自然言語処理 128−20
(1998. 11. 6)

要約文生成のための単語抽出方法

窪田 健一　　山下 浩一　　吉田 敬一

{kubota, yamasita, yoshida}@cs.inf.shizuoka.ac.jp

静岡大学大学院理工学研究科

概要

本論文では、文章の要約に必要な単語の抽出手法を提案する。これまでの単語抽出手法は、出現頻度を中心に考えられていた。しかし、文章中には、出現頻度は低いが要約には必要な語もある。そこで、単語間に連想関係という概念を導入し、文章中に使用された単語から作成した単語間ネットワークを利用して、文章の意味を考慮しながら、文意を表していると思われる単語を抽出する手法を述べる。また、この手法を用いて抽出された単語に対する妥当性の評価をおこなった。評価の方法は、上記の単語をもとに抽出された文が、原文の要旨として適当か否か、人により判定した。

A Word-Extracting Method for an Abstract

Kenichi Kubota, Kouichi Yamashita, Keiichi Yoshida

Graduate School of Science and Engineering, Shizuoka University

Abstract

In this paper, we propose a method for extracting words to make an abstract. Most methods are based on a word frequency. In composition, several words have a low frequency, but some of them play an important role in making an abstract. We introduce a tool of word-network constructed by connecting two or more words. Using this idea, we propose a word-extracting method for an abstract.

1　はじめに

文章の要約には、文章中に用いられている単語の重み付けを行い、それを基に重要と思われる文や段落を抽出する手法がある。単語の重み付け計算は、キーワード密度法やTF*IDFなど、単語の出現頻度を基本にしたものが多い。しかし、単語の重要性は、全てが出現頻度にしたがっているわけではなく、出現頻度が低くとも要約文には必要な単語は存在する。これらの単語は、文章の表す意味を考慮して抽出する必要がある。これに対して、文脈を利用する方法[7]も提案されているが、対象となる文章の分野が確定されていなければならず、また、あらかじめ用意する知識が必要である。

そこで、本論文では文章中に用いられる単語間に連想関係という概念を導入し、単語間ネットワークを作成することにより、文章の意味を考慮しながら、表層から得られる情報のみに基づいて、文意を表していると思われる単語を抽出する手法を提案する。さらに、この手法を用いて抽出された単語を利用して各文の評価値を計算し、この評価値に基づいて文を抽出することにより抄録を作成する実験をおこなった。この抄録が、原文の要約文として適当か否かを人により判定した。

以下、第2節で本論文で扱う語を定義し、第3節で連想関係の概念を導入し、文章の意味の捉え方を説明

する。第4節で、本論文で提案する手法を述べ、第5節で実験とその評価を述べる。

なお本論文では、対象を話題が比較的まとまっている短い文章とし、複数の話題が含まれる文章は対象から外す。また、照応関係、否定表現の問題にはここでは触れない。

2 語の定義

本論文で使用する要約文に関しての主な語を定義する。

単語 (word) 一つの意味のまとまりをなし、文法上一つの機能を持つ最小の単位。

文 (sentence) まとまった意味を持ち、句点で区切られた一つ以上の単語の集まり。

文章 (composition) まとまった意味を表現するため、順序づけられた一つ以上の文の集まり。

基点語 (given word) 文章中に表れ、文章全体の主題や文意を特定するのに重要な役目をもつ単語。

関連語 (relational word) 基点語に付随して、より詳細に文意を表す単語。基点語自身は含まれない。

文意 (meaning) 文章が表現する意味。

本論文で使用する単語間ネットワークに関しての主な語を定義する。

単語間ネットワーク (word-network) 2つ組 (W,P) で与えられるネットワークをいう。ただし、Wは単語の集合、Pはある関係で結ばれた2単語の非順序対集合。
2単語間をある関係で結ぶことを、パスを張るという。

