対訳辞書を利用した同義語辞書作成手法

3G－10

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

我々は人間の持つ柔軟な理解や判断を計算機上に実現する研究をしている[1]．理解，特に自然言語理解を実現するためには，数十万語規模の体系づけられた知識が必要とされる．そこで，我々は自然言語処理で使われるシソーラスを知識として利用することを考えている．しかし，大規模なシソーラスでは数千語程度の主な語に関して階層関係を構築し，その下に残りの数十万語が分類されているのが現状である．特に，最下位の階層には沢山の語が配置されてしまい，各語の間には，同義，類義の関係にあるものや全く関係ないものもある．従って，このようなシソーラスをそのまま意味理解のための知識として利用することはできない．そこで，同じ訳語を持つ単語同士は意味が等しいと仮定し，対訳辞書を用いて分類の不十分な単語群を同義語の集合に分類し，その集合間の関係を整理する手法を提案する．これは，シソーラス上の単語群を更に細かく分類し，シソーラスを整理することに相当するので，同義語分類という観点でシソーラスを細分化することと捉えることができる．本稿では，このように細分化された辞書を同義語辞書と呼ぶ．

2 シソーラス

2.1 対象とする意味的関係

語1と語2が同義であるとき，両者は等価関係にあると呼ぶ．語1が語2の性質を持っているとき，語1を下位語，語2を上位語と呼び，この上位下位関係を階層関係と呼ぶ．例えば，“犬”は“動物”の性質を持っているので，“犬”を下位語，“動物”を上位語とすることができる．シソーラスは，このように語と語の意味的関係を等価関係および階層関係などで表現したものである．シソーラスには両関係以外に全体部分関係や連想関係などもあるが[2]，本稿では等価関係と階層関係のみを細分化する手法を提案するので，以降，この2つの関係についてのみ述べる．

A Method for Subdividing Thesaurus with Bilingual Dictionaries
Shouichi HORIGUCHI, Toshiyuki IIDA
NTT Communication Science Laboratories

2.2 シソーラス上の単語の意味

一般に単語は複数の意味を持つ．例えば，広辞苑（第4版）では，“車”は，(1)「軸に貫いて回転する仕組みの輪」と(2)「車輪の回転によって動く仕掛けのものの総称」という2つの意味を持つ．シソーラス上に単語を位置づけることにより，その単語は上位語の性質を持つ意味を表すと考えることができる．“車”を“輪”の下位語として位置づけることにより，(1)の意味のみを表すようにすることができる．同様に，“乗り物”の下位語として位置づけることにより，(2)の意味のみを表すようにすることができる．従って，シソーラス上の単語の意味は一意に決定されると考えることができる．

3 シソーラスの細分化

3.1 対訳辞書の利用

対訳辞書（和英辞典や英和辞典など）では，見出し語に対して意味別に訳語の集まりを対応させている．例えば，研究社新和英中辞典（第3版）では，見出し語“車”は，(1)「(車輪の意味で) wheel」，(2)「(ピアノなど重い家具につける意味で) castor」，(3)「(乗り物の意味で) vehicle」，(4)「(自動車の意味で) car, motorcar, automobile」などとなっている．しかし，見出し語の意味分類は普遍的ではなく，シソーラスの分類と必ずしも一致しない．そこで，見出し語の意味分類をはずし，1つの見出し語に異なる意味がある場合でも訳語を区別せずに利用する．

3.2 細分化の方法

1つの上位語しか持たない見出し語と，1つの訳語としか対応しない見出し語に関しては，その訳語は見出し語の上位語の性質を必ず持つと考えられる．これより，この訳語を用いて他の見出し語に対応する複数の訳語を意味別に分類することを考える．図1に対訳辞書を用いてシソーラス上の単語に訳語を対応させた例を示す．ただし，矢印は階層関係を，＊印は1つの上位語しか持たない単語を，□枠に囲まれた訳語は上位語の性質を必ず持つことを，／印の付いた訳語は他の上位語の性質を持つことを表している．“木材”は上位語を1つしか持た

