4Q－2 自動ターム抽出における重み付け方法の比較

斉藤 貴也　中川 裕志　森 辰則
横浜国立大学 工学部

1 はじめに

自動ターム抽出は，膨大なテキストがネットワークを介して入手可能になった現在，テキストデータの有効利用のために必須の技術である．Kageura96[1]によれば，ターム抽出の基準になるのは，ある表現(Collocationなど)がテキストデータベース中で安定して使用される度合を表すUnithood基準と，ある表現が対象分野固有の概念をどれだけ強く表現するかを表すTermhood基準が重要である．この報告で比較検討する二つの自動ターム抽出法は，主としてこの両基準に対応する方法である．ひとつはNested Collocation方式[2]である．この方法は，後に述べるように基本的にはUnithood 基準による方法と言える．もうひとつは我々が提案した方法[3]である．詳細は次節で述べるが，これはTermhoodを反映した方法である．このような方法で順位付けされた候補が得られると，次にこれらから望ましいタームを選択するプロセスが必要になる．我々が，既に提案した選択方式[3]を両方法で得られた候補語リストに適用した結果について最後に報告する．

2 タームの頻度ランキング

本節では候補語のランキングの方法について述べる．

2.1 連接情報を考慮した重要度計算法(連接方式)

この計算法は，単名詞Nがたくさんの複合語を構成するほど重要であるという考えに基づいている．前に連接する単名詞の種類数を前方連接数$Pre(N)$，後ろに連接する単名詞の種類数を後方連接数$Post(N)$とする．ここで重要なのは，連接の頻度ではなく種類数を用いている点である．つまり，ある単名詞からどれだけ多くのの複合語を生成できるかを測っているのであり，これはその単名詞が，対象のテキストデータベースにおける中心的概念である度合を示している．よってこの方法は Termhood を基本にする方法であると言える．次に，複合名詞$N_1N_2N_3\cdots N_K$の重要度の尺度$Imp_{1or2}(N_1N_2N_3\cdots N_K)$は$Pre$，$Post$の積，相乗平均を使った次式で表される．

$$Imp_1(N_1N_2\cdots N_k) = \prod_{i=1}^{k}((Pre(N_i)+1)\cdot(Post(N_i)+1))$$

$$Imp_2(N_1N_2\cdots N_k) = (\prod_{i=1}^{k}((Pre(N_i)+1)\cdot(Post(N_i)+1)))^{\frac{1}{2k}}$$

2.2 Nested Collocationによる重要度計算法

この計算法は，collocationの入れ子構造に着目しcollocationを構成する単名詞数，collocationの出現頻度，入れ子になっているcollocationの種類数を基に順位付けを行なっている．重要度の尺度として以下に示すC-valueを用いる．

$$\text{C-value(a)} = (length(a)-1)(freq(a)-\frac{t(a)}{c(a)})$$

ただし，aはタームの候補のcollocationである．$length(a)$はaを構成する単名詞数，$t(a)$はaを含む候補collocationの頻度，$c(a)$はaを含む候補collocationの種類数である．よって，aが多数の文脈で安定して多数使用される場合には C-valueは大きくなる．しかし，aの使用頻度が高くても，一定の文脈でしか使用されないなら，aはより大きな安定した collocation の一部であるとみなされ，C-valueは，小さくなる．よって，C-value は，その collocation がテキストデータベース中で安定して使われる度合を示すと考えられるから，この方法は Termhood 基準による方法でみなせる．ただし，本実験では，単名詞を考慮するので，$length(a)-1$の所は$length(a)$として計算を行なう．また，n-grams ($length = n$) のC-valueを単名詞相当に正規化するため，C-value×nとする．

3 窓方式によるターム選択

前節の計算法より得られる候補語から望ましいタームを選択する方法として窓方式を提案している．窓方式とは，重要度順にソートされた候補語リスト上を一定幅の窓を移動させ，その時の窓内の複合名詞の割合(複合

Comparative Study of Ranking Methods in Automatic Term Extraction
takaya Saitoh, Tatsunori Mori and Hiroshi Nakagawa
Division of Electrical and Computer Engineering,
Faculty of Engineering, Yokohama National University