距離 (distance) 単語間ネットワーク上における2つの単語 w_i,w_j 間の最小のパス数を距離といい、$d(w_i,w_j)$ で表す。

連想語 (associated word) 単語間ネットワーク上で、単語 w に対する距離 k の連想語とは、単語 w からの距離が1以上 k 以下の単語をいう。すなわち、単語 w に対する距離 k の連想語の集合 A^k は、

$$A^k = \{x | 1 \leq d(w,x) \leq k\}$$

で表される。

交接語 (intersectional word) 単語 $w_i(i = 1,2,\cdots,n)$ に対する距離 k の交接語とは、単語 w_i に対する距離 k の各連想語の集合に共通して含まれる単語をいう。すなわち、単語 w_i に対する距離 k の交接語の集合 C^k は、

$$C^k = \{x | x \in \bigcap_{i=1}^{n} A_i^k\}$$

ただし、A_i^k は単語 w_i に対する距離 k の連想語の集合

で表される。

3 文の意味

3.1 連想関係

いくつかの単語を並べると、そこからある単語を想い浮かべたり、ある単語を特定することができる。例えば、

1. ボール、ドリブル、シュート、ゴール
2. ボール、投げる、打つ、走る

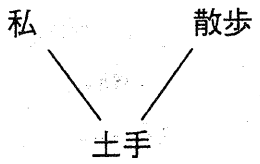

図 1: 単語の連想的意味表現の例

と単語を並べれば、1. のボールはサッカーボール、2. のボールは野球のボールを想い浮かべることができる。「ドリブル」、「シュート」、「ゴール」、「投げる」、「打つ」、「走る」といえば、サッカーと野球に共通するものとして、単に「ボール」を想い浮かべる。このように、いくつかの単語を人に与えると、人はこれまでに習得した知識に基づき、単語と単語の結びつきからある単語を想い浮かべることができる。このような単語間の関係を連想関係ということにする。連想関係から想い浮かべることのできるものは、与えられる単語群を受ける人の経験に大きく影響するが、本論文では、「単語の意味は、その単語を連想させる単語の集合で定義される。」とする。これを連想的意味と呼ぶことにする。

例えば、「私は、土手を散歩します。」という文を考える。この文だけからでは、「私」、「散歩」という単語が「土手」を連想させるので、「土手」の連想的意味を、「私」、「散歩」の単語の組が表現する。

3.2 文意

文章に対する意味にも、連想的意味を導入することができる。つまり、文章の中に使用されている単語、特に内容語 (content word) の組に連想的意味を応用すれば、文章に意味を割り当てることができる。すなわち、「文章の意味は、文章に使われている単語の集合が表している」と仮定する。この文章に対する連想的意味は、文章に用いられる単語数が少ない程、例えば、3 単語程度であれば、文意を連想しやすい。[9]。文意をより詳細に連想させるために、単語数を増やしすぎると、単語の組合せ数が急激に増え、逆に、単語の集合から意味を連想することが困難になるで、単語数は 10 単語程度が妥当であると考えられる。

3.3 文、段落、節

文章は、著者の特性が内在する。あるいは、節、時には段落が変わると、時間や背景、内容が大きく変わることが多い。これは、著者が表現しようとしている内容を変えようとしているからで、このような場合、節や段落をまたがって文意を考えることは意味がない。また、全体を対象に抽出される単語をもとに抄録文を作成すれば、まとまりのないものとなる。したがって、節や段落など、まとまりがある単位ごとに処理する。

4 提案する単語抽出方法

4.1 単語抽出方法の概要

準備として、与えられた文章を形態素解析した後、対象とする単語を文章中に現れる単語から、名詞 (形式名詞、数詞を除く)、形容詞、動詞 (形式動詞、いう、行う等を除く) の自立語に限定し、以下の手順に従って単語を抽出する。なお、手順の各段階の詳細については、続く節で順に説明する。

1. 与えられた文章から、単語間ネットワークを作成する。
2. 単語間ネットワーク上に基点語を指定する。
3. 単語間ネットワークにおいて、各基点語に対して、距離 1 の連想語の集合を求める。

