和英辞書を用いたシソーラス細分化手法

堀口 賞一　　飯田 敏幸
NTTコミュニケーション科学研究所

1 はじめに

我々は、人間のように柔軟な理解や判断のできるコンピュータの実現を目指し研究を行なっている[1]。自然言語を柔軟に理解するためには、文法情報などの知識以外に語の意味を扱った膨大な知識が必要である。この知識としてシソーラスの利用が考えられる[2]。しかし、既存のシソーラスは意味別に分類され階層化されているものの、分類された単語群には同義、類義、上位下位関係(階層関係)にある単語が混在しているという問題があった。そこで、英語の同じ訳語又は訳語群を持つ日本語の単語は同義であり、訳語群が包含関係にある単語間には階層関係があると考え、上記問題を解決するために和英辞書を用いてシソーラスを細分化することを試みている。既に、机上評価によりこの方法の実現性の見通しを得ている[3]。この手法を大規模なシソーラスに適用し、得られたシソーラスについて評価したのでその結果について述べる。また、シソーラスと和英辞書の複数の組合せに対して評価実験を行ない、この手法が特定のシソーラスや和英辞書に依存しない手法であることを示す。さらに、今回明らかになった手法の改良点についても述べる。

2 シソーラスの細分化手法

2.1 基本的な考え方

シソーラスはノードとそのノードを表す単語および、階層関係にあるノード間のリンクから構成されている。ここでは特に、最下位のノードはそのノードを表す単語が1つのためにノードが単語に置き換えられているものとする。各ノードは単語を意味別に区分けしたものなので、以降ノードをカテゴリと呼ぶ。カテゴリ間の階層関係および各カテゴリを表す単語は正しいと考え、最下位のカテゴリに含まれる単語群に対してのみ、同じ意味を持つ単語を1つにまとめ階層化することによりシソーラスを細分化することを考える。

同じ意味を持つ単語を1つにまとめ階層化するためには、同義関係や階層関係を抽出しなければならない。同義関係や階層関係を抽出する手段として、国語辞書の語義文を利用することが考えられる。しかし、一般の国語辞書では同じ意味の見出し語の語義文でも表現の異なるものが多く、意味理解の技術が確立していない現状では、語義文を理解して同義関係や階層関係を抽出することは不可能である。また、単語の字面から単語同士の同義関係や階層関係を推定することも考えられるが、同一表記で複数の意味を表すもの(“身長”と“駅長”は同じ“長”を持つ単語であるが意味は全く異なる)も多く存在し単純に扱うことはできない。そこで、日本語と英語の対応関係を表している和英辞書に着目し、同じ訳語又は訳語群をもつ日本語の単語は同義であり、訳語群が包含関係にある単語間には階層関係があると考えた。

図1は細分化の考え方を示したものである。①は既存のシソーラスの一部で、“木材”、“材木”、“木”はカテゴリ<資材>に含まれる単語群を表している。②は和英辞書で、“木材”の訳語は“wood, timber, lumber”であることを表している。①と②を見比べた結果、“木”の5つの訳語の内“a tree”は“高木”の訳語、“a shrub”は“低木”の訳語、“wood, timber, lumber”は“木材”や“材木”の訳語であるので、<植物>に含まれる“高木”や“低木”の訳語“a tree, s shrub”と、<資材>に含まれる“木材”や“材木”の訳語“wood, timber, lunmer”とに“木”の訳語を分類できると考えられる。この考えに基づいて訳語を対応づけたものが③である。“木材”と“材木”と“木”は同じ訳語群“wood, timber, lumber”を持つので同義関係にあると考えられる。更に、“a tree, a shrub”を訳語とする“木”は“a tree”を訳語とする“高木”と“a shrub”を訳語とする“低木”とで表されるので、“高木”と“低木”は“木”を上位とする単語であると考えられる。この考えに基づいてシソーラスを細分化した結果の一部が④である。

