複数の概念を有する単語セットからの類似概念の抽出

神山 義之*　上原 徹三**　石川 知雄**
*武蔵工業大学大学院工学研究科　**武蔵工業大学工学部

1 はじめに

自然言語処理に広く利用される電子化辞書の作成には膨大な労力とコストがかかるため，作成対象の言語は限られている．特に概念情報の作成には人手を介入せざるを得なく容易ではない．そのため概念情報の付与された電子化辞書を持たない言語については，計算機処理による文法解析が十分に行なえないという問題がある．古文の分野においても単語の概念情報は必要であるが，概念付き電子化辞書は存在しない．機械可読形式の古文の対訳辞書 [1] が存在することから，対訳辞書を用いた概念の自動獲得方法を検討した．

計算機上に用意する対訳辞書は，概念情報を持たない言語を見出し語，概念情報を十分にもつ言語を訳語とする．対訳関係と訳語側言語の概念情報を利用することで，見出し語側の言語に概念情報を自動的に付与できる可能性がある．一般に見出し語の1つの語義に対する訳語は複数個あり，更にこれらの各訳語は多義語であるため，訳語間の共通する語義を抽出し，見出し語の概念として最適な解を得ることができれば良い．そこで，複数の訳語，つまり複数の多義語を比較して，それらの間の類似概念を推定する方法を提案する．

本研究では，作成対象を概念付き英語単語辞書とした．英単語の概念は，EDR 電子化辞書 [2] に記載されているため，それと本手法によって作成した概念付き英語単語辞書を比較することで，本手法の方法自体の評価を行なう．

即ち，本稿では，類似概念推定法の提案と，その評価について述べる．

2 類似概念推定法の考え方

複数の多義語を比較して，その共通概念を抽出する上で，**単語セット**という用語を定義する．単語セットとは，図1に示すように1つの単語表記が持つすべての語義の集合である．単語セットの要素である概念を，**概念終端**と呼ぶ．

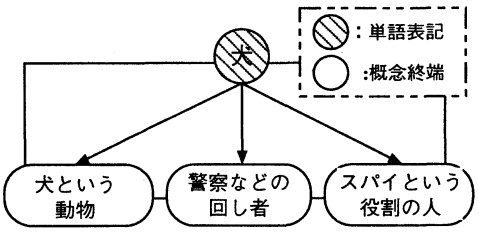

図 1: 単語セット

本手法の目的は，複数の単語セットを入力にして，その共通概念を抽出することであり，**類似概念推定法**と呼ぶ．

ここで，抽出すべき概念について考える．概念は，シソーラス上のノードとして存在するが，複数の単語セットの共通概念を抽出するためには，各単語セットの要素である概念終端が，シソーラス上でどのように配置されているかが重要となる．求めるべき概念が，全ての単語セットの共通部分であることから，シソーラス上に配置される概念終端が密に集合している部分に存在すると考えられる．本手法では，単語セットの概念終端のシソーラス上での位置関係に着目した類似度の計算方法を定義する．

すなわち，対訳辞書の訳語から単語セットを作成し，各訳語の単語セットの共通概念を本手法により推定することで，見出し語の概念の自動獲得を行う．

3 類似概念推定法の詳細

複数の単語セットから類似概念を抽出する実際の処理としては，参照するシソーラスの全ノードに対してある一定の規則にしたがった得点付けを行い，その得点によって類似概念を選択する．ここで計算機内部に実現するシソーラスを，概念をノードに持つ構造であることを強調する意味で，以降，**概念ツリー**と呼ぶことにする．類似概念推定法は，図2に示すように，入力を対象となる複数の単語セット，出力をそれらの類似概念とし，以下の3ステップから成る．

