7H-06

文章に対する重要度付与処理における視点情報の有効性の分析*

永松健司　田中英彦†
東京大学 工学部

1　はじめに

文章読解において読み手が文章のどこに重要性を見出すかは、彼 (彼女) の興味・目的に大きく影響される。従って、読み手に応じた要約処理等を考える場合には、それらを考慮しないことには十分な結果を得ることはできないはずである。我々の研究では、読み手が解釈に先だって踏まえているそのような立脚点を大きく “視点” と呼ぶこととし、テキストの理解に対して重要な前提として機能するものと見做している。

本稿では、テキストに対して重要度を付与する処理において、読み手が保持する視点情報がどの程度、その結果に寄与するかを評価実験を通して明かにする。

2　視点情報の適用

2.1　視点情報に求められる効果

本研究での “視点” は、読み手が文章を解釈する際に、その前提として保持している立脚点を指すものである。本稿では、視点に対し要請される効果として

- 概念間の距離の算出において、視点情報を前提とした経路が採用されること
  ⇒ 視点に依る概念距離関係の変化

だけを考えることとし、ここからどのような結果が得られるかに注目する。

概念間の距離は、概念の関連性や解析処理における選好情報等に対し重要な役割を果たすため、解釈の前提として保持する立脚点という意味での視点にとって適切な指標となると判断した。

2.2　概念間の距離

概念間の距離の定義は次の式に依った (図 1参照)。この式で使われている概念体系は EDR 電子化辞書での概念辞書を利用している。

$$\mathrm{Dis}(c_1,c_2)=\frac{\text{概念 } c_1,c_2\text{を結ぶ kind-of リンクの数}}{c_1,c_2\text{の共通な上位概念 } c_0\text{の概念体系内の深さ}} \quad (1)$$

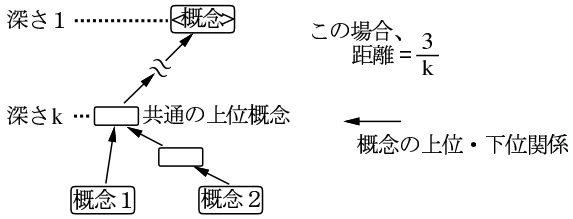

図 1: 概念間の距離の定義

また、視点情報を前提とした場合の距離の式は次のように定義する。

$$\mathrm{Dis}_{\mathrm{pov}}(c_1,c_2)=\mathrm{Dis}(c_1,c_{\mathrm{pov}})+\mathrm{Dis}(c_{\mathrm{pov}},c_2) \quad (2)$$

ここで、c_{pov}は現在の視点を指す概念とする。

3　重要度の付与処理

3.1　重要度の算出プロセス

今回の評価実験において採用した重要度付与処理の手法について述べる。今回の実験は視点情報の有効性を見るという目的のため、重要度付与処理は実装の容易さを考えて次の手順からなる方法を採った。

1. テキスト全体での中心概念の決定
2. 各々の文を構成する概念 (単語) 群と中心概念との距離を算出
3. 同じ文内にある概念に対する中心概念との距離の最小値をその文と中心概念との距離と見做す

そして、平均距離の近い文ほど、テキスト内での重要度が高いものとする (今回の実験では、重要度を平均距離の逆数で定義した)。

3.2　テキストの中心概念の決定

テキストの中心概念は、概念間の距離の定義式 (1) に従い、テキストを構成するすべての概念 (単語) 間の距離の分布を調べ、もっとも分布の密接している場所にある概念をテキストを代表する中心概念c_cと定義する。すなわち、

$$c_c=\max_{c_i\in text}\mathrm{Imp}(c_i) \quad (3)$$

ここで、

$$\mathrm{Imp}(c)=\sum_{c_i\in text}(\mathrm{Dis}_{(\mathrm{pov})}(c,c_i)^{-1}\cdot\mathrm{Frq}(c_i)) \quad (4)$$

$\mathrm{Frq}(c_i)$は概念c_iのテキスト内での出現頻度を示す。

---

*Analysis of point-of-view informations' effectiveness in the task of giving importance values for sentences

†Kenji Nagamatsu　Hidehiko Tanaka
{naga,tanaka}@mtl.t.u-tokyo.ac.jp
Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113, Japan

概念 (ID)	重要度	概念 (ID)	重要度
年 (1e85d1)	170.6	元 (3cebb9)	98.2
数 (0e423c)	160.6	気 (0ebab7)	90.1
メートル (3c08d5)	159.5	遊び心 (0e3145)	86.1
毎年 (102cd6)	118.6	苦 (3d0207)	85.1
縁起 (3ce5aa)	117.3	···	···

表 1: 各概念のテキスト全体に対する重要度 (視点なし)

