2C－3

並列名詞句と連体修飾語の曖昧性解消*

細井 貴晴　山口 昌也　乾 伸雄　小谷 善行　西村 恕彦
(東京農工大学　工学部　電子情報工学科)

1　はじめに

様々な曖昧性解消における研究が、過去に成されてきた。本研究では、並列名詞句と連体修飾語の曖昧な文を扱うが、とりわけ並列構造の曖昧性に関する研究報告としては、[4] などの文献がある。本研究では、シソーラス [2] の類似度を用いて並列名詞句と連体修飾語の選択に関する曖昧性解消を行う。これに加えて、シソーラスにない単語をシソーラスに割り当てる方法を提案し、その単語を含んだ文節の係り先の曖昧性を解消する。本稿では、新聞における「「名詞 I (A)」「と」「名詞 II (B)」「の」「名詞 III (C)」」の文節を対象とし、その文節の前後の情報は考慮しないこととする。

2　並列名詞句、連体修飾語

日本語で、「A < 接続助詞 > B < 連体助詞 > C」という句が現れたとき、一般に、A と B が名詞句になるかどうかで、'構文的および意味的曖昧さが生じる'。本論文では、下記の二つの仮定に基づき、並列助詞「と」、連体助詞「の」の場合について扱う。

仮定 1 並列される二つの単語は、シソーラスにおいて類似度が高い

仮定 2 名詞句「B の C」は、B と C の類似度が高いと一つの名詞とみなせる

という二つの仮定に基づき、シソーラスを用いて、並列名詞句と連体修飾語の作る曖昧性解消を行う。

3　シソーラス

本システムで用いたシソーラスは、図 1のような構造になっており、これを用いてシステムを作成した。

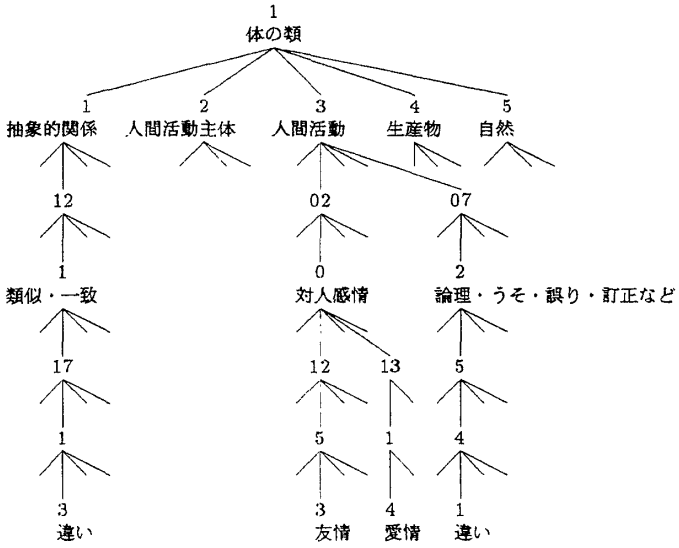

図 1: シソーラスの概念図

3.1　本研究におけるシソーラスの活用法

システムを作成する上での幾つかの定義を示す。

定義 1 二つの語の類似度は、枝分かれしているところとする

例えば、図 1から、**'友情'** と **'愛情'** の類似度を調べると、階層 4 (上から 4 番目) の **'対人感情'** のところで枝分かれしていることにより、類似度は 4 となる。

このことにより、本研究においては単語間の類似度は、数字が大きいほど高いということが言える。

次に、多義語の扱いについて述べる。例えば **'違い'** という単語は、図 1から分かるように二つの意味を持っている。このとき、候補を 1 つに絞ろうと考えると

定義 2 単語間の距離は、より近いものを選好する

ことにする。例えば、**'友情'** と **'違い'** の類似度を調べると、**'違い'** の 1 つは、階層 2 の **'人間活動'**、もう 1 つは、階層 1 の **'体の類'** のところから枝分かれ