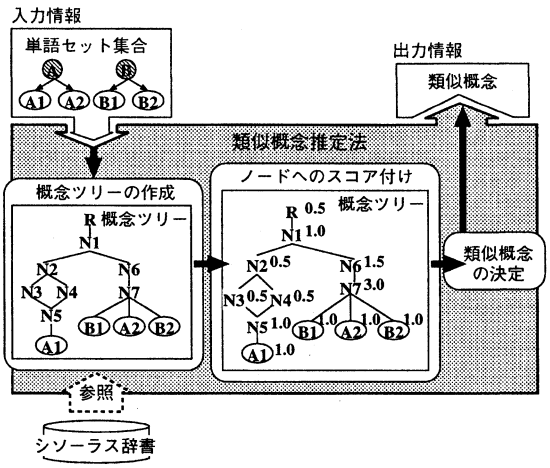

図 2: 類似概念推定法の流れの例

(STEP1) 概念ツリーの作成

入力の各単語セットの概念終端をノードに含む概念ツリーを, シソーラス辞書を参照して計算機内部に作成する.

(STEP2) 概念ツリーのノードに対するスコア付け

本手法の核となるステップであり, STEP1 で作成した概念ツリーの全ノードを対象に, 概念終端の位置を考慮した得点付けを行う. 概念終端が密に集合している部分を抽出するために, 各ノードの周りにどれだけ概念終端が存在しているかを数値によって表現する. そこで概念終端を下位概念に多く保持するノードには, 高い得点を与える計算方法を提案する. つまり高得点を与えられるノードは, 概念終端が密に集合している部分だと考える. 得点付けの対象は STEP1 で作成した概念ツリーの全ノードであり, 以下の 2 段階で行われる.

[2-1] 単語セットスコアを全ノードに付ける

得点付けの対象ノードから見て, 下位に存在する概念終端から受ける影響を単語セット別に求める. この単語セット別に与える値を**単語セットスコア**と呼ぶ. 単語セットスコアは, 対象ノードから概念終端までの距離によって決定され, 距離が小さい程, 大きい値をとる. 以下の 2 値により算出される.

1. **対象のノード自身が概念終端であるか否かによって与えられる値**
   対象ノードが所属する単語セットの単語セットスコアに一定値 S_{score} を与える.
2. **直下に存在する子供がもつ単語セットスコアから算出される値**
   子供ノードの各単語セットスコアの Dw 倍した値を対象ノードの各単語セットスコアに継承する. 継承は子供から親への方向のみである.

以上をまとめると式 (1) になる. ただし n は類似概念を求める対象である単語セットの種類数, j は対象ノードが持つ子供の数, $WS_{score}i$ は i 番目 $(i = 1, \cdots, n)$ の単語セットスコア, $C_{score}ki$ は直下の k 番目 $(k = 0, \cdots, j)$ の子供が持つ i 番目 $(i = 1, \cdots, n)$ の単語セットスコアとする.

$$WS_{score}i = S_{score} + \sum_{k=0}^{j} C_{score}ki * Dw \quad (i = 1, 2, \cdots, n) \tag{1}$$

現時点では, 経験的に S_{score} は, 対象ノードが概念終端である場合を 1, 概念終端でない場合を 0 とし, また Dw を 0.5 としている.

[2-2] ノードスコアを全ノードに付ける

単語セットスコアを基に対象ノードのスコアを決定する. これをノードスコアと呼び N_{score} で示す. ノードスコア付けには, 3 種類の方法を準備した.

方法 (1) 全単語セットスコアの積をとる

$$N_{score} = \prod_{i=1}^{n} WS_{score}i \tag{2}$$

方法 (2) 2 つの単語セット間でセットスコアの積を取り, 全組み合わせの和で決定する

$$N_{score} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (WS_{score}i * WS_{score}j) \tag{3}$$

方法 (3) 全単語セットスコアの和をとる

$$N_{score} = \sum_{i=1}^{n} WS_{score}i \tag{4}$$

(STEP3) 最大スコアを持つノードの選択

全てのノードスコアの中で最大値をもつノードの概念識別子を，類似概念推定法の出力とする．ただし同スコアのノードが複数存在する場合は，全てを結果とする．

ノードスコア付けの方法により，最終的に選択されるノードは異る．方法 (1) は，全単語セットの影響を受けているノードのみを類似概念とし，方法 (2) は，より多くの単語セットの影響を受けているノードを類似概念とし，方法 (3) は，単語セットの種類による影響を特には考えず，概念終端が密に集合しているノードを類似概念とする．つまり方法 (1) から方法 (3) の順に，条件が緩やかになっており，どういった概念を類似概念として抽出するかによって，ノードスコア付けの方法が選択できる．

