大規模な英文の分析

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

〔 0 〕はじめに

英語で書かれた新聞、雑誌などがCD－ROMとして売り出されている。これを自然言語処理に利用しようとすると幾つかの大きな問題に出くわす。

大規模な英文を分析するにあたりどのように処理すべきかという問題について述べる。

文の認定、単語の認定という単純そうでめんどうな作業について分析した方法と分析結果を示し、さらに、この方法で分析した文や単語の応用について述べる。単語数別の文の統計は幾つかの研究発表はあるが、それらをどのように意味付けし、実用化に結びつけるという研究はない。ここではこの点に注目した。

〔 1 〕大規模な英文

英語で書かれた新聞、雑誌などがＣＤ－ＲＯＭとして売り出されている。これを自然言語処理に利用しようとすると幾つかの大きな問題に出くわす。具体的に検討を進めるために対象としたデータについて述べる。

ペンシルベニア大学が出しているウォールストリートジャーナルの1年分の記事データである。これは新聞データをそのまま入力したものである。ここで最初に問題になることは、文の認定である。

英文の文末には「・」が使われているが、これ以外にも多くの「・」が使われている。

想定しうることをプログラムにして試行した。この部分は九大助教授の菅沼明助教授に依頼した。

1．ピリオドの後に空白または改行文字がある箇所を文の句切りの第一候補とする。

2．文区切りの第一候補のうち、以下の文字列である場合には該当のピリオドは文の句切りとしない。

- 「No. x」「no. x」「Ch. x」「ch. x」「Ca. x」「ca. x」（xは数字）のピリオド
- 「Mt. 」「St. 」「Dr. 」「Mr. 」「Mrs. 」「Prof. 」のピリオド
- 「Fig. 」「Rev. 」「Eng. 」のピリオド
- 「Mt. 」「St. 」「Dr. 」「Mr. 」「Mrs. 」「Prof. 」「Professor」に続くアルファベット1文字の後のピリオド　{名前の頭文字と判断}

しかし、上にあげた条件だけでは文の認定には不十分であった。これと同時に使用されている単語（文字列）の分析を行ってみた。この分析には在庫管理で使用するＡＢＣ分析を行った。

また、文の長さでデータ件数を分析した。ただ、単純に数値で表わすだけでなくグラフ表示を行った。

個々の文の長さ別のデータを分析する中で「・」とその予想した例外だけでは不充分なことが分かった。

〔 2 〕実験と分析結果

2－1）文と単語のデータ件数

文と単語の抽出を行ったところ次のような結果を得た。

(1)文の数	1,188,528文
(2)単語の総数	26,848,148単語
(3)単語の種類	298,737単語（文字列）
(4)平均1文の単語数	22.6
(5)単語の分布中央値	18
(6)一文中の最少の単語数と最多の単語数	1～342

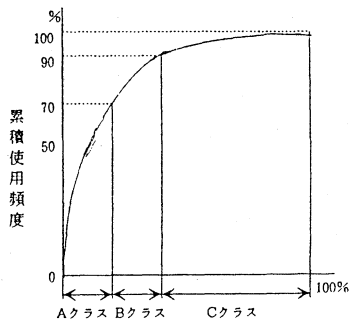
2－2）単語のＡＢＣ分析

単語をＡＢＣ分析すると次のようになる。

	種類	累積頻度
Aランク－－－	943	18,794,335（ 70%）
Bランク－－－	4,496	24,163,102（ 90%）
Cランク－－－	293,298	26,848,148（100%）
合計	298,737	

ABC分析を単語の分析に使うには次のようにする。単語の使用頻度の高い順に分類する。そして分類された単語の累積使用頻度と累積使用頻度の比率を計算する。

さらに累積使用頻度の比率が７０%までの単語をＡクラス、７０～９０%までの単語をＢクラス、残り９０%を超えるものをＣクラスとする。このような状況をグラフに表すと次のようになる。



