自 然 言 語 処 理 153－19
（2003．1．21）

国語辞典とシソーラスの統合

正津康弘　徳永健伸　田中穂積
東京工業大学 大学院情報理工学研究科

自然言語処理において，言語知識は非常に重要な役割を果たす．しかし，言語知識を人手によって構築するには多大な労力が必要である．そこで本論文では，既存の言語知識を組み合わせることによって，言語知識を自動的に拡張する手法を提案する．具体的には，日本語の国語辞典における名詞の語義とシソーラスの意味クラスの間の対応付けを自動的に行う．この手法では，語釈文から抽出した上位語の出現頻度と語釈文中のキーワードのTF-IDFの２種類のスコアを用いて，対応関係の尤もらしさを見積もる．この手法で実験を行った結果，被覆率は99.3%，精度は59.2%となった．

The integration of Japanese dictionary and thesaurus

Shoutsu Yasuhiro　Tokunaga Takenobu　Tanaka Hozumi
Department of Computer Science, Tokyo Institute of Technology

Linguistic knowledge plays crucial role in natural language processing. Constructing large linguistic knowledge requires a lot of human effort and much cost. This paper describes algorithms to enlarge existing linguistic knowledge automatically by combining with another linguistic knowledge. More concretely, this algorithm links a word sense defined in a monolingual dictionary to semantic classes in a thesaurus. Experiments showed that the precision of linking was 59.2% and its coverage was 99.3%.

1. はじめに

情報検索や翻訳などの自然言語処理アプリケーションにおいて，言語知識は非常に重要な役割を果たす．しかし，言語知識を人手によって構築するには多大な労力が必要となる．

そこで本論文では，既存の言語知識の間の対応関係を自動的に獲得し，それらを組み合わせることによって，言語知識を拡張する手法を提案する．具体的には，既存の言語知識として日本語の国語辞典と日本語の語を分類したシソーラスを利用し，国語辞典で定義された名詞の語義とシソーラスの名詞の意味クラスの対応関係を自動的に付けることによって言語知識を拡張することを目的とする．

国語辞典において，見出し語の語義は自然言語によって表現されており，それは見出し語自身に関する記述が中心である．一方，シソーラスにおいては，語はあらかじめ定められた意味クラスに沿って分類されており，他の語との関係を中心に整理されているという特徴を持つ．このように，国語辞典とシソーラスはそれぞれ別の観点から語に関する記述をしている．このような異なる性質の言語知識の間の対応関係を同定することによって，語に関する知識を豊かにすることができる．

東京工業大学 大学院情報理工学研究科 田中研究室
Tanaka Laboratory, Department of Computer Science, Tokyo Institute of Technology.
〒152-8552 東京都目黒区大岡山2-12-1
E-mail: shotsu@cl.cs.titech.ac.jp

2. 言語データ

本論文では，国語辞典として RWCP によって形態素タグが付与された岩波国語辞典第５版のデータ[5]を使用した．また，シソーラスとして NTT によって作成された日本語語彙体系[6]を使用した．

岩波国語辞典は 49742 語の名詞の見出し語を持っている．各見出し語には１つ以上の語義が定義されていて，それらの総数は 60194 個となっている．

一方，日本語語彙体系は，2710 の名詞意味クラスを定義しており，48970 語の名詞をこれらの意味クラスに分類している．各語は１つ以上の意味クラスに所属しており，語が所属する意味クラスはその語の上位概念に相当する．また，日本語語彙体系は意味クラスをノードとする木構造を成しており，親の意味クラスは子の意味クラスよりも上位の概念を表している．日本語語彙体系の一部を図１に示す．

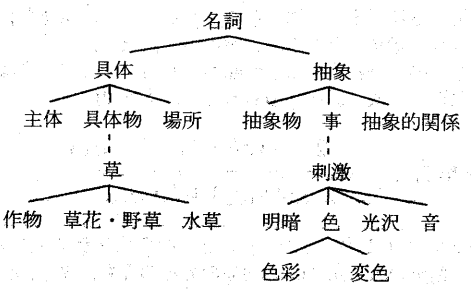

図１．日本語語彙体系

3. 上位語の抽出

本論文で提案する手法は，国語辞典の語釈文から見出し語の上位語を抽出し，それを利用して対応付けを行う．本節では，語釈文から上位語を抽出する方法について述べる．

国語辞典において，語釈文の第1文目の末尾に現れる語は見出し語の上位語になっていることが多い[7][8][9]（例：「アーク灯」の語釈文：[放電を利用した電灯。]）．場合によっては，末尾ではなく，その付近の語が上位語になっていることもある（例：「アイシャドー」の語釈文：[まぶたに青色・茶色などの化粧をすること。]）．また，名詞ではないが語義の核となっているような語が末尾付近に現れることもある（例：愛飲：[好んで飲むこと。]）．