4 評価実験

本実験では EDR 電子化辞書を用いて，対訳辞書による概念の自動獲得を検証する．対訳辞書に EDR 英日対訳辞書，訳語側の辞書として EDR 日本語単語辞書，シソーラスには，EDR 概念体系辞書をそれぞれ用いて，英単語の概念の自動獲得を行う．シソーラスを使用するに当たって，従来は 2 概念間の概念に注目した計算方法が定義されている [3][4][5] が，本稿では，先に定義した類似概念推定法を用いる．EDR 概念体系辞書では，概念は**概念識別子**と呼ばれる記号で表されている．

まず，対訳辞書の訳語である日本語の単語表記をキーに，EDR 日本語単語辞書から訳語の概念識別子を獲得して，単語セットを作成する．対訳辞書の見出し語に対して，複数の訳語が記載されているため，訳語数分の単語セットが作られる．本手法により訳語の単語セットを入力として類似概念を抽出し，これを見出し語の概念として与える．

ここで EDR 英日対訳辞書は，一般の英和辞典と同様に，英語の単語を見出し語として，その日本語訳語を記した辞書であり，見出し語には概念識別子が付与されている．そこで，本手法により日本語の訳語を用いて類似概念を求め，この類似概念と見出し語に付与される概念識別子を比較することで評価を行う．

4.1 評価に使用するデータ

EDR 英日対訳辞書の全レコードの訳語情報を対象に，概念識別子を付与して単語セットを準備する．概念識別子は，訳語の単語表記を基に EDR 日本語単語辞書を引いて求めた．このとき，訳語の単語表記から概念識別子が得られない単語については，形態素解析などの処理はせずに概念識別子の獲得失敗として扱った．また本研究では，複数の訳語情報から類似概念を推定することが目的であるため，概念識別子が付与できた訳語を 2 つ以上持つレコードのみが評価対象である．EDR 英日対訳辞書の全 289,882 レコードに対して，訳語を 2 つ以上保持するものは 92,220 レコードであった．そのうち概念識別子を付与できた訳語を 2 つ以上持つものは 28,975 レコードであり，これを評価に使用した．

4.2 評価方法

EDR 英日対訳辞書の見出し語にあらかじめ与えられている概念識別子 (**正解**と呼ぶ) と，本手法で推定した概念識別子 (**推定概念**と呼ぶ) を比較する．正解と推定概念の近似度は，概念ツリー上における距離で評価する．2 概念間の距離について，両ノードが一致する場合は 0，一方が他方の先祖である場合は両者の間の辺数，それ以外の場合は共通の親までの辺数の和であるとする．つまり家系図の親等の数え方に等しい．図 3 に例を示す．(例 1) では，推定概念から，正解と推定概念の共通概念まで 2 つ上り，そこから正解へ 1 つ下がっているため，距離は 3 となる．また (例 2) は正解と推定概念が親子関係の場合であり，その距離は 1 である．配置が逆の場合も同様である．(例 3) は，正解と推定概念が，同じノードの場合であり，距離は 0 とする．したがっ

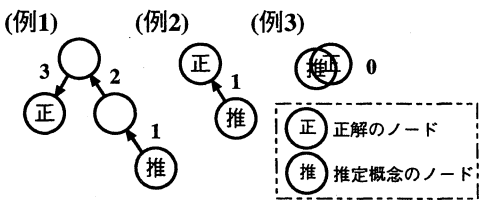

図 3: 評価方法

て距離が小さい程, 正解と推定概念は近い関係にある.

4.3 評価結果

ノードスコア付けの方法 (1)(2)(3) を使用したときの評価を行った. 表 1 は評価に使用するデータ数と, EDR 概念体系辞書の不備による類似概念の獲得失敗数, および類似概念が概念ツリーのルートになった数を示している.

