社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－170 (18)
2005／11／22

関連性理論を用いた文章の自動要約

石塚 隆男
亜細亜大学経営学部

本研究では、テキスト自動要約を行う際のひとつの観点あるいは評価規準として関連性理論を活用することを目的とする。関連性理論によれば、人間は最小の処理労力でできるだけ多くの認知効果が得られるように発話や文章を理解するとされているが、説得力や発展性のある理論であるにも拘らず、定量的な研究はほとんどなされていない。本研究では、文章をパラグラフを要素とする構造体として認識し、各パラグラフの情報量を定量化することにより情報量が増加したパラグラフのみを抽出することにより文章全体の要約が行えることを示した。自動要約の問題は、関連性理論により新しいパラダイムが登場することが期待される。

Text Autosummarization by Relevance Theory

Takao Ishizuka
Faculty of Business Administration, Asia University

We propose a new method of text autosummarization by Relevance Theory. Text summarization is often formulated as optimization problem with outer or physical constraints, e.g. within a given number of words. However, in case we want to follow the context or intention of the author from a text by means of minimal labor, our method, that is, extraction of paragraphs with high information by Relevance Theory is very useful. Relevance Theory is expected to a new paradigm of text summarization norm or criterion.

1. はじめに

本研究は、テキスト自動要約を行う際のひとつの観点あるいは評価規準として関連性理論の活用可能性について検討を行うことを目的とする。

テキスト自動要約は、自然言語処理の実用・応用領域として既に多くの研究がなされている（奥村・難波(2005)）。多くの自動要約のタスクは、与えられた字数内で要約を作成する条件付最適化問題として定式化され、原文から見て要約字数の上限が適切かについて議論の余地はない。また、一般的な読者を想定しての要約か、あるいは特定の視点からの要約か等のオプションが存在し、原文に対して唯一最適な要約は存在しない。

このように、テキスト自動要約の問題は物理的かつ外的な必要条件からの制約のもとでの情報抽出・再構成（あるいは再配置）を行う最適化問題であり、必ずしもテキスト本来の文脈や意図を反映した要約ではない。私たちは、テキストに書かれている文章内容が重要かどうかを判断する前提として当該テキストがそれなりの文字数を費やして何について書かれているのかを知りたいことが多い。したがって、自動要約以前の課題として、与えられた文章全体を抽象化し、俯瞰できるように可視化する＝図解化することが必要であると考える。図解化は、要素間の関連に重点を置いた“要約”の一手法であり、テキストの構成要素として本研究ではパラグラフに着目する。

今回、従来の語用論を越えたシステム的発想に基づく Sperber&Wilson の関連性理論を用い、自動要約への活用可能性を検討し、いくつかの知

見が得られたので報告する。

2. パラグラフによる構造の同定

一般に文章は複数の段落（パラグラフ）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。

しかしながら、新聞記事等の文章やWeb上の文章では、読みやすさを考慮してか、1段落1文の文章も数多く存在し、形式的な段落から本来の意味的なパラグラフの同定が必要となっている。

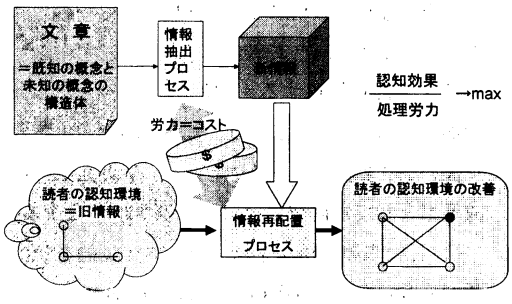
本研究では、従来から新聞記事の構造として知られている"逆ピラミッドモデル"仮説の検証を試み、形式的段落から意味的なパラグラフへの自動的な集約方法を検討する。

3. 関連性理論の解釈

関連性理論は、「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) ことを前提とした理論であり、高い説得性をもち、今後の発展が注目されている。関連性理論は、人間の情報処理行為を合理性の観点から説明したものであり、コストの概念が導入されている。しかしながら、処理労力、すなわち、コストをどのように定量化すべきかについては答えておらず、本理論の数量的な扱いに関する研究はほとんど見当たらない。

図1は、関連性理論に基づく認知プロセスを図示したものである。

図1. 関連性理論による認知プロセス



人間の認知行為には、情報抽出と情報再配置の2つのプロセスが関与しており、時間や労力を惜しまなければ無限のコンテキストや仮説を想定あるいは生成しうるが、フレーム問題に陥るため現実にはサイモンのいう限定された合理性あるいは満足度規準が働いていると考えられる。