そこで，語義の語釈文の第1文目の末尾が次のような形をしているとき，〈〉の部分に現れる名詞，動詞，形容詞，副詞を上位語として抽出する．

- 〈 〉。
- 〈 〉こと。
- 〈 〉すること。
- 〈 〉をすること。
- 〈 〉の一つ。
- 〈 〉の一種。
- 〈 〉の略。
- 〈 〉の～称。

抽出した語の中には同義語や下位語，または名詞ではない語なども含まれているが，本論文ではこれらもまとめて（語釈文から抽出した）上位語と呼ぶことにする．

4. パス長によるリンク

本節では，見出し語と上位語の意味クラス間のパス長を基にした対応付けの手法について述べる[1]．

4.1 名詞の上位語によるリンク

語釈文から抽出した上位語が名詞の場合，対応付けの手法として，見出し語の所属する意味クラスと上位語の所属する意味クラスが意味的に最も近くなるような意味クラスの組を見つけ，その見出し語側の意味クラスを語義に対応付ける，という手法が考えられる．

意味クラス間の意味的な近さを見積もる方法としてまず考えられるのが，2つの意味クラスをつなぐパスの長さを利用する方法である．しかし，日本語語彙体系は葉の意味クラスの深さが均一ではないため，パス長をそのまま利用すると問題が生じる．そこで，意味的な近さを次の式で見積もることにする[2]．

$$Sim(c_1,c_2)=\frac{2\times level(MSCA(c_1,c_2))}{level(c_1)+level(c_2)}$$

level(c)は根の深さを1としたときの意味クラスcの深さ，MSCA(c_1,c_2)はc_1,c_2共通の先祖のうち，最も深い位置の意味クラスを表す．この Sim(c_1,c_2)を用いて，語義dに対応付ける意味クラス $Link_n$(d)を決定する．

$$Link_n(d)=\arg\max_{c_1\in Class(h_d)}\left(\max_{c_2\in Class(n_d)}S(c_1,c_2)\right)$$

$$S(c_1,c_2)=\begin{cases}0 & if\ level\ (MSCA\ (c_1,c_2))\le 3\\ Sim\ (c_1,c_2) & otherwise\end{cases}$$

h_dは語義dの見出し語，n_d はdの語釈文から抽出した上位語の名詞，Class(x)は xの所属する意味クラスの集合を表す．S(c_1,c_2)の最大値が0の場合は対応付けを行わない．MSCA(c_1,c_2)の深さが 3 以下の場合にS(c_1,c_2)を0にしたのは，語義dに対応付けられるべき意味クラスが Class(h_d)内に存在しないケースに対処するためである．

4.2 動詞の上位語によるリンク

語釈文から抽出した上位語が動詞の場合，4.1 節の手法では対応付けを行うことができない．なぜなら，日本語語彙体系では名詞と動詞は全く別の木に分類されているからである．

しかし，その動詞の意味に対応した名詞の意味クラスを見つけることができれば対応付けは可能である．共通の上位語の動詞を抽出できる語義を持つ見出し語のグループは，共通の意味クラスに所属していることが多い．例えば，語釈文の末尾が［飲むこと。］となっている見出し語のほとんどは，“飲み”という意味クラスに所属している．この意味クラスの分布を利用して対応付けを行う．具体的には，次の式で語義dに対応付ける意味クラス $Link_v$(d)を決定する．

$$fr(v,c)=\left|\{d\mid v=v_d\wedge c\in Class(h_d)\}\right|$$

$$P(c\mid v)=\begin{cases}0 & if\ fr(v,c)\le 1\\ \dfrac{fr(v,c)}{\sum_{c\in Class}fr(v,c_i)} & otherwise\end{cases}$$

$$Link_v(d)=\arg\max_{c_1\in Class(h_d)}\left(\max_{c_2\in Class}P(c_1\mid v_d)\cdot S(c_1,c_2)\right)$$

h_dは語義dの見出し語，v_d はdの語釈文から抽出した上位語の動詞，Class はシソーラスの全意味クラス，Class(x)は x の所属する意味クラスの集合を表す．S(c_1,c_2)は4.1節と同じ式で求める．P(c|v)は動詞vを上位語として抽出できる語義を持つ見出し語が意味クラスcに所属する確率を表す．意味クラスの近さを表すS(c_1,c_2)にP(c_1|v_d)を掛け合わせることによって，v_dの意味に対応するとは言い難いような意味クラスの優先度を下げることができる．

