1E3-01

エンティティの自動抽出による英文記事の構造の可視化

A Visualization of the Structure of English Newspaper Articles by extracting Entities

石塚 隆男
Takao Ishizuka

亜細亜大学経営学部
Faculty of Business Administration, Asia University

We propose a new visualization method of document structure by extracting entities. Entities are defined as key phrases of each paragraph composing the target document, and for example, subjective words, proper nouns, etc. Idioms and words having single or unique word class except for noun were extracted from an English electronic dictionary file and were listed for stop words matching. Consequently noun phrases were selected into entities by some criteria. Entities are aligned according to paragraph process and co-occurrence relationships between entities are indicated. Thus, documents can be transformed into visual charts, and our method is expected as a tool of diagramatization of documents.

1. はじめに

本研究では、文章データの内容を容易に把握するために文書構造を可視化することを目的とする。われわれは、パラグラフに着目し、パラグラフ間の関係を中心とした図解化を試みてきた[石塚 2004]。しかし、形式的パラグラフに基づいているため、過度に細分化され、パラグラフだけでは全体像が見えにくいことが明らかとなった。そこで、本研究では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。すなわち、パラグラフを考慮しつつ、文章の中心的な概念であるエンティティを抽出し、エンティティ間の関連を可視化することを試みる。具体的には、文章から固有名詞を含む名詞句を自動的に抽出し、それらの中からいくつかの基準を満たすものをエンティティとし、それらをもとにワードマップに相当するエンティティ関連図を自動作成するプログラムを開発し、英語の新聞記事文章へ適用した。その結果、記事文章の流れと大意の可視化が可能であることが確認されたので報告する。

2. エンティティによる構造化へのアプローチ

本研究では、当該文章における重要な単語として、主語を構成する名詞（句）、固有名詞（句）、高頻度の名詞（句）を取り上げ、これらをエンティティ（実体）と呼ぶことにした。いわゆるキーワードやキーフレーズもエンティティの一部であろうが、文中における役割や関係に注目し、エンティティと名づけることにした。

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

e-mail address: ishizuka@asia-u.ac.jp

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。また、単語の頻度に基づく TF＊IDF 法は、新聞記事のように出現頻度が１回の単語が多い場合にはあまり効果がない。そこで、適切なエンティティを抽出し、パラグラフの流れの中にマップすることにより文章の構造が可視化できると考えられる。

3. 名詞句の抽出とエンティティの判別

本研究では、英文記事を解析の対象とする。

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 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、本研究では英文を単語列とみなし、各単語は名詞句の可能性に応じたウエイトをもつとする。ウエイト値を以下に述べる手順によりプログラムで自動的に設定する。

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 ver.1』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。抽出したストップワードの数を表１に示す。

表１．抽出された品詞別ストップワード数

不規則動詞	349	助動詞	**31**
～ly 以外の副詞	485	ｂｅ動詞	**10**
前置詞	90	形容詞＋to do	**173**
代名詞	124	イディオム	**12205**
接続詞	50		

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１～４に解析例を示す。対象とした記事は、２パラグラフから成る。第１パラグラフから抽出された名詞句の一覧である。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。図３は、エンティティがどのパラグラフの何番目の文に出現したか、共起エンティティは何かを示している。＊のついた単語が見出しと一致したエンティティであり、第１パラグラフに出現した単語であることがわかる。

図４は、逆ピラミッド構造であることを前提に第１パラグラフ（＝lead 文）のエンティティを中心に関連を可視化したものである。

図１．英文記事の例（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.（以下、省略）

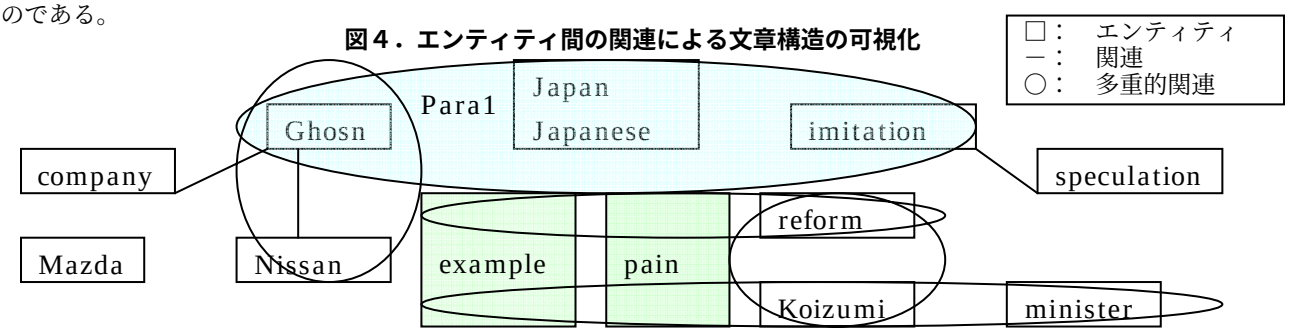
図２．図１の記事から抽出された名詞句

a Brazilian-born automotive executive／Renault／Carlos=Ghosn／Japan's=Nissan=Motor=Company／last week／the biggest annual profit／history／a staggering loss the year／imitation／the purest form of flattery／Japan／speculation／other／Japanese companies／a foreign chief／Mr_=Ghosn

図３．抽出されたエンティティとパラグラフの関連図

パラグラフ / エンティティ	P1 ①	P1 ②	P2 ①	P2 ②	P2 ③	P2 ④	P2 ⑤	P2 ⑥	P2 ⑦	P2 ⑧
＊Ghosn	◆	◆					◆	◆		◆
＊Japan	◆	◆	◆					◆	◆	
*imitation		◆								
speculation		◆								
Japanese		◆	◆					◆		
company	◆			◆						
mazda					◆					
example								◆	◆	
pain								◆	◆	
reform								◆		◆
Koizumi									◆	◆
minister									◆	
Nissan	◆								◆	◆