	Imp_2 の値	候補語
	19.90	辞書
⇓	17.18	形態素辞書
	14.83	形態素
	13.52	形態素辞書ファイル
⇓	13.25	形態素の連接
	12.90	辞書ファイル
	12.20	活用辞書
⇓	12.18	形態素コスト

図 1: 候補語リスト Imp_2 の例，及び窓の移動：窓の幅 5

語率）によって窓の中央の候補語を望ましいタームであるのか選択していく方法である．例として図1のような窓幅 5 の場合は，複合語率を 0.3 とすると，窓内の複合語率が 0.8 なので窓の中央の候補語（形態素の連接）をタームとして選択する．複合語率を用いた理由として正解語も複合名詞も候補語のランキングの上位に集まっている点，複合語率がマニュアルの長さに依存しないという実験結果が挙げられる．ここで述べた正解語とはマニュアルから前もって人手で望ましいタームであると判断されたものである．

4 実験と評価

本実験では，5 本の日本語マニュアルとして，JUMAN・SAX・たまご（ソフトウェア），HV-F93（三菱電機のビデオ），PS（SONY のゲーム機）を使用した．また，英文に対する実験も行なっている．文章には Cognitive Science Conference (http://www.cse.ucsd.edu/groups/geuru/cogsci96/accepted.html) に採録された言語学分野の論文のアブストラクトを用いた．ただし，正解語は用意していない．
評価には適合率，再現率を用いた．実験手順は，

1. マニュアルを JUMAN で形態素解析する。
2. 上で出た結果からタームの候補語を抽出し 2.1 と 2.2 で述べた 2 つの方法で順位付けを行なう．
3. 窓方式により候補語からタームを選択する．

選択時の窓幅は 5,10,20,30 とし，複合語率は 0.1 ～ 0.9 まで 0.1 刻みとした．最も良い結果を図 2に示す．また JUMAN のマニュアルについて窓幅 20，複合語率 0.3 の時の抽出正解語の例を図 3に示す．英語については抽出語を図 4に示す．この結果から連接方式も Nested Collocation 方式も良い結果が出たが，二つに大きな違いはでなかった．しかし，連接方式は termhood を反映している分，単名詞を選択しやすく，Nested Collocation 方式は unithood 基準の方法である分，複合名詞を選択しやすい傾向が見られた．

連接方式

窓幅	複合語率	適合率	再現率
20	0.3	0.362	0.458

Nested Collocation 方式

窓幅	複合語率	適合率	再現率
30	0.3	0.420	0.424

図 2: 両方式の最良値

連接方式のみが抽出した正解語
エントリ / オプション定義 / グラフ / ハッシュテーブル / 活用形 / 基本形 / 語尾 / 表層 / 変換 /
Nested Collocation 方式のみが抽出した正解語
オプション定義ファイル / 形態素文法 / 後接情報 / 構造 / 束状 / 田窪文法 / 連接可能性 /
両方式で抽出した正解語
C 版 / JUMAN システム / Prolog 版 / グラフ構造 / コスト / コスト計算 / コスト幅 / システム辞書 / システム標準辞書 / システム標準文法 / ユーザ辞書 / 意味辞書 /

図 3: JUMAN についての抽出正解語の例

連接方式のみが抽出した語
word/semantic/detection/phonological/network/reading/gender/anaphoric/sentence:processing/
Nested Collocation 方式のみが抽出した語
constant:cue/outcome:base:rate/past:tense:mapping/model/variable:cue/frequency:targets/lexical:concepts/
両方式で抽出した語
model/processing/data/dual-lexicon:model/semantic:memory/results/Lexical/detection:times/interactive-activation:model/connectionist:network/

図 4: 英語における抽出語の例

参考文献

[1] Kageura,K.and B.Umino,1996,Methods of automatic term recognition:A review, Terminology 3 (2) 259-289

[2] Frantzi,K.T.,S.Ananiadou and J.Tsujii,1996, Extracting Nested Collocations,COLING'96,41-46

[3] Nakagawa,H.,1997,Extraction of Index Words from Manuals,RIAO'97,598-611