4.3 見出し語・上位語に対する操作

対応付けの精度と被覆率を高めるため，実験の際には見出し語と上位語に対して次のような操作を行った．

(1) 国語辞典の見出し語に異なる表記方法がある場合

は，全ての表記方法の意味クラスの中から最適なものを決定する（例：明（か）り→明かり，明り）．

(2) 本来は漢字で表記すべき上位語がひらがなで表記されている場合は，岩波国語辞典によって漢字に変換する．

(3) 名詞の上位語が〈こと〉〈事〉〈もの〉〈物〉〈略〉〈一つ〉〈一種〉〈方〉〈語〉〈～称〉ならば対応付けを行わない．また，動詞の上位語が〈する〉〈ある〉〈いる〉〈なる〉〈せる〉ならば対応付けを行わない．

(4) シソーラスに存在しない複合名詞が上位語になる場合は，複合名詞を構成する名詞の一番左の名詞から削除し，シソーラスに存在するかどうかをその都度調べる（例：〈一年|生|植物〉→〈生|植物〉→〈植物〉）．

4.4 実験結果

実験の結果，国語辞典の31672個の語義に対応付けを行うことができた．国語辞典に存在する名詞の語義の総数は60194個なので，被覆率は52.6％となる．

対応付けの結果から1000個の語義を選出し，2名の判定者によって対応付けの正解不正解を判定し，手法の精度を求めた．その結果，判定者Aによる評価では精度91.0%，判定者Bによる評価では精度91.3%，平均すると91.2%となった．また，判定者A,Bの少なくとも一方が正解だと判定したものは93.7%，判定者A,Bの両方が正解だと判定したものは88.6%となり，κ統計量（$\kappa=(P_a-P_e)/(1-P_e)$）[10]は0.684となった．

4.5 問題点

この手法では高い精度での対応付けが可能である．しかし，実験結果を見ればわかるように，この手法で対応付けを行うと，被覆率が非常に悪くなってしまう．これには4つの原因が考えられる．

原因(1)：この手法では，シソーラスに登録されていない見出し語に対して対応付けを行うことができない．例えば，「アート紙」という見出し語はシソーラスに登録されていないため，語義に対応付ける意味クラスの候補（Class(「アート紙」)）を得ることができない．

原因(2)：抽出した上位語がシソーラスに登録されていない場合には対応付けを行うことができない．上位語の所属する意味クラスの集合が得られないと，候補の意味クラスの中から対応付ける意味クラスを選択することができない．

原因(3)：抽出した上位語が4.3節の(3)に当てはまる場合には対応付けを行わない．4.3節の(3)の語は確かに見出し語の上位語だが，候補の意味クラスの中から対応付ける意味クラスを選択するために使うには，あまりにも抽象的過ぎる．

原因(4)：見出し語の所属する意味クラスの中に，語義に対応する意味クラスが1つも無い場合がある．例えば，「藍」という見出し語の語義の1つに［秋、穂状の赤い小花をつける，たで科の一年生植物。］というものがある．しかし，Class(「藍」)={“染料”，“色彩”}であるため，この語義に対応付ける意味クラスの候補は“染料”と“色彩”の2つのみとなってしまう．この手法では，このような場合，対応付けを行わないことが最良の選択となる．そのため，被覆率が下がる原因となる．

少ないデータで高い精度を得るためには，この被覆率の低さは避けられない事柄なのかもしれない．しかし，言語知識の構築という観点から見れば，既存の知識からの選択（つまり，見出し語が所属する意味クラスの中からの選択）よりも新たな知識の獲得の方が，より意義のある事柄だと考えられる．そこで第5節では，見出し語・上位語のシソーラスへの登録の有無に関わらず，シソーラス中の全意味クラスの中から語義に対応する意味クラスを選択する手法について述べる．