文番号	1	2	3	4	5	6
重要度	0.86	2.67	4.50*	1.40	0.67	4.50*
文番号	7	8	9	10	11	12
重要度	0.14	2.67	0.15	0.17	4.50*	2.67
文番号	13	14	15	16	17	18
重要度	0.23	0.86	0.75	2.67	2.67	2.67
文番号	19	20	21	22	23	24
重要度	0.13	0.75	0.13	2.67	0.14	0.13

表 2: 各文に対する重要度 (視点なし)

4 評価実験

4.1 実験方法について

節 2の方法に基づき、文章への重要度付与の処理において視点情報を前提とした場合の評価実験を試みた。

入力テキストには朝日新聞天声人語 [1] を使用し、視点情報を与えない場合と与えた場合に、各文に対して付与される重要度を求めた。ただし、節 3の手法に対し、簡単化のため、入力とする概念を名詞相当語句に対するもののみに限定した。

このテキストの内容は、毎年、その年の西暦の数字と同じ高さの山に登る人達がいることを紹介し、今年 (1995年) に当たる山がかろうじて見つかったことを述べている。そして、最後に数字が行動を縛ってしまう迷信に言及している。また、このテキストは、24 文、76 概念 (のべ 169 概念) (ただし、名詞相当語句に限定) から構成されている。

4.2 視点情報を与えない場合

各概念に対する重要度 Imp を求めると、表 1に挙げたようになり、このテキストの中心概念は「年 (1e85d1)」となる。

次に、この中心概念を前提として、テキストを構成する各文の重要度を算出したものを、表 2に示す。

ここで最も重要度が高いのが

- 文 3「郷里の最高峰だということもあるが、高さが西暦の年と同じ数字の一、九八二メートルだということに遊び心を誘われたそうだ。」
- 文 6「二十万分の一の地図で見るかぎり、山頂が一、九九五メートルという山が無い、というのだ。」
- 文 11「標高は一、九九四・五二メートル、四捨五入すれば一、九九五メートルの山となる。」

の 3 文である。

概念 (ID)	重要度	概念 (ID)	重要度
縁起 (3ce5aa)	56.9	メートル (3c08d5)	43.1
毎年 (102cd6)	47.4	行動 (3cf6aa)	39.2
元 (3cebb9)	44.6	気 (0ebab7)	39.2
年 (1e85d1)	44.6	遊び心 (0e3145)	38.8
数 (0e423c)	44.6	···	···

表 3: 各概念のテキスト全体に対する重要度 (視点あり)

文番号	1	2	3	4	5	6
重要度	0.55	0.74	0.55	0.74	0.43	0.67
文番号	7	8	9	10	11	12
重要度	0.13	0.74	0.13	0.14	0.67	0.74
文番号	13	14	15	16	17	18
重要度	0.19	0.55	0.50	1.00*	0.74	0.74
文番号	19	20	21	22	23	24
重要度	0.12	0.50	0.11	0.74	0.13	0.11

表 4: 各文に対する重要度 (視点あり)

4.3 視点情報を与えた場合

次に視点情報として、概念「前兆 (3cfbf9)」(テキスト中には陽に現れない) を与えた場合での中心概念、および各文の重要度を表 3、表 4に示す。

視点情報を与えた場合、テキスト内で最も重要度の高くなる文章は文番号 16 「日本では九や四を縁起が悪いと言って嫌う人が多い。」になる。

5 考察 – 視点情報の有効性

視点情報を与えない場合、テキスト内での概念の出現頻度とシソーラスのみに依存して重要度の高い文が決定されるが、文 3,6,11 は、このテキストの要旨とも関連性は大きいものと判断できる。

一方、視点情報を与えた場合、視点情報との距離が大きな影響をもってくる。表 3での上位にきた概念や、重要度が最も高くなった文 16 を見ると、概念間の距離のみに依った視点情報でも関連性の判断に対して適切な指標となりうることが推察される。

6 おわりに

概念間の距離のみに対する視点情報の適用において、十分適切な結果が得られることが分かった。しかし、シソーラスの上下関係だけを拠り所とする距離は概念一般に対して利用可能という利点がある一方で、より意味の深い関連性が要求される場合に必ずしも妥当な結果が得られないことも分かった。

今後は、概念間での意味の共起関係を見るなどの改良とともに、サンプル数を増やして人間の判断との相違についても調べていきたい。

本研究では、日本電子化辞書研究所から EDR 電子化辞書評価版 2.1 版を利用させていただいた。

参考文献

[1] 朝日新聞社. 天声人語. 1994 年 6 月 10 日.
[2] 宮崎清孝, 上野直樹. 視点. 東京大学出版会, 1985.