2V-1

自然言語テキストにおける注視関数を用いた注視語抽出方式の提案

Proposal of a Method of Extracting Focused Words that Uses a Focusing Function on Texts of Ntural Langrage

齋木貴博 Takahiro Saiki
鈴木 寿 Hisasi Suzuki

中央大学理工学部
Fac. of Science and Engineering, Chuo Univ.

1 はじめに

本研究の目的は自然言語テキストから意味論的に重要な情報を含む名詞の抽出である．本研究では, そのような名詞を注視語と呼ぶ.

本研究では新たに注視関数を定義し, 自然言語テキストから注視語を抽出する．本論文では複数のテキストコーパスから注視関数を求めたのち, 新たに自然言語テキストが与えられたとき, 注視関数を用いて注視語を抽出する方式を提案する.

2 用語

2.1 格文法

格文法 [1] の格には構文的な “表層格” と意味的な “深層格” がある．表層格は格助詞から判断し, 例文 “私が投げた。” 中の文節 “私が” の表層格は “ガ格” となる．深層格は名詞の役割から判断し, 同じく文節 “私が” では “動作主格” となる．本研究では 12 種類の深層格を定義する．深層格とその役割の例を表 1 に示す.

表 1　本研究で定義した深層格とその役割の例

深層格	役割
動作主格 (Agent)	動作を引き起こす者
場所格 (Location)	動作がおこなわれる場所

2.2 意味素

意味素とは名詞に対して与えれる意味の基本単位である．本研究では日本語語彙体系 [2] から意味素の判断をおこなう．また, 本研究で用いる意味素の例を表 2 に示す.

表 2　本研究で用いる意味素の例

意味素	例
人	私, 少年, 先生, 幽霊
施設	学校, 工場, 神社, 公園

2.3 注視に関する用語の定義

本論文において名詞の抽出に関し, 新たに用語を定義する．注視とは, 自然言語テキスト中の名詞を注視語の候補として注目することとする．また注視語とは, 自然言語テキストにおいて意味論的に重要な情報を含む名詞とする．例文 “家にいます。” から場所を示す単語を注視すると, “家” が注視語となる.

3 注視関数

3.1 定量化手法

注視関数を用いた注視度の定量化手法について述べる．注視関数とは, 名詞の表層格と意味素, およびその深層格の関係を統計的に定量化するものである．注視関数を非負係数 a, b を用いて

$$\text{注視度} = \text{a 表層注視度} + \text{b 意味注視度}, \quad (1)$$

$$\text{a} + \text{b} = 1 \quad (2)$$

と表わす．注視度の定量化は意味注視度と表層注視度の 2 つに分けておこなう.

3.1.1 表層注視関数による表層注視度の定量化手法

表層注視関数による表層注視度の定量化手法について述べる．表層注視関数は, ある表層格と任意の深層格が共に出現する頻度に基づいて設計する．手順は以下のとおりである.

Step 1
テキストを形態素解析し, 格助詞と名詞を抜き出す.

Step 2
抜き出した名詞に手動で深層格を割り当てる.

Step 3
Step 2 の結果から表層注視関数

$$\text{表層注視度} = \frac{\text{表層格 A と対応する深層格 B の個数}}{\text{表層格 A の個数}} \quad (3)$$

を用いて表層注視度を算出する.

なお, **Step 3** の表層格 A とは, 表層注視度を求める任意の表層格とし, 深層格 B はそれに対応する深層格とする．表層注視度は 0.0 から 1.0 の間の値をとり, 値が 1.0 に近づくほど深層格 B になりやすい.

3.1.2 意味注視関数による意味注視度の定量化手法

意味注視関数による意味注視度の定量化手法について述べる．意味注視関数は, ある意味素が任意の深層格が共に出現する頻度に基づいて設計する．手順は以下のとおりである.

Step 1
テキストを形態素解析し, 名詞と格助詞を抜き出す.

Step 2
抜き出した名詞に手動で意味素と深層格を与える.

Step 3
Step 2 の結果から意味注視関数の式

$$\text{意味注視度} = \frac{\text{意味素 A と対応する深層格 B の個数}}{\text{意味素 A の個数}} \quad (4)$$

を用いて意味注視度を算出する.

なお, **Step 3** の意味素 A とは, 意味注視度を求める任意の意味素とし, 深層格 B はそれに対応する深層格とする. 意味注視度は 0.0 から 1.0 の間の値をとり, 値が 1.0 に近づくほど深層格 B になりやすい.