単語数比

(1) Aクラス

自然言語処理の単語にあてはめて考えると、数量が少ないわりに使用頻度の高いもの、重要な用語、個々に詳細な分析が必要なもの。

(2) Bクラス

単語数もかなりあるし、使用頻度も高いもの、Aの次に重要な用語、個々の分析も必要。

(3) Cクラス

単語数が多いわりに使用頻度が少ないもの。

これらも極力辞書に登録はしなければならない。

脚　注　ＡＢＣ分析について

ＡＢＣ分析は在庫管理ではごく普通に使われている概念であり、大量の在庫品目を管理するために考え出された方法である。

重点管理のレベルをA、B、Cの三つのクラスに分けることからＡＢＣ分析と呼ばれている。

個々の単語としては使われる割合は少ないが、これら単語がまとまると頻度の高い語と同様に重要になるため登録が必要である。

2－3）1文の単語数と分布

単語数別の文の数をグラフにすると下記のようになる。（図1参照）

2－4）文の長さの分布の詳細な分析

1単語、2単語、4単語、6単語が突出しているがこれは新聞の中によく使われている図、表等の説明の表示によるものである。

この図、表、等について述べる。

数値が1の場合

〔数〕fig.〔数〕figure.〔数〕drawing.〔数〕app.
〔数〕reference.〔数〕ref.〔数〕tab.〔数〕table.
〔数〕photo.〔数〕photograph.

数値が2以上の場合

〔数〕drawings.〔数〕Figs.〔数〕figs.
〔数〕figures.〔数〕graphs.〔数〕photographs.
〔数〕photos.〔数〕references.〔数〕refs.
〔数〕tables.〔数〕tabs.〔数〕ills.〔数〕illus.
〔数〕schemes.

これが多量に出現する。また、これらが複数個組み合わさっている場合もある。前置詞としては〝with〟をともなって現われる。

このようなものを除き修正するとポアソン分布をなしていることが分る。

2－5）文の認定誤りの略語や記号

「・」についての簡単な規則だけでは多くの例外が出現する。これらは1単語、2単語、3単語等の中に多くあらわれる。

そこでこれらを詳細に分析し、略号を辞書にしたり、文認定のプログラムのテーブルを追加すること等により、よりよい文認定を行うことができる。

この部分には例外がかたまって出現するので単語の断片だけではわからないので文章のどのような部分で発生しているか、データの中で個々に詳細に調べてゆかねばならない。

単なるテーブルや規則の追加、略号等の辞書への追加では処理できないものもある。

例えば名前、姓等である。

〔3〕分析結果の応用

1）抽出した単語と辞書作成

抽出した単語はＡＢＣ分析してある。このデータを既に開発済の英語辞書と見出しで照合する。一致しなかったものを手作業によってよく内容を検討し追加する。

このようなコーパスによる分析結果は頻度情報が付いているため判断の基準として利用できる。

しかし、単語の中には原形化されていないものもあるので個々に検討しなければならない。

一方、市販の辞書を参照することも見出し語の追加として利用できる。しかし、頻度情報が付いていないため追加等についての判断が付きにくいものがある。

複数の辞書を参照し、その辞書の規模によって点数付けを行い、追加の基準を作ることができる。

大規模の辞書の照合で不一致になったものには小さい得点を与え、小規模の辞書で不一致になったものには大きな得点を与えればよい。

しかし、市販の辞書は独特の形式で入力しているものが多いため、参照できるものは少ない。

次の二つのＣＤ－ＲＯＭは簡単に参照することが可能である。

(1) COLLINS COBUILD on CD－ROM
(2) Merriam Webster's Collegiate Dictionary Tenth Edition　　Zane Publishing

しかし、あまり多くの情報を抽出することは著作権法違反になるので注意しなければならない。

もう一つの方法は、WWWの中で公開されている一般辞書を見つけ出す方法である。

2）文の抽出結果と応用

1単語から7単語で構成される文の中で単なる記号列の集まりであるものを除く。

- 文と認められるものは機械翻訳のテストデータとして利用する。
- 対訳データを付け用例ベースの機械翻訳のサンプル・データとする。
- マルチリンガルのコーパスを作成するための基礎データとする。自然言語研究の基礎データとなる。
- 文の中に使われている数を〔数〕としてパターン文を作成することができる。

例えば　This book contains〔数〕papers.
　　This book contains〔数〕chapters.

以上のようなパターン文が考えられる。

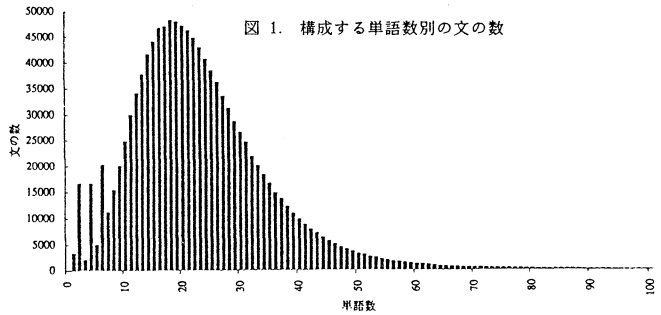


図 1.　構成する単語数別の文の数

単なる記号列の集まりを含んだ文の中から使用頻度の高い文を抽出する方法としては次の方法がある。

(1) まったく同一の文は頻度を付け集約する。

頻度2以上のものを抽出

頻度2以上の文をウォール・ストリート1年分の記事から抽出すると次のようになった。

単語数	文 数 (a)	延文数(b)	b／a
3	24	129	5.38
4	110	354	3.22
5	165	919	5.57
6	132	340	2.58
7	263	627	2.38
合　計	694	2,369	3.41

