単語間ネットワークを利用した関連語の抽出*

4R－9

窪田 健一† 山下 浩一‡ 吉田 敬一§

静岡大学大学院理工学研究科¶

1 はじめに

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。文章の要約には、文意に意味的に重要な影響を与える単語をいくつか抽出した上で、これらの単語を手がかりにして文章全体の要点を表す文のみを抜粋する方法や、新たに要約文を生成する方法がある。いずれの方法においても、これらの単語の抽出が最初の作業となる。いままでのこれらの単語を抽出する方法は、頻度を主体としたものが中心である。しかし、意味的に影響をおよぼす単語の中には、頻繁に現れないものもあり、また、意図的に同じ意味を持つ他の単語や表現に言い換えられる単語もある。これらの単語を切り捨てることは、作成される要約文が文意を十分に反映しないものとなってしまう可能性がある。

そこで、本論文では、「文章には、主題や文意を特定する単語と、その単語に付随してさらに詳細に文意を表す単語が存在する。」と仮定し、単語間ネットワークを利用して、頻度が低くとも、文意を表わしていると考えられる単語を抽出する手法を提案する。この手法により、より文意を反映した要約文の作成が可能になると思われる。

なお本論文では、対象を比較的話題がまとまっている短い文章とし、いくつかの話題が含まれる文章を対象から外す。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) あるまとまった意味を持ち、句点で区切られた言語単位。一つ以上の単語からなる。

文章 (composition) ある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。

単語間ネットワーク (word netowark) 複数個の単語をある関係に基づいて関連づけたもの。

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。基点語は含まない。

文意 (meaning) 文章が表現する意味。

3 提案する手法の概要

文章を形態素解析をした後、文章中に現れる単語を、名詞 (形式名詞、数量名詞を除く)、形容詞、動詞 (形式動詞を除く) に限定し、以下の手法により単語を抽出する。なお、手法の各段階については、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で、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、パス 1 にある単語を抽出する。
4. 各基点語に対して 3 で抽出した単語の集合に、共通する語を関連語として抽出する 。

以上により抽出さ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 単語間ネットワークの作成

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、適当なウィンドウサイズ内にある単

---

*A Method for Extracting Relational Word Using Word Network

†Kenichi Kubota

‡Kouichi Yamashita

§Keiichi Yoshida

¶Graduate School of Science and Engineering, Shizuoka University

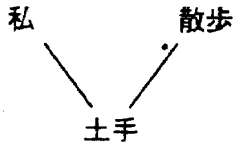

図 1: 単語間ネットワークの例

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単語の組により、意味的に表現されると仮定している。例えば、「私は、土手を散歩する。」という文からは、「私、土手、散歩」が抽出され、ウィンドウサイズ 1 で作成される単語間ネットワークは、(私, 散歩) が「土手」を意味的に表す事を示す。

5 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の中から指定する。

特に、このための手法を限定しないが、現在のところ、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] を仮定し、頻度により 2 ～ 4 語を機械的に指定している。2 ～ 4 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、わずかな単語が抽出できない場合は、名詞を優先に人が決めることにする。

6 関連語の抽出

作成された単語間ネットワークにおいて、すべての基点語 $w_i(i = 1, 2, \cdots, n)$ に対し、これらの基点語を中心としてできる単語の集合 $S_i(i = 1, 2, \cdots, n)$ の各々を考える。すなわち、

$\forall w_i(i = 1, 2, \cdots, n)$に対して、

$S_i = \{x | x$は基点語からパスの長さ 1 にある単語$\}$

のとき、

$$\{x | x \in \bigcap_{i=1}^{n} S_i, x \notin G\}$$

ただし、G は基点語の集合

を満たす単語 x を抽出する。抽出された単語は、各基点語 w_i を関連づける単語、もしくは、関係を表している単語として考えることができる。つまり、これらの単語は、各基点語に付随して、文意をより詳細に表している単語とみなせる。

7 単語抽出の問題点

現段階では、何人かの被験者にこの手法で抽出された単語群から元の話を推測してもらった結果から考えて、要約文作成に必要な単語がおおむね抽出されていると考えられる。

この手法で抽出される関連語数は、ネットワークを作成する際の距離と、関連語を抽出する際に指定する基点語数とにより変動する。すなわち、ネットワーク作成時に指定する距離を小さくする、または、指定する基点語数を多くすれば、抽出される関連語が少なくなる。逆に、ネットワーク作成時に指定する距離を大きくする、または、指定する基点語が少なくすれば、抽出される関連語は少なくなる。抽出される関連語数が多すぎても少なすぎても、文章を要約するには不適である。

8 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。現段階では、おおむね適切な単語が抽出されるていると思われるが、単語間ネットワーク作成の時に指定する距離と基点語の数により、関連語の数が変動するので、この 2 つを適切に定める方法を考究したい。このことにより、適合率と再現率の向上に期待できる。

参考文献

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

[2] 長尾, 佐藤, 黒橋, 角田 : "自然言語処理," 岩波書店, 1996

[3] 井口, 往住, 岩山 : "文学を科学する," 朝倉書店, 1996

[4] 益岡, 田窪 : "基礎日本語文法," くろしお出版, 1992

[5] "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, 福本, 鈴木 : "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, 1997