5. 頻度によるリンク

第4節で述べた手法は，既に見出し語が所属している意味クラスの中からしか語義に対応付ける意味クラスを選択できなかった．これは，新たな言語知識の構築という観点から見れば問題である．

そこで本節では，語釈文中の語の頻度を基にした対応付けの手法を提案する．この手法で対応付けを行えば，既存の知識に縛られることなく，新たな言語知識を獲得することができる．

この手法では，上位語の頻度と語釈文中のキーワードのTF-IDFという2つのスコアを用いて，対応付けの候補の尤もらしさを見積もる．

5.1 上位語の頻度

共通の上位語を抽出できる語義を持つ見出し語のグループは，共通の意味クラスに所属していることが多い．そして，この共通の意味クラスは，共通の上位語を抽出できる語義に対応する意味クラスであると予想できる．例えば，〈植物〉という上位語を抽出できる語義を持つ見出し語の多くは“草花・野草”という意味クラスに所属している．このことから，〈植物〉という上位語を抽出できる語義は“草花・野草”に対応する語義であると予想できる．この情報を用いれば，「藍」の語義［秋、穂状の赤い小花をつける，たで科の一年生植物。］に対して，「藍」が所属していない意味クラス“草花・野草”を対応付けることができる．このような尺度で，上位語wと対応付け候補の意味クラスcの間の関係の強さWF(w,c)を見積もり，対応付けの指標とする．

$$WF(w,c)=\left|\{d \mid w=w_d \wedge c \in Class(h_d)\}\right|$$

h_dは語義dの見出し語，w_dはdの語釈文から抽出した上位語，Class(x)はxの所属する意味クラスの集合を表す．つまり，WF(w,c)はcに所属する見出し語の語義のうち，上位語としてwを抽出できるものの数を表している．WF(w,c)の値が大きいほど，wとcの関連は強い．

〈植物〉のように頻繁に現れる上位語ならば，WF(w,c)だけでも適切な対応付けが可能かもしれない．しかし，上位語の多くは数個の語義からしか抽出できないため，WF(w,c)だけで上位語と意味クラスの関係の強さを正確に見積もることは難しい．そこで，語釈文から抽出した上位語 w をシソーラスで抽象化し，上位語の所属する意味クラス c'と対応付け候補の意味クラス c の間の関係の強さ CF(c',c)を見積もることによって，w と c の関係の強さを求める．

$$CF(c',c)=\sum_{d\in D(c',c)}\frac{1}{|Class(w_d)|}$$

$$D(c',c)=\{d\mid c'\in Class(w_d)\wedge c\in Class(h_d)\}$$

CF(c',c)の値が大きいほど，c'と c の関連は強い．

この 2 つの値（WF(w,c)と CF(c',c)）を用いて，上位語 w と対応付け候補の意味クラス c の間の関係の強さを表すスコア Sc1(w,c)を求める．

$$Sc1(w,c)=WF(w,c)+\frac{\alpha}{|Class(w)|}\sum_{c'\in Class(w)}CF(c',c)$$

CF(c',c)の方が WF(w,c)に比べて大きな値になることが多いため，値 α を用いてその差を補正する．今回は α=0.3 で実験を行った．

5.2 語釈文中のキーワードの TF-IDF

第 4 節の手法や 5.1 節のスコアは，語釈文から抽出した上位語のみを見て尤もらしさを見積もり，対応付けの指標としている．しかし，語釈文の中には上位語以外にも語義の要素としての働きを持つ語がたくさん含まれている．対応付けの精度を上げるためには，これらの語に関する情報も利用したい．そこで，語釈文中の各語と意味クラスの間の関係の強さを見積もり，それを対応付けの指標として利用する．

ある意味クラスに所属する見出し語の語釈文中に頻繁に現れる語は，その語義を特徴付ける語だと予想できる．例えば，“草花・野草”という意味クラスに所属する見出し語の語釈文中には〔花〕という語が頻繁に現れる．このことから，〔花〕という語を含む語義は“草花・野草”に対応する語義である可能性が高いと判断できる．このような尺度で，語釈文中の語 k と意味クラス c の関係の強さ tf(k,c)を求める．