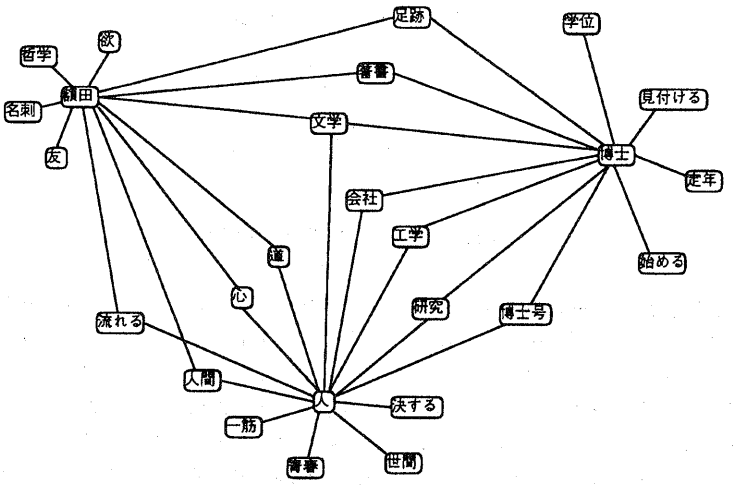

図 2: 単語間ネットワークの例

4. 3 で求めた各基点語に対する交接語を求める。この集合の要素を関連語とする。

以上により求められた単語を関連語とみなし、基点語とともに用いることにより、文意を表現していると考える。

4.2 単語間ネットワークの作成

単語間ネットワークの作成は、文章中に用いられる単語の集合 W と、文章中の共起関係にある 2 単語との 2 つ組 (W,P) を用いる。

与えられた文章から、名詞、形容詞、動詞を対象に出現順を変えずに抽出する。そして、これら抽出された各単語に対して、適当なウィンドウサイズ内にある単語へパスを張ることにより、単語間ネットワークを作成する。

ここで、単語は、ウィンドウサイズ内の自分以外のいくつかの単語の組により、連想的に意味が表されていると考えられる。例えば、図 2. の単語間ネットッワークから、「会社」、「工学」、「研究」に注目すると、この 3 単語からは、「会社で工学の研究をしている人」、「博士は、会社で工学の研究をしている」等、解釈することができる。

4.3 基点語の指定

基点語として、文章の主題を的確に表す単語や文意を特定する単語を、文章中に用いられている単語 W の中から指定する。

このための手法は、現在、よりよいものを模索中であるが、現段階では、Luhn の「一つの文献において、主題と関係が深い語は概して文献中に繰り返し出現する」[6] に従い、頻度により 2 ～ 4 単語を機械的に指定している。2 ～ 4 単語にしたのは、まとまりのある一つの文章の中で扱われる主題は、そう多くないと考えられるからである。

単語の頻度に差がなく、2 ～ 4 単語が抽出できない場合には、名詞を優先に人が決めるものとする。

4.4 連想語と関連語の抽出

作成された単語間ネットワークにおいて、各基点語 $w_i(i=1,2,\cdots,n)$ に対し距離 1 の連想語を抽出することにより、連想語の集合 $A_i^1(i=1,2,\cdots,n)$ を求める。単語間ネットワークの作成方法からみて距離 1 の連想語は、距離 2 以上の連想語より基点語を連想しやすい。

次に、各基点語に対する交接語を求める。すなわち、先に求めた距離 1 に対する連想語の集合に共通して含まれる単語を抽出する。つまり、

$$C^1 = \{x | x \in \bigcap_{i=1}^{n} A_i^1\}$$

の集合の各要素を、与えられた文章の関連語とみなす。この交接語は、C^1 の各要素は各基点語の連想語となっているので、各基点語 w_i と連想関係が存在する。すなわち、各基点語とともに文意を連想させる単語として考えることができる。したがって、C^1 に属する単語は、各基点語に付随して文意をより詳細に表す関連語として用いることができる。

5 実験と評価

5.1 実験の形態素解析処理

形態素解析には、JUMAN を使用した。JUMAN の解析結果に対して、解析エラーや未知語に対しては人手で修正を行った。また、JUMAN は人名に対し、姓と名を分けて出力するので、これらを合わせて一単語とした。複合語については、構成する各単語に意味があると考えられるので、JUMAN の処理結果をそのまま利用した。本システムが対象とした品詞は、名詞 (普通名詞、サ変名詞、形容詞、動詞、人名、地名、組織名、固有名詞)、形容詞、動詞 (する、なる、ある、いう、行う、等を除く) である。