約690文程度の慣用文を抽出できた。これらのうち特殊な文、誤りを除くと、常識的な新聞に出てくる文であることを確認した。

10年分程度の記事又は10地区1年分程度の記事を集め、同じような実験を行えば慣用的な文は頻度も多くなるし、文の種類も増える。筆者の経験的な推測では約3倍の約2，000文程度のものが集められる。また、単語数7で打ち切ったが、10年分程度の記事の場合には単語数8，9のものも入れるべきだと考える。

(2) 構成されている単語を調べ高頻度語を含む文を抽出する。

例えば　results, conclusion, presented, gives, discussed,samples,conclusions 等で抽出する。

少し誤りを含むが、文の認定ができるようになった。今後はこの情報を利用し、エデイターを使って一つ一つの文を調べ、検査し、誤りを正し、SGML等の形式に移してゆかなければならない。このようにすることで多くの人々が正確な文として取り扱うことができる。統計的方法や、プロブラムの処理から、人間の確認によって正確さが保障されたデータに移してゆくべきである。

(3)新聞記事の特徴

新聞の記事文はごく普通の文のように思われるが、次のような特徴があることが分かった。

1）主語に人称代名詞が表れない。受身形が多い。
2）疑問文が極端に少ない
3）人名等の固有名詞が多い。
4）図、表、参照等の表示が多い。
5）新聞特有の表現、文の書き方がある。前述(2)を参照、特徴ある語が出現する。

これらの特質を充分知った上で機械翻訳のテスト・データや自然言語処理の各分野で利用しなければならない。

3）専門用語の抽出

コーパスに出現する単語に対してABC分析を行い、単語のクラス分けを行った。

用語 ＝ 一般用語 ⊕ 専門用語

上記式から

専門用語 ＝ 用語 ⊖ 一般用語

このような等式が一般的に成り立つ。

一般用語としてABC分析した単語のA、Bの二つのグループを割り当てる。すると専門用語はCグループの単語で構成されている単語列となる。このようにして専門用語または専門用語らしきものを抽出できる。

さらに詳細に調べるには、Cグループの連続した単語のKWICを作成し調べればよい。これは専門用語を抽出する簡易な一つの方法である。この方法については実験を行うとしている

〔4〕この文抽出方法の改良

この分析方法は1回の試行によって得られたものであるが、誤って文と認識しているところ等を詳細に調べプログラムやテーブル、辞書を更新することによりより良いものになることを期待する。そのために〔9〕の付録にその内容の一部を示した。

〔5〕おわりに

英字新聞のCD－ROMから、

① 文の認定
② 単語の抽出
③ 専門用語の抽出

これらの応用について一つの方法を確立することができた。これにより英語の機械可読辞書の追加、強化が可能になった。また文の長さにより英文の機械翻訳への利用方法について一つの見とおしを得ることができた。

大規模な新聞コーパスの中から文を切り出し、単語数別に分解してしまうとこれは単なる文の集合体となり、著者の創造物というものとは関係が薄くなったものになっている。多くの人が同じような表現を使っているとしたらこれら一文、一文に著作権はあまり感じられなくなる。しかし、私は著作権が完全に無くなっていると言うのではない。

〔6〕データについて

この分析に用いたCD－ROMは次の所から入手できる。

(1) Liberman, M.（ed.）（1991）. "CD－ROM I" Association for computational Linguistics Data collection Initiative, Univ. of Pennsylvania. US$25.

