複数の概念を有する単語のセットからの共通概念の抽出*

5M−05

池田 直　神山 義之　上原 徹三　荒井 秀一†
武蔵工業大学‡

1. はじめに

自然言語処理において単語の概念情報が利用できれば有用である. EDR 電子化辞書 [1] の単語辞書は単語の概念識別子を与える. これは階層ツリー上に保持する概念体系辞書 (シソーラスの一種) のノードの識別子である. しかし, このような概念情報を含む辞書は多くない. 英和辞典や古語辞典等の多くの辞書は, 言語 A の見出し語の語義を言語 B の訳語で与えるという対訳辞書:A → B の形式を持つ. 訳語を理解する人は, この対訳辞書から見出し語の概念を得ている. そこで, 訳語の概念情報が存在する場合に, 対訳辞書から見出し語の概念を獲得するという課題を考える. 一つの問題は, 自然語の単語は多義であるのが普通であり, 対訳辞書の訳語も概念を複数個持つことである. 従って上の課題は, 「複数の概念を有する単語 (訳語) のセットからの共通概念の抽出」による見出し語の概念の推定, と言える.

2. 手法の概略と用語の定義

対訳辞書:A → B の見出し語に対する複数の訳語から, 見出し語の概念を抽出することが課題である. 一般に, 見出し語に対して複数個の訳語が挙げられ, その各訳語がまた複数個の概念を持つ. 言語Aの単語も多義であり得るが, ここでは, その語義ごとに一つの見出し語を立てることとする.

見出し語と訳語の概念の間には, 次のような関係が有ると考えられる.

- 各訳語の概念は見出し語の概念 (の一部) を含む.
- 各訳語は一般に多義であり, 複数概念を持つが, その内の少なくとも一概念は見出し語の概念 (の一部) を含む.
- 見出し語の概念は, 全訳語の共通概念である場合もあり, 複数の訳語の併合概念である場合もあり, それらの複合の場合もあると考えられる.

上のように各訳語の概念と見出し語の概念の関係は多様だが, ここでは簡単に, 推定する見出し語の概念を, 各訳語の概念の共通概念と呼ぶ. 本手法では, 各訳語の持つ複数の概念のシソーラス上での配置を調べることで, それらの共通概念を求め, 見出し語の概念を推定する. シソーラス上で近傍にある概念は近い意味を持つので, シソーラス上で複数の訳語の持つ概念が集中している部分があれば, そこが共通概念を表すと考えられる.

ここではシソーラスとして,EDR 概念体系辞書を用いる. 訳語が現代日本語であると,EDR 日本語単語辞書よりその概念 (概念識別子) が得られる. 一つの訳語と, それが持つ (複数の) 概念とをまとめて**概念セット**と呼ぶ. 見出し語は訳語の個数分の概念セットを持つ. 各概念セットの要素である個々の概念を**概念終端**と呼ぶ. 本手法は複数の概念セットから, それらの共通概念を求めるものであり, これを**共通概念推定法**と呼ぶ.

従来,2 概念間の類似性を扱う研究 [2][3][4] があるが, 本手法は, 概念間の類似性の扱いを含みつつ, 複数の概念セットの指示する共通概念を推定するものである.

*Extracting a common concept from a set of words each of which has several concepts

†Sunao IKEDA, Yoshiyuki KAMIYAMA, Tetsuzou UEHARA, Shuichi ARAI
‡Musashi Institute of Technology

3. 共通概念推定法

共通概念を推定する実際の処理は, シソーラスを参照し, 概念終端群を葉とするサブツリーを作り, 全ノード (概念) に対してスコア付けを行い, 最大スコアのノードを選択することで行う. 最大スコアのノードが複数となる場合は, それら全てを共通概念とする. 下記のように, スコアには概念セット別に求める**概念セットスコア**と, それを基に与える**ノードスコア**がある.

[1] 概念セットスコア付け

概念セットスコアは, 対象ノード自身が概念終端であるか否かの区別と, 下位に存在する概念終端及び余分な子概念の数を考慮し, 次式より概念セット別に求める.

$$W_i = S + \sum C_i * (j + c)/2j \quad (i = 1, 2, \cdots, n)$$

対象ノードが概念終端である場合は所属する概念セットのスコア W_iに一定値 Sを与え, 下位概念の概念セットスコア C_iを子概念数に応じて継承する. 但し n は概念セットの種類数,jは対象ノードの持つ子概念数,cはスコアを持つ子概念の数である.$(j+c)/2j$の部分は継承の割合を表しており, 他にいくつかの代替案を用意した.