$$tf(k,c)=\frac{|DK(k)\cap DC(c)|}{|DC(c)|}$$

$$DK(k)=\{d\mid k\in KW(d)\}$$

$$DC(c)=\{d\mid c\in Class(h_d)\}$$

h_d は語義 d の見出し語，KW(d) は d の語釈文中の名詞・動詞・形容詞・副詞の集合，Class(x)は x の所属する意味クラスの集合を表す．

しかし，語釈文中に頻繁に現れる語だからといって，必ずしも語義を特徴付ける語であるとは限らない．例えば，“草花・野草”という意味クラスに所属する見出し語の語釈文中には〔つける〕という語も頻繁に現れる．しかし，〔つける〕は“草花・野草”に所属する見出し語の語釈文中だけではなく，あらゆる語の語釈文中に現れる語である．つまり，〔つける〕という語を含む語義だからといって，“草花・野草”に対応する語義である可能性が高いとは言えない．このような尺度で語釈文中の語 k の重要性 idf(k)を求める．

$$idf(k)=\log\frac{|Class|}{|\{c\mid DK(k)\cap DC(c)\neq\phi\}|}$$

Class はシソーラスの全意味クラスを表す．

この 2 つの値（tf(k,c)と idf(k)）の積を，語釈文中の語 k と対応付け候補の意味クラス c の間の関係の強さを表すスコア Sc2(k,c)とする．

$$Sc2(k,c)=tf(k,c)\times idf(k)$$

5.3 対応付け

5.1 節と 5.2 節で述べた 2 つのスコア Sc1(w,c)と Sc2(k,c)を組み合わせて，語義 d に対応付ける意味クラス Link(d)を決定する．

$$Sim(d,c)=\left(Sc1(w_d,c)+\theta_1\right)\times\left(\sum_{k\in KW(d)}Sc2(k,c)+\theta_2\right)$$

$$Link(d)=\arg\max_{c\in Class}Sim(d,c)$$

w_d は語義 d の語釈文から抽出した上位語，KW(d) は d の語釈文中の名詞・動詞・形容詞・副詞の集合，Class はシソーラスの全意味クラスを表す．θ_1, θ_2 は Sc1(w_d,c)，ΣSc2(k,c)の一方が 0 のときに，もう一方のスコアで対応付けを行うための値である．今回は θ_1=1，θ_2=0.1 で実験を行った．

5.4 見出し語・上位語に対する操作

対応付けの精度と被覆率を高めるため，実験の際には見出し語と上位語に対して次のような操作を行った．

(1) 国語辞典の見出し語に異なる表記方法がある場合は，見出し語の所属する意味クラスの集合として，全ての表記方法の意味クラスの和集合を用いる．（例：Class(「明（か）り」)=Class(「明かり」)∩Class(「明り」)）

(2) 本来は漢字で表記すべき上位語の名詞がひらがなで表記されている場合，岩波国語辞典によって漢字に変換し，CF(c',c)を求める．

(3) 上位語が複合語の場合，複合語を構成する一番左の単語から削除し，それぞれの段階で Sc1(w,c)を求める．対応付けの際にはそれらの平均を用いる（例：Sc1'(〈一年|生|植物〉,c)=(Sc1(〈一年生植物〉,c)+Sc1(〈生植物〉,c)+Sc1(〈植物〉,c))/3)．Sc2(k,w)に関しては，複合語は構成語に分解して，

それぞれについてスコアを計算する.

5.5 実験結果

実験の結果, 国語辞典の 59799 個の語義に対応付けを行うことができた. 被覆率は 99.3%となる.

対応付けの結果から 2000 個の語義を選出し, 2 名の判定者によって対応付けの精度を求めた. その結果, 判定者 A による評価では精度 57.5%, 判定者 B による評価では精度 60.8%, 平均すると精度 59.2%となった. また, 判定者 A,B の少なくとも一方が正解だと判定したものは 63.6%, 判定者 A,B の両方が正解だと判定したものは 54.6%となり, κ は 0.814 となった.

Sim(d,c)の大きさで意味クラス c をソートしたときに, 上位 r 位以内に正解と判定された意味クラスが存在する語義 d の数を求め, その割合(%)を表 1 と図 2 にまとめた. or は少なくとも一方が正解と判定したもの, and は両方が正解と判定したものの割合を表している.

r	A	B	or	and
1	57.5	60.8	63.6	54.6
2	67.0	69.6	72.9	63.0
3	73.3	74.5	78.3	68.4
4	77.4	77.6	81.8	71.8
5	79.8	79.3	83.8	73.7
6	81.3	81.0	85.3	74.9
7	83.5	82.3	86.9	76.5
8	84.7	83.1	87.8	77.4
9	85.8	83.7	88.7	78.0
10	86.7	84.4	89.4	78.7