そこで、文脈効果や情報処理労力の定式化について検討する。文脈効果は、新規に追加される情報と既存の知識とで張られる部分空間の説明力を表わしている。一方、処理労力はペナルティであり、それだけの説明力を実現するのに必要な変数や次元の数で定量化が可能であろう。このように考えると、関連性理論に基づく認知とは、オッカムの剃刀やケチの原理に匹敵する思考法であり、それらに科学的な根拠を与えたということができよう。その意味で少数の次元や因子により説明力を高める主成分分析や因子分析等のデータ解析手法は理にかなっているといえよう。

テキスト自動要約に関連性理論を実装化する方法にはさまざまな方法が考えられる。目的関数や評価関数の定式化に関連性理論をどう組み込むかについても同様である。たとえば、生産性や効率のように比率により目的関数を表現するか、あるいはAICやMDL規準のようなコスト関数を表現することも考えられるが、導出するためにはテキストあるいは相当するベクトル空間データを確率モデルとして表現する必要がある。そこで、本研究では具体的な目的関数を用いず、情報量の観点から関連性理論の実装化を行う。

4. 関連性理論に基づく文章構造の可視化

関連性理論によれば、人間は文章の理解を通じて自己の認知環境を改善したいという希望をもつ。しかし、文章の理解に多くのコスト＝前提知識を要し、それに見合うだけの改善効果が期待できない場合には努力を放棄する。関連性理論における文脈効果＝認知効果とは、読者の想定あるいは保有する旧情報を更新する新情報との相互作

用であり、処理労力は新情報である未知の概念数に比例すると考えられる。

そこで、文章の処理労力を最小にし、かつ文脈効果が最大になるようにするためには、当該文章上のあらゆる未知の概念をとりこむのではなく、絞ることにより労力を抑え、同時に最も文脈効果のある部分に焦点を当てそこを中心に読めばよいことになる。

任意の文章について読者の知識レベルは予め想定できないので、絶対的な知識レベルではなく、相対的な知識レベルを考えよう。すなわち、文章をパラグラフ順に読み進めていくうちに新情報の量がどのように変化するかを定量化する。各パラグラフの文章は、第1パラグラフ～直前のパラグラフまでの旧情報と当該パラグラフにおける新情報が共起した相互作用の結果として表現されている。

新情報は、文章中の新語によってもたらされる。文脈効果上、有意なパラグラフとは新情報の追加が顕著なパラグラフであり、そうしたパラグラフを抽出することは読者の処理労力を最小にすることにつながると考えられる。

5．単語×段落マトリクスの情報構造

対象とする文章の形態素解析を行い、各単語がどのパラグラフに何個出現したかの集計を行う。図2は、このマトリクスを示したものである。

図2．単語×段落マトリクス

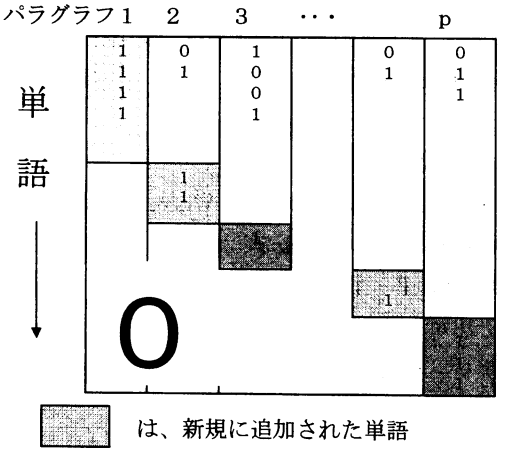


図2の列ベクトルであるパラグラフベクトルは、パラグラフ間に共起語がなければ、すなわち、新語だけを見れば直交していることがわかる。パラグラフiの情報量をHiとする。重要度が情報量によって測定できると仮定すれば、逆ピラミッドモデルのもとでは情報量は単調減少し、

$$H_1 > H_2 > H_3 > \cdots > Hp$$

が成立する。

各パラグラフの新語数と総単語数を調べ、情報量を次式により定義する。

$$H_i = -\log(1 - p_i)$$

ここで、p_i：　パラグラフiの新語率

p_i＝パラグラフiの新語数／パラグラフiの総単語数

とする。

新聞記事の場合、H_iの値は単調減少することが期待できる。そこで、H_iの符号変化を調べることによりH_iが減少から増加に転じたパラグラフでは、新情報の追加が有意であり、当該パラグラフで話題の転換があったと考えられる。また、情報量の増加のみられたパラグラフだけを抽出することにより、処理労力を最小にし、しかも文脈効果のある新情報のみを得ることができよう。また、情報量が単調減少にある連続するパラグラフをグループ化し、意味パラグラフ化することも可能である。情報量の変化を調べることにより文章構造の同定と可視化が可能になる。

