言語処理学会 第 4 回年次大会発表論文集, pp.572–575, 1998

ウェブ英文ページの速読支援

奥西 稔幸　吉見 毅彦　山路 孝浩　福持 陽士
シャープ (株) ソフト事業推進センター
大和郡山市美濃庄町 492 番地
{okunishi,yoshi,yamaji,fuku}@isl.nara.sharp.co.jp

1　はじめに

WWW（ウェブ）上に急速に増加し始めた英語情報を読解支援するツールとしてＰＣ用の英日機械翻訳ソフトが市販され、主に、英文ページの概要把握の目的で利用されている。我々は、この用途をさらに進め、ウェブ上の英文ページから必要な情報を素早く抽出するという速読支援を目的として、1. ウェブページのタイトルを基に重要な文を選択し翻訳する機能、2. 各段落に小見出しを付与することで全体の概要を素早く把握する機能、3. 固有名詞や日付表現などの特徴的な表現を含む文を抽出し翻訳する機能、を英日機械翻訳ソフト上に試作している。本稿では、紙面の都合上、最初の 2 機能の実現手法について報告する。

2　重要文選択

本節では、表層的な情報を手がかりとして文と文のつながりの強さを評価し、その強さに基づいて文の重要度を決定する手法を提案する。提案手法では文の重要度に関して次の仮定を置く。

1. タイトルはテキスト中で最も重要な文である。
2. 重要な文とのつながりが強ければ強いほど、その文は重要である。

この仮定に基づいて、文のタイトルへのつながりの強さをその文の重要度とする。文と文のつながりの強さの評価は、1) 人称代名詞と先行 (代) 名詞の前方照応と、2) 同一辞書見出し語による語彙的なつながりを検出することによって行なう。

重要文を選択するために文間のつながりを解析する従来の手法としては、1) 接続表現を手がかりとして修辞構造を解析し、その結果に基づいて文の重要度を評価する手法 [8, 9] や、2) 提案手法と同じく、語彙的なつながりに着目した手法 [1, 4, 7, 10] がある。文と文をつなぐ言語的手段には、照応・代用・省略・接続表現の使用・語彙的なつながりがある [6] が、接続表現の使用頻度はあまり高くない。このため、前者の手法には、接続表現だけでは文間のつながりを解析するための手がかりとしては十分でないという問題点がある。後者の手法では、語彙的なつながりとほぼ同じ頻度で見られる照応を手がかりとして利用していない。

2.1　文の重要度

テキストを構成する文 $S_1, S_2, \cdots, S_n$ の間で次の条件が成り立つと仮定する。

1. 冒頭文 S_1 はどの文にもつながらない。
2. S_1 以外の各文 S_j について、S_j が直接つながる先行文 $S_i (i < j)$ が唯一つ存在する。

この仮定に従えば、文が同時に複数の先行文に直接つながることはないので、図 1 に示すように、テキスト構造は冒頭文 S_1 を根節点とする木で表される。

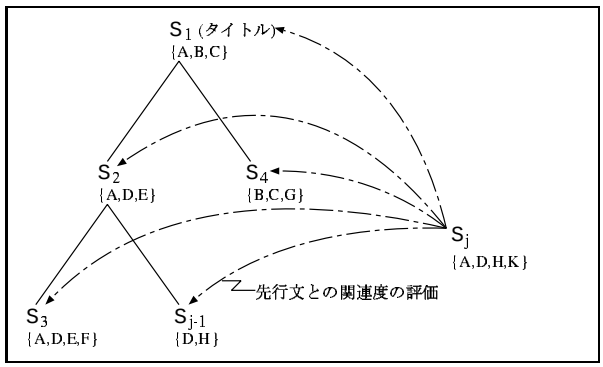

図 1: 文の重要度の評価

テキストの冒頭文 S_1 は、多くの場合、そのテキストのタイトルであるので、S_1 にはテキスト全体で最大の重要度を与える。冒頭文 S_1 以外の文 S_j の重要度は、S_j から先行文 S_i へのつながりの強さ (関連度) と S_i の重要度によって決まると考え、文 S_j の重要度を求める

式を次のように定める。

$$S_j\text{の重要度} = \max_{i<j}\{S_i\text{の重要度} \times S_i\text{と } S_j\text{の関連度 }\} \quad (1)$$

文の重要度を (1) 式で求めるためには、まず二つの文の間の関連度を求めなければならない。

2.2 二文間の関連度

文 S_j の先行文 S_i へのつながりの強さ (関連度) を求める式を次のように定める。

$$S_i\text{と } S_j\text{の関連度} = \frac{M_{i,j}}{N_i} \quad (2)$$

ただし、$M_{i,j}$ は、文 S_j 中の重要語[1] のうち、先行文 S_i の題述中の重要語につながるものの重みの和であり、N_i は、先行文 S_i の題述中の重要語の数である。(2) 式の意味は以降で説明する。

