大規模コーパスを用いた共起関係の抽出

田中康仁
兵庫大学
Email : yasuhito@humans−kc. hyogo−dai. ac. jp

〔１〕はじめに

各新聞社から数年分の記事データが安価に販売されはじめるようになった。これは自然言語処理、計算言語の研究者にとっては大変喜ばしいことである。又、ＣＤ－ＲＯＭ化された辞書も入手可能である。これも良い研究材料である。

このような新しい時代には新しい考えで色々な分野の共起関係データの抽出を試みるべきである。ここではこのような考えに立ち共起関係データの抽出を試みる。

〔２〕大規模コーパスを用いての今まで行ってきた分析について

大規模コーパスを用いての今まで行ってきた分析について検討してみる。

(1)格関係を中心とした共起関係

格関係を中心とした共起関係はコーパスからKWICを作りそのKWICから人手によって格を中心とした共起関係を抽出し、入力して整理し、頻度付けを行った。しかし、これは形態素関係の精度や統語解析の曖昧性が解決しない時にはこの方法しか手段がなかった。しかし、共起関係データが少しずつ集積されるなかで形態素解析の精度、統語解析の曖昧性が解消されてきた。

今後は自動抽出を試みても良い環境ができあがった。

(2)四文字漢字列、五文字漢字列等からの語の共起関係の抽出

大規模コーパスからの四文字漢字列、五文字漢字列からの語の共起関係の抽出は既に自動的抽出分析が行われるようになってきた。しかし、一部のデータについては、まだ、手作業に頼らなければならない。

その他、２文字漢字列、３文字漢字列の分析も分析済データの充実と共に分析の精度を上げている。

(3)共起関係のデータの利用と研究について

共起関係のデータは筆者が姫路短大に在職中、学生達と共同でKWICから多量の共起データを人手と半自動的方法で抽出した。これは文部省の科学研究費特定研究「言語情報処理の高度化」（統括班長、長尾真）の援助も得て行った。これらのデータはその後ソフトウエア会社や電気機器メーカが持帰り仮名漢字変換のＡＩ変換へと利用された。また機械翻訳の訳語選択に利用された。しかし、筆者のデータは日本語の分析データであり、訳語が対になっていないため機械翻訳にはうまく利用されるにはいたっていない。

この研究について最初に発表したものは1987年3月27日情報処理学会自然言語研究会（60−3）
田中康仁、吉田将　　知識データ（語と語の関係）による多義性の解消である。

又、文部省科学研究費特定研究「言語情報処理の高度化」総括班（班長　長尾真）「語と語の関係解析用資料－“を”を中心とした－」解説編、資料編Ⅰ、Ⅱ昭和61年（1986）度科学研究費特定研究（Ⅰ）課題番号61120005である。

これにより9万件のデータを資料として公表した。

EDRは1986年4月26日に設立されたが
1988年8月1日付の発表した資料には単語辞書と概念辞書だけを作成しようとしていた。

その後の多くの研究者の共起関係（語と語の関係）の研究成果により、EDRでも重要性の認識と概念辞書の研究の中から共起関係の辞書を独立させることがなされた。

しかし、共起関係辞書は日本語、英語独立に作られたため機械翻訳への応用には今、一歩の感じがある。

〔３〕共起関係データの抽出

日本経済新聞５年分の中から共起関係の抽出

Extraction of Co−occurrence data using large corpus.
Yasuhito Tanaka

を行った。その具体的内容を述べる。

3．1　共起関係の抽出

3．1．1　並列関係の抽出

	種類	延べ件数
□や□	413,057	604,178
□と□	480,978	783,659

“や”、“と”の前と後にある漢字、片仮名文字列で抽出した。

並列関係の抽出は日本語解析の中で重要である。今まではデータによる並列関係の同定ではなく、何らかの規則で並列関係の抽出を行う方法であった。この方法ではある程度のところまではできるが、より精度を上げるためには並列データの抽出と整理が重要である。

