実例からの語義共起関係の自動抽出

5F−1

小林義行 佐伯元司 田中穂積

東京工業大学 工学部

1 はじめに

実用的な自然言語解析を実現するためには、十分に規模の大きな計算機用意味辞書が必要である。そのような辞書の構築は、人手によって行うのは困難であるため、計算機を用いて(半)自動的に行うことが望ましい。(半)自動的に辞書を構築する場合、深い意味解析を必要としない語の共起関係による意味表現が適していると考えられる。また、共起関係による意味解析の有効性もいくつかの研究で示されている[2]。意味を語の共起関係で表現する辞書を以下では共起辞書と呼ぶ。

共起辞書構築に必要なことは実際の文章から共起データを抽出することである。共起データの抽出は、(1)語と語の意味のある関係を抽出すること (2) 語義の曖昧性を解消することと考えられる。辞書を構築するためには大量のデータを扱わなくてはならないので、これらの処理は簡単なものであることが望まれる。

本研究では、実際の文章としてオンライン・コーパスを利用し、語義の曖昧性解消と意味のある共起関係の抽出をMRD(Machine Readable Dictionary:機械可読辞書)の用例とシソーラスを利用したパターンマッチで行う方法を提案する。

2 MRDとシソーラスを用いた曖昧性解消

文章中で近い位置に高い頻度で現れる語の組を意味のある共起関係とする方法がある[1][3]。この方法では、意味的に正しくても頻度の低い共起関係は失なわれてしまい、決まりきった用法での語の組ばかりが収拾される可能性がある。これを避けるため、意味のある共起関係のみを選択し抽出しなくてはならない。本研究では、語Aが語Bに係っている場合に、AとBには意味のある共起関係があると考える。そして、コーパス内の文に対して係り受けの曖昧性を解消する処理を行い、係り受け関係を意味のある共起として抽出する。

また、共起関係に多義語が含まれている場合、語義の曖昧性を解消しておく必要がある。語義が曖昧なままの共起関係では、意味情報として十分な効果が期待できないからである。本研究では、語義の曖昧性解消をMRDので対象語項目中の語義番号をtaggingすることと考える。

2.1 語義の曖昧性解消

語Wの語義の曖昧性解消に使える情報は、MRDのWの項目に記載されている語釈文と用例である。語釈文を利用するためには、深い意味解析が必要である。一方、用例を使った語義の区別は、用例と対象文のパターンマッチであり、深い意味解析は不要である。そこで、各語義項目に記載されている用例を用い、この用例に一致するかによって語義の曖昧性解消を行う。

しかし、MRDに記載されている用例は限られている。そこで、MRDから求め得る関係語やシソーラスで同じ分類に入る語で用例内の語を置き換え、利用できる用例を増やす。

MRDから関係語を求める場合、陽に記述されている関係語以外に、語釈文からも関係語を求め得る[4, 5]。ただし、対象としている語の語釈文から分かる関係語であり、語義の曖昧性のないものに限定する。

Extracting semantic collocations between words from a on-line corpus

Kobayashi Yoshiyuki, Saeki Motoshi, Tanaka Hozumi
Faculty of Engennering, Tokyo Institute of Technology

2.2 係り受けの曖昧性解消

正しい係り受け関係と処理対象文とのパターンマッチで係り受けの曖昧性解消を行うことができる。従って、語義の曖昧性を解消する処理は、語義の曖昧性解消と同時に係り受けの曖昧性を解消していることになる。

対象としている語が、他の項目の用例に含まれる場合や、語釈文に含まれる場合も用例の一種と見做せる。これらの用例は語義を人間が分析しなければ、語義の曖昧性解消には利用できないが、係り受けの曖昧性解消には利用できる。語釈文を利用する場合には、正しい共起関係を抽出する必要があるので、形態素解析と簡単な規則で正しい共起関係を抽出できるもののみ利用することにする。