人称代名詞と先行 (代) 名詞の照応の検出　しばしば指摘されるように、代名詞との間で照応が成り立つ先行 (代) 名詞は、代名詞を含む文 S_j と同一文中、あるいは S_j の直前の文 S_{j-1} に現れることが多いので、先行 (代) 名詞の検索対象文を S_j と S_{j-1} に限定する。検索は S_j、S_{j-1} の順で行ない、S_j 中の (代) 名詞との照合が成功した場合は、S_{j-1} に対する処理は行なわない。

重要語の語彙的なつながりの検出　二つの文に現れる重要語が文字列として一致するとき、両者の間に語彙的なつながりがあるとみなす。二つの文がある一定の距離以上離れていると、それらに含まれる重要語の文字列照合が成功しても、二つの文の間に直接的なつながりはないと考えられる。このため、直接的なつながりの有効範囲を文 S_j から五文前までの先行文 $S_i(j-5 \leq i < j)$ とする。

タイトル語への重み付け　テキストのタイトル中に現れる重要語 (以降、タイトル語と呼ぶ) は、そのテキストにおいて重要な情報を伝えると考えられる。従って、タイトル語を含む文の重要度を大きくするために、他の重要語に与える重みの値よりも大きな値を与える [2, 8, 11] のが適切である。ここでは、タイトル語に対する重み付けを行なう際に、テキスト中でのタイトル語の出現頻度を考慮する。タイトル語を含む文の重要性は、タイトル語がテキスト中に頻繁に現れる場合は、タイトル語を含まない文の重要性に比べて特に高いわけではないが、タイトル語がテキスト中に希にしか現れない場合には、タイトル語を含まない文に比べて特に高くなると仮定する。実験では、タイトル語を含む文の数がテキストの総文数の 1/4 以下である場合に限り、タイトル語の重みを 5 とした。

[1] 品詞が名詞・人称代名詞・動詞・形容詞・副詞のいずれかである辞書見出し語を重要語と呼ぶ。

先行文の題述へのつながり　テキストは、通常、ある文 S_i での題述 (rheme) が後続文 S_j での主題 (theme) となり、それに新たな情報が付け加わるという形で展開する [5]。従って、文 S_j の先行文 S_i へのつながりの強さの評価を、S_j が S_i の題述をどれだけ多く主題として受け継いでいるかに基づいて行なう。主題と題述は、文の前半部分が主題、後半部分が題述というように文中の位置で区別されることが多い [3] が、ここでは、文中の位置ではなく、関連文[2] とのつながりに基づいて区別する。具体的には、文 S_j 中の重要語のうち、S_j の関連文中の重要語との文字列照合が成功するものが文の S_j の主題を構成し、成功しないものが S_j の題述を構成するとみなすことである。関連文を持たない冒頭文 S_1 では、それに含まれる重要語すべてが題述を構成するとみなす。例えば、図 1 において、括弧内の英大文字を各文に現れる重要語とすると、各文の主題と題述は表 1 のように分けられる。

表 1: 図 1 の各文の主題と題述

文	関連文	主題	題述
S_1	—	—	A,B,C
S_2	S_1	A	D,E
S_3	S_2	A,D,E	F
S_4	S_1	B,C	G
:	:	:	:
S_{j-1}	S_2	D	H

2.3 実験

実験には英文テキスト 80 編を用いた。テキストの総文数は、最も短かいもので 12 文、最も長いもので 64 文、一テキスト当たりの平均では 29.0 文であった。各テキストについて、第三者によって重要と判断された文を、選択すべき正解文とした。正解文の数は、平均で元テキストの総文数の 17.9% であった。

まず、各テキストについて、正解文と同じ数だけ文を選択するように設定して重要文選択実験を行なった。この場合の精度 (再現率と適合率は同じ値となる) は、72.3% であった。次に、提案手法の精度と、WWW 上で試用可能なシステム A、市販されている三つのシステム B、C、D の精度を比較した。それぞれの平均再現率と平均適合率を表 2 に示す。システム A、B、C、D の文選択率は、各システムの既定状態で選ばれた文の数とテキストの総文数から逆算したものである。提案手法の文選択率は、四システムの文選択率とほぼ同じである 25% とした。表 2 より、比較的短いテキストに対して提案手法が有効であると考えられる。

