自動ターム抽出における候補順位付けと選択方式の評価

中川 裕志　　森 辰則　　斉藤 貴也
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．すなわち，a.テキストに自動的にインデクスを付与する，b.ある分野のテキスト群から自動ターム抽出を行なって，その分野固有の用語抽出を行なう，など重要な応用分野を持つ．この論文では，まず自動ターム抽出の基本になる termhood と unithood の考え方を紹介し，次にこれらの考え方に基づくふたつのターム抽出法を比較する．ターム抽出とは，タームの候補をなんらかの重要度に従ってランキングすることに加え，ランク付けされた候補リストから適性なタームを選択する方法が必要になる．これについては，我々は窓方式 [3] を提案しており，その性能の評価も併せて記す．

2 termhood と unithood

[1] によれば，ターム抽出の基準になるのは，(1) ある表現 (Collocation や複合語など) 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，(2) 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** 基準の二つである．この報告で比較検討する二つの自動ターム抽出法は，各々，基本的に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．これらは，いずれもタームの候補をランキングする重要度の定義法に関する提案であり，次節で詳述する．

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

3 タームの頻度ランキング法

3.1 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，あるテキスト T においてある単名詞 N がたくさんの複合語を構成するほど，N は T において (より正確には T の表す分野，領域において) 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．例えば，あるテキストで，「システムファイル」「辞書ファイル」「補助ファイル」という後に「ファイル」を接続する複合語が3つあれば，$Pre(\text{ファイル}) = 3$ になる．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，対象のテキストが記述する分野において，その単名詞が基本語彙であり，それ使って多くの概念を生成するような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 $Imp(N_1N_2N_3\cdots N_K)$ は Pre, $Post$ の関数として定義する．これには無数の定義法があるが，ここでは積，相乗平均を使った次式で表される定義 Imp_1 と Imp_2 を比較してみた．

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

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

この定義によれば，termhood を近似的にせよ測定する一方法である Pre と $Post$ を組み合わせているので，複合語の Imp 関数もまた termhood に基づく尺度であるといえる．

3.2 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．

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

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．

4 ターム選択

以上で述べた方法でタームの候補を順位付けると，次にそのように順位付けられたターム候補のリストから望ましいタームを選択する処理を行うことになる．望ましいタームを選択する方法として我々は，1) 閾値によるふるい分け，2) 我々が[3]に提案した窓方式について検討する．

4.1 閾値方式

閾値方式は，順位付けられた候補語リストにおいて適当な閾値を定め，その閾値より大きいImp関数あるいはC-valueを持つ候補語をタームとして選択する方式である．ここで問題になるのは，個別のテキストにできるだけ依存しないような閾値の決定法である．ここでは，対象とするテキストのC-value ないしImp関数の値の分布の平均と標準偏差を使う．具体的には，閾値として

平均値 $+\alpha\times$ 標準偏差

を用いる．ただし，αの値は，各テキスト毎にRijesbergenのEを最適にする値を求め，それを全テキストで平均した値を用いることにした．これは，テキスト毎のC-valueやImp関数の値の分布の差異に左右されにくい閾値を選ぶという方針で選んだものである．

4.2 窓方式

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（複合語率）によって窓の中央の候補語を望ましいタームとして選択するかどうか判断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，複合語率の閾値を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．

	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

複合語率を用いた理由として人間が抽出したいターム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．ここで述べた抽出したいタームは，与えられたテキスト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．

窓方式にも窓幅と複合語率の閾値という二つのパラメターがある．ここでの実験では，各テキスト毎にRijesbergenのEを最適にする値を求め，そのうち最も多くのテキストで高い値を得られるパラメター値を用いることにした．ただし，窓幅は5,10,20,30とし，複合語率の閾値は0.1 ～ 0.9まで0.1刻みで実験した．