5.2 実験における単語抽出処理

このあと、アルゴリズムにしたがい単語を抽出する。このとき、基点語の数は 3 を原則とし、本手法で抽出する単語数を、基点語と関連語合わせて 10 語程度が抜き出せるまで、ウィンドウサイズを 3 から順に 1 ずつ増やしていく。

5.3 実験に用いた文書

実験に用いた作品は 6 編で、文書のサイズは以下の通りである。

作品 1. 全文　著者 A　原文 約 6300byte　異なり単語数 約 840 単語

作品 1.§2　著者 A　原文 約 1400byte　異なり単語数 約 120 単語

作品 2. 全文　著者 A　原文 約 9600byte　異なり単語数 約 570 単語

作品 2.§4　著者 A　原文 約 2500byte　異なり単語数 約 200 単語

作品 3.　著者 B　原文 約 1800byte　異なり単語数 約 210 単語

作品 4.　著者 B　原文 約 2700byte　異なり単語数 約 140 単語

作品 5.　著者 C　原文 約 4100byte　異なり単語数 約 200 単語

作品 6.　著者 C　原文 約 5000byte　異なり単語数 約 450 単語

基点語	博士, 人, 額田
関連語	民俗学, 文学, 分野, 年, 道, 動機, 足跡, 人生, 人間, 人, 心, 時代, 私, 工学, 結び, 学位, 会社員, 会社, 一筋, 送る, 生きる

図 3: ウィンドウサイズ 7 の基点語と関連語

被験者 1.	民族学の分野で博士号をとった人がいた。彼は会社員として会社一筋に生きてきたが、ふと人生に疑問を感じ脱サラをして前から興味のあったこの分野で残りの人生を送ろうと思った。
被験者 2.	会社一筋に生きてきた人が、ある日突然、文学と工学と民族学に目覚めて、博士号をとった。人生とは、こんなものだ。
被験者 3.	額田さんは文学に傾倒していたが、時代の流れで父の会社を経営をすることになった。したこともない電気工学の仕事をし、新しい環境を受け入れなければならなかった。しかし、努力を重ね、実績を上げ、会社をたばねるようになった。人とは、努力すればそれなりの成果を得ることができる。好きなことでなくとも、あたえられたことをやれば、それなりの成果をあげることができるし、喜びや人生の意義を見いだすことができる。

図 4: 図 3. の基点語と関連語から連想される話

5.4 抽出された単語群から連想される文意 (評価方法 1)

作品 1. 全文に対し、本手法で抽出された単語群 (図 3.) から連想される意味が、原文が表す意味を十分に反映しているか否かを確認するため、抽出された単語群から連想することを 3 人の被験者に答えてもらった。その結果が、図 4. である。

被験者 1. は、おおむね話の筋を追ってくれている。被験者 3. は、話の筋は全く異なっているが、著者の意図したことが伝わっている。被験者 2. は、著者の意図に反し、人生に対して悲観的に解釈したと思われる。

5.5 従来の手法との比較 (評価方法 2)

本手法を用いて抽出された単語と頻度順に抽出された単語を用いて抄録を作成して評価をおこなった。抄録の方法は、各文に対し、以下の式で重要度を計算し、全文の重要度の平均値 μ と分散値 σ を求め、重要度 $\geq \mu + \sigma$ となる文を抽出したものを、要点のまとまり具合と文の完成度という観点から、3 人に評価してもらった。その結果を表 3. に示す。表の中の値は、良いと答えた人の人数である。

$$\text{一文に抽出された単語を含む割合} = \frac{\text{一文中の抽出された単語数}}{\text{一文中のシステムの処理対象となった単語数}}$$

