E-035

共起漢字ベクトルを用いた同義語抽出方式

Synonym Extraction using Co-occurring-Kanji Vectors

秋良 直人† 森本 康嗣†
Naoto Akira Yasutsugu Morimoto

1．はじめに

テキストマイニングや文書検索などの文書処理システムでは，同義語を同じ語として扱う必要がある．一般語に対する同義語は，汎用の同義語辞書の利用がある程度有効であるが，専門語に対する同義語は，汎用辞書では対処できない．しかし，辞書に未登録の同義語を人手で網羅的に作成することには限界があり，コーパスから同義語を抽出する技術が必要とされている．

テキストマイニングでは，未発見の低頻度事例を発見したいというニーズがあり，文書検索では比較的頻度が低い単語が検索に貢献する傾向がある．低頻度語の例としては，新商品や新技術の出現に伴う新語があげられる．大規模なコーパスにおいても低頻度語は必ず存在する．そこで，本稿では低頻度語の同義語抽出精度を向上する方法を提案する．

2．従来方式とその課題

コーパスから同義語を抽出する従来技術は，同義語が類似の文脈に出現することを利用する．すなわち，単語の共起情報を基に，共起単語の集合が類似する語を同義語とする[1][2]．具体的には，共起頻度の重みをつけた共起単語から成るベクトル，すなわち共起単語ベクトルを生成し，共起単語ベクトルが類似する単語を同義語とする．図 1の例では，「自動車」と「車」の共起単語ベクトルが類似するので「自動車」の同義語として「車」が抽出される．しかし，低頻度語では共起単語が少ないために，同義語間で共通する共起単語を持たず，同義語の抽出が困難であるというデータスパースネスの問題があった．

		走る	乗車	走行	飛行	転倒	…	転覆
自動車	= (	50	20	30	0	0	…	0)
車	= (	40	15	20	0	0	…	0)
飛行機	= (	1	0	0	20	0	…	0)
船	= (	0	0	1	0	0	…	30)
自転車	= (	10	2	2	0	10	…	0)
…								

図 1　共起単語ベクトルの例

3．提案方法

3.1　共起漢字ベクトル方式

本稿で提案する共起漢字ベクトル方式は，コーパスに含まれる単語各々に対して，共起関係にある単語に含まれる漢字の頻度を要素とするベクトル，すなわち共起漢字ベクトルを用いて単語の出現文脈を表す．同義語を抽出する単語(注目単語と呼ぶ)と，共起漢字ベクトルが類似する単語を同義語とする．図 2に共起漢字ベクトルの例を示す．

†（株）日立製作所 中央研究所

共起漢字ベクトル方式によれば，「走行」と「走る」のように別の単語として扱われたものが，漢字では「走」という共通要素が発生するため，同一の次元として扱われる．したがって，共起単語が少ない低頻度語の同義語抽出に効果があると考えられる．漢字は母集団が既知で異なり数が少ない(JIS 第一水準漢字の場合 2965 文字)ため，共起ベクトルを容易に低次元化できるという特徴もある．しかし，単語を文字に分解することにより，曖昧性を生じて精度が低下する可能性があり，実験により影響を確認することとした．

		走	乗	行	飛	借	…	転
自動車	= (	90	40	50	0	8	…	0)
車	= (	80	25	30	0	9	…	0)
飛行機	= (	1	10	8	50	0	…	0)
船	= (	1	15	3	0	0	…	35)
自転車	= (	20	20	5	0	10	…	20)
…								

図 2　共起漢字ベクトルの例

3.2　処理手順

(1) 係り受け単語ペアの生成

コーパス中のテキストを形態素解析して単語列に変換した上で，単語の品詞情報を用いて係り受けの関係にある単語ペア(主語 - 述語)を生成する．ただし，主語は名詞または未知語とし，複合語も対象とする．述語は，動詞または形容詞とする．例えば，「車が走行する」という文からは，「車 - 走行する」という単語ペアが抽出される．

(2) 共起漢字ベクトルの生成

生成した単語ペアの主語となる単語各々に対して，ペアになる述語に含まれる漢字(共起漢字)を計数し，共起漢字の頻度を要素とする共起漢字ベクトルを生成する．

(3) 同義語候補の抽出

注目単語に対して共起漢字ベクトル間の類似度が高い単語を，同義語候補として出力する．類似度の計算式は，予備実験にて既知の尺度の組み合わせをいくつか検証した結果，単純で高い精度が得られた以下の式を用いた．ここで，$J(\mathbf{a}',\mathbf{b}')$は，ベクトル間の非ゼロ要素の一致度を示す 2 値 Jaccard 係数である．

$$Sim(\mathbf{a},\mathbf{b}) = J(\mathbf{a}',\mathbf{b}') \cdot \cos(\mathbf{a},\mathbf{b})$$

$\mathbf{a} = (a_1, a_2, \cdots a_N)$, $\mathbf{b} = (b_1, b_2, \cdots b_N)$: 共起ベクトル

$\mathbf{a}' = (a_1', a_2', \cdots a_N')$, $\mathbf{b}' = (b_1', b_2', \cdots b_N')$

$$a_i' = \begin{cases} 0 & a_i = 0 \\ 1 & a_i \neq 0 \end{cases}, \quad b_i' = \begin{cases} 0 & b_i = 0 \\ 1 & b_i \neq 0 \end{cases}$$

$$J(\mathbf{a}',\mathbf{b}') = \frac{\sum_i \min\{a_i', b_i'\}}{\sum_i \max\{a_i', b_i'\}}$$

