1J-4

エンティティの自動抽出による英字新聞記事文章の構造解析

石塚　隆男

亜細亜大学経営学部

１．はじめに

本研究は、英字新聞の記事文章の構造を実体概念（エンティティ）の自動抽出によって明らかにすることを目的とする。

よく知られているように、英字新聞記事は逆ピラミッド構造にしたがって書かれており、第１パラグラフが最も重要であるとされている。

英文テキストから単語を抽出し、パラグラフ単位の出現頻度を調べたのでは一般的な単語の頻度に埋もれてしまい、重要な実体概念を抽出することは難しい。本研究では、記事文章から名詞並びに名詞句を抽出し、その中からエンティティの抽出を行う。そのために、まず、副詞、接続詞、助動詞、動詞など品詞をひとつに特定できる単語を削除し、残った単語列から名詞句の構成を行った。

今回、いくつかの知見が得られたので報告する。

２．構造化へのアプローチ

文章データの構造化には以下のようなアプローチが考えられる。

①エンティティのグルーピング＋エンティティのラベリング

ｅｘ．ＫＪ法（親和図法）

②個々のエンティティの機能／役割、エンティティ間の関係を記述・保存した図的抽象化

ｅｘ．ＥＲダイアグラム

ここで、エンティティ（実体）とは、文章を構成するキー概念であり、主語に相当する名詞句や高頻度の単語はエンティティの候補になりうる。

構造化の問題は、いかにグルーピング並びにラベリングを行うかの問題としてとらえることができる。数量化３類等の多変量データ解析手法は、反応表等のデータ行列をもとに少数の構造因子を抽出し、大局的なマップを描くことはできるが、エンティティの位置情報や尺度情報を用いていないため、文章構造を適切に要約・再現するのは難しい。

対象データとして新聞記事文章を選択した理由は、莫大な量のテキストが電子化・データベース化され、増大する一方であり、自動構造化への期待が高いと判断されること、新聞記事文章は以下に述べるように一定の構造にしたがって書かれていることから一般の散文を扱うよりも構造を活かした分析が可能であると判断されるためである。

欧米の新聞記事は、一般に各パラグラフが重要度の高い順に並んだ“逆ピラミッド”の構造をしていることが知られている。

しかし、重要度は記事を執筆した記者や編集者が判断したものであり、読者から見て上述のモデルが妥当かどうかは一概には言えない。

３．名詞句の抽出とエンティティの判別

日本語文章は、分かち書きされていないことからテキスト・マイニングの前処理として形態素解析を行うのが通例である。形態素解析には、茶筌等のツールが用いられる他、ＪＩＳ２バイト系の文字コードの大小比較により漢字熟語、カタカナ語等の抽出は容易に行うことができる。

一方、英文は単語単位に分かち書きされており、日本語文よりもはるかに論理的であるとはいえ、主語や述語の判別が困難な英文も日常的に存在する。英文の形態素・構文解析ツールとして、いくつかの tagger や parser が開発されているが、ひとつの単語が複数の品詞をもつため前後関係から品詞を判断しなければならない場合が多く、これらのツールも完璧ではない。

そこで、情報検索や知識抽出の精度を上げるためには構文解析よりも key phrase としての名詞句（noun phrase）の抽出に重点を置いた方が効率がよいと考えられる。たとえば、PHRASER(http://tamas.nlm.nih.gov/phrasercgi3.html)は医学領域の文書から名詞句を抽出することを目的として開発されているが、ネット上でのデモ体験はできても一般に提供されてはいない。

そこで、本研究では英文を構成するウエイト付きの単語列と見なすことにする。ウエイトは、以下に述べる方法により、名詞句を構成する単語であれば１、それ以外であれば０をセットする。

Structure Analysis of English Newspaper Articles by Extracting Entities

Takao Ishizuka

Faculty of Business Administration, Asia University

１）電子辞書ファイルから単一品詞の単語リストの作成

『英辞郎 CD-ROM 版』にはテキスト形式の辞書ファイルが付加されており、その中から名詞以外で単一の品詞をもつ単語を品詞別に抽出し、ファイル化した。作成されたのは、副詞、接続詞、動詞、助動詞、形容詞＋to＋動詞、代名詞、名詞でないイディオム（慣用句）の各リストである。

２）英字新聞記事と単一品詞リストとのマッチング

英文から名詞句を抽出するために、名詞句を構成しない、あるいは名詞句の一部の可能性が低い単語のウエイトを最小値に設定した。ウエイト最小値の単語は、ストップワードであり、名詞句と名詞句の区切り（delimiter）として用いる。

３）英文中において大文字で始まる単語のウエイトの最大化

固有名詞の単語はエンティティの候補になりうるのでウエイト値を２にセットする。日本語に限らず英語においても個々の単語は一般語でも連続することにより固有名詞化する。そこで、大文字で始まる単語が連続する場合にはこれらを＝で連結し、１語とみなすようにした。

４）抽出された各名詞句がエンティティであるかの判定

以下の基準のいずれかを満たす名詞句をエンティティとして判定を行った。

（基準１）主語であること

（基準２）固有名詞句であること

（基準３）出現頻度が２回以上であること

（基準４）１文中に複数の名詞句が存在する場合には、文頭に近いほど高いウエイトを与える。

４．解析例

CD-ROM 版の New York Times の記事データベース並びに Daily YOMIURI の英文記事をテキスト保存し、上述の処理を行った。図１に解析例を示す。大文字の固有名詞は＝で連結され、１語の扱いになっている。また、Mr.等の省略語は文末のドットと区別するために、『英辞郎 CD-ROM』から省略語リストを作成し、一致したものはドットを＿に自動的に変換している。

図１．英字新聞記事から名詞句、エンティティを抽出した解析例（NYT，May 20，2001）

```
Imitating Mr. Ghosn in Japan
it was a Brazilian-born automotive executive
from Renault named Carlos=Ghosn who finally
turned around Japan's=Nissan=Motor=Company
which last week reported the biggest annual
profit in its history after a staggering loss
the year before and because imitation is often
the purest form of flattery in Japan there is
now speculation that other troubled Japanese
companies may embrace a foreign chief like
Mr_=Ghosn

a Brazilian-born automotive executive ,
59 ,  4 ,  1
Renault ,  57 ,  1 ,  1
Carlos=Ghosn ,  54 ,  1 ,  2
Japan's=Nissan=Motor=Company ,  46 ,  1 ,
4
last week ,  43 ,  2 ,  0
the biggest annual profit ,  38 ,  4 ,  0
history ,  35 ,  1 ,  0
a staggering loss the year ,  29 ,  5 ,
0
imitation ,  25 ,  1 ,  0
the purest form of flattery ,  18 ,  5 ,
0
Japan ,  16 ,  1 ,  1
speculation ,  12 ,  1 ,  0
other ,  10 ,  1 ,  0
Japanese companies ,  7 ,  2 ,  1
a foreign chief ,  2 ,  3 ,  0
Mr_=Ghosn ,  66 ,  1 ,  2

----- ENTITY LIST -----
  ENTITY FREQUENCY>1
Ghosn ,  2
Japan ,  3
  SUBJECTIVE ENTITY
imitation
speculation
  DOMINANT ENTITY
imitation
```

５．考察及び今後の課題

今回、単一品詞リストの外部知識を用い、単語にウエイトを与え、名詞句の抽出を行い、さらに記事文章中の中心的な概念としてエンティティの判別を試みた。完璧には程遠く、試行錯誤によりプログラムの修正を行っているが、本手法を何らかの形で評価を行いたいと考える。