3H－2

語彙的結束性に基づいた語義曖昧性解消の観点から見たシソーラスの比較

望月 源, 本田 岳夫, 奥村 学
北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

語義曖昧性解消などに使用される代表的な機械可読の日本語シソーラスには、国立国語研究所の分類語彙表[4](以下、[分類])と、角川書店の角川類語新辞典[5](以下、[角川])が挙げられる。これらは人手により、異なる語彙分類基準で構築されている。本稿では我々の語彙的結束性[1]に基づいた語義曖昧性解消アルゴリズム[3]を両シソーラスそれぞれを用いて実装し、実際のテキストで曖昧性解消の実験を行なう。その結果から、語彙的結束性に基づく語義曖昧性解消の観点からシソーラスの比較を行なう。

2 語彙的結束性

テキスト内の語の間には意味的な関係があり、そのような単語の集まりを語彙的連鎖[2]と呼ぶ。語彙的連鎖は談話のつながりを明示する語彙的結束性を表すものであり、多義語の語義決定に文脈情報として利用できる。我々はこの語彙的連鎖をシソーラス上の同一カテゴリに属する単語として計算し、語義曖昧性解消を行なう手法を提案している。この計算に異なるシソーラスを用いれば、同一のテキストから生成される語彙的連鎖も異なる。例えば、

> 膨張を続ける宇宙の中で数多くの星が誕生、消滅を繰り返しました。そして宇宙の誕生から約100億年後、他の星と同じ様にして、原始太陽を中心にして原始太陽系星雲と呼ばれるガスの円盤を作りました。

という文章に[角川]、[分類]のそれぞれを用いると、生成される語彙的連鎖はそれぞれ、{ 星, 星, 星, 太陽, 太陽系, 星雲 }[分類]、{ 星, 星, 星, 星雲 }[角川]と異なる。また、「星」という単語は[分類]では非多義語だが、[角川]では3語義を持つ多義語となる。[角川]の「星」は「星雲」と語彙的連鎖を生成する時点で、「運命」や「点・線」関する意味ではなくて、「星」に関する意味に決定されている。[分類]の方も例えば「地球」という多義語に{ 星, 星, 星, 太陽, 太陽系, 星雲 }が文脈情報として与えられれば、「地表」に関する意味ではなくて「天体」に関する意味に決定できる。

どちらのシソーラスによる場合でも、語彙的連鎖を文脈情報として語義の決定に利用できるが、生成される語彙的連鎖や単語の分類は異なる。この相違が実際のテキストでの語義の曖昧性解消でどのような差を引き起こすのか調査を行なった。

3 実験

実験は我々の語義曖昧性解消アルゴリズム[3]を使用し、[角川]、[分類]を用いて実際のテキストでの評価を行なった。

実験で使用した2つのシソーラスの基本的構成を表1に示し[1]、両シソーラス中の見出し語の共通・固有関係の領域図を図1に示す。

表1 シソーラスの基本的構成　*() 内の数字は多義語

項目	角川類語新辞典		分類語彙表	
見出し語数	47146	(6401)	51864	(6414)
語義数	55490	(14733)	59649	(14191)
分類数	995	(963)	832	(819)
平均語義数	1.20	(2.30)	1.15	(2.21)
1分類当り語数	55.77	(15.31)	71.69	(17.34)
分散	1442.96	(192.03)	2737.07	(158.62)

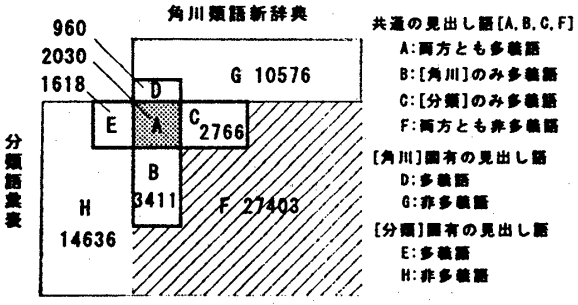

図1. 見出し語どうしの基本的構成

実際の4つのテキストに対する曖昧性解消結果を表2に、図1の見出し語の各領域の割合と実際のテキストに出現する単語の割合との対応を表3にそれぞれ示す。

Comparative study of Japanese thesauri by word sense disambiguation based on lexical cohesion.
Hajime Mochidzuki,Takeo Honda,Manabu Okumura
Japan Advanced Institute of Science and Technology
15 Asahidai, Tatsunokuchi, Ishikawa 923-12, Japan

[角川]は見出し語数47315語, 語義数約57000語,[分類]は見出し語数51915語, 語義数約60000語である。このうち[角川],[分類]とも"～"や"・"などの記号や単位などを取り除いたものを今回の調査対象とした。語義の分類は[角川]が上位3桁,[分類]は上位5桁としている。

