E-009

パラグラフの抽象化による新聞記事文章の自動構造化

An Automatic Structurization of Newspaper Articles by Abstraction of Paragraphs

石塚隆男† 新行内康慈‡ 大友拓也§ 高嶋啓介§ 山本久志§
Takao Ishizuka Kouji Shingyouchi Takuya Ohtomo Keisuke Takashima Hisashi Yamamoto

1. はじめに

本研究では、文章データの内容を著者の意図や文章の流れを保存・再現しながら構造化する方法について検討を行う。文章の表題や見出しは究極の要約といえるが、見出しだけでは著者がそれだけの文字数を費やして言わんとしていることはわからないことが多い。本のタイトルを見れば、その本が何について書かれたものかはわかるが、目次や章立てを読むことにより初めて著者の枠組みが見えてくる。目次は本の構造を知るための要件であるが、新聞記事のように数百字の文章であっても、現状では本文を読まない限り、文章の流れとしての構造は読み取れない。

文章データの構造化は、文章からの知識発見や文章の図解化のための前処理として位置づけることができる。私たちは目的に応じてさまざまな図＝モデルを使い分けているように、一般に図解化は目的に依存し、図解の答えはひとつではない。そこで、具体的な図解をする前に、図解モデルに中立的な文章の構造化を行う必要があると考えられる。

文章の構造化を行う上で、パラグラフ（段落）は意味のあるまとまりを示す単位であり、文章を構成するコンポネントあるいはサブシステムとみなすことができる。通常の文章はパラグラフを意識して書かれているが、1文で構成されているパラグラフも存在し、いくつかの隣接パラグラフをグループ化した方が文章全体の構造をつかみやすい場合もある。

このように、文章全体を構造化するためには、ボトムアップ的アプローチとしての各パラグラフの内容の縮約とともにトップダウン的アプローチとして文章全体のパラグラフ群への分割が必要となる。本研究では、これら2つのアプローチにより文章の構造化する具体的な方法を提案し、新聞記事文章に適用した結果、いくつかの知見が得られたので報告する。

2. 関連する研究

本研究は、コンピュータを用いて膨大な文章データ＝テキストデータからの知識の発見を行うテキスト・マイニングの1分野とみなすことができる。テキスト・マイニングは、自由記述型アンケートデータの分析等に供されているが、文章中の単語の出現頻度や関連性あるいは共起性のある単語のマップ出力が中心であり、本研究の目的である「著者の意図や文章の流れを保存・再現した構造化」は実現されていない。

各パラグラフを縮約するためには、各パラグラフの特徴を表わすキーワードの抽出が必要であり、情報検索や自動要約の分野とも関係する。テキスト・マイニングも含め、これらを実行するためには自然言語処理技術が不可欠である。単語の正確な頻度を得るためには、シソーラス（類語集）や概念辞書の整備も必要となり、大がかりな道具立てを要する。自然言語による文章表現は多様であり、あらゆる表現に対応した自然言語解析システムを構築することは困難である。

しかしながら、本研究の対象としている文章の長さは、新聞記事の本文や書籍の中のある章の1節であり、人間が短時間で読める程度の長さの文章である。したがって、完全性よりも簡易性や低コスト性を優先し、しかも利用者によって修正可能な構造が得られることが望ましい。本研究では、こうした観点からいくつかの提案を行う。

3. パラグラフの縮約化・抽象化

ひとつの文章データを読み込み、各パラグラフの特徴を表わすキーワードを抽出する方法を検討する。

一般にキーワードは名詞（句）であり、ほとんどの名詞（句）は文章中に漢字、カタカナ、アルファベット、数字等の文字種が2文字以上連続した文字列として存在する。そこで、文章データから名詞（句）の単語を抽出し、単語×パラグラフの頻度行列を作成し、TF･IDF 値が高いキーワードを各パラグラフ単位に抽出する。

新聞記事本文を対象にした場合、DF（＝document Frequency）が1件の単語が複数あり、TF･IDF 法は同点の単語に関しては無差別でキーワードを十分に絞りきれないことがわかる。TF･IDF 法はベクトル空間モデルに基づいており、文中における単語の統語解析的情報は用いていない。

本研究では、統語解析を行う代わりに文中における単語の位置（location）情報や文の文字数といった尺度（scale）情報を用い、キーワードの“モーメント”を計算することによりキーワードのウエイトとした。以下にその方法について説明する。

日本語の文は、句読点により区切られている。「。」は文の終わりを示す。一方、「、」の使い方には個人差や状況差が見られ、英語の「,」のような文法的必然性は見られない。しかし、読点は副詞節や挿入句といったまとまりの区切りとして用いられる他、主語を強調したい場合にも用いられる。