[2] ノードスコア付け

概念セットスコアを基に, ノードとしてのスコア Nを決定する. これは概念セットに跨る共通概念を取り出すための処理である. 方法は以下の 3 種類準備した.

1. 全概念セットスコアの積をとる方法 ($N_1 = \prod W_i$)
2. 概念セット対の間で概念セットスコアの積をとり, 全ての対の和をとる方法 ($N_2 = \sum\sum(W_i * W_j)$)
3. 全概念セットスコアの和をとる方法 ($N_3 = \sum W_i$)

これら 3 種類の方法により, 推定する概念は異なる. 方法 1 は全概念セットの影響を受けている概念を選択し, 方法 3 は異なる概念セットに共存するという条件を置かず, 単純に概念終端が集合している部分から選択する. 方法 2 は方法 1 と 3 の中間に当たる方法である.

4. 評価

本手法の評価には, 予め見出し語に概念情報を持つ EDR 英日対訳辞書を用いて評価する.

4.1 評価方法

EDR 英日対訳辞書は, 日本語訳を示すとともに見出し語の概念を持つ. この概念を**正解概念**と呼ぶ. 見出し語に対して複数の訳語が与えられていることが多く, これらより複数の概念セットを作成し, 本手法により共通概念を推定する. これを**推定概念**と呼ぶ. この推定概念と正解概念を比較することで, 本手法を評価する. 比較方法としては, 以下の 3 種類を考える.

1. シソーラス上で正解概念と推定概念を結ぶ経路の枝数を概念間距離とする. これが小さいほど 2 概念の意味は近い.

2. シソーラス上で, 正解概念と推定概念から共通上位概念までの距離を d_1,d_2とし,$r_1 = 2/((d_1 + 1) + (d_2 + 1))$ を 2 概念間の類似度 r_1とする. この値が大きいほど 2 概念の意味は近い.

3. シソーラス上で, 正解概念と推定概念, 及びこれらの共通上位概念のルートからの深さをそれぞれ d_1,d_2,d_cとし, $r_2 = (d_c * 2)/(d_1 + d_2)$ を 2 概念間の類似度 r_2とする. この方法は「深くに位置する概念ほど意味が細か

くなる」というシソーラスの特徴を考慮した方法である [5]. r_2の値が大きいほど2概念の意味は近い.

評価実験に用いるデータは,EDR 英日対訳辞書の全289,882 レコードの内, 概念セットを2つ以上作成出来た28,763 レコードである.

4.2 評価結果

本手法の評価として, 正解概念と推定概念との類似度の統計的評価, ノードスコア付けの方法1〜3の評価, 評価方法の検討の三点を述べる.

・正解概念と推定概念との類似度の統計的評価

本手法の有効性を示すためには, 本手法で得た推定概念と正解概念との類似度が, シソーラス上で無作為抽出した2概念間の類似度よりも有意に大きいことが, 統計的に言える必要がある. そこで前者を求める予備実験を行った. シソーラス中の任意の2概念のペアを, 無作為に500組を選び,2概念間の類似度の平均を求めた. この作業を200回繰り返し,500組毎の平均距離を200個得てヒストグラムを作成した. また, 全10万ペアについて類似度別の頻度グラフを作成した. 本手法の評価実験についても同様のグラフを作成した. 図1は, 評価方法1について,500組毎の平均距離200個に対するヒストグラムであり, 図2,3,4は, それぞれ評価方法1〜3についての類似度別頻度グラフである. いずれもノードスコア付けの方法2を用いた結果である.