5 実験と評価

本実験では，まず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（三菱電機のビデオデッキ），PlayStation（SONYのゲーム機）を使用した．日本語のテキストのターム抽出の評価には適合率，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3.1と3.2で述べた2つの方法で順位付けを行なう．
3. 閾値方式および窓方式により候補語からタームを選択する．

まず，選択方法として4.1で述べた閾値方式の場合について連接方式でImp_1，Imp_2を用いた場合，およびNested Collocationの場合の各々について，先に述べた方法ででた最も良い結果を以下に示す．

連接方式 Imp_1

α	適合率	再現率	最適E値
0.16	0.407	0.423	0.597

連接方式 Imp_2

α	適合率	再現率	最適E値
-1.67	0.403	0.435	0.590

Nested Collocation 方式

α	適合率	再現率	最適 E 値
0.44	0.439	0.393	0.602

次に，選択方法として 4.2 で述べた窓方式の場合について連接方式で Imp_1，Imp_2 を用いた場合，および Nested Collocation の場合の各々について，先に述べた方法ででた最も良い結果を次に示す．

連接方式 Imp_1

窓幅	複合語率	適合率	再現率	最適 E 値
30	0.1	0.360	0.395	0.604

連接方式 Imp_2

窓幅	複合語率	適合率	再現率	最適 E 値
20	0.3	0.362	0.458	0.606

Nested Collocation 方式

窓幅	複合語率	適合率	再現率	最適 E 値
30	0.3	0.419	0.396	0.599

この結果を見ると，候補の順位付けには連接方式で Imp_2 関数を用い，ターム選択には閾値方式を用いる組合せが一番良い結果を与える．2 番目によいのは，候補の順位付けには Nested Collocation 方式で，ターム選択には窓方式を用いる組合せである．また，この実験で用いたテキストでは，Nested Collocation 方式は窓方式と相性がよく，連接方式は閾値方式と相性がよい．もっとも，各組合せにおいて，E 値の差は微小であり，テキストによっては違った結果も予想される．大規模なテキストで実験するのが望ましいのだが，テキストが大規模になると，人間から見て望ましいタームを抽出する作業が困難であり，実験が難しくなるという問題点がある．逆に言えば，ここで調査した方法は，順位付け方式，選択方式のいずれも，そこそこの結果が期待できるわけで，実際の応用にあたって利用者が適切と思うタームを抽出できる方法を同定し，使用することが実際的である．また，2. で述べた termhood と unithood の差については，極端に異なる性能を導くわけではないことも推測される．

ちなみに，JUMAN のマニュアルについて窓幅 20，複合語率 0.3 の時の抽出タームの例を次に示す．

連接方式のみが抽出したタームの例

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タームの例

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 /

両方式で抽出したタームの例

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 /

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．

また，英文に対する実験も行なった．英文のテキストには 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，英文テキストの場合は抽出したいタームは用意できなかった．英語についての実験結果の抽出タームを以下に示す．なお，Stoplist のみを用いて，二つの Stoplit の 語の中間を複合語とみなした．選択方式としては窓方式を用いた．

連接方式のみが抽出したタームの例

word/semantic/detection/phonological/network/
reading/gender/anaphoric/sentence:processing/

Nested Collocation 方式のみが抽出したタームの例

constant:cue/outcome:base:rate/past:tense:mapping/
model/variable:cue/frequency:targets/lexical:concepts/

両方式で抽出したタームの例

model/processing/data/dual-lexicon:model/
semantic:memory/results/Lexical/detection:times/
interactive-activation:model/connectionist:network/

英語においても，Nested Collocation 方式が複合名詞を，連接方式が単名詞を抽出しやすい傾向は見られる．

6 おわりに

テキストからの自動ターム抽出の方法として termhood に基づく連接方式と unithood に基づく Nested Collocation 方式の比較実験を行なった結果，両方式とも数値的には隔たりがなかったものの，抽出されたタームの性質には差が見られることが分かった．英語についての実験結果も記載したが，これについては POS Tagger を用いた比較実験も行なっており，機会を改めて発表したい．

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