大規模コーパスの分析

田 中 康 仁
兵 庫 大 学
Email : yasuhito@humans−kc. hyogo−dai. ac. jp

0）はじめに

自然言語処理に必須の機械可読辞書の充実をめざすには、コーパスの分析により共起関係の抽出と見出し語の追加が重要である。ここでは朝日新聞記事データの'を'の分析結果と日本経済新聞5年分の記事データの分析について述べる。日本経済新聞データについては3文字，4文字，5文字の結果についてまとめた。

1）朝日新聞記事からの共起関係

朝日新聞記事1年分の中から'を'を中心とした共起関係の抽出を行い終了したのでその結果を述べる。

朝日新聞一年分の記事から'を'を中心としたKWICを作り、'を'の右側を20文字程度分類した。これにより右側には同じ動詞等が集まり収集することができた。

共起関係の種類	延べ件数
201,745	437,557
'を'の前の語	'を'の後の語
55,416	12,911

一つの種類に対する延べ件数 2.17

'を'の前の語に対する種類 3.64

'を'の後の語に対する種類 15.63

一つの種類に対する延べ件数は2.17であるがこれはもっと多くすべきであると思う。

'を'の前の語に対する種類は3.64でありこれは'を'の前の語（目的語に対する）動詞の割合は少ない。10以上にすべきである。'を'の後の語に対する種類は10以上であり十分である。

このため新聞記事5年分程度の分析が必要と考えられる。約50万行のKWICから人手により3年かけて行った。しかし、資金があれば分散して抽出でき、もっと簡単にできたと考えられる。多くの人は完全自動化を考え何も行わずに過ごしているが人手によっても共起関係を抽出できる。

2）読売新聞記事からの共起関係

朝日新聞の記事データと同様に読売新聞についても'を'の共起関係の抽出を行っている。約70万行のKWICを作成し、'を'の右側で分類し抽出は手作業で行っている。これも、3年程度で終りそうである。1年分の作業は既に終了した。

今後に期待していただきたい。筆者としても形態素解析が完全に自動化することを期待し、さらに、係り受け解析が完全に自動化され共起関係が自動抽出することをのぞみたい。このデータがこれら自動化の基礎として、また文の格構造抽出の基礎材料となることを期待している。

3）日経新聞記事からの漢字列の抽出と分析

日経新聞記事データ5年分の記事を分析して次のような結果を得た。

3−1）文字列の抽出

文を解析するためには形態素解析のプログラムを使うのが普通考えられるが、ここでは文字種の変り目で抽出するという単純な方法を採用した。（但し、ひらがなカタカナは2文字以上とした）

この結果をまとめると次のようになる。

	異り件数	延べ件数
2文字漢字列	93,047	26,817,980
3文字漢字列	318,472	8,382,228
4文字漢字列	757,432	8,532,996
5文字漢字列	557,240	3,022,650
6文字漢字列	546,318	2,117,339
7文字以上漢字列	1,147,804	2,467,383
2文字以上ひらがな	1,050,800	27,098,906
2文字以上カタカナ	194,576	7,826,426

3−2）用例数について

新聞の5年分の記事データであれば大量であると単純に考えがちであるが必ずしもそうともいえない。

延べ件数/ 異なり件数を計算してみる。

2文字漢字列	288.22
3文字漢字列	26.32
4文字漢字列	11.26

Analysis of very large corpus−using NIKKEI News Papers（5 years volumes）−
Yasuhito Tanaka
Hyogo University

5文字漢字列	5.42
6文字漢字列	3.87
7文字以上漢字列	2.15
2文字以上ひらがな	25.79
2文字以上カタカナ	40.22

筆者の考えであるが一つの文字列について平均10ケの用例があればと考えている。

これからみると5文字漢字列以上の漢字列については不十分といえる。10年分以上の新聞記事が必要と考えられる。どのような知識データを抽出するかに応じて考えが変る。これも平均でのことで個々のものについては実際の頻度をもとに考えなければならない。用例が10件あったとしても全ての場合が出つくしているというものでもない。ここでも偏りがある。

3－3） 文字例の累積頻度との割合

各文字列を累積頻度と異り文字列の関係を調べてみた。これは頻度の高い順に分類し、ある一定の頻度の高い順に分類し、ある一定の頻度の%を満たす異なりを調べたものである。

累積頻度の割合	3文字漢字列	4文字漢字列	5文字漢字列
10%	37	104	130
20%	132	430	578
30%	325	1,113	1,521
40%	681	2,425	3,423
50%	1,313	4,865	7,450
60%	2,452	9,582	16,545
70%	4,761	19,879	38,338
80%	10,259	47,226	94,585
85%	16,439	80,561	154,500
90%	29,696	155,088	254,975
95%	70,097	332,609	406,108
98%	157,335	586,773	496,787
100%	318,472	757,432	557,240