図3．記事文章のパラグラフ別情報量の変化

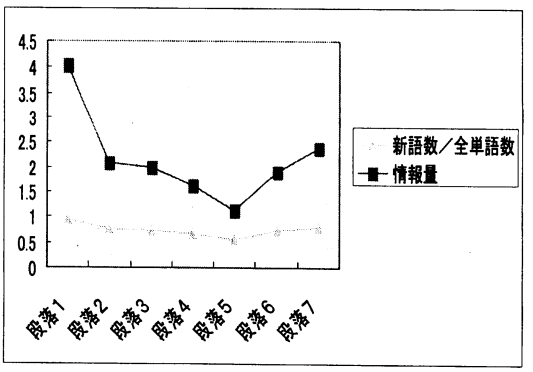


図４．図３で用いた新聞記事文章を自動図解化した例

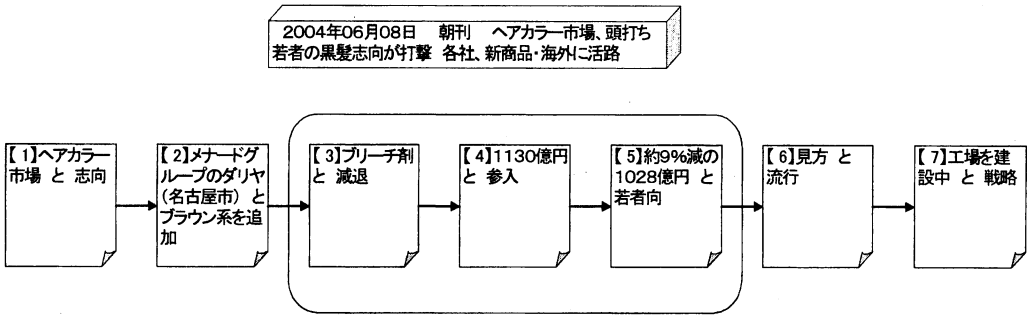


図５．英文記事文章の情報量の分布の解析例

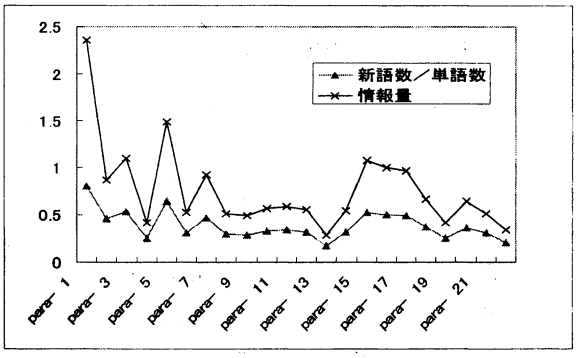
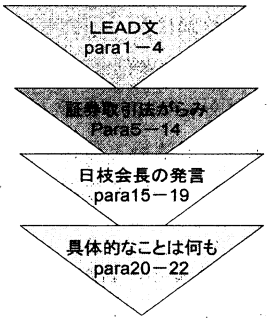


図６．図５の英文記事の意味的パラグラフの図解化（手作業）



６．解析例

以上の方法により、新聞記事文章から情報量の変化を調べ、それをもとにパラグラフ間の構造を図解化した例を図３～６に示す。

図３は、７パラグラフから成る日本語新聞記事の情報量の変化を示している。この記事では、第６パラグラフで話題が転換していることにより情報量が増加していることが確認された。したがって、第１・第６・第７パラグラフを抽出することにより文脈効果のある新情報だけに絞ることができると考えられる。図４は、同記事の各パラグラフ内の単語の位置情報を考慮することにより見出しを付与し、Excel の VBA スクリプトとして出力したものである。

図５、６は英文記事を対象に解析を行ったものである。意味的パラグラフの構造を把握することができ、これらの意味的パラグラフ単位に自動要約を行うことにより記事文全体の文脈を伝達することができる。

７．考察並びに今後の課題

本研究では、人間の認知行為を説明する理論である関連性理論が文章の自動要約に合理的な根拠を与えうるのではないかとの仮説をもとに検討を行った。今回は、処理労力を小さくし、しかも文脈効果を高めることを同時にねらい、新語数の割合を情報量とみなし、図解化を行った。

本来、逆ピラミッドモデルならば、第１パラグラフだけ読めば十分なはずであるが、現実には、図６に示すように多重逆ピラミッド構造であり、自動要約の前処理といてこのような可視化は十分意味があると考える。

参考文献

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)

奥村学・難波英嗣(2005)『テキスト自動要約』オーム社