2.2 訳語のシソーラスへの対応づけ

前節で示した細分化の考え方の中でキーとなるのは訳語をシソーラスへ対応づけることである。これは次の手順により実現できる。

(手順1) 1カテゴリにしか含まれない単語の訳語と1訳語しか持たない単語の訳語からなる訳語集合をカテゴ

A Method for Subdividing a Thesaurus with a Japanese-English Dictionary
Shouichi HORIGUCHI and Toshiyuki IIDA
NTT Communication Science Laboratories

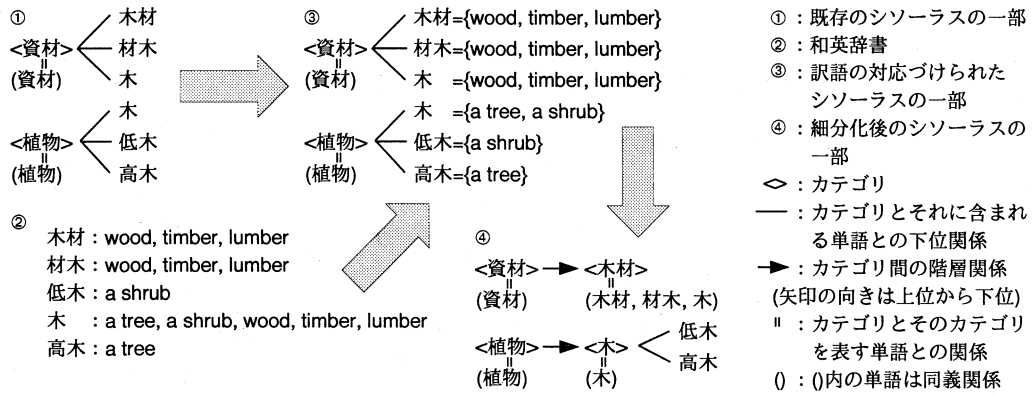

図 1: 細分化の考え方

リ毎に作成する。カテゴリ c に関する訳語集合を E^c とする。

(手順 2) 複数カテゴリに含まれる単語の訳語と (手順 1) で得られる訳語集合とを比較して、その訳語を含む訳語集合のカテゴリの単語とその訳語を対応づける。

カテゴリ c に含まれる単語の集合を W^c とすると、(手順 1) は具体的に次のように表される。

【E^c 作成アルゴリズム】

< step1 > E^c を空集合にする。
< step2 > W^c に含まれる全ての単語 w に対して次の処理を繰り返す。
$||C(w)|| = 1$ または $||T(w)|| = 1$ ならば、E^c に $T(w)$ を追加する。

但し、$C(w)$:w を含むカテゴリの集合
$T(w)$:w に対応する訳語の集合
$||*||$: 集合 $*$ の要素数

単語 w の訳語の集合 $T(w)$ をシソーラスに対応づけることは、対応づけられるべきカテゴリ c に含まれる訳語の集合 $T^c(w)$ を作成することに相当する。(手順 2) は具体的に次のように表される。

【$T^c(w)$ 作成アルゴリズム】

W^c に含まれる全ての単語 w に対して $T^c(w)$ を以下により求める。

$$T^c(w) = T(w) \cap E^c$$

2.3 同義関係および階層関係の抽出

前節で求めた $T^c(w)$ を用い、S を同義語の集合、B を上位語の集合、N を下位語の集合とすると、単語 w の同義関係および階層関係を抽出するアルゴリズムは次のように表される。

【w の同義関係および階層関係抽出アルゴリズム】

< step1 > B, N, S を空集合にする。
< step2 > W^c に含まれる全ての単語 $w_k(\neq w)$ に対して次の処理を行なう。

$T^c(w) = T^c(w_k)$ ならば、
S に w_k を追加する。

$T^c(w) \subset T^c(w_k)$ ならば、
B に w_k を追加する。

$T^c(w) \supset T^c(w_k)$ ならば、
N に w_k を追加する。

3 評価実験

3.1 評価実験で用いたシソーラスと和英辞書

シソーラスとして機械翻訳用に開発されたもの [4]、学研の国語シソーラス電子辞書及び角川の類語国語辞典 (第四版) を用い、和英辞書として学研の和英電子辞書と研究社の新和英中辞典 (第三版) を用いて実験した。各々のシソーラスおよび辞書の特性を表 1、表 2に示す。尚、計算機可読でない 2 つの辞書については人手で辞書引きし机上で評価実験を行なった。