図４．エンティティ間の関連による文章構造の可視化



５．考察及び今後の課題

１）エンティティの抽出について

本研究では、外部知識として単一品詞のストップワードリストを与えることにより名詞句の抽出を行い、抽出された名詞句からエンティティを判定した。

本方法により名詞句を高い精度で抽出できることが確認できたが、多くの英文を対象にして定量的な評価を行う必要があろう。

エンティティの抽出では、固有名詞の抽出はできたが、現実には主語の判別が難しい英文も多い。そこで、エンティティとして文頭により近い名詞句を選択することが考えられる。名詞句の位置情報をどのように用いるべきかが今後の課題である。また、新聞記事のエンティティと見出しを構成する語句との関係についても定量的に評価する必要があろう。

２）文章構造の可視化への関連性理論からのアプローチ

一般に、文章は、著者が伝えたい情報を意図的に表現したものとみることができる。

文脈とは、著者の意図を複数のパラグラフまたは文間の関連性により表現したものである。

そこで、文章に関して以下のような前提を置くことは無理がなかろう。

①文章は、著者の意図を目的とし、各パラグラフをサブシステムとするシステムである。
②各パラグラフは、文脈上の目的をもち、ひとつのまとまりである。
③各パラグラフは、文脈にしたがって配置されており、出現順序・位置はパラグラフの基本的な属性のひとつである。
④各パラグラフは、文脈にしたがい、隣接する前後のパラグラフの一方または両方と関連をもつ。
⑤文章は、利用目的にしたがい、特定の構造パターンにしたがって書かれたものである。

文章構造のパターンとしては、起承転結型、結承転提型（ビジネス文書）、逆ピラミッド型（新聞記事）、序論・本論・結論型（論文）等が知られている。

一般に文章の構造は、人間が読めば自然に了解されるものである。コンピュータによる自然言語処理においては、話題の転換点を見つけるためのいくつかの試みがなされている他、接続詞を解析することにより文やパラグラフを抽象化する試みもなされているものの、文章構造のパターンを完全に自動認識することは現状では難しい。

人間は、対象とする文章の構造に関して何ら情報をもたない場合、文章の先頭から末尾まで順に読み進めることにより著者の意図を理解できるだろう。もしも、当該文章の構造パターンが事前に明らかになっている場合には、著者の意図が書かれている箇所を効率的に見つけ、理解することができる。このように、文章構造のパターンを知ることは当該文章がどういう種類の文章なのかを知ることでもある。本研究では、通常、“逆ピラミッド構造”にしたがって書かれているとされる英文新聞記事を対象に、構造が既知であることを前提に解析を行ったことになろう。

文章構造の可視化には、何らかの合理性や客観性が求められよう。以下では、「人間は情報処理にあたって最大の関連性を目指す」、つまり「最小の処理労力で、できるだけ多くの認知効果を得ることを目的とする」［Dan Sperber 1995］［西山 1999, p.37］ことを前提条件とする関連性理論の活用の可能性を中心に考察を行う。

関連性理論［Dan Sperber 1995］や談話分析における整合性や結束性の理論［亀山 1999］にしたがえば、

① 読者は、できる限り労力をかけずに著者の伝えたいことを文章から読み取ろうとする。
② 文章は、特定の構造パターンにしたがうだけでなく、できるだけ読者にとって読みやすい＝著者の意図を理解しやすいように書かれている。

ことを前提とする。

実際に私たちが日常的に目にするあらゆる文章にこのような前提が成立するかどうかは疑わしいが、文章を書くことに慣れている職業人が著した文章であればこれらの前提は概ね妥当であろう。

そこで、文章に関して以下の仮説が考えられる。

［仮説１］著者の意図はエンティティによって構成される。

当該文章のエンティティを明示することは、読者が文章を理解する労力を最小限にする意味がある。

［仮説２］文章の文脈は、エンティティ間の関連によって明示される。

エンティティ間の関連が過度に複雑な文章やエンティティが孤立した文章はわかりにくく、“最適な関連性”が求められる。

すなわち、関連性理論にもとづく規準として、たとえば、以下のような規準が考えられる。

文章の最適な可視化＝Min f（エンティティの数，エンティティ間の関連の数）

関連性＝可視化の効果／可視化された結果を読む労力

このような考え方は、MDL（Minimum Description Length）規準や“Simple is best！”を旨とするケチの原理にも見ることができ、生産性＝産出／投入や価値分析＝機能／コストも同様の考え方である。そこで、何らかのコスト関数を構成し、それを最小にする図解化や可視化が最適であるとすることは妥当な考え方であろう。図解化という自由度無限大の問題を定量的に扱える可能性がでてきたことにより今後、コスト関数の理論的な構成を検討していきたい。

参考文献

［Dan Sperber 1995］Dan Sperber and Deirdre Wilson：*Relevance: Communication and Cognition*(2nd ed.), Blackwell（内田聖二他訳：『関連性理論 －伝達と認知－第２版』研究社，1999）

［石塚 2004］石塚・新行内・大友・高嶋・山本：キーワード・モーメント法を用いた新聞記事文章の自動構造化，平成 16 年度秋季大会予稿集，社団法人日本経営工学会，p.110-111

［亀山恵 1999］亀山恵：第３章 談話分析：整合性と結束性，『談話と文脈（岩波講座言語の科学７）』岩波書店

［西山佑司 1999］西山佑司：第１章 語用論の基礎概念，『談話と文脈（岩波講座言語の科学７）』岩波書店

PHRASER(http://tamas.nlm.nih.gov/phrasercgi3.html)