表 1.推定概念獲得状況

評価対象のレコード数		28,975 (100%)
概念体系辞書の不備による評価失敗数		212 (0.7%)
類似概念が概念ツリーのルートになった数	方法(1)	1,783 (6.2%)
	方法(2)	1,008 (3.5%)
	方法(3)	0 (0.0%)

(単位：個)

EDR 概念体系辞書の不備とは, 正解として扱った概念識別子が概念体系辞書に記載されていない場合を指す. 評価対象の 28,975 レコードに対して, 概念体系辞書の不備は, 212 レコードであった. 方法 (1) による推定法では, 類似概念が概念ツリーのルートになる確率が 6.2%と最も多くなった. これは, 全訳語の概念終端を下位ノードに含むという厳しい条件を付けたためであり, 訳語として他の訳語と概念が大きく離れた単語が 1 つでも存在すると, 推定概念はルート側に上昇してしまう. 方法 (2) では同条件の確率が 3.5%, 方法 (3) では 0.0%となっている. 方法 (3) は単語セット間で共通する概念でないものを含んでおり, 1 つの単語セット中で, より類似した概念が存在する場合はそれらの親概念を結果として出力する. 中間に位置する方法 (2) は距離 4 以内が全体の約 50%という結果になった.

正解と推定概念間の距離の頻度を, 距離別に測定したものを図 4に示す. それぞれの平均距離を求めると方法 (1) は 4.9, 方法 (2) は 5.0, 方法 (3) は 5.1 となった.

5 おわりに

対訳辞書の訳語情報を利用して, 最適な見出し語の概念を推定する手法を提案した. 概念を推定する上でシソーラスを参照するが, 単語表記とその概念終端を組みにした単語セットを対

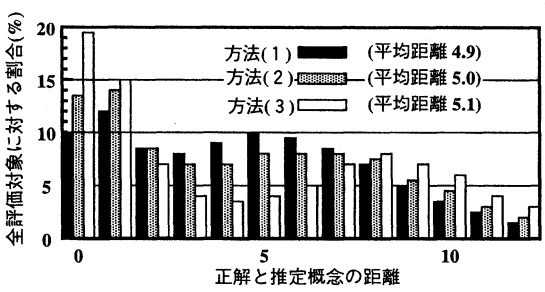

図 4: 距離別の出現確率

象とし, その概念終端の位置関係に注目することで類似概念を抽出する方法を考案した. さらに訳語の解釈の違いを考慮して計算方法を 3 つを用意した. 方法 (1) は全訳語の概念終端を下位ノードに含む部分を抽出する手法, 方法 (3) は単に概念終端が集合している部分を抽出する手法, 方法 (2) はその中間に位置する手法である.

既存の英単語の概念と比較することで本手法の評価を行った. その結果, 方法 (2) において正解との距離が 4 以内であるものが約 50%であることが分かった.

今後は現在, 経験的に与えている単語セットスコア付けの初期値に関する検討, ノードスコアの 3 つの方法 (方法 (1)(2)(3)) を組み合わせた利用方法の検討, 実応用への適用などが考えられる.

なお, 本研究の一部は文部省科学研究費補助金 (基盤研究 C2 No.11680422) によって実施したものである.

参考文献

[1] 金田一 春彦:全訳用例古語辞典, 学研, 1998.
[2] EDR 電子化辞書仕様説明書, 日本電子化辞書研究所, 1996.
[3] 堀口 賞一, 飯田 敏幸:和英辞書を用いたシソーラス細分化手法, 言語処理学会第 1 回年次大会発表論文集 pp.209-212, 1995.
[4] 大井 耕三, 隅田 英一郎, 飯田 仁:単語間の意味的類似度に基づく文書検索手法, 言語処理学会第 2 回年次大会発表論文集 pp.109-112, 1996.
[5] 中山 聡, 峯 恒憲, 東 優, 谷口 倫一朗, 雨宮 真人:EDR コーパスを利用した動詞の語義分類, 信学技報, NLC95-43, 1995.