日経新聞のデータから日本語処理に必要なデータを単純におしはかることはできないが参考にしていただきたい。

次の二つの点には注意しなければならない。

(1) 文字種の変わり目で抽出しているため誤った抽出がある。

(2) 日経新聞5年分の記事といえども日本語の使われている世界から見ると小さなサンプルでしかない。

他の統計データもあるが紙面の都合で省いた。

3－4） 分析済データとの照合

朝日新聞一年分、読売新聞一年分のデータをもちいて日経新聞記事データからのものとどの程度照合するか調べてみた。

	一致したもの		不一致のもの	
	件数	延べ件数	件数	延べ件数
3文字漢字列	49,386	6,197,082	269,086	2,185,146
4文字漢字列	141,112	7,013,716	478,402	1,519,280
5文字漢字列	71,109	1,859,041	486,131	1,163,609

実数でみると上記のようになる。少し、わかりにくいので%で表示してみると次のようになる。

	一致したもの%		不一致のもの%	
	異り	延べ	異り	延べ
3文字漢字列	15.5	73.9	84.5	26.1
4文字漢字列	18.6	82.2	81.4	17.8
5文字漢字列	12.8	61.5	87.2	38.5

(但し、3文字漢字列は読売新聞のデータはまだ分析中であり含まれていない)

これからわかるように一致した件数は12%～18%程度であるが延べ件数でみると61%から82%にまでなっていることがわかる。少しばらつきはあるが朝日新聞、読売新聞各一年分の分析データから重要な文字列の抽出は行われていることがわかる。しかし、年代、新聞社が異なると文字列の異なりが多く発生することがわかる。

これは自然言語処理のむつかしさを表している。

3－5） 規則による分析

既に参考文献1)、2)、3)によって詳細を発表しているがその内容を充実させて実施した。

この規則を簡単に述べると

(1) 4文字漢字を構成している2文字漢字列を抽出し、2文字・2文字に分割できるものをコード付けした。同様に5文字漢字列を構成する3文字・2文字,2文字・3文字を抽出し、分割に使用した。

(2) 地名、人名、企業名、数詞に使われる漢字、用語をもちいてコード付け等に使用した。

(3) 独立して使われる用語がたまたま他の漢字と結合したものを抽出した。

これらをまとめコード付けられたものの割合を次に%表示をもちいて示す。

	コード付けがなされたもの %		コード付けがなされなかったもの %	
	異り	延べ	異り	延べ
3文字漢字列	69.3	95.7	30.7	4.3
4文字漢字列	81.8	96 9	18.2	3.1
5文字漢字列	80.6	93.0	19.4	7.0