表 1. 頻度によるリンクの結果（表）

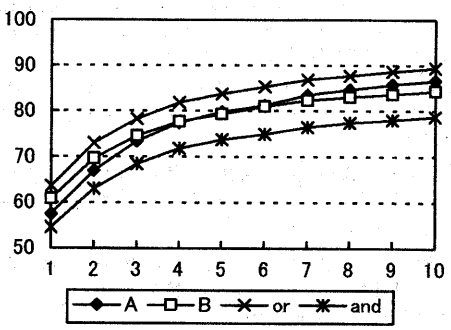

図 2. 頻度によるリンクの結果（グラフ）

6. 手法の比較と考察

本節では, 第 4 節のパス長による対応付けの手法と, 第 5 節の頻度による対応付けの手法の比較を行う.

まず, 両手法の精度と被覆率を比較する. 表 2 はパス長による対応付けの結果(P)と, 頻度による対応付けの結果(F)をまとめた表である。F の 1・5・10 は, それぞれ頻度による対応付けのスコア Sim(d,c)が上位 1・5・10 位以内のものを見たときの結果である. 精度は A,B の結果の平均である.

手法(r)		被覆率	精度	Or	and	κ
P		52.6	91.2	93.7	88.6	.684
F	1	99.3	59.2	63.6	54.6	.814
	5		79.6	83.8	73.7	
	10		85.6	89.4	78.7	

表 2. 手法の比較

表を見れば分かるように, パス長による手法は, 低い被覆率ながら, 高い精度での対応付けが可能である. 一方, 頻度による手法は, 100%に近い被覆率での対応付けが可能だが, その精度は低い. これは, パス長による手法では見出し語の所属する意味クラス数個の中から対応付ける意味クラスを選択しているのに対し, 頻度による手法ではシソーラス中の全意味クラス 2710 個の中から対応付ける意味クラスを選択しているからである. よって, 頻度による手法で得られた見出し語と意味クラスの対応関係の中には新規の情報も多く含まれるため, 単純に精度のみで性能を比較することはできない. また, 頻度による手法でも, 結果の上位 5 位・10 位まで見れば, その中に正解となる意味クラスが含まれる確率は格段に高くなる. これらの特徴から, 頻度による対応付けは, 新たな語をシソーラスに登録する際にその語が所属すべき意味クラスの候補を提示するなど, 人手による言語知識構築の労力軽減にも役立つと考えられる.

次に, 両手法で対応付けできた語義に対する精度と, 一方の手法でしか対応付けできなかった語義に対する精度を比較する. 表 3 の PF は両手法で対応付けできた語義, PO はパス長による手法でしか対応付けできなかった語義, FO は頻度による手法でしか対応付けできなかった語義を表す. 割合は（対応付けできた語義／国語辞典中の全語義×100）を表す.

	手法(r)		割合	精度	Or	and	κ
PF	P		52.6	91.2	93.7	88.6	.684
	F	1		74.7	78.4	70.9	.802
		5		92.1	94.4	88.8	
		10		95.0	96.7	91.7	
PO	P		0				
FO	F	1	46.7	41.8	47.1	36.4	.780
		5		65.9	72.5	57.1	
		10		75.7	82.5	65.2	

表 3. 重複部分の比較

パス長による手法で対応付けできた語義は, 全て頻度による手法でも対応付けできた. また, パス長による手法で対応付けできる語義に対しては, 頻度による手法でも比較的高い精度で対応付けできた（上位 5 位以内の精度では逆転している）. 反面, パス長による手法で対応付けできなかった語義に対しての頻度による手法の精度はかなり低い. これには 3 つの原因が考えられる.

原因(1)：実験結果を見たところ，上位語の頻度Sc1(w,c)が 0 になる語義に対しては精度が低かった．語義の核となる語から情報を得ることができないと，その語義が表すものが何なのか判断するのは難しい．

原因(2)：抽出した上位語が抽象的な語の場合の精度が低かった．パス長による手法では，抽象的な上位語を抽出した場合には対応付けを行わなかった．一方，頻度による手法では，そのような上位語からも何らかの情報が得られるものと予想し，対応付けを行った．しかし，そのような上位語から得られる情報は不十分な情報であったため，原因(1)と同じ結果になった．

