言葉の意味に関する概念ベースの精錬法

2ZA－05

町田　厚†　　石川　勉†　　笠原　要‡
拓殖大学工学部情報工学科†　　ＮＴＴコミュニケーション科学基礎研究所‡

1．はじめに

言葉の意味に関する概念ベースの構築を進め[1]、現在、約４万語の基本概念ベース、約２５万語の拡張概念ベース、造語や新語に対処するための漢字概念ベースからなる大規模概念ベースを構築している。[2]

本論文では、この大規模概念ベースにおける類似性判別能力を向上させるための精錬法について提案する。

2．概念ベース

概念ベースは、言葉の意味の類似性判別を主な用途に想定したもので、二つの概念間の類似性を類似度として定量化（０～１の数値）しようとするものである。この概念ベースでは、各概念の属性は、国語辞書等の単語（概念に相当）の語義文に含まれる独立語であり、各属性の属性値はその独立語の出現頻度をもとに算出している。

この概念ベースの構築手順は、まず概念に対して属性と出現頻度を取得する。例として、概念【馬】に対して取得すると以下のようになる。

【馬】＝{(家畜：１)，(草食：１)，(動物：２)，・・・・}

これを全ての概念に対して行うと、概念$Word_i$は以下のように表わされる。

$Work_i=\{(p_{i1},q_{i1}),\cdots,(p_{ij},q_{ij}),\cdots\}$

ここで、p_{ij}は属性、q_{ij}は属性値である。従って、概念ベースWは、総概念数をNとし、すべての概念が属性になりうるとすると、以下のようにN×Nの行列で表わされることになる。

$$W=\begin{pmatrix} q_{11},\cdots,q_{ij},\cdots,q_{1N} \\ \cdot \quad \cdot \quad \cdot \\ q_{i1},\cdots,q_{ij},\cdots,q_{iN} \\ \cdot \quad \cdot \quad \cdot \\ q_{N1},\cdots,q_{Nj},\cdots,q_{NN} \end{pmatrix}$$

ここで、行列の要素q_{ij}はその属性が獲得されていなかったとき“０”とする。

次いで、この概念ベースの属性を、日本語シソーラスに対応した２７１５のカテゴリーに分別することにより圧縮する。最後に、各概念の属性値q_{ij}を、ｔｆ・ｉｄｆ[4]の考えに基づいて再計算し、二乗和が“１”となるように正規化する。

概念間の類似度Ｓは、こうして得られた２７１５次元のベクトル間の内積として以下のように算出する。

$$S=Word_A\cdot Word_B=\sum_{j=1}^{2715} q_{Ai}\times q_{Bi} \quad (1)$$

ここで、$Word_i=(q_{i1},\cdots,q_{ij},\cdots,q_{i2715})$

3．精錬手法

先の構築法には以下の問題がある。

ｉ）属性の重みは出現頻度を基としているため、一つの概念に複数の語義文があるとき、構成単語の多い語義文ほど属性値の合計が大きくなってしまう。

ii）類似度計算を内積により求めているため、意味合いの近い属性同士でもシソーラス上で異なったカテゴリーに属していれば類似度に算入されない。

iii）不必要属性も、他の属性同様に孫引きを行っている。

これらの問題に対して以下の改良を行う。

3－1　語義文長による重み付け

概念【家】の辞書中の記述が、“人が住むための建物。すまい。”だとすると、従来の方法ではその概念は以下のよう表現される。

（人，１．０）（住む，１．０）（建物，１．０）（すまい，１．０）

しかし、この場合「すまい」という単語は他の単語に比べ【家】との関連度が高いと思われる。このため、一語義文の重みを“１．０”とすることにより、短い語義文と長い語義文との情報量の違いを適正化した。先の例では以下のように修正される。

（人，０．３）（住む，０．３）（建物，０．３）（すまい，１．０）

3－2　上位カテゴリー取得

従来、下記のような概念間で類似度を計算した時、類似度の計算に内積を用いるため、同じカテゴリーに属していない属性「男」と「女」は類似度に算入されない。

【夫】＝{(夫婦，０．２)，(男，０．４)，・・・・}

【妻】＝{(夫婦，０．２)，(女，０．４)，・・・・}

ここでは、図１のようなシソーラス構造の時、上位に存在する全てのカテゴリーを属性として取得することにより、カテゴリー「男女」から「男」と「女」の類似性を類似度に算入されるようにした。算入方法としては、上

Refining Method of A Knowledge Base for Measuring Semantic Similarity between Words.

