ISS-SP-056

係り受け構造に着目した専門用語自動抽出手法の提案

木村優介 † 楠和馬 ‡ 馬場睦也 ‡ リュウショウウ ‡ 波多野賢治 †
† 同志社大学文化情報学部 ‡ 同志社大学大学院文化情報学研究科

1 はじめに

ユーザの知識量に応じた検索には，専門用語に基づいたテキストの難易度推定が必要になる [1]．専門用語をテキストから自動抽出する既存手法として，大半の専門用語は複合名詞であることから，その構成要素である単名詞の連接頻度に着目した FLR が提案されている [2]．FLR で用いられている統計量は経験則に基づいて選択されているが，テキストから抽出可能なあらゆる統計量を調査した上で，FLR に適用したわけではない．そのため，専門用語抽出に有効な統計量を取り逃している可能性が残されている．

そこで本研究では，専門用語の持つ特徴であるにも関わらず，今まで扱われていなかった特徴から考えられ得る係り受け構造から得られる統計量に着目する．専門用語に関する統計量を用いた一般化線形モデルを作成し，そのモデルによって専門用語らしさを算出することで専門用語を抽出する手法を提案する．

2 提案手法

本研究では専門用語に関する統計量を用いた一般化線形モデルを作成し，そのモデルで算出される値を専門用語らしさとして捉えることで，専門用語を抽出する手法を提案する．一般化線形モデルとして，二値分類でありながらその値域 $[0, 1]$ の確率を算出できるロジスティック回帰モデルを用いる．その回帰モデルを用いて，専門用語の候補語が専門用語である確率を算出し，確率が高いほど専門用語らしい語として判断する．

説明変数として，先行研究で扱われた統計量である出現頻度，単名詞の連接頻度と先行研究の専門用語に関する記述から単名詞か否か，ある専門用語の候補語を含む文節がどれだけ他の文節に影響を与えているかを係り受けの数から求める．その数を説明変数として採用した理由は，専門用語には専門用語を説明する文があり，その中には関連する他の専門用語がある [3]，という専門用語の特徴から文中のある文節がそれより後方の文節を修飾する特徴 [4] をもつ係り受け構造を用いることで，ある専門用語の候補語を含む文節を説明している文節を明らかにできると考えたからである．

図 1 は，ある文の係り受け構造を示しており，各アルファベットは文節を示している．ある専門用語の候補語を含む文節 b を中心にみると，文節 a から説明されていることを示す白矢印を 1 本受けている．また，文節 b は文節 c を説明していることを示す黒矢印を 1 本出している．さらに，文節 b によって説明がなされた文節 c は文節 d を説明する黒矢印を 1 本出している．そのため，文節 b 内のある専門用語の候補語の統計量は他の文節から説明を受ける数が一つ，他の文節を説明する数が二つとカウントする．図 1 の文に，文節 b 内のある専門用語の

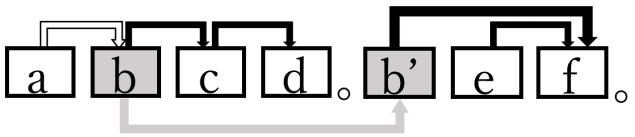

図 1　文をまたいだ係り受け構造

候補語と灰矢印で照応関係にある語句が含まれる文節 b' があった場合，前述の文節 b と同様の方法で計算を行う．ただし，係り受け構造と照応関係は異なるため，専門用語の候補語 b にその照応関係にあたる b' のカウントを加算しない．

3 評価実験

専門用語をどれだけ正確に抽出できるかを評価するために，提案手法と先行研究の FLR で専門用語の候補語を確率，スコア順にそれぞれ並び替え，式 (1) の適合率 P，再現率 R で評価を行う．

$$P = \frac{TP}{TP + FP} \quad , \quad R = \frac{TP}{TP + FN} \tag{1}$$

ただし，TP は正しく専門用語と判断した件数，FP は専門用語でない名詞を誤って専門用語であると判断した件数，FN は専門用語を専門用語ではないと誤って判断した件数である．評価用データとして，人工知能分野の論文抄録に対し，どの語が専門用語であるかをタグ付けしている NTCIR-1: 情報検索/用語抽出研究用テストコレクションを使用する．

4 おわりに

本研究では，これまでの研究では取り扱ってこなかった専門用語に関する統計量である係り受けによる影響度を用いたロジスティック回帰モデルで専門用語の自動抽出手法を提案した．今後の課題として，本研究で提案した手法は未知語に対応できないため，専門用語に関するより汎化した特徴を統計量として用いる必要がある．

謝辞　NTCIR1 の用語抽出研究用テストコレクションは，国立情報学研究所より提供された．ここに記して謝意を表する．

参考文献

[1] 内山, 鈴木, 相澤：“専門用語の専門度の指標に関する一考察”, 言語処理学会第 16 回年次大会発表論文集, pp. 571–574 (2010).

[2] 中川, 湯本, 森：“出現頻度と連接頻度に基づく専門用語抽出”, 自然言語処理, **10**, 1, pp. 27–45 (2003).

[3] 佐藤, 佐々木：“ウェブを利用した関連用語の自動収集”, 情報処理学会研究報告自然言語処理（NL）, 第 2003 巻, pp. 57–64 (2003).

[4] 内元, 村田, 関根, 井佐原：“後方文脈を考慮した係り受けモデル”, 自然言語処理, **7**, 5, pp. 3–17 (2000).