表 1: 評価実験に用いたシソーラス

	機械翻訳用	学研国語	角川類語
登録語数	約150,000	約85,000	約60,000
カテゴリ数	約2,800	約5,500	約1,000
辞書形態	MRD	MRD	書籍
備考	一般名詞	全品詞	全品詞

注) MRD は計算機可読辞書。

表 2: 評価実験に用いた和英辞書

	学研和英辞書	新和英中辞典
見出し語数	約45,000	約75,000
辞書形態	MRD	書籍

3.2 評価法

評価実験は3つのシソーラスと2つの和英辞書の各組合せの中から4組を選んで行なった (表4参照。但し、翻訳用のシソーラスと研究社の新和英中辞典との組は [3] で行なったもの)。シソーラスから5つのカテゴリをランダムに取り出し、それらのカテゴリに含まれる単語の中で訳語を持つものについて、同義関係および階層関係を人手で抽出したものと本手法を適用して抽出したものとを比較し、適合率と再現率により評価した。

3.3 訳語のシソーラスへの対応づけの評価

表3から約98 %の適合率で訳語がシソーラスへ対応づけられることが分かる。しかし、複数カテゴリに含まれる単語の訳語に着目してみると、対応づけられない訳語が約50 %もある。これは対応づけられるべきカテゴリに同じ訳語を持つ単語が存在しないためである。また、訳語をシソーラスに対応づけられないために誤った同義関係を抽出した例が見受けられた。例えば、

T(犬) = {a dog, doggie, a spy}
T(小犬) = {a doggie}

などである。"(動物の) 犬" の訳語は "a dog" であるが、動物のカテゴリに "a dog" を訳語とする単語が存在しないためにシソーラスへ対応づけられない。一方、"a doggie" は動物のカテゴリ、"a spy" はスパイのカテゴリへ各々対応づけられる。従って、"犬" と "小犬" は階層関係として抽出されるべきであるにもかかわらず、$T^{動物}$(犬) = {a doggie} と $T^{動物}$(小犬) = {a doggie} より誤って同義関係として抽出されてしまうことになる。そこで、"a dog" を "(動物の) 犬" の訳語としてシソーラスへ対応づける必要がある。

動物のカテゴリの中には "a house dog" などの修飾語を伴った訳語が多く存在する。この "a house dog" のうち、中心的な意味を担う単語 "dog" を取り出し[4]、中心的な意味を担う単語同士を比較することにより訳語をシソーラスへ適切に対応づけられると考えられる。この改良によりシソーラスへ対応づけられない訳語数を減少させることができる。

3.4 同義関係および階層関係抽出の評価

表4において抽出された同義関係は約90 %の適合率であり、本手法が同義関係の抽出に有効であることを示している。しかし、再現率については約60 %と低い。その原因を調べてみると、同じ意味の訳語であるにもかかわらず、省略形や口語表現、語形変化などの訳語表記の違いが原因であることが分かった。代表例を以下に示す。

【省略】 T(常盤木) = {an evergreen}
T(常緑樹) = {an evergreen tree}
【口語表現】 T(母親) = {mother}
T(ママ) = {mama}
【語形変化】 T(広葉樹) = {a broadleaf tree}
T(闊葉樹) = {a broad-leaved tree}

これらに対する対策は今後の課題である。

一方、抽出された階層関係は適合率も再現率も共に低く、特に再現率は10 %未満と非常に低い。このことから分かるように、今回提案した手法のみでは階層関係を十分に抽出することはできない。階層関係を抽出するためには他の方法 [4]、[5] なども検討する必要があり、これについても今後の課題である。