原因(3)：対応する意味クラスがシソーラスに存在しない，もしくは人の目でみても対応する意味クラスを判定しづらい語義というものも存在する．例えば，「合縁奇縁」の語義［人の交わりには互いに気がよく合う合わないがあって、それは不思議な縁によるのだということ。］に対応する意味クラスを判断するのは難しい．

7. おわりに

7.1 まとめ

国語辞典の語釈文から抽出した上位語の頻度と，語釈文中の語の TF-IDF を用いて，国語辞典の語義とシソーラスの意味クラスの間の対応関係の尤もらしさを見積もり，尤もらしさが最大になる意味クラスを語義に対応付ける手法を提案した．その結果，この手法によって 59799 個の語義に意味クラスを対応付けることができた．被覆率は 99.3%，精度は 59.2%となった．また，尤もらしさが 10 位以内の意味クラスの中に正しい対応関係の意味クラスが含まれる率は 85.6%となった．これらの対応関係は，見出し語が所属する意味クラスの枠に縛られることなく得た情報である．そのため，この手法で獲得できる情報は，新たな言語知識として非常に価値のあるデータだと言える．

7.2 今後の課題

- 上位語の頻度情報が得られない場合や，上位語が抽象的過ぎる語である場合には，頻度による対応付けの手法の精度は低くなってしまう．よって，対応付けの精度を高めるためには，上位語をもっと広く拾う工夫をするか，もしくは語釈文中の他の語をもっと有効に利用する必要がある．前者の方法としては，「第 2 文目以降からも上位語を抽出する」「上位語の抽象化の度合いを工夫する」などといったものが考えられる．後者の方法としては，「語釈文の統語的・意味的構造に関する情報を利用する」などといったものが考えられる．

- 語義に対応する意味クラスがシソーラスに存在しない場合や，人の目で見ても対応する意味クラスを判定しづらいような場合でも，ほぼ全ての語義に対して何らかの意味クラスを対応付けていた．しかし，このような語義には対応付けを行わないのが望ましい．このような対応付け不要の語義を判定するために，対応付けの信用度を表すスコアを導入する必要がある．そして，対応付けの情報を利用する際に，利用者の要求する精度に応じて信用度の閾値を変化させ，対応付け情報の取捨をすると良いだろう．

- 本手法では，形態素解析された国語辞典と，ある程度の規模のシソーラスの 2 種類のデータがあれば対応付け可能である．しかし，この 2 種類のデータのみを使って対応付けの精度を上げるには限界がある．さらに精度を上げるためには，新たなデータを導入する必要があるだろう．新たのデータの導入の仕方としては，「意味タグ付きコーパスから各語の中心的な意味を獲得し，上位語の頻度や語釈文中の語の TF-IDF に反映させる」「正解不正解の判定が済んでいる対応付けのサンプルデータを用いて，スコアの重み付けにフィードバックする」などが考えられる．

参考文献

[1] 正津康弘，白井清昭，徳永健伸，田中穂積．国語辞典の語釈文の解析と語義のシソーラスへのマッピング．第15回人工知能学会全国大会論文集, 2B2-01, 2001

[2] Changki Lee, Geunbae Lee, Seo Jung Yun. Automatic WordNet mapping using word sense disambiguation. In (EMNLP/VLC-2000) Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 142-147, 2000.

[3] Changhua Yang, Sue J. KerConsiderations of linking WordNet with MRD. Proceeding of COLING-2002, 2002.

[4] Jen Nan Chen, Jason S. Chang. Topical clustering of MRD sense based on information retrieval technique. Computational Linguistics, 24(1):61-93, 1998.

[5] RWCP テキスト・サブ・ワーキンググループ. RWC テキストデータベース，RWC-DB-TEXT-96-2, 1996.

[6] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦．日本語語彙体系．岩波書店, 1997.

[7] 鶴丸弘昭，内田彰，日高達，吉田将．国語辞典からの情報抽出とその構造化．情報処理学会自然言語処理研究会，43(6), 1984.

[8] 鶴丸弘昭，水野浩司，内田彰，日高達，吉田将．単語の釈語文を利用した単語間の階層関係の抽出について．情報処理学会自然言語処理研究会，45(4), 1984.

[9] 鶴丸弘昭，兵頭竜二，松崎巧，日高達，吉田将．語義を考慮した単語間の階層関係の抽出について．情報処理学会自然言語処理研究会，64(2), 1987.

[10] Carletta J. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2):249-254, 1996.