テキストデータの自動抽象化に関する研究

石塚 隆男
亜細亜大学経営学部

１．緒言

近年、テキストマイニングや機械学習等の領域において自然語処理技術へのニーズが高まっている。自然語解析は、そのプロセスから形態素解析や構文解析、意味解析、文脈解析等に分けることができ、最近では具体的な応用課題としてテキスト自動要約が注目されている。

しかしながら、自然語表現の多様性に対応するためには、膨大なコーパスや辞書的知識が不可欠であり、容易に利用できるまでにはなっていない。

本稿では、文書データを簡易的に抽象化する方法について検討を行う。自動要約は、文章中の重要な箇所をある規準により自動抽出し、さらに必要な大きさに自動縮約するための技術であるが、ただの要約あるいはエッセンス以上に、文章中に何が書かれているのか、タイトルだけでなく、流れを知りたい場合も多々ある。たとえば、教材作成において文献を紹介する資料を作成する場合、結論だけではもの足りず、図解化や大まかに文脈をフォローしたいことも多い。

現状では、人間が文章を読み、自分にとって重要な箇所を抽出し、要約や図解を行うしか方法はないが、一般に文章データは冗長であり、抽象化することが求められる。抽象化は、人間の思考活動においてきわめて重要な機能であり、文章の意味や知識を理解することの本質は抽象化による概念間のネットワークを構成することにあると考えられる。

今回、文章を構成する段落単位に抽象化する方法について検討を行い、いくつかの知見を得たので報告する。

２．文章データの構造と抽象化

ひとまとまりの文章データは、一般にいくつかの段落により構成されている。書籍等の文献の文章には、章や節、項の見出しが付与されているが、段落単位に見出しはついていないため、私たちは斜め読みにせよ、自分の読みたい節や項は字面を目で追う必要がある。

一般に文章は複数の段落から成り、各段落は１つ以上の文により構成され、文章全体は段落をサブシステムとするシステムとしてみなすことができる。各段落は１字下がりで始まり、前後のつながりはあるものの著者の書きたいことのまとまりの最小単位であると考えられる。

文章の自動抽象化とは、簡単に言えば、「何がどうした」あるいは「何について書いてあるのか」を把握することを容易にするために、各段落に適切な見出しを自動付与する問題として捉えることができる。

抽象化は、知識や道具立ての観点からいくつかのレベルに分けることができる。

Level１：名詞句の文内語彙による抽象化

各段落において名詞句を抽出し、抽出した名詞句の上位概念を表す語彙を文中より探し、is-a 関係や part-of 関係として代表させる。

Level２：述語（動詞句）の抽象化

動詞は、語尾変化を伴うため、終止形に変換する作業が必要となり、活用形に関する文法や動詞辞書を必要とする。

Level３：文外語句による抽象化

通常、人間が行っている抽象化に相当し、Level１、２で抽出した名詞句や動詞句について意味を踏まえ、文外語彙の中から適切な上位概念を選択し抽象化する。大規模な概念辞書やシソーラスを必要とし、しかも、意味を考慮した最適なレベルの抽象化を行う必要がある。

本稿では、Level１の抽象化を行う具体的な方法について検討を行う。Level１は完全な抽象化ではないが、コーパスや辞書を必要とせず、しかも文中の語彙を用いた抽象化であることから最適性の問題を回避することができる。本研究では、形態素解析や構文解析等の大がかりな道具立てを用いず、日本語文章の文字種とてにをはを識別することにより名詞句の抽出を行った。

３．方法

A Study on the Automatic Abstraction of Text Data
Takao Ishizuka, Faculty of Business Administration, Asia University

対象とした文書データは、新聞の社説記事である。以下の作業を行うプログラムの開発を行った。

1）段落単位に記事データを読み込み、「。」により構成する文単位に分ける。
2）文字単位に文字コードの範囲から文字種をと「てにをは」を識別し、分かち書きを行い、切り出した語句に文字種パターンを付与し、基本語彙とする。その際、カタカナ＋漢字、数字＋漢字等はひとつの語句として扱う。
3）文字列の包含関係や語尾の文字によりくくり、共通語尾の文字を語彙に追加する。
4）文中から「～の…」、「～が…する」、「～を…する」（～と…は名詞句）の表現を抽出し、「～の…」の形に統一し、語彙に追加する。
5）以上の語彙について包含関係を考慮した重みつき頻度を計算する。
6）段落内の最終文の最後の名詞句を Key term とする。
7）段落内語彙の中から最高頻度の語句を抽出し、キータームと合せ、見出しを、
最高頻度の語句＋“と”＋Key term
として表現し、これらの語句に4）で追加した「～の」の修飾語がついている場合には、それらをつけて見出しとする。

なお、語句の頻度のレンジが基準値以下の場合には、段落内各文の主格部分の名詞句を併記することにより見出しを構成した。

段落内の文は並列に並んでいるのではなく、著者の言いたいことは最後の文に集約されることが多い。逆に、最初の文で結論を述べ、以下に続く文で事例を述べることもある。Key term の抽出においては、機械的に最終文の最後の名詞句ではなく、段落内文の数、文の時制（過去形かどうか）等を元にヒューリスティックに判別を行った。

以上のアルゴリズムは、各段落内最適化であり、文章全体として見た場合、その見出しが必ずしも適切とは言えない場合もある。そこで、文章全体の語彙頻度をもとに以下の修正頻度により見出しを構成し、比較を行った。

①TF-IDF 値
$fij \cdot (\log(n / \Sigma \delta ij) + 1)$
②割引頻度Ⅰ
$fij \cdot (1 - (\Sigma \delta ij - 1) / n)^{\wedge} r$
③割引頻度Ⅱ
$fij \cdot (1 + r^{\wedge}(fi. - fij)) / 2$
ただし、
fij： 語句ｉの段落ｊにおける頻度
fi.： 語句ｉの文章全体での頻度
ｎ ： 段落数
δij： 語句ｉが段落ｊに存在する⇒1
nor ⇒0
ｒ ： 割引率のパラメータ

４．結果並びに考察

毎日新聞 2001 年版 CD-ROM に収載された社説記事データを対象にプログラムを実行した。図１に 2001 年１月１日の社説について行った結果を示す。段落内最高頻度をそのまま用いた場合と、①～③の修正を行った頻度による場合とで見出しの比較を行ったところ、ほとんどの段落で一致した。したがって、計算の簡便さを考慮すれば、パラメータの調整等のヒューリスティクスを避けるためにも段落内最適化でも十分実用になりうると判断された。

文章データからこのような形で見出しの候補が容易に生成できれば、見出し間の関連によりさらに大きなくくりでくくることもでき、文章の自動図解化も可能になると考えられる。

図１．自動抽象化により作成された見出しの例（毎日新聞 2001 年１月１日社説の場合）

［社説］縦の秩序から横の秩序へ　求められる国民の自発性

◆ 新しい・世紀
◆・ 昨年・ 今年と {普通} の新年
◆ {世界共通} の暦と現実
◆ 区切り・機会
◆ {既存} の組織と問題解決
◆ 組織と {個人} の復権
◆ 手工業時代と政府
◆ 個人と工業化
◆国家と {２度} の大戦
◆ {通信・輸送技術} {コンピューター} の発達とサービス業
◆ 組織と閉塞状態
◆ {ジョイントベンチャー} の組織と問題解決
◆ 国家と { {横} の秩序} の重視
◆／ {インターネット} の発達／ {地球市民} の連帯／ {団体} の組織化
◆／ {縦} の秩序感覚／ {自治体} のシステム
◆ 一人一人と２１世紀