2.3 heuristics

以下のheuristicsを導入することによって用例を増やしている。

H1 用例に「サ変名詞+する」が用いられている場合、「サ変名詞」と共起しているか調べる。あるいは格助詞なしで共起しているかどうか調べる。例えば、「景気」の用例には「景気を刺激する」があるので「景気を刺激」や「景気刺激」についても処理を行う。

H2 「名詞A+の+用言+名詞B」という形の節では、名詞Aをガ格に、名詞Bをヲ格とする。

H3 用例中の動詞を受け動態変形したものや自動詞化、他動詞化したものも格を調整して処理する。例えば「持つ」を処理している場合、「Aヲ持つ」という用例があれば「Aガ持たれる」「Aヲ持たせる」のような用例があるとして処理する。

H4 複合名詞が例文に使われている場合、複合名詞中の名詞で、複合名詞を構成するのに共通に使われる名詞を利用して処理する。例えば、「持つ」の例文「説得力を持つ」から「名詞+力」の解析をする。

H5 動詞を処理対象としている場合、それを他動詞あるいは自動詞形に変形した語がMRDに載っているなら、その用例も利用する。例えば、他動詞「進める」を処理する場合、自動詞「進む」の用例「Aが進む」を「Aを進める」に変形し利用する。ただし、機械的に語義の対応がとれる場合に限る。

2.4 処理手順

曖昧性解消の処理は以下の順番で行う。番号の小さい処理ほど優先度が高い。ある処理で曖昧性が解消された文は、それよりも優先度の低い処理の対象にはならない。(1)(2)(3)の処理は語義の曖昧性を解消でき、(4)(5)(6)の処理は、係りの曖昧性のみ解消できる。H1,H2,H3は、全ての処理で用いる。

用例は、名詞または動詞から成る文に限定する。動詞の解析の場合を使い手順を述べる。名詞の解析もほぼ同じ手順である。処理対象の動詞をVとする。

1. MRDのVの項目に記載されている用例NVと一致するか
2. 用例内の名詞Nを関係語である名詞N'と置きかえる

3. heuristics H4,H5 を用いる

4. V 以外の項目で V を含む用例 N V と一致するか

5. 辞書内の語釈文で V を含むものを利用する

6. V の関係語である動詞 V' の用例を利用する。名詞の置き換えはしない。

7. 処理できなかった文に対しては人間が共起関係を抽出する

関係語は、(1) 分類語彙表の分類で番号の上位 3 桁が同じ語 (2)MRD に関係語として記載されている語 (3) 処理語の語釈文の末尾にあり品詞が同じ語 (4) 語釈文末尾が機能語である場合はその前の語で品詞が同じ語とする。

3 実験

使用した言語資源は、コーパスとして日本経済新聞 1982 年 1 月から 3 月の約 3 万文 (別ち書き処理と品詞 tagging 処理済み)、MRD として新明解国語辞典、シソーラスとして分類語彙表である。

3.1 結果

名詞は「景気」と「技術」、動詞は「進める」と「持つ」について実験した。どの語も MRD での語義は 2 つ以上あり、コーパスに良く現れる語である。実験の結果を表 1 から表 5 に示す。左欄は、上記処理手順の番号を示す。ただし、処理 1+H1 は処理 1 に heuristics H1 を組み合わせた処理を意味する。

表 1: 景気, 技術

	景気	技術
文の数	669 文	717 文
subentry の熟語	0 文	121 文
熟語以外	669 文	596 文
処理 1	245 文 (37%)	70 文 (12%)
処理 1+H1	214 文 (32%)	0 文 (0%)
処理 2	82 文 (12%)	37 文 (6%)
語義の曖昧性解消できる	327 文 (49%)	107 文 (18%)
処理 4	108 文 (16%)	48 文 (7%)
処理 5	0 文 (0%)	18 文 (3%)
係り受けの曖昧性解消できる	425 文 (64%)	173 文 (29%)