4. 評価実験

4.1 実験方法

(1) コーパスと注目単語

社内のコールセンタに蓄積された約 13 万件の問合せ履歴テキストから同義語を抽出する実験を行った．このコーパスに含まれる頻度 10 以上の単語 3103 語を注目単語として選定した．

(2) 同義語候補の正誤判定

各々の注目単語に対して，類似度が上位 N 個の同義語候補を出力し，人手で正誤判定を行った．同義語候補には，上位下位語，関連語などが多く存在するが，注目単語と可換である単語のみを同義語として選択した．

(3) 注目単語あたりの平均同義語個数

抽出すべき同義語の数を正確に把握するのは困難である．そこで，コーパスに含まれる単語のペアをランダムに選び，そのペアが同義語であるかどうか判定することによって，注目単語あたりの平均同義語個数を推定した．例えば，1000 語の単語を含むコーパスには，499500 個の単語ペアが含まれるが，ランダムに選んだ 10000 個の単語ペアのうち 50 ペアが同義語であれば，注目単語あたりの平均同義語個数は 50×(499500／10000)／1000=2.5 語と推定される．この方法で推定された注目単語あたりの平均同義語個数は 0.765 語であった．

(4) 評価尺度

同義語の抽出精度を再現率(Recall)と適合率(Precision)を用いて評価した．再現率は，抽出すべき同義語の数に対する抽出できた同義語の割合を示す．ただし，抽出すべき同義語の数は，(3) で得られた平均同義語個数と注目単語数の積で計算される．適合率は，出力した同義語候補の個数に対する抽出された同義語の割合を示す．

4.2 結果

注目単語を低頻度語に限定しない場合と低頻度語に限定した場合の各々において，共起単語ベクトル方式と共起漢字ベクトル方式の精度を比較した．

(1) 高頻度を含めた場合の精度の比較

全注目単語(3103 語)を対象として，同義語候補の個数 N を 1 から 20 まで変化させて計算した再現率と適合率のグラフを，図 3に示す．グラフより，共起漢字ベクトル方式と共起単語ベクトル方式の精度は，ほぼ同等であるという結果が得られた．N =20 の再現率は，共起単語ベクトル方式の 63%に対して，共起漢字ベクトル方式は 65.7%であり，若干向上している．N =20 の適合率は両方式ともほぼ 2.6%である．

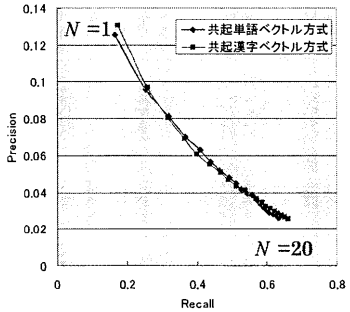
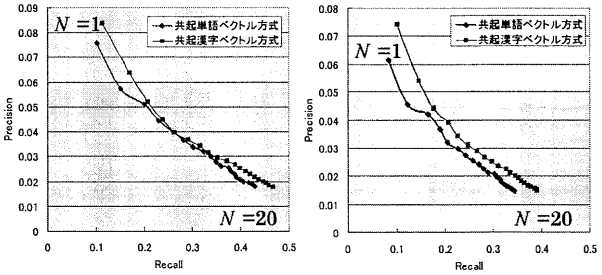


図 3　高頻度語を含めた精度の比較

(2) 低頻度語に対する精度の比較

頻度が 50 以下の注目単語(2021 語)と，頻度が 30 以下の注目単語(1512 語)の同義語抽出精度をそれぞれ図 4(a), (b) に示す．グラフより，低頻度の注目単語ほど，共起漢字ベクトル方式の精度が共起単語ベクトル方式の精度よりも相対的に高いことが確認できる．頻度が 30 以下の注目単語の再現率は，共起単語ベクトル方式の 34%(N = 20) に対し，共起漢字ベクトル方式は 38.6%(N = 20)である．N = 20 の適合率は両方式ともほぼ 1.5%である．また，N の値に拘わらず共起漢字ベクトル方式の精度が共起単語ベクトル方式よりも高いことが分かる．



(a)頻度 50 以下の注目単語　　(a)頻度 30 以下の注目単語

図 4　低頻度語の同義語抽出精度

5. 考察

低頻度語に限定した場合に共起漢字ベクトル方式の精度が高く，限定しない場合に両方式の精度が同等であることは，高頻度語の同義語抽出では共起単語ベクトル方式の精度が高いことを意味している．十分な量の共起単語が存在する場合には，データスパースネスに対する効果が少なく，単語を漢字に分解することによる曖昧性が影響しているためであると考えられる．

提案方法は，低頻度語に対する同義語抽出に有効であることを示したが，上位 20 個の候補を出力する場合でも再現率が 50%以下と低い．これは，注目単語の共起漢字と，同義語の共起漢字に，同じ漢字がない同義語は抽出できないという方式の限界があるためであると考えられる．実際に，抽出できなかった同義語には，注目単語と共通した共起漢字を持たないものが確認できた．

6. まとめ

共起漢字の頻度を要素とするベクトルの類似性に基づき，コーパスから同義語を抽出する方法を開発した．

コールセンタのコーパスを用いた評価実験を行ない，提案方法が低頻度語に対する同義語抽出に有効であることを示した．

今後は，他のコーパスを用いた場合の精度を検証すると共に，再現率の向上を計る予定である．

参考文献

[1] D.Hindle. 1990. Noun classification from predicate-argument structures. Proceedings of the 28th Annual Meeting of the ACL.

[2] 稲子他: 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 (2000/8).