作品 3.、作品 4. については、頻度によって抽出した単語を用いた方がよいようである。これは、著者により使われる言葉の特性が影響しているものと思われる。

表 1: 抄録文の評価

	作品 1§2	作品 2§4	作品 3	作品 4	作品 5	作品 6
頻度による手法	0	0	3	3	0	0
どちらとも言えない	0	1	0	0	1	0
本手法	3	2	0	0	2	3

6 考察

本手法では、単語間ネットワークを利用したため、出現頻度が低い単語も抽出の対象となる。これにより、文意をより正確に反映する単語の抽出が可能となった。

3 人の被験者に、本手法で抽出された単語群から原文を推測してもらった。その結果から考えて、文意を伝える要約文作成に必要な単語がおおむね抽出されてきていると考えられる。

また、この手法で抽出された単語を用いた文の重要度は、頻度によって抽出した単語を用いて計算した場合より、差がはっきりと現われる。これは、関連語が基点語に付随して現われるためと考えられる。そのため、頻度による方法と比べて、文を抽出しやすい。そして、本手法で抽出した単語を基に作成した抄録は、人が判断して、頻度で抽出された単語を用いたときと比べて要点がよりまとまっており、話の流れもより滑らかになる傾向があることがわかった。

しかし、本手法では、基点語の選び方により抽出されてくる単語が異なってくるために、十分な配慮が必要である。主題をよく表している、または、主題に大きく関係している語を指定することができれば、文の重要度は、より真値に近付くと思われる。

また、異なる単語が、ほぼ同じ意味に使われる場合があり、これらを同一の単語と扱うことができれば、抽出される単語数が増えると予想されるので、重要度の精度が変わる。また、抄録を作成する際の文の重要度の閾値は、現在のところ確定しておらず、これらは、今後の課題である。

7 まとめ

与えられた文章から単語間ネットワークを作成し、これに基点語を与えることにより、要約文作成に必要な単語の抽出手法を提案した。抽出された単語が、原文の意味を反映しているものであるか人に原文を予想してもらったところ、3 人中 2 人から適当と思われる回答を得た。また、抽出された単語と頻度により抽出した単語を用いて抄録文を作成し、6 作品中 4 作品が本手法で抽出した単語を用いたほうが良いという回答を得た。

今後、基点語の指定方式、文を抽出する基準を定め、より意味を反映した要約文の作成を試みたい。

参考文献

[1] Charniak, E. : Statiscal Language Learning, Mit Press, 1993

[2] 長尾真 編 : 自然言語処理, 岩波書店, 1996

[3] 井口時男, 往住彰文, 岩山真 : 文学を科学する, 朝倉書店, 1996

[4] 益岡隆志, 田窪行則 : 基礎日本語文法, くろしお出版, 1992

[5] 黒橋禎夫, 長尾真 : 日本語形態素解析システム JUMAN V3.5 マニュアル, 京都大学工学部, 1998

[6] Luhn, H. P. : The Automatic Creation of Literature Abstract," IBM JOURNAL, 1958

[7] 福本文代, 福本淳一, 鈴木料弥 : 文脈依存の度合いを考慮した重要パラグラフの抽出, 自然言語処理, 第 4 巻 2 号, 1997

[8] 鈴木康弘, 栃内香次 : キーワード密度方式の自動抄録法の改良, 情報処理学会論文誌, 第 29 巻 3 号, 1988

[9] 深谷昌弘, 田中茂範 : コトバの意味づけ論, 紀伊ノ国屋書店, 1996

A　作品 1. 全文の要旨

栄光を呼んだ挫折

額田さんは、高校受験を失敗したことにより心に傷を負うが、大学時代の教授の一言により心を動かされ、会社に入ってから仕事の中からテーマを見つけ、工学博士となった。そして、たまたま興味をもった「結び」について、休日や夜の時間だけでこの研究を続け、長い年月の経て文学博士になり、工学と文学という2つの異色の組合せの博士になった。しかし、メーカーという会社の性格を考えて、名刺には工学博士しか刷り込まなかった。定年後の名刺には、工学博士、文学博士と刷り込まれている。八十歳になっても、作詞という新しい分野に挑戦している。