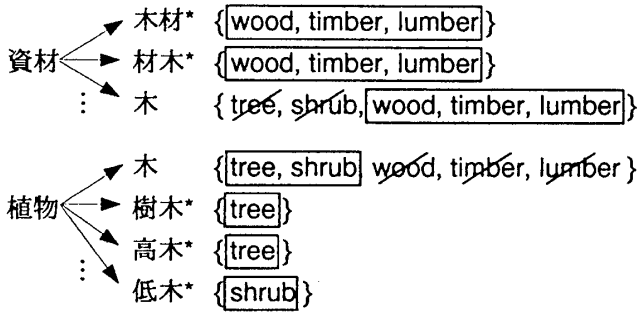

図 1: 対訳辞書の利用

ない単語であるため，“wood”，“timber”，“lumber”は“資材”の性質を持つ訳語であることが分かる，同様に，“tree”，“shrub”は“植物”の性質を持つ訳語であることが分かる．これより，“資材”の下位語の“木”に“wood”，“timber”，“lumber”を，“植物”の下位語の“木”に“tree”，“shrub”を対応させることができる．このようにして，訳語を意味分類し，他の上位語の性質を持つ訳語をはずす．

シソーラス上の単語の持つ意味は一意に決まるので，同じ上位語を持ち，同じ訳語と対応する単語同士は等価関係にあると考えることができる．例えば，“木材”と“材木”と“木”を等価関係としてまとめ，シソーラスを細分化する．

また，複数の訳語に対応する単語では，各訳語がその単語の性質を持つと考えることができる．この訳語の中の１つの訳語のみと対応する単語があれば，その単語も必ず元の単語の性質を持つと考えられる．これは，両者が階層関係を持つことを意味しているので，複数の訳語に対応する単語を上位語とすることにより，階層関係の細分化をする．例えば，“木”を“樹木”，“高木”，“低木”の上位語とすることができる．なお，“樹木”と“高木”は等価関係となる．

4 実験結果

機械翻訳用に開発されたシソーラス [3] を対象として，シソーラスより５個の上位語をランダムに抽出し，その下位語からなる単語群に対して本手法を適用する実験を行なった．対訳辞書としては研究社新和英中辞典 (第３版) を利用した．評価方法は以下の通りである．

1) 等価関係に関しては，人手により細分化した関係と本手法を適用して細分化した関係とを比較する．

2) 階層関係に関しては，本手法を適用 して細分化した関係の適否を人手によりチェックする．

その結果を表１に示す．

表１より等価関係の適合率が非常に高いことが分かる．これは，同じ訳語を持つ単語同士は意味が等しいとした仮説が正しいことを意味している．一方，等価関係の再現率は低い．その原因に訳語の省略や口語表現がある．例えば，“母”と“ママ”は同一の上位語“女性”を持っているが，対訳辞書では“母”は“mother”のみに対応し，“ママ”は“mama”のみに対応する．従って，両者は同義であるにもかかわらず，等価関係は作成されなかった．

適合率が高いため，作成された等価関係および階層関係はそのまま利用できることが分かる．

表 1: 等価関係と階層関係の評価結果

上位語	下位語数	等価関係			階層関係	
		α	β	γ	α	β
案	114	64	93.8	86.5	90	86.7
企業	199	39	97.4	79.6	0	—
女性	234	99	96.0	56.9	12	83.3
獣	161	25	100.0	65.8	0	—
電気機器	102	14	92.9	56.0	0	—
計	810	241	96.0	69.0	102	85.0

α: 関係数，β: 適合率 (%)，γ: 再現率 (%)

5 おわりに

評価実験の結果，提案した手法では，等価関係の再現率は低いものの，非常に高い適合率が得られることが分かった．これより，等価関係と階層関係についてシソーラスを細分化できる見通しが得られた．

今後，シソーラス全体に適用し，本手法の検証を行なう．

参考文献

[1] Iida,T., Shimada,S., Ohta,M., and Kawaoka,T.: Artificial Intelligence for Semantic Understanding, Proceedings of the IFIP Congress94, Vol.2,1994(予定).

[2] 内藤, 中倉, 影浦 他訳: シソーラス構築法, 丸善,1989.

[3] 池原他: 日英機械翻訳のための意味解析用の知識とその分解能, 情報処理学会論文誌,Vol.34,No.8,1993.

[4] 大野, 柴田 編: 日本語 (語彙と意味), 岩波書店,1977.

[5] 徳永他: 対訳辞書からの概念項目の自動抽出, 人工知能学会誌,Vol.6 No.2,1991.