コーパスからの同義語の獲得(2)
－スパース性への対処－

4Q－4

伊藤 山彦　相川 勇之　鈴木 克志
三菱電機（株）情報技術総合研究所

1．はじめに

電子化文書の増大に伴い、蓄積した文書を必要に応じて迅速に検索したいという要求が高まっている。特に近年多くの分野で導入が盛んなヘルプデスク業務においては、エンドユーザからの問い合わせに対し、関連する過去の問い合わせ記録を効率よく検索する必要がある。我々は、ヘルプデスク業務における過去の問い合わせ記録の検索処理を高度化する目的で、コーパスから同義語辞書を自動構築する研究を行っている。

コーパスからの同義語自動獲得における課題の1つに、自然言語の表現が多様であるため、同じ内容に対して同じ表現がコーパス中に出現しないというスパース性への対処を挙げることができる。本稿では、シソーラスを外部知識として用いることによって、スパース性を補完する手法を提案する。

2．同義語獲得の課題

同義関係の判定は、判定対象となる語と共起する語を手がかりにして行う(文献[1,2])。この方法によると、2つの語は、それぞれの語に共起する語の集合の中で共通の語の割合が多いほど類似していると判定される。

しかし実際のコーパス中では、自然言語の多様性のために、同じ意味を表すために同じ語が用いられるとは限らない。例えば、「放送する」と「放映する」は、EDR((株)日本電子化辞書研究所)の概念辞書においては、図1のように「不特定の相手に伝える」という概念を共通の親に持ち、近い意味を持つ語である([]内の符号は概念識別子)。しかし、語としては異なるため、「テレビで放送する」と「ＴＶで放映する」という文における「テレビ」と「ＴＶ」の同義性を検出することができない。

本稿では、判定対象語の近傍に出現する単語間のの類似性をシソーラスによって検出することによってコーパスのスパース性を補完し、近傍単語中に一致する語が少ない場合でも、同義語獲得を可能にする手法を提案する。

文献[1]においても同義性の判定にシソーラスを用いるが、文献[1]は判定対象語をシソーラスに配置するのに対し、本方法は判定対象語の近傍に現れる語の類似度を計るためにシソーラスを用いる。

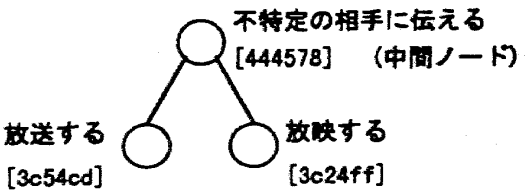

図1 「放送する」と「放映する」の関係

3．処理方法

3.1 近傍共起単語頻度ベクトル

同義性の判定は、判定対象語の近傍に出現する自立語の頻度値を重みとしたベクトル(近傍共起単語頻度ベクトル)を用いる。類似度はベクトルの内積をとることによって判定する。

3.2 ベクトル中の単語間の類似度

ベクトルの要素となる語の類似度は、シソーラス上で共有する親ノードのレベルに従って表1のように定義する。レベルが異なる語同士は、下のレベルの語を基準とする。シソーラス上の位置関係による類似度の例を図2に示す。

表1 ベクトル中の単語間の類似度の定義

親ノード	一致	1つ上	2つ上	3つ以上上
類似度	1	0.5	0.1	0

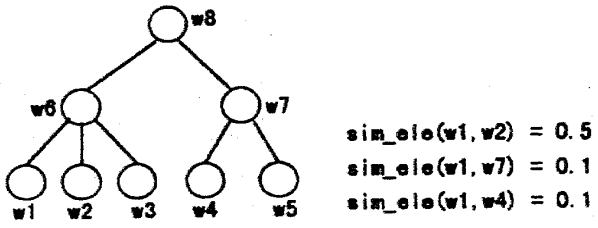

図2 ベクトル中の単語間の類似度の例

3.3 ベクトル間の類似度の補正

シソーラスに基づく単語間の類似度の定義を用いて判定対象語ＡとＢの類似度を求める処理について説明する。Ａの近傍に出現する語の集合をＳ

Acquisition of Synonyms(2)
- A Method to Deal with Sparseness -
Takahiro ITO, Takeyuki AIKAWA, Katsushi SUZUKI
Mitsubishi Electric Corporation.
5-1-1 Ofuna, Kamakura, Kanagawa 247, JAPAN