また、これら並列関係はシソーラスの内容を充実させるためにも役立つ。

例えば

アナウンサーとミュージシャン
アナウンサーと女優
アナウンサーと名解説者

これらは同じ性格をもつ語であり、多くのものは同一のグループに属するものである。

しかし、機械的抽出には誤りも発生するので注意しなければならない。

例

“室内の温度と庭の温度の差が”
手作業によるデータの検査と精度向上は
重要である。

3．1．2　修飾関係の抽出

～する～、～した～、これも同じように“する”、“した”の前と後にある漢字、片仮名文字列で抽出した。

	種類	延べ件数
□する□	255,479	665,779
□した□	228,079	382,052

しかし、この関係は単語の関係ではなく“する”、“した”の前にあるものは文である場合が多い。期待するほどのものは少なかった。

しかし、“する”、“した”の前にある文字列は用言である。これを整理すると次のようになった。

	語教	組織	頻度
“する”、“した”両方にあるもの	5,954	440,389	991,184
“する”だけにあるもの	9,813	17,935	21,370
“した”だけにあるもの	8,114	25,231	35,277
合　計	23,881	483,555	1,047,831

抽出例

アクセス	53
アタック	5
アピール	296
アプローチ	33
アレンジ	74
イメージ	506

これらの語は述部には出てこず、修飾関係にだけ表われるものもある。

次のような文字列でも抽出できる。

～れる～、～られる～、～される～、
～れた～、～られた～、～された～、
～かれる～、～がれる～、～たれる～、
～かれた～、～がれた～、～たれた～、
～なれる～、～ばれる～、～まれる～、
～なれた～、～ばれた～、～まれた～、
～われる～、～らされる～、～さされる～、
～われた～、～らされた～、～さされた～、
～かされる～、～がされる～、～たされる～、
～かされた～、～がされた～、～たされた～、
～なされる～、～ばされる～、～まされる～、
～なされた～、～ばされた～、～まされた～、
～わされる～、～わされた～、

形容詞語尾による抽出

～きい～、～さい～、～しい～等も、前回と同じように“きい”、“さい”、“しい”の前と後にある漢字、片仮名文字列で抽出した。

	種類	延べ語数
～しい～	26,199	103,435
～きい～	2,714	5,503
～さい～	1,243	2,471

例えば“大きい”の訳語を調べてみると次のような例がある。大きいの次にくる語により英語の訳語が異ってくる。

1）大きな家　　a big（or large）house
2）大きな間違　　a big（or great）mistake
3）大きな声で話す　　speak in a loud voice
4）口を大きく開ける　open one's mouse wide
5）(人が)大きくなる　grow older, be grown up
赤尾好夫編綿貫陽補訂　和英基本単語熟語集
大学入試4000語　4訂版　1996年旺文社

形容動詞による修飾関係

～的な～、～的に～等も、同じように"的な"、"的に"の前と後にある漢字、片仮名文字列で抽出した。

	種類	延べ語数
～的な～	85,170	210,821
～的に～	53,211	147,664

これはおもしろい関係である。

抽出例

具体的な内容	1,027
基本的な考え	832
積極的な姿勢	743
具体的な計画	578
積極的に取る	2,689
積極的に進む	1,949
圧倒的に多い	1,025
本格的に取る	945

さらに次のような用語の抽出も可能である。

漢字＋ひらがな＋漢字　の関係である。

例えば　"流れ星"、"忘れ物"

このひらがな文字列は"え列"、"い列"の一文字とする。これらは用言の連体形による体言への修飾関係で一語として取り扱ってよいものや複合動詞等が抽出できる。

抽出例

下支え効果	145
下支え要因	663
稼ぎ時	101
共稼ぎ世帯	15
利下げ観測	32
値下げ攻勢	30
寄せ木細工	16
混ぜ合せ	283
手持ち物件	1
撃ち方	33
持ち分	265
重ね着	33
思ひ出	7