表 2 語義曖昧性解消結果

[分類]	text1	text2	text3	text4	平均
候補語数	130	163	433	212	234.5
多義語数	52	78	118	53	75.8
未解消語数	10	11	19	13	13.3
正しい解消語数	22	42	71	34	42.3
正当率	52.4	62.7	71.7	85.0	68.1
適用範囲	80.8	85.9	83.9	75.5	82.4
[角川]	text1	text2	text3	text4	平均
候補語数	98	157	408	200	215.8
多義語数	65	88	145	68	91.3
未解消語数	16	16	19	8	14.8
正しい解消語数	23	48	85	42	49.5
正当率	46.9	66.7	67.5	70.0	64.5
適用範囲	75.4	82.8	86.9	88.2	83.9

表 3 見出し語の各領域の割合

	領域	図 1(%)		4text 平均 (%)	
		角川	分類	角川	分類
多	両方とも多義 (A:A)	4.3	3.9	24.3	23.2
	多方は非多義 (B:C)	7.2	5.3	20.0	9.1
義	他方がない (D:E)	2.0	3.1	0.6	2.3
非	他方は多義 (C:B)	5.9	6.6	9.9	18.5
多	両方とも非多義 (F:F)	58.1	52.9	43.1	40.5
義	他方がない (G:H)	22.5	28.1	2.1	6.4

4 考察

実際のテキストでは、共通の見出し語 [A+B+C+F] が候補語の 91.3% [分類] ～97.3% [角川] を占めており、シソーラス上の割合がそれぞれ 68.7%,75.5% であることから、共通の見出し語は出現頻度の高いことがわかる。また両シソーラスとも実際のテキストの候補語内の多義語の割合が 34.6%[分類,A+C+E] ～44.9%[角川,A+B+D] と、シソーラス上の 12.4%,13.6% に比べてかなり高いことから、一般にテキストにおける多義語の出現頻度が高いことを伺わせる。

[分類] に比べて [角川] の共通部分での出現頻度が高く、かつ多義語の割合も高いことから実際のテキストでは、シソーラスでの割合以上に [角川] の方が多義性の高い候補語をもとに語彙的連鎖を生成している。

これは語彙的連鎖内の単語の 3 つの組合せ、1.{ 多義の単語のみの連鎖 },2.{ 多義の単語と非多義の単語による連鎖 },3.{ 非多義の単語のみの連鎖 } の内、1. の割合が大きくなる確率を向上させる。表 4 は語彙的連鎖内の多義語の割合と、その多義が 1. になる割合を示している。

表 4 語彙的連鎖内の多義語の割合

	多義語の割合		多義語のみの連鎖となる多義語	
	[角川]	[分類]	[角川]	[分類]
1 text1	47.6	43.0	94.0	76.7
text2	49.0	58.0	75.3	62.0
text3	37.0	30.7	29.7	31.0
text4	43.0	40.6	76.1	40.6

表 4 で相対的な多義語と非多義語の割合は語義曖昧性解消の観点からあまり重要でない。非多義語の割合が高くても、上記の 3. の { 非多義語の単語のみの連鎖 } の割合が高ければその連鎖は語義曖昧性解消のための文脈情報としては機能しないからである。生成された連鎖の特徴として重要なのは、1.{ 多義の単語のみの連鎖 } となる多義語の割合の方だろうと考えられる。

一般に語義数が多いものほど正しい語義の決定が困難である。また多義語同士の一致する語義によって語彙的連鎖を決定するよりも、非多義語と多義語の語義の一致によって語義を決定する方がより正確性を増すものと考えられる。

表 4 で [角川],[分類] ともにテキスト 1 の多義語のみの割合が他のテキストに比べて高い。このことが両シソーラスともに結果のテキスト 1 の正当率の低下につながっているものと思われる。

また、テキスト上の多義語の平均語義数が [角川]3.03 > [分類]2.73 という関係にあり、平均正当率の [分類] > [角川] という結果に関連していると思われる。

一方、多義語の語義数が多いことは、他の単語と同じ語義を持つ可能性があがるために連鎖しやすいことにつながる。つまり、連鎖する単語のいない未解消語を減らすことに貢献すると考えられる。このことは若干ではあるが、平均適用範囲の [角川] < [分類] という結果に関連しているものと思われる。

5 おわりに

基本構成では両シソーラスの語義数、多義語数、見出し語数ともそれほど大差はない。しかし今回調査した実際のテキストでは、両シソーラス共通の見出し語および多義語の出現頻度が非常に高いことがわかった。また、[角川] の方が候補語中の多義語の割合が高く、多義語の平均語義数も多いので、[分類] に比べて正当率は低くなるが、適用範囲は高くなる傾向にあることがわかった。

謝辞 本研究で「角川類語新辞典」を使用させていただいた、(株) 角川書店に深謝する。

参考文献

[1] Michael Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, 1976.

[2] Jame Morris and Graeme Hirst. Lexical Cohesion Computed by Thesaurual Relations as an Indicator of the Structure of Text. *Computational Linguistics*, Vol. 17, pp. 21–48, 1991.

[3] 本田岳夫, 奥村学. 語義曖昧性を考慮した有意な語彙連鎖の生成. 情報処理学会自然言語処理研究会 NL97-14,pp.95-102, 1993.

[4] 国立国語研究所. 分類語彙表. 秀英出版,1964.

[5] 大野晋, 浜西正人. 角川類語新辞典. 角川書店,1981.