3.5 シソーラスと和英辞書への依存性の評価

表4から分かるように、シソーラスと和英辞書のどの組合せの評価実験においても抽出された同義関係は高い適合率である。このことは、特定のシソーラスや和英辞書に依存せずに高い適合率で同義関係を抽出できることを示している。

4 おわりに

評価実験の結果、提案した手法により高い適合率で同義関係を抽出できることを示した。これにより、同じ意味を持つ単語同士を1つにまとめることができる。また、特定のシソーラスや和英辞書に依存せずに高い適合率で同義関係を抽出できることを示した。しかし、複数カテゴリに含まれる単語の訳語のうち約50 %がシソーラス

表 3: 訳語のシソーラスへの対応づけの評価結果

シソーラス		学研国語	角川類語	角川類語
和英辞書		学研和英	学研和英	研究社
適合率 (%)		96.1	99.2	98.7
再現率 (%)		76.7	94.5	93.3
一意に対応づけられた訳語数		138(58.5)	362(87.4)	507(85.1)
複数カテゴリに含まれる単語の訳語数		98(41.5)	52(12.6)	89(14.9)
	適切に対応づけられた訳語数	39(39.8)	23(44.2)	41(46.1)
	誤って対応づけられた訳語数	7(7.1)	5(9.6)	11(12.4)
	対応づけられなかった訳語数	52(53.1)	24(46.2)	37(41.6)
合計		236(100.0)	414(100.0)	596(100.0)

注) () 内の数値は百分率。

表 4: 同義関係および階層関係抽出の評価結果

シソーラス		学研国語	角川類語	翻訳用	角川類語
和英辞書		学研和英	学研和英	研究社	研究社
対象語数		148	264	810	231
平均語数 (語 / カテゴリ)		29.6	52.8	162.0	46.2
同義関係	関係数 (組)	19	33	241	13
	適合率 (%)	89.5	91.0	96.0	92.3
	再現率 (%)	54.8	65.2	69.0	66.7
階層関係	関係数 (組)	13	38	102	22
	適合率 (%)	30.7	39.5	85.0	63.6
	再現率 (%)	4.8	8.8	——	17.1

注) 対象語数はカテゴリに含まれる単語の中で訳語を持つ単語数。

へ対応づけられないことや、省略形、口語表現、語形変化などの訳語表記の違いが原因で抽出した同義関係の再現率は約 60 %しかないという問題点が顕在化した。更に、階層関係の抽出に関しては再現率が 10 %未満と低く、提案した手法のみでは階層関係は抽出できないことが分かった。2.1 節において国語辞書の語義文の意味を理解して同義関係や階層関係を抽出することは難しいと述べた。しかし、語義文の意味を理解するのではなく、語義文中に現れる「～の旧称。」、「～の総称。」などといった特徴的な表現に着目し、見出し語とその見出し語を説明する中心的な単語との同義関係や階層関係を抽出することは考えられる [5]。

今後、本稿で提案した手法とその他の手法 [4]、[5]、[6] などを併用し、各々の手法の特徴を利用したシソーラスの細分化を行なう予定である。

参考文献

[1] Iida,T. et al:Artificial Intelligence for Semantic Understanding, Proceedings of the IFIP Congress 94, Vol.2, pp.137-142, 1994.

[2] 田中他: 上位 / 下位関係シソーラス ISAMAP1 の作成 [I][II], 情報処理学会研究会報告, 87-NL64-4, 1987.

[3] 堀口他: 対訳辞書を利用した同義語辞書作成手法, 情報処理学会第 49 回全国大会, 3G-10, 1994.

[4] 池原他: 日英機械翻訳における利用者登録語の意味属性の自動判定, 言語処理学会誌, Vol.2, No.1, pp.3-17, 1995.

[5] 鶴丸他: 語義を考慮した単語間の階層構造の抽出について, 情報処理学会研究会報告,87-NL64-2,1987.

[6] 鶴丸他: 国語辞書に基づくシソーラスの構築に関する一考察, 電子情報通信学会技術研究報告, NLC93-58, 1993.