（3文字漢字列の分析の結果が悪いのは、朝日新聞の分析結果だけをもちいていることと考えられる。）

規則による分析には例外が発生し、誤りも含まれているが大きく区分を付けるためには大変役立っていることがわかる。

これら規則については付録として付け加えているので参照されたい。

4）今後の課題

日経データの分析からさらに次のようなことを目ざして研究を深めなければならない。

(1)コードの付かなかったデータについてコードを付ける。

(2)規則によってコード付けされたものの中にも誤って付けられたものもあるので全体を再度見直す。

(3)「戻り値」「バス停」「現代っ子」などの用語についても、今までの考え方を発展させて抽出する。

(4)'を'の共起関係についてはもっと大量に抽出する方法と自動化を行わなければならない。

(5)共起関係のデータ、単語の抽出、単語間の関係を分析し、共起関係への発展、応用を目ざしての研究を進めなければならない。

5）おわりに

大規模な新聞記事データの分析を行い、良い分析結果（データ）を得ることができた。今後はさらに詳細な分析や別のデータとの照合により質の高いデータにしてゆきたい。

分析の自動化を進めることも重要であるし、別の見方による分析も考えなければならない。

朝日新聞、読売新聞の記事データ分析に比べ5倍もデータが多かったにもかかわらず10倍程度早くできた。これは経験の積み重ねである。

しかし、自動化が行えないので人手による分析が無意味であるとか、その努力を認めないとするような考えは無くさねばならない。

この論文の一部は情報処理学会第52回全国大会に「大規模コーパスの分析」－日経新聞5年分をもちいて－と題して発表したものである。

6）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。

また、朝日新聞一年分と読売新聞一年分の記事を用いた。

7）参考文献

1．田中康仁　自然言語の解析による知識獲得と拡張－四文字漢字列を用いて－自然言語処理94－9,情報処理学会　1993.3

2．田中康仁　自然言語の解析による知識獲得と拡張－五文字漢字列を用いて－自然言語処理94－10情報処理学会　1993,3

3．田中康仁　自然言語の知識獲得－五文字漢字列の自動コード化について－信学技報　NCL94－7　1994.7

付録資料

日本経済新聞社のデータ分析に用いた規則をここにまとめて述べる。これらの規則の適用順序は利用者にまかせたい。

規則No.1

既に分析済データ。辞書による照合

規則No.2

4文字漢字列については既に分析済の4文字漢字列の2文字・2文字で分析出来るものから2文字漢字列を抽出し、2文字・2文字に両方とも照合できるものを抽出する。

5文字漢字列についても同様に分析済みの2文字・3文字，3文字・2文字から2文字，3文字を抽出し2文字・3文字,3文字・2文字と両方とも照合できるものを抽出する。

規則No.3

独立性の強い用語による抽出。

案外、以上、以下、以内、以前、以後、以降、以東、依然、一応、一括、一見、一言、一向、一時、一時期、一瞬、一生、一切、一層、一体、一度、一回、一般、一晩、一番、一部、一歩、一面、一目、一律、一昨年、一人、一体化、何人、何度、過去、回、独自、極力、現行、午前中、月末、現在、現代、午前、午後、今回、今後、今月、今春、今秋、今週、今度、今期、今月末、今、今世紀、今大会、今年、再三、再度、最近、最終、最大、最大限、最初、最低、最高、昨年、昨年春、昨年夏、昨年秋、昨年度、昨年末、氏（先頭文字）

時折、時代、若干、終日、事件後、事実上、
終了後、十分、十数回、将来、深夜、週間、
週末、将来、順次、数週間、数日間、数日後、
数年間、数年、世界一、双方、場合、常時、
随分、数日、世紀、世代、絶対、先月、先週、
先日、先月末、先週末、戦後、戦前、前回、
前後、前日、前期、前年、前夜、全国、全身、
全然、全部、前面、全員、早速、早期、相当、
即刻、即時、多少、多数、対応、対抗、断然、
直接、程度、当時、当初、当然、突然、当分、
当日、当面、到底、同年、同時、同日夜、
同日朝、年間、年度、半年間、比較的、本件、
本年度、本来、毎月、毎週、毎日、唯一、
来年、来年度

規則No. 4

三文字漢字列で接尾語的な漢字、用語による用語の抽出

～案、～員、～屋、～下、～化、～家、～課、
～画、～会、～界、～外、～学、～額、～株、
～感、～管、～間、～館、～器、～機、～協、
～業、～局、～型、～形、～系、～権、～後、
～語、～向、～工、～行、～号、～合、～座、
～債、～済、～祭、～剤、～材、～作、～策、
～産、～時、～式、～室、～者、～車、～社、
～手、～出、～集、～受、～所、～署、～書、
～上、～状、～色、～図、～制、～製、～場、
～職、～人、～水、～性、～税、～席、～跡、
～説、～前、～組、～荘、～側、～層、～族、
～体、～帯、～代、～中、～団、～通、～的、
～展、～点、～店、～度、～党、～東、～堂、
～届、～内、～入、～派、～版、～費、～品、
～付、～部、～物、～法、～北、～本、～面、
～薬、～用、～率、～流、～料、～量、～力、
～林、～類、～歴、～連、～炉、～例、～話、
～論

規則No. 5

三文字漢字列で接頭語となる語による用語としての抽出

不、再、末、元、現、前

規則No. 6

三文字漢字列で××氏、××君、××様の付くものは××は姓であるとする。

××の付くものを集めて姓とみなす。妻○○は名前とみなす。最後の文字が’子’のものは固有名詞とみなす.

規則No. 7

次の文字を含む漢字列は地名である。

県、市、区、郡、町、村、字、大字、地区、
地域、丁目、丁、地帯、省、また最後の文字が
（浦、駅、沖、河、海、街、岳、狭、橋、国、
郷、局、圏、原、湖、江、港、崎、山、坂、寺、
州、沼、川、線、台、沢、谷、地、島、峠、灘、
岸、線、帯、道、岬、野、路、湾）である。

規則No. 8

次の文字を含む漢字列は数表現である。

「一、二、三、四、五、六、七、八、九、十、拾
百、千、万、億、兆」

規則No. 9

固有名詞に特有な用語を含む文字列を抽出する。

本名、氏、女史、様、会長、社長、最後の文字が子である。営業所、病院、小学校、中学校、
大学、議長、社長、投手、会長、総理、弁護士、
容疑者、君、空港、協会、教授、協会、会館、
医師会、工業会、村長、報道官、支部、支部長、
農協、出版社、警察署、本店、店長、課長、
所長、編集長、美術館、博物館、裁判長、船長、
市役所、病院、資料館、校長、医師、判事、
監督、名人、専務、代表、署長、教諭、支店長、
教頭、相談役、部長、学長、県議、都議、選手、
夫人、頭取、理事長、委員長

これ以外のものもある。これらは既に分析済の４文字、５文字の固有名詞を右から分類し、後から（右から）２文字、３文字列に機械的に分割し、同じものを集めてまとめ、その中から手作業で抽出した。

これらの規則・適用には例外があるので、再度データの検討が必要である。