「技術」の欄の割合は熟語以外を基にする数にして計算

表 2: 持つ, 進める

	持つ	進める
文の数	432 文	362 文
処理 1	64 文 (15%)	31 文 (9%)
処理 2	157 文 (37%)	47 文 (13%)
処理 3	0 文 (0%)	36 文 (10%)
語義の曖昧性解消できる	221 文 (51%)	114 文 (31%)
処理 4	1 文 (1%未満)	27 文 (7%)
処理 5	47 文 (11%)	0 文 (0%)
処理 6	8 文 (2%)	12 文 (3%)
係り受けの曖昧性解消できる	277 文 (64%)	153 文 (42%)

4 考察

今回の実験では、用例を用いて 20%から 50%語義の曖昧性を解消できた。処理できなかった原因の一つに MRD に記載されている用例の偏りがある。例えば、最も一般的に用いられる語義に対する用例は、使われる頻度に比べ少ない。また、広い範囲で使われるので用例があまり特徴的でない。処理できなかった文で使われている語義は、最も一般的な語義のものが多いと予想される。これについて、「持つ」を対象に調査したので簡単に結果を述べる。

処理できなかった 211 文から 50 文を抜き出し、どの語義で使われているか調査した。その結果と、処理できた 221 文の語義分類結果とを比較した (表 3)。「持つ」は、自動詞と他動詞にまず分けられ、自動詞の語義は 1 つ、他動詞の語義は 7 つである。

表 3: 持つの分析

語義番号	用例	処理できた文	処理できなかった文
自 1	3 文	0 文 (0%)	0 文 (0%)
他 1	2 文	6 文 (3%)	5 文 (10%)
他 2	7 文	10 文 (5%)	40 文 (80%)
他 3	12 文	103 文 (47%)	2 文 (4%)
他 4	4 文	23 文 (10%)	2 文 (4%)
他 5	5 文	50 文 (23%)	0 文 (0%)
他 6	14 文	27 文 (12%)	1 文 (2%)
他 7	2 文	2 文 (1%)	0 文 (0%)
合計	49 文	221 文	50 文

処理できなかった文の多くは、語義「他 2」であった。この語義は「所有」などを表すもっとも一般的な意味である。用例の数は少なくはないが、使用範囲が広いため、MRD の用例が不十分なものであると考えることができる。「他 1」は「手に持つ」意味で使われる場合の語義である。用例が少なかったことと語義の用いられる範囲が広いため、処理できなかったものであろう。

5 おわりに

既存の言語資源を用いて共起辞書を構築するために、簡単な処理で共起データを収拾する方法を提案し、評価実験を行なった。

今後さらに、処理できなかった文について、原因や必要な情報などを分析しなくてはならない。これについては、係り受けの非交差や同じ格は存在しないなどの知識を利用して共起候補を自動的に抽出すれば処理を効率的に行えるであろう。また、他動詞の場合、用例に出てくる格はヲ格が多い。そのため、ヲ格が省略されている場合は語義の曖昧性を解消することはできない。利用可能な格を増やす方法として語釈文を利用することが考えられる。語釈文の内で用例を用いて語義の曖昧性が解消できるものがある。これらの語釈文に含まれるヲ格以外の格を利用するのである。

ここで述べた方法は共起辞書構築のための共起データ収拾であるが、この方法で収拾した共起データは言語分析や辞書編纂、辞書評価などにも役立つであろう。

謝辞

本研究に対してさまざまな助言をして下さった東京工業大学工学部田中研究室の皆さまに感謝いたします。また、コーパスを提供して下さった NTT 情報通信処理研究所に感謝いたします。

参考文献

[1] F. Smaja. Macrocoding the Lexicon with Co-occurence Knowledge. Lexical Acquisition chap.7.

[2] P. Velardi, et. How to Encode Semantic Knowledge: A Method for Meaning Representation and Computer-Aided Acquisition. vol.17-2 p153.

[3] U. Zernik. Train1 vs. Train2: Tagging Word Sense In Corpus. Lexical Acquisition chap.5.

[4] 鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田将. 国語辞典を用いたシソーラスの作成について情報処理学会 自然言語処理研究会, NL83-16, 1991.

[5] 富浦洋一, 日高達, 吉田将. 語義文からの動詞の上位ー下位関係の抽出. 情報処理学会 論文誌, 31-1, 1991.