1P-5

自然言語テキストから効率よく注視語を抽出するための注視関数の提案

Proposal of a Focusing Function for Efficiently Extracting the Focused Words on Texts of Natural Language

齋木貴博 Takahiro Saiki
鈴木 寿 Hisasi Suzuki

中央大学大学院理工学研究科
Department of Information and System Engineering
Chuo University

概要: 本研究の目的は, 所望の意味役割を示す名詞である注視語を効率よく抽出するための注視関数を開発することである. 注視関数を表層注視関数と意味注視関数の線形和とする. 従来手法ではすべての意味役割に対し一定であった表層注視関数に対する荷重を学習用テキストにより学習する手法を提案する.

注視語の抽出実験の結果, 従来の注視関数と比較し提案した注視関数にて良好な結果が得られた.

キーワード: 注視語, 注視関数.

1 格文法, 意味素, 注視について

1.1 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. 本研究では意味役割を 12 種類の深層格で定義する. 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

1.2 意味素

意味素とは名詞に対して与えら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] によって意味素を判断する. 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	分類される名詞の例
人	私, 少年, 先生, 幽霊
施設	学校, 工場, 神社, 公園

1.3 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。” から深層格が場所格となる名詞を注視すると, “家” が注視語となる.

2 注視関数

注視関数は, 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し, その定量結果を注視度とする. 本研究では, 注視関数を表層注視関数と意味注視関数の線形和とし, 表層注視関数の定量結果を表層注視度, 意味注視関数の定量結果を意味注視度とする. また, 従来手法 [4] では各注視度に対する荷重が抽出する注視語の深層格にかかわらず一定であったが, 本研究では注視語の深層格ごとに異なる荷重 a_d を用いる.

注視関数を表層注視度に対する荷重 a_d, および意味注視度に対する荷重 $(1-a_d)$ を用いて

$$\text{注視度} = a_d \text{表層注視度} + (1-a_d)\ \text{意味注視度} \quad (1)$$

と表わす. 注視度の定量化は表層注視度, 意味注視度, 各深層格の表層注視度に対する荷重の 3 つに分けておこなう.

2.1 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, 格文法における表層格と深層格との関係度合である表層注視度を統計的関係に基づき定量化する.

ここで, 表層格集合を C, 深層格集合を D とし, テキスト t 上で, 表層格が $c \in C$ である名詞の個数を $N(c|t)$ と表す. また, 表層格が $c \in C$ かつ深層格が $d \in D$ である名詞の個数を $N(c,d|t)$ と表す. このとき, 表層注視度 $P(d|t,c)$ を以下の表層注視関数の式により定量化する.

$$P(d|t,c) = \frac{N(c,d|t)}{N(c|t)} \quad (2)$$

2.2 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, 格文法における深層格と任意の意味素との関係度合である意味注視度を統計的関係に基づき定量化する.

意味素集合を S, 深層格集合を D とする. テキスト t 上において, 意味素が $s \in S$ である名詞の個数を $N(s|t)$ と表す. また, 意味素が $s \in S$ かつ深層格が $d \in D$ である名詞の個数を $N(s,d|t)$ と表す. このとき, 意味注視度 $P(d|t,s)$ を以下の意味注視関数式により定量化する.

$$P(d|t,s) = \frac{N(s,d|t)}{N(s|t)} \quad (3)$$

2.3 各深層格の表層注視度に対する荷重の定量化手法

各深層格の表層注視度に対する荷重 a_d の定量化手法について述べる. テキスト t 上において, 深層格が $d \in D$ である名詞の個数を $N(d|t)$ と表す.