*Disambiguation between coordinate-noun phrases and noun modifiers,
Takaharu HOSOI, Masaya YAMAGUTI, Nobuo INUI, Yoshiyuki KOTANI, Hirohiko NISIMURA,
Tokyo University of Agric. and Tech., Dept. of Computer Science

しているので、それぞれは、類似度 2 、類似度 1 となる。この場合は、類似度 2 の方を選ぶことになる。

4 解消方法

4.1節に示すアルゴリズムにより、「A と B の C」の文節の解析を行う。もし、{ A,B,C } のうち一つの語が、シソーラス上になければ 4.2節での評価を行う。

4.1 アルゴリズム

1. 助詞「と」に着目し、単語「A」と「B」、「A」と「C」、「B」と「C」の類似度を調べる
2. 類似度 (A,B) > 類似度 (A,C) で、類似度 (A,B) >= 類似度 (B,C) ならば、「(A と B) の C」とする
3. 類似度 (A,B) > 類似度 (A,C) だが、類似度 (A,B) < 類似度 (B,C) ならば、「A と (B の C)」とする
4. 類似度 (A,B) = 類似度 (A,C) で、類似度 (A,B) >= 類似度 (B,C) ならば、「解析不可能」とする
5. 類似度 (A,B) = 類似度 (A,C) で、類似度 (A,B) < 類似度 (B,C) ならば、「A と (B の C)」とする
6. 類似度 (A,B) < 類似度 (A,C) ならば、「A と (B の C)」とする

例えば、

例 1　友情と愛情の違い

を解消しようとすると、類似度 (A,B) は 4、類似度 (A,C) は 2、類似度 (B,C) は 2 である。ゆえに、類似度 (A,B) > 類似度 (A,C) で、類似度 (A,B) >= 類似度 (B,C) なので、結果は、「(友情と愛情) の違い」となる。

4.2 学習

A、B、C のうち 1 語がシソーラスにないものの扱いをそれまでの結果から、候補を得るという形で行った。16,346 の { A,B,C } の組合せのうち階層 3 までを比べ、一致するものを一つにまとめ、10,110 の学習のためのデータを得た。

定義 3 シソーラス上にない単語の最もらしさは、既知の単語の組合せのうち、最も候補が多いものとする

例えば、

例 2　東京と大阪の中間

のような文で、大阪がシソーラス上にない場合について考える。東京と中間のシソーラス上での番号は、東京は、階層 3 までだと { 1,2,59 } で中間は、{ 1,1,65 } 、{ 1,1,74 } 、{ 1,1,76 } である。この時の (A,C) の組合せで、最もらしい B は、{ 1,2,59 } となる。**B** がこのように決まると、4節により、「(東京と大阪) の中間」となる。

5 実験と考察

4 章で述べたアルゴリズムで、「A と B の C」の 885 文の解析を行った。その解析結果を表 1にまとめ、正解率を表 2にまとめる。

この実験結果の段階では、学習における結果が出てなく、{ A,B,C } のうち一つの語がないものは、表 1の 'シソーラスにない' のところに含めた。

表 1: 解析結果

文	解の数	解析不能	シソーラスにない
885	937	112	318

表 2: 正解率

解析文	正解	正解率 (%)
417	367	88.0

表 2では、表 1において '解析不能' と出力されたものや、シソーラス上にないもの、実験としてのデータとして不適切なものを除いている。

以上のように、「A と B の C」の解析を行うにあたって適当な文においては、本手法のアルゴリズムでの解析で、88 % の正解率が得られることが分かった。しかし、多義語が存在するもの全てが定義 2で、一つの候補にしぼられてはいないので、今後、システムの改良を行う。

参考文献

[1] 長尾真 偏 : 自然言語処理, 岩波講座ソフトウェア科学 15 (1996).

[2] 国立国語研究所 :「分類語彙表」形式による語彙分類表 (1996).

[3] 益岡隆志, 田窪行則 : 基礎日本語文法, くろしお出版 (1992).

[4] 黒橋 禎夫, 長尾真 : 長い日本語文における並列構造の推定, 自然言語処理 91-11 pp.79-86, (1992).