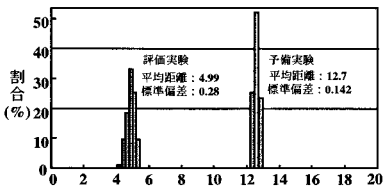

図 1: 評価方法1の平均距離のヒストグラム

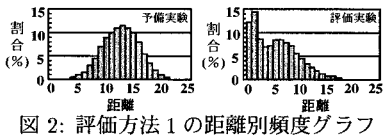

図 2: 評価方法1の距離別頻度グラフ

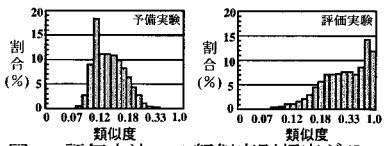

図 3: 評価方法2の類似度別頻度グラフ

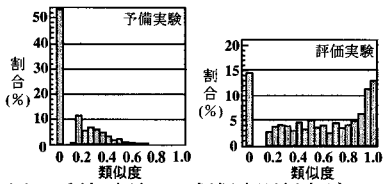

図 4: 評価方法3の類似度別頻度グラフ

図1を見ると, 予備実験で得た基準値(平均距離)が12.7であるのに対して, 評価実験で得た本手法の結果(平均距離)は4.99であり, 距離が7.7だけ(39 %に)縮まった. また図2を見ると, 予備実験の距離別分布は平均距離12.7の正規分布であるが, 本手法の評価実験では距離0と1が最も頻度が高く, 全体的に左上がりの分布となった. 無作為抽出した場合に比べ, 本手法により推定した共通概念は, 大きく正解概念に近付いたと言える. 同様に評価方法2と3についても, 基準値(平均類似度)が0.14と0.13であるのに対し, 本手法の結果では0.41と0.56であり, 大きく類似度が上がった(それぞれ2.9倍,4.3倍に上昇). また図3,4からも分布が大きく類似度の高い方向に寄っていることが判る. 図4において, 類似度0が約53 %もあるが, これは評価方法3では対象の2概念の共通上位概念がルートとなった場合は2概念間の類似度を0としているからである.

・ノードスコア付けの方法1〜3の評価

図5はそれぞれの方法による, 評価方法1の距離別頻度をグラフにしたものである. 図5を見ると, 平均距離は方法1が最も小さく, 方法3が最も大きい. しかし, 正解概念と推定概念の完全一致率(距離0)は方法3が最も高い. 従って完全一致率を重視するか, 平均距離を重視するかにより, 方法1〜3を使い分けることが出来る.

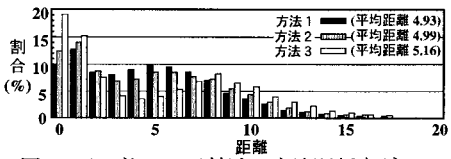

図 5: ノードスコア付けの方法別頻度グラフ

・評価方法の検討

評価方法1と2は考え方は同じで,2概念の類似性を距離と類似度で表している. 共に2概念がツリー上のどこに配置されていても, 概念間距離が同じならば,2概念間の類似性は同等に評価される方法である. これに対して評価方法3は, 概念間距離が同じでも2概念が深い位置に配置されている方が類似度が高くなる方法であり, シソーラスの特徴をよく捉えた評価方法だと言える. 但し, もともと概念終端群が浅い位置に密集している場合は, 類似度は高くなりにくくなるという欠点もある.

5. まとめ

対訳辞書の見出し語の概念を, 複数の概念を持つ訳語のセットから推定する手法を提案した. また, 見出し語の概念が与えられている英日対訳辞書と, 訳語の概念を獲得するために日本語単語辞書を用いて本手法の評価を行った. ここでは推定概念と正解概念の類似度を3種類の評価方法で比較した. その結果, いずれの評価方法についても推定概念が正解概念の近似解を与えていることを確認した. さらにノードスコア付けの方法1〜3と,3種類の評価方法についても検討した.

今後は, 本手法の改良とともに, 訳語が一つしかない場合の見出し語の概念推定法の追加を予定している.

なお, 本研究の一部は文部省科学研究費補助金(基盤研究 C2No.13680492)によって実施したものである.

参考文献

[1]EDR 電子化辞書仕様説明書. 日本電子化辞書研究所(1996).

[2] 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,pp23-30(1995).

[3] 大井耕三, 隅田英一朗, 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).

[4] 山西公一朗, 小島一秀, 渡部広一, 河岡 司 : 国語辞書の意味分類を利用した概念ベースにおける多義概念の分割. 情報処理学会自然言語処理研究会報告書 2001-NL-145,pp37-44(2001).

[5] 長尾 真 : 自然言語処理. 岩波書店(1996).