(2) COLLINS COBUILD on CD－ROM
Harper Collins Electronic Reference
14 Steep Lane Findon Worthing West Sussex BN14 OUF UK
E-mail rhcc @compuserve.com
http:／／www.collins.cobuild.co.uk

(3) Merriam Webster's Collegiate Dictionary Tenth Edition
For More Information:
Zane Publishing Customer Service:
customer@zane.com
Technical Support:
Internet: http:／／www.zane／com

詳細な情報は利用者が入手していただきたい。

その他、アメリカの新聞等についてのCD－ROMは丸善（株）の "CD－ROM CATALOG" から選び簡単

に入手することができる。

〔７〕謝辞

この研究に協力して下さった九州大学大学院システム情報科学研究科菅沼明助教授と徳島大学工学部北研究室の方々に心から感謝致します。

また、ペンシルベニア大学のCD－ROMの入手と手続に協力して下さった文教大学前田英明教授に感謝致します。

〔８〕参考文献

１）田中康仁　機械可読辞書の更新について
自然言語処理112－17　情報処理学会　1996. 3
２）田中康仁　機械可読辞書の見出しについて
自然言語処理117－5　情報処理学会　1997. 1
３）CD－ROM　CATALOG　1997　MARUZEN
４)経営科学　電子開発学園　1988

〔９〕付録

(1) 一単語の分析結果を示す。

〔数〕年号と思われる数　１７××、１８××
１９××

〔数〕.〔数〕　〔数〕).　〔数〕%
〔数〕%)〔数〕㎜.〔数〕]〔数〕－〔数〕

Am.	Appl.	Arg.	Appendix.
Appendices.	Applications	Approx.
Astrophys,	Chap.	Chem.
DM.	Ⅰ,Ⅱ,Ⅲ,Ⅳ,Ⅴ～	Index.
Ltd.	Mon.	Congress.
Commun.	Class.	Cryst.
Figs.	File.	Gen.
Instrum.	Introduction.	Inst.
i,e,	Jan.	Sept. ……
Lett.	Letter－to－the－editor.
Mon.		
Phys.	Pro.	References.
Refs.	Rep.	Soc.
Sect.	Tables.	Tabs.
Vol.		

(2) 二単語の分析結果

〔数〕が１の場合

〔数〕fig.〔数〕drawing.〔数〕app.〔数〕illus.
〔数〕reference.〔数〕ref.〔数〕tab.〔数〕table.
〔数〕photo.数〕photograph.

〔数〕が２以上の場合

〔数〕.〔数〕	〔数〕drawings.	〔数〕Figs.
〔数〕figs.	〔数〕figures.	〔数〕photos.
〔数〕graphs.	〔数〕photographs.	〔数〕references.
〔数〕refs.	〔数〕tables.	〔数〕schemes,
〔数〕tabs.	〔数〕ills.	〔数〕illus.
〔数〕footnotes.	〔数〕pages.	〔数〕percent.
〔数〕years	〔数〕months.	〔数〕weeks.
〔数〕days.	〔数〕hours.	〔数〕min.
〔数〕sec.	〔数〕ms.	〔数〕pm.
〔数〕app.	〔数〕maps.	〔数〕deg.
〔数〕KHz.	〔数〕MHz.	〔数〕Hz
〔数〕KW.	〔数〕KA.	〔数〕GV.
〔数〕MW.	〔数〕MeV.	〔数〕GeV.
〔数〕KeV.	〔数〕V.	〔数〕KV.
〔数〕para.	〔数〕mm.	〔数〕cm.
〔数〕nm.	〔数〕m.	〔数〕km.
〔数〕mg.		

2単語専門用語

Contains res.　Includes res.
Majority decision　Rev.　Lett.
Uncertain remains.　Section organization

(3) 三単語の分析結果

高頻度の文と文型

With〔数〕refs.　With〔数〕figs.
With〔数〕maps.　With〔数〕photos.
Figs and tabs.
Contains author index.
Contains subject index.
Includes subject index.
Refs and figs.
Refs and tabs.
Refs, figs, tabs.
Results are discussed.
Results are given.
Results are presented.
Results are discribed.
Results are summerized.
Results are graphed.

(4) 五単語の分析結果

五単語の文を先頭文字から順に並べてみると

１）８割から９割が普通の文である。

２）数個の文がほぼ同じものであるものがある。
但し数値を変えただけのものである。

例 This book contains of〔数〕papers.
This book contains of〔数〕chapters.
This book contains of〔数〕selections.
This book contains over〔数〕chapters.
This book contains over〔数〕papaers.
This book contains over〔数〕selections.
With〔数〕figs.,〔数〕tabs.
With〔数〕refs.,〔数〕figs.
With〔数〕tabs.,〔数〕figs.
With〔数〕figs.,〔数〕refs.

３）同じ表現が多い文

The results are as follows.
The results are as following.
The following results were obtained.
The following conculusion were obtained.
The agreement is very good.
The Board approved the applications.
Results to date are discussed.
Results are given and discussed.
Relevant papers are abstracted separately.
Progress to date is described.
Published in summary from only.
No major complications were observed.
New approaches are briefly discussed.
Includes author and subject index.