提案手法があまり有効に働かないテキストは、複数のサブトピックから構成されているテキストであった。一般に、トピックが切り替わると、それまでとは異なっ

[2] (1) 式において、$\{S_i$ の重要度 $\times S_i$ と S_j の関連度 $\}$ の値が最大となるときの先行文 S_i を S_j の関連文と呼ぶ。この値を最大にする先行文が複数存在する場合は、S_j との距離が最も近いものを関連文と呼ぶ。

表 2: 提案手法と他のシステムの精度比較

	再現率	適合率	文選択率
提案手法	78.2%	57.7%	25%
システム A	72.3%	52.6%	26%
システム B	61.7%	39.5%	29%
システム C	61.4%	40.9%	29%
システム D	57.5%	42.2%	27%

た語彙が用いられるようになる。このため、提案手法のように語彙的なつながり (と人称代名詞による前方照応) に基づいて文と文のつながりを評価する手法では、トピックが切り替わる文から先行文へのつながりが弱いと判定され、トピック切り替わり文に対して与えられる重要度は小さくなる。従って、トピック切り替わり文が正解文であるようなテキストでは、高い精度を得ることが難しくなる。

3 段落見出し付与

英語があまり得意でない人が英語文書を読む時の手間の１つに、辞書引きや翻訳などの苦労の末に読解できた文章から必要な情報を得るため数日後に読み直そうとした場合、どこに何が書いてあったかすらわからなくなりもう一度同じ苦労（辞書引きなど）を繰り返さないといけないというもどかしさがある。本稿では、このような状況で、一度読んだ英語文章のどこに何が書いてあったかが容易に思い出せる程度に適切な見出しを付与することを目的とする。

文書中の各段落に振られた見出しには

1. 段落内のキーポイントを的確に表したもの
2. 興味を誘うショッキングなもの
3. 文書全体中のその段落の位置づけを表したもの

など様々な役割がある。１はある意味では段落の究極の要約であり、２は読者の興味／関心など更に一歩進んだものである。いずれの見出しも意味処理などが必要で容易に作り出せるものではない。本稿で提案する見出しは３に近い役割を果たすもので、表層的な情報だけで生成することとする。

3.1 文見出しの生成

段落中の各文から文見出しを生成した上で、それらを比較し最も適切な見出しを段落見出しとして付与する。

一文の見出しは、文の主要素と特徴語に基づき以下の手順で生成する。なお、速読支援という目的上、軽い高速処理が必須なため構文解析はもちろんのこと、辞書引きすら行なわずに簡単な形態素処理（誤基に戻す処理）結果だけから生成する。

文の主要素からの見出し生成　以下の３つの処理から見出しを生成する。

1. 各単語の品詞を推測する。
   (a) 語基変形 (stemming) の適用
   (b) 機能語（接続詞、前置詞等)、副詞の登録
   (c) 単語の前後関係を使った規則
   (a) の語基変形（stemming）とは、例えば、”naturally”の末尾の ”ly” を取り除く処理を言うがこの規則が適用された単語は副詞であると推測する。(b) は機能語、副詞（但し ”ly” で終るものは除く）を各々約 500 語登録した。この２つだけでは、原形や不規則変化形の単語の品詞が推測できない。このため「the の次の品詞不明語は名詞[3] 」などの簡単な規則を組み込んでいる。
2. 文の主要素（主語、主動詞、目的語）を判定する。基本的には文の最初に出現する名詞と動詞を各々主語と主動詞、主動詞の後に最初に出現する名詞を目的語とする。但し、文頭の前置詞句／副詞節や特殊構文 (it is ～to 不定詞/that, there is) には対応している。
   これだけでは、関係節を含む文や疑問文、更には構文解析でも難しい倒置、省略表現を含んだ文など、主語述語の位置が大きく変わる文に対して誤る可能性はあるが、そのような文からは見出しを生成しない。
3. 見出し変形規則
   主語や主動詞に応じた変形規則を設けている。
   - ”(主語) ＋発言を表す動詞 (say,talk,speak 等)” → ”(主語)’s words” 『(主語) の言葉』
   - ”You can (主動詞)” → ”How to (主動詞)”『～する方法』

   これらの規則に当てはまらなかった場合は、主動詞を名詞化したものを主語と組み合わせものや主語だけを文の見出しとする。

文の特徴語からの見出し生成　以下の表層的な情報を手がかりに文の特徴語を抽出し、見出しとする。

- 頻出語、タイトル語を含む文でそれら以外の語
- 初出語
- 引用句
- 数字表現を含む語句
- 強調の副詞や接続詞

見出し表現　生成する見出しは翻訳することが前提なので見出しらしい表現（be 動詞欠け、他）ではなく、逆に素直に翻訳できるものとした。

見出し度　このようにして生成した見出し候補には以下の調整により評価値（見出し度）を与えており、その最大値をもつ見出し候補を文の見出しとする。

- 見出し変形規則自身が見出し度を持つ。
- 品詞推測や文の主要素認識に誤った可能性のある文、また否定語を持つ文の見出し度は低くする。
- 逆接の接続詞を持つ文の見出し度は高くする。

