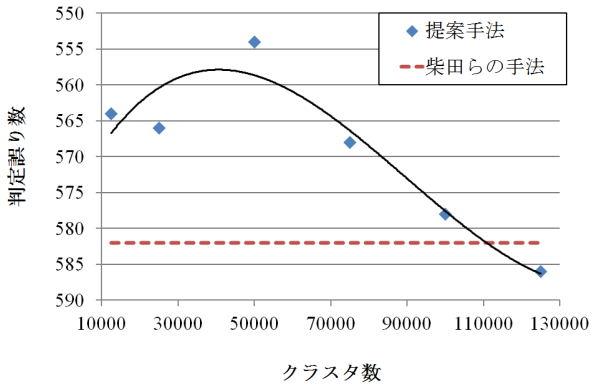

図 1: クラスタ数と判定誤りの関係：「強＋中」セット

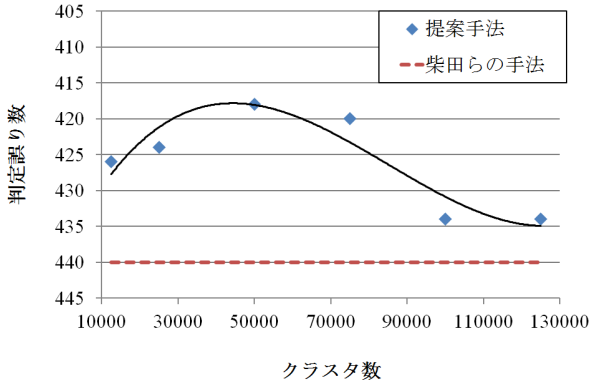

図 2: クラスタ数と判定誤りの関係：「中＋弱」セット

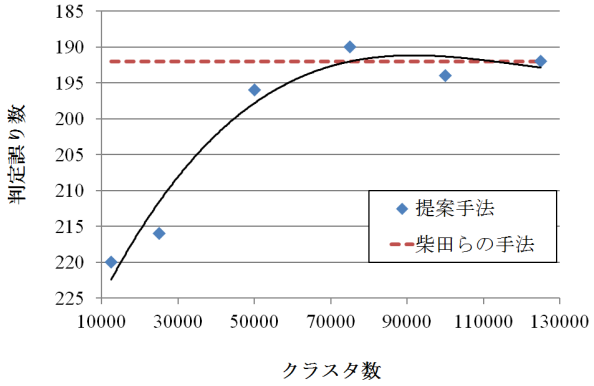

図 3: クラスタ数と判定誤りの関係：「弱＋非」セット

5 考察

5.1 クラスタリング結果の分析

実際のクラスタリング結果の一例を以下に示す。

クラスタ A ⇒ の:南禅寺, の:銀閣寺
クラスタ B ⇒ の:青い, の:海兵

クラスタ A に見られるような特定の単語の特徴を表す共起要素同士がまとめられることが共起要素のクラスタリングを行った大きな目的である。しかし、クラスタ B に所属する「の:青い」のように多義性を持つ共起要素が他の共起要素とまとまることは、類似性のない単語同士の類似度を不当に上昇させることにつながるため望ましくない。クラスタ B のように、多義性を持つ語や関連性の低い共起要素同士がまとまることはクラスタ数の設定が小さすぎる場合に発生し易いため、クラスタリングの効果を最大限に生かすためにはクラスタ数の設定が重要であるといえる。

表 2 に設定したクラスタ数ごとの、クラスタリングの結果に 2 件以上の共起要素が所属する割合を示す。

表 2: クラスタ数と所属する共起要素数の関係

クラスタ数	共起要素が 2 件以上所属するクラスタ数の割合
125,000	4.62%
100,000	11.74%
75,000	25.08%
50,000	53.20%
25,000	99.55%
12,500	100.00%

共起要素が 2 件以上所属しているクラスタの割合がクラスタ数を 25,000 とした場合からほぼ飽和している。実験において最も判定誤りが減少したクラスタ数は 50,000 と 75,000 の場合であることから、複数の共起要素を含むクラスタの割合が精度と関係していると考えられる。

共起要素をどの程度のクラスタ数とするか決定する際には複数の共起要素を含むクラスタの割合が飽和しない程度のクラスタ数とすることが望ましいといえる。

5.2 誤り分析

クラスタリングを行った場合とそうでない場合の判定誤りには共通しているものが多く存在した。共通している判定誤りを除いた場合の改善例には国名や地域名が多く見られ、「強＋中」のセットにおいては、提案手法による改善例 95 件のうち 37 件と 4 割近くを占めている。国名や地域名には人名や建物などクラスタリングの効果が現れやすい共起要素が多いことが要因として考えられる。

5.3 今後の課題

今回用いた手法では収集した共起要素は全てクラスタリングの対象とした。その結果、全体としての精度は上昇したが、クラスタリングを行ったことで判定誤りが起こってしまう例があった。そのため、精度向上に寄与するクラスタのみを使用するなど、素性選択の手法についての改善が求められる。

6 まとめ

分布類似度計算の精度向上のために、これまで有効に活用されていなかった共起要素に注目し、クラスタリングを用いて似た意味をもつ共起要素をまとめ類似度計算を行う手法を提案した。実験の結果、クラスタリングを行わない場合よりも高い精度となり、手法が有効であることを示した。

使用した言語資源及びツール

(1) IPA 品詞体系辞書 IPADIC,Ver.2.7.0, 奈良先端科学技術大学院大学 松本研究室, http://sourceforge.jp/projects/ipadic/

(2) 構文解析器 CaboCha, Ver.0.53, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/ taku/software/cabocha/

(3) データクラスタリングツール bayon, Ver.0.0.11, Mizuki Fujisawa, http://code.google.com/p/bayon/

(4) 日本経済新聞全記事データーベース 1990-2004 年度版, 日本経済新聞社

(5) 分類語彙表増補版, 国立国語研究所

参考文献

[1] 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.

[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] 朝倉剛史, 山本和英. 素性の相対性による分布類似度計算. 言語処理学会年次大会, A3-1, pp.688-691, 2010.

[4] 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 11, 2007. http://yans.anlp.jp/symposium/2007/paper/hagiwara.pdf

[5] Maayan Zhitomirsky-Geffet, Ido Dagan. Bootstrapping Distributional Feature Vector Quality. Computational Linguistics, Volume 35, Issue 3, pp.435-461, 2009.