	件数	延べ語数
い	78,714	420,191
え	9,671	62,861
き	24,971	320,177
ぎ	2,494	9,267
け	24,947	160,606
げ	6,816	66,796
し	55,233	225,878
じ	10,794	33,948
せ	780	4,042
ぜ	46	357
ち	5,585	159,216
て	14,306	51,222
で	790,876	1,226,218
に	1,491,805	3,379,981
ね	376	3,159
ひ	43	54
び	13,337	94,266
へ	24,651	35,765
べ	12,674	56,934
み	11,559	146,101
め	12,793	49,000
り	40,816	592,673
れ	6,189	57,591

1万件以下のものから分析してみるのも1つの方法である。

近い

1）近い将来　　in the near future
2）公園に近い　be near(or close to)the park
3）50歳に近い　be nearly fifty years old
4）近く来日する　come to Japan soon(or before long)

赤尾好夫編綿貫陽補訂　和英基本単語熟語集
大学入試4000語　1996　旺文社

次のような共起関係も抽出できる。

～の～の～、～の～等も同じように"の"の前と後にある漢字、片仮名文字列で抽出することができる。

	種類	延べ
～の～の～	488,300	549,797
～の～	4,128,730	8,295,939

このような関係を機械的に抽出することができた。しかし、期待するほどのものでないものもあるし、人手によって個々のデータを分析し整理しなければならない。これらのデータを分析する中で共起関係の抽出ができる。

大量の抽出ができるのが良い点である。

例1	～の～	頻度
	三分の一	4,434
	企業の割合	3,518
	取引の中心	3,150
	事業の再構築	3,136
	景気の先行	2,434
	三分の二	2,084
	世の中	1,998
	四分の一	1,833
	小口の売り	1,758
	閣議後の記者会見	1,748

例2	～の～の～	頻度
	全体の三分の一	149
	三分の二以上の賛成	123
	法の下の平等	109
	全体の三分の二	97
	国民生活の質の向上	91
	縁の下の力持	89
	世界の中の日本部会	87
	ベルリンの壁の崩壊	85
	生活の質の向上	78
	全体の四分の一	78

〔４〕平仮名列を用いた共起関係の抽出

これまでに述べた共起関係の抽出は代表的な平仮名列と一部漢字まじりの平仮名列による抽出であるがこれは一つの例として述べたものである。

まず網羅的に考えるにあたっては次の手順を取るとよい。

1）コーパスから平仮名列を全て抽出する。

2）同じ平仮名列は集め頻度付を行う。

3）桁数別にまとめる。5文字以下の平仮名別について何か特徴をみつけグループ化する。

4）3）のグループごとに仮名文字列の前後の漢字列、片仮名列を抽出して、用語として、共起関係として分析する。

このような分析により全ての平仮名列の分析ができる。

〔５〕今後の課題

共起関係のデータ抽出について一つの方法を提案することができたが、今後次のようなことに注目してゆきたい。

(1)日本語とその対訳語（英語）についてデータの抽出方法の検討を行ってゆきたい。

(2)大規模の日本語とその対訳語（英語）のテキスト・コーパスを収集したいと考えている。

これらのことが機械翻訳ならびに自然言語処理の研究を進める上で重要な点である。

〔６〕参考文献

1)赤尾好夫編綿貫陽補訂和英基本単語塾語集 大学入試4000語 4訂版 1996旺文社

2)田中康仁、吉田将 知識データ（語と語の関係）による多義性の解消 情報処理学会自然言語研究会（60－3） 1987. 3

3)文部省科学研究費特定研究「言語情報処理の高度化」総括班（班長長尾真）「語と語の関係解析用資料－"を"を中心とした－」解説編、資料編Ⅰ、Ⅱ 昭和61年（1986年）度科学研究費特定研究（Ⅰ） 課題番号 61120005

4)EDRカタログ 1988年8月 説明会配布資料

〔７〕分析データ

この分析で用いたものは日経総合販売㈱から購入した「日本経済新聞CD－ROM 1990, '91, '92, '93, '94年版」である。