3.2 複数のテキストコーパスにおける定量化実験

本実験は朝日新聞記事データベース CD-HIASK'94[3] に収録されている記事 6 件について定量化をおこなう. 実験結果の例を表 3, 表 4 に示す.

表 3 表層注視度の例

深層格	表層格	表層注視度
動作主格	ガ格	0.59783
場所格	デ格	0.35849

表 4 意味注視度の例

深層格	意味素	意味注視度
動作主格	人	0.47826
場所格	施設	0.11111

4 注視語の抽出手法

4.1 注視語の抽出アルゴリズム

注視関数を用いて自然言語テキストから注視語を特定し, 抽出するアルゴリズムについて述べる. 注視語の抽出は以下の手順でおこなう.

Step 1

各注視度に対する係数 a, b としきい値 c の値を決定する.

Step 2

各自然言語テキストに関し, 特定する注視語の深層格を選択する.

Step 3

抽出対象となる自然言語テキストに対して形態素解析をおこない, 格助詞とそれに付随する名詞を抜き出す. また, 抜き出した名詞に意味素を割り当てる.

Step 4

抜き出した名詞に対し意味注視度と表層注視度を与え, 注視度を算出する.

Step 5

注視度がしきい値 c 以上となった名詞を注視語として抽出する.

4.2 自然言語テキストにおける注視語の抽出実験

事前に深層格を指定した新聞記事の名詞 100 個に対し, 注視語の抽出実験をおこなう. 指定した深層格のとおりに抽出された名詞の個数を正解数, 指定以外の深層格で抽出された名詞の個数を重複数と判断する. 例文 "人に元気者がいるように、国にも元気国がある。" に対し, 係数 a=0.5, b=0.5, しきい値 c=0.1 としたときの抽出結果を表 5 に示す. 表中の "○" は各深層格を示す名詞として抽出された場合であり, "×" は抽出されなかった場合である. 表 5 の "人" は指定した対象格以外に動作主格でも抽出される. このように指定以外の深層格で抽出された名詞が 1 つ出現するたびに重複数をカウントする. 表 5 の場合, 正解数 4, 重複数 4 となる. また, しきい値 c=0.1, c=0.3 の場合の実験結果を図 1, 図 2 に示す. この実験結果からしきい値を大きくすると重複数が減少することがわかる. しかし, 同時に正解数も減少することになる. また, 表層格のみ, 意味素のみを考慮し名詞の抽出をおこなうよりも, 表層格と意味素の両方に重みをかけて抽出をおこなう方が正解数が大きくなる.

表 5 実験結果の例

名詞	指定した深層格	対象格	動作主格	場所格
人	対象格	○	○	×
元気者	動作主格	○	○	×
国	場所格	○	○	○
元気国	動作主格	○	○	○

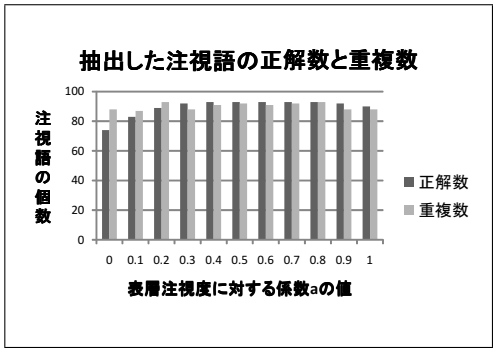

図 1 しきい値 c=0.1 の場合

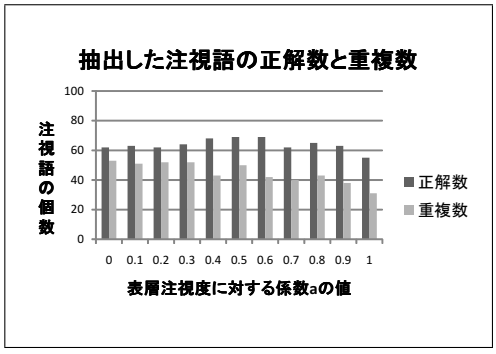

図 2 しきい値 c=0.3 の場合

5 おわりに

本稿では注視関数を用いた注視語の抽出方式について述べ, その方式に基づく実験の結果を示した. 今後の課題は定量化手法, 抽出手法を検討し, この方式の精度をより高めていくことである.

参考文献

[1] 長尾 真, **岩波講座ソフトウェア科学 15 自然言語処理**, 岩波書店, 2005.

[2] 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦, **日本語語彙大系 CD-ROM 版**, 岩波書店, 1999.

[3] 朝日新聞社東京本社版, **CD-HIASK'94**, 紀伊国屋書店・日外アソシエーツ, 1996.