(A)、Bの近傍に出現する語の集合をS(B)とする。このとき、S(A)とS(B)に含まれる語の数の関係を図3に示す。

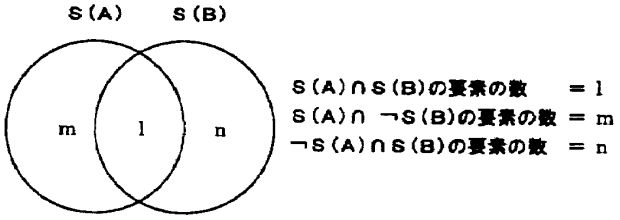

図3 S(A)とS(B)に出現する語の数の関係

S(A)及びS(B)に現れる語をw(i)からw(1+m+n)で表す。それぞれの語に対応して、単語Aの近傍単語の頻度値をf(i)、単語Bの近傍単語の頻度値をg(j)で表す。このとき、語Aと語Bの近傍共起単語頻度ベクトルV(A)、V(B)は、図4のように表すことができる。

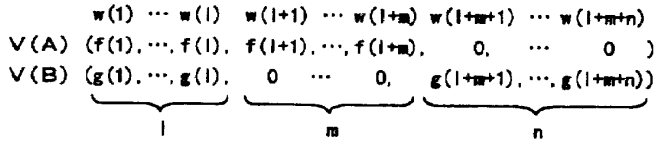

図4 頻度ベクトルV(A)とV(B)の関係

ここでベクトルの要素となる語に対して、次の(1)～(5)の処理を施す。

(1) g(i)の値が 0 に対応する語(w(1+1)～w(1+m))に対して、g(j)の値が0以外に対応する語(w(1)～w(l)または w(1+m+1)～w(1+m+n))との類似度が最大の語を取り出す。類似度最大の語が複数ある場合は、それらのうち最も頻度値の高い(g(j)の値が大きい)語を選択する。

(2) 上記(1)の処理により w(p1)に対して、類似度最大の語w(q1)が見つかった場合、次の式に従ってV(A)の成分値を変換する。

$$f(q1) \leftarrow f(q1) + sim\_ele(w(p1), w(q1)) * f(p1)$$

(3) f(i)の値が 0 に対応する語に対して、f(j)の値が 0 以外に対応する語との類似度が最大の語を取り出す。類似度最大の語が複数ある場合は、それらのうち最も頻度値の高い語を選択する。

(4) 上記(3)の処理により w(p2)に対して、類似度最大の語 w(q2)が見つかった場合、次の式に従ってV(B)の成分値を変換する。

$$g(q2) \leftarrow g(q2) + sim\_ele(w(p2), w(q2)) * g(p2)$$

(5) 次の式に従ってV(A)とV(B)の内積を計算することにより、類似度を算出する。

$$sim = \frac{\sum f(i) * g(j))}{\sqrt{\sum f(i) * f(i))}\sqrt{\sum (g(j) * g(j))}}$$

4. 実験

以下の手順で同義語獲得の実験を行った。コーパスには当社お客さま相談センターの問い合わせ記録(約9万文)を用い、シソーラスにはEDRの概念辞書を用いた。

(1) コーパス中に現れる全ての自立語の前後 2 語以内に出現する自立語から類似度計算に不要な語(接続詞、連体詞など)を削除して近傍共起単語頻度ベクトルを作成する。

(2) カタカナ語、及び英字列のみの未知語 455 語に対し、3節の方法に従って、頻度 10 以上の自立語との類似度を計算し、上位 30 語を取り出す。

(3) 同義性の高い単語同士は共起しにくいというヒューリスティックを適用してフィルタリングを行い、最終的に得られた上位 10 語から目視チェックにより同義語及び類義語を取り出す。

上記の実験を継続中であるが、実験結果を評価したところ、期待した効果は現れていない。原因として、以下の理由により不適切な語の組み合わせが高い類似度を持つと判定されるためと考えられる。

・語の多義性に基づく原因

「コントローラー」は、「電動機の制御装置」という意味で使用されているが、「企業経営を管理する役割」という意味も持つため、「所長」と類似した語と判定される。

・階層の構成に基づく原因

シソーラス上の位置の近さが必ずしも意味の近さに対応していない。例えば「有料」「こんなだ」「少ない」など、意味的に関係のない語が「いろいろな抽象物の属性」という概念を共通の親に持つ。

5. おわりに

本稿では、シソーラスを利用してコーパスのスパース性を補完した同義語獲得方法を提案した。今後4節で述べた考察に基づき、シソーラスによる語の類似度の適用方法を改良して実験を継続する。

参考文献

[1] 浦本ほか: コーパスに基づくシソーラス-統計情報を用いた既存シソーラスへの未知語の配置 情報処理学会論文誌, Vol. 38, No. 12, pp. 2182-2189 (1997).

[2] 相川ほか: コーパスからの同義語の獲得(1)-近傍単語頻度統計によるアプローチ-, 情報処理学会第 56 回全国大会予稿集掲載予定(1998).