そこで、本研究では以下の経験的仮説に基づいて各パラグラフを構成する文の句読点で区切られた節（clause）単位にキーワードの候補を抽出し、各パラグラフ内でウエイト値が上位の単語を当該パラグラフのキーワードとする。

［仮説1］日本語文では、修飾語句等の係り受けは後続の名詞（句）に係っているため、各節の末尾に近い名詞（句）ほど重要である。

［仮説2］重要なキーワードは、一般的な名詞よりは名詞の熟語であり、キーワードの文字数が長いものは特定の対象を指す固有名詞化したものが多く、重要である。

† 亜細亜大学
‡ 十文字学園女子大学
§ 東京都立科学技術大学

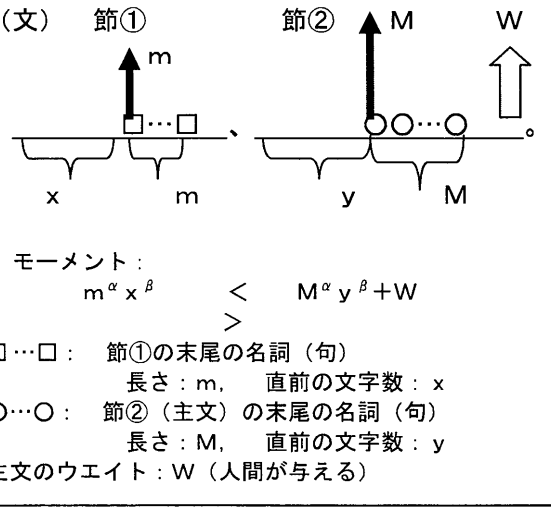
［仮説３］読点により複数の節から成る文において、句点「。」の前の節が主文であり、他の節に比べて圧倒的に重要である。

以上の仮説に基づき、各節の末尾の名詞（句）について、

名詞（句）の文字数 $^{\alpha}$ × 節の先頭から名詞（句）の直前までの文字数 $^{\beta}$ ＋ 節が主文の場合のウエイト

によりその名詞（句）のウエイトを算出した（図１）。

図１．モーメントによるキーワードのウエイト



各パラグラフ内のすべての節について、上述の方法でウエイトを計算し、ウエイト値の大きいものから上位２個の名詞（句）を選択し、

キーワード１ と キーワード２

を当該パラグラフの見出しとする。

４． パラグラフのグループ化

ひとつの文章を構成するパラグラフは出現順に並んでおり、前後関係には意味がある。また、パラグラフ数がいたずらに多い文章は全体の構造をとらえにくい。そこで、隣接するパラグラフ間に距離や類似度を定義し、パラグラフのクラスタリングを行うことが考えられる。

TF･IDF 値によりパラグラフベクトル間のコサイン尺度を計算することが考えられるが、ベクトルの次元が大きいため有意な結果を得られない。

そこで、本研究では、あるパラグラフで新規に追加された単語数をパラグラフの文字数で除した値（新語数／文字）の系列を計算し、差分の符号を調べることにより話題の転換点を探索し、符号の変化がないパラグラフのグループ化を行った。

５．方法

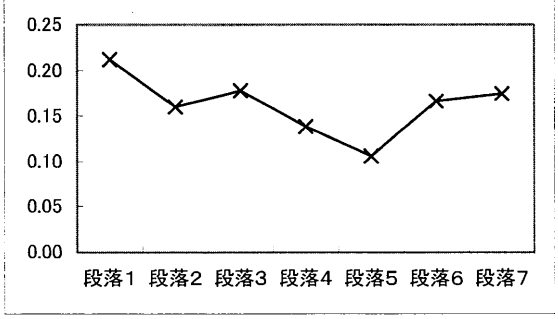
新聞記事本文のテキストを与え、以上の方法によりパラグラフの抽象化を行い、文章全体の構造を図解するプログラムを作成した。

図解の方法は、パラグラフの流れを示すプロセス・チャートとし、グループが検出されたパラグラフを枠でくくった図を作成した。図は、修正可能性を重視し、Microsoft Excel のＶＢＡマクロの自動出力により作成している。

６． 結果

朝日新聞の記事(2004/6/8)に本方法を適用した例を図２，３に示す。新語数／文字数の区分的変化率の符号により起承転結構造が見てとれる。各パラメータの見出しは、主文のウエイトや次数α、βの与え方により変化する。

図2. 新語数／文字数の段落別変化



７． 考察並びに結語

キーワード・モーメント法のパラメータの与え方に根拠を与えるとともに作成された見出しの妥当性について評価する必要がある。今回の構造化は、プロセス順を示したにすぎず、多元的な構造化を行う必要があると考える。

図３．新聞記事本文の自動図解化プログラムの出力例