任意の深層格 d における表層注視度の期待値を x_d, 意味注視度の期待値を y_d とすると,

$$x_d = \sum_c (P(d|t,c) \times \frac{N(c,d|t)}{N(d|t)}) , \quad (4)$$

$$y_d = \sum_s (P(d|t,s) \times \frac{N(s,d|t)}{N(d|t)}) \quad (5)$$

と求められる. これより, a_d を以下のように定量化する.

$$a_d = \frac{x_d}{x_d + y_d} . \quad (6)$$

2.4 注視関数による注視度の定量化手法

3.1 節, 3.2 節, 3.3 節の結果を用いた注視関数による注視度の定量化手法について述べる. 仮想テキスト u 上において, 任意の名詞 n の深層格が $d \in D$ となる注視度 $Q(d|u)$ は

$$Q(d|u) = a_d P(d|t,c) + (1 - a_d) P(d|t,s) \quad (7)$$

となる. この定量結果を用いて文に含まれる名詞から指定する深層格を示す名詞を注視語として抽出する.

3 注視語の抽出手法

3.1 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.

ステップ 1
抽出対象の自然言語文に関し, 特定する注視語の深層格を選択する.

ステップ 2
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, 抜き出した各名詞に意味素を割り当てる.

ステップ 3
抜き出した各名詞に対し表層注視度, 意味注視度および指定した深層格に対する荷重 a_d を与え, 注視度を算出する.

ステップ 4
注視度が最大となった名詞を対象文における指定した深層格を示す注視語として抽出する.

3.2 自然言語テキストにおける注視語の抽出実験

注視語の抽出実験に際し, 本研究では朝日新聞記事データベース CD-HIASK'94 に収録されている記事に含まれる格助詞が付随する名詞 1036 個に意味素および深層格を割り当て, 表層注視度, 意味注視度および各深層格に対する荷重 a_d を学習する.

次に注視語の抽出実験について述べる. 事前に CD-HIASK'94 に収録されている新たな記事内の格助詞が付随する名詞 100 個に意味素および深層格を割り当てる. 選択した名詞 100 個に対し, 前述の学習データを用い, 4.1 節で述べたアルゴリズムに沿って抽出実験をおこなう. 実験結果の例を表 3 に示す. 表中に見られる数は学習データより得られた各深層格に対する注視度である. 例えば, 動作主格の欄を見ると, "首相" に対する注視度が 0.596 と最大になっていることから "首相" が抽出される. さらに "首相" に割り当てた正解の深層格と一致したことにより, 動作主格を示す注視語は正しく抽出された.

最後に, 抽出実験における従来手法と新手法の比較を図 1 に示す. 従来手法では荷重 a_d に抽出対象の深層格に関係なく 0.0 から 1.0 までの値を 0.1 ずつ代入していき, 抽出実験をおこなった. 図 1 を見ると, 従来手法での最大値が $a_d = 0.5$ のときの 88 に対し, 新手法では正しく抽出された注視語の数が 90 となり, 良好な結果が得られた.

表 3 実験結果の例

名詞	正解の深層格	動作主格	時間格	対象格
首相	動作主格	0.596	0.005	0.263
午後	時間格	0.021	0.515	0.042
会見	対象格	0.030	0.001	0.732

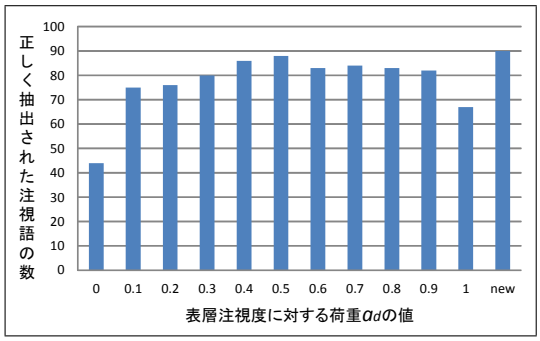

図 1 注視語の抽出実験結果

4 結論

本稿では従来手法を拡張した注視関数を提案し, その方式に基づく従来手法との比較実験をおこなった. 実験の結果として従来手法と比較しより効率的に注視語を抽出できるという良好な結果が得られた.

今後の課題は意味素の深度を変更することによる本方式の精度の変化を確認し, より効率よく注視語を抽出する注視関数を開発することである.

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.

[4] 齋木 貴博, "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," **中央大学情報工学科卒業研究論文**, 2012.