Atushi Machida†, Tsutomu Ishikawa†, Kaname Kasahara‡
†Department of Computer Science, Faculty of Engineering, Takushoku University
‡NTT Communication Science Laboratories

位のカテゴリーは下位のカテゴリーをすべて含むため、下位カテゴリーの最大値を上位カテゴリーの属性値として取得する。先の例では以下のように修正される。

【夫】

(人間関係, 0．4) (夫婦, 0．2) (男女, 0．4) (男, 0．4) …

【妻】

(人間関係, 0．4) (夫婦, 0．2) (男女, 0．4) (女, 0．4) …

なお、上位カテゴリーに下位の最大値と同じ属性値を付けるのは重すぎるように思えるが、上位カテゴリーほど出現頻度が多くなるので、ｔｆ・ｉｄｆの考えに基づく再計算で値が小さくなる。

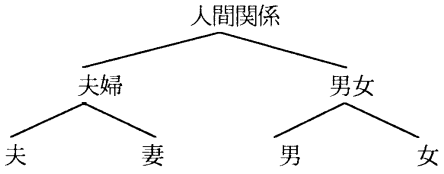

図１　「人間関係」部分のシソーラス構造

３－３　孫引き

孫引きとは、概念に含まれる属性を概念として再帰的に参照することで、概念の属性を増加させる手法である[1]。具体的には、“$Word_i$:{(p_{i1}, q_{i1}), …, (p_{ij}, q_{ij}), …}”に対し、孫引きベクトルM_iを以下のように求めている。ここで$p_{ij}=Word_j$とする。

$$M_i=q_{i1}\times Word_1+\cdots+q_{ij}\times Word_j\cdots \quad (2)$$

ここでは、重要な属性ほど情報量を大きくして孫引きをする方法を提案する。具体的には、まずカテゴリー化する前の概念ベースWから、属性の情報量を算出する。すなわち、ｊ番目の属性の情報量d_jを下式から求める。

$$dj=-\log\{Num(j, W)/N\} \quad (3)$$

ここで、Ｎｕｍ（ｊ，Ｗ）とは概念ベースW中でｊ番目の属性の属性値が“０”でない概念の数であり、Ｎは総概念数である。こうして算出した情報量を平均値が“１．０”となるようにしてベクトル表現する。

$$D=(d_1, d_2, \cdots, d_j, \cdots, d_N)$$

このベクトルの要素を、(２) 式の各項に乗じて孫引きベクトルM_iとする。即ち、M_iは以下の式で表わされる。

$$M_i=\sum_{j=1}^{N} q_{ij}\cdot d_j\cdot Word_j \quad (4)$$

４．評価

従来の拡張概念ベースに対し、《語義文長による重み付け》・《上位カテゴリー取得》・《孫引き》の順で精錬を行い、拡張概念ベースを再構築した。このなかで、孫引きベクトルは “０．２”倍してもとの概念ベクトルに加算している。この値は予備実験の結果、最も評価が高くなった値である。

この新たな概念ベースに対して、概念ベースの特性として望まれる以下の点を考慮して評価をおこなった。

ｉ）類似する概念間の類似度と全く類似しない概念間の類似度の差が大きい。

ii）二つの候補概念があったとき、どちらが対象概念に似ているかを識別可能。

具体的には、対象概念に対して、これと“類似する概念”、“比較的類似する概念”、“非類似概念”の組を２００個用意し、以下の式により評価した。

$$Fd=F1*F2 \quad (5)$$

ここで　$F1=(R1-R3)/(\phi 1+\phi 3)$

$F2=1.0/(1+wg1+wg2)$

ここで、Ｒ１・Ｒ３は対象概念に対する類似概念・非類似概念の類似度の平均値であり、ϕ１・ϕ３は、同じくこれらの標準偏差である。また、ｗｇ１は類似概念・比較的類似概念の類似度の大小関係が反転した数であり、ｗｇ２は、比較的類似概念・非類似概念の大小関係が反転した数である。上式のＦ１・Ｆ２はそれぞれ特性ｉ），ii）に対応するので、Ｆｄの値が高いほど、類似性判別能力が高いことになる。評価結果を表１に示す。同表に示すように、大幅な性能向上が実現できた。

表１　評価結果

	Ｆ１	Ｆ２	Ｆｄ
従来の概念ベース	２．１３	０．０２３	０．０４９
精錬後概念ベース	３．３３	０．０４３	０．１４５

５．　まとめ

本論文では、従来の概念ベースの構築上の問題点を解決した精錬法を提案した。具体的には、語義文長を考慮した重み付け、上位カテゴリーの属性化、孫引きの適正化をおこなった。さらに、実験的にこの精錬法の有効性を確認した。

参考文献

[1]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　Ｖｏｌ３８－７（１９９７－０７）.

[2]帆苅　譲、石川　勉、笠原　要：言葉の意味に関する階層化大規模概念ベースの構築，信学技報　ＡＩ９８－６５（１９９９－０１）.

[3]石川　勉、井澤　潤次郎、Ｎｇｕｙｅｎ　Ｖ．Ｈ、笠原　要：単語の意味に関する概念ベースの類似性判別能力からの最適構成，人工知能学会誌　Ｖｏｌ１３－３（１９９８－０３）.

[4]G. Salton, and J. Allen: Text Retrieval Using the Vector Processing Model, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.