[3] 形容詞や動詞分詞形もありえるが見出し生成の目的からは区別の必要がない

- タイトル語を持つ文、段落先頭文の見出しは高い。

図 2 に見出しの生成例をあげる。文中のイタリックの単語を基に太字部分の見出しを生成している。

Sharp's words: "We have developed the HR-TFT Super Mobile LCD" *said Sharp.*

Start of production: *Production* will *begin* on the HR-TFT Super Mobile LCD at Sharp's Tenri Plant in Nara Prefecture in January 1998.

図 2: 見出しの生成例

3.2 段落見出しの選択・修正

段落各文の見出しの中で見出し度が最大のものを段落見出しとする。段落の区切りは基本的には HTML タグ情報 (<p>,
,<li>他) に基づき決めているが、段落見出しの内容によっては付与しない場合もあり（小さい段落見出しや見出し度が低い見出し)、この場合見かけ上は、２つの段落を併合したことになる。

また、見出し付与の目的（読み直し時の想起キーワード）上、同一見出しや類似見出しが多く現れるのは好ましくない。同一見出しが連続する場合は 2 つめ以降の見出しを付与せず、類似した見出しが連続する場合は差異部分だけを付与する。

3.3 付与見出しの有効性

以上の規則に基づく段落見出し付与機能を試作してみたところ、次のようなことがわかった。

- 主要素以外（修飾語句）が重要な場合もある
- 並列（文並列含め）を含む文は対比が重要である
- 初出語を見出しとした場合に良い見出しが多い
- 「〜の発言」のような見出しではなく発言内容を知りたい場合もある

ウェブページはそもそも HTML タグ等により読み易さが工夫されており、最初から見出しを持つページも多い[4] 。が、英語が得意でない人にとっては見出しの数が少ないこと、更には同じウェブページでも速報性を重んじるニュース記事ではそういった見出しすらないこと、などを考えると、本手法で提案する見出しも有用ではないかと考える。

4 おわりに

HTML のタグ情報を利用した重要文選択では良好な結果が得られた。また、段落見出し付与は、この重要文選択（抄録）と真の意味での要約文章生成との中間に位置付けできる技術ではないかと思う。

今後は各機能とも速読支援という観点からどのような効果があるかを十分に評価をした上で精度向上を図るとともに、今回は列挙／保存するだけであった抽出結果に対しても、整理の観点からの何らかの操作（既に保存している内容は抽出しない、など）も考え合わせ、積極的な情報収集活動に有用な支援形態を提案していきたいと考えている。

謝辞 日頃よりご指導を頂きましたソフト事業推進センター 仲川與万所長に感謝致します。

[4] 原文にある見出しとの関係をどうするかは課題である

参考文献

[1] A. Collier. A System for Automating Concordance Line Selection. In *Proceedings of NeMLaP*, pp. 95–100, 1994.

[2] H. P. Edmundson. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, Vol. 16, No. 2, pp. 264–285, 1969.

[3] 福地肇. 談話の構造, 新英文法選書, 第 10 巻. 大修館書店, 1985.

[4] 福本淳一. 文の結合度に基づく内容抽出法. 言語処理学会第 3 回年次大会発表論文集, pp. 321–324, 1997.

[5] T. Givon. From Discourse to Syntax: Grammar as a Processing Strategy. In T. Givon, editor, *Discourse and Syntax*, Vol. 12 of *Syntax and Semantics*, pp. 81–112. Academic Press, 1979.

[6] M. A. K. Halliday and R. Hasan. *Cohesion in English.* English Language Series 9. Longman, 1976.

[7] M. Hoey. *Patterns of Lexis in Text.* Describing English Language. Oxford University Press, 1991.

[8] 間瀬久雄, 大西昇, 杉江昇. 説明文の抄録作成について. NLC89-40, 電子情報通信学会, 1989.

[9] K. Ono, K. Sumita, and S. Miike. Abstract Generation based on Rhetorical Structure Extraction. In *Proceedings of COLING*, pp. 344–348, 1994.

[10] 佐々木一朗, 増山繁, 内藤昭三. 語彙的結束性に着目した文章抄録法の提案. NL98-9, 情報処理学会, 1993.

[11] H. Watanabe. A Method for Abstracting Newspaper Articles by Using Surface Clues. In *Proceedings of COLING*, pp. 974–979, 1996.