ISS-SP-049

係り受け解析を用いた ドメイン固有のキーワード抽出手法の提案

木村 優介 †　　楠 和馬 †　　馬場 睦也 †　　波多野 賢治 ‡

† 同志社大学大学院文化情報学研究科　　‡ 同志社大学文化情報学部

1　はじめに

専門用語のようなドメイン固有のキーワードの翻訳は一貫したルールがないと，その対訳にばらつきが生じてしまう．翻訳における対訳のばらつきをなくすためには，そのドメインの専門家がドメイン固有のキーワードをテキストから抽出し，翻訳する必要がある．ただし，その抽出作業にはドメイン固有のキーワードか否かの判断に時間がかかるため，抽出の自動化が求められている．しかし，ドメイン固有のキーワードを自動で抽出するために有効な特徴量は明らかにされていない．

そのため，既存の手法が有効な特徴量を取り逃している可能性があり，この問題に対処するためにはあらゆる特徴量を用いる必要がある．

2　提案手法

本研究ではドメイン固有のキーワードに関するあらゆる特徴量を用いた手法を提案する．その特徴量を用いた教師あり分類学習手法のうち，最も精度が良かったモデルで算出された分類確率をドメイン固有のキーワードになりやすい重要度として用いることで抽出を行う．

先行研究では扱っていないドメイン固有のキーワードの特徴として，その語の説明や定義の存在が挙げられている [1]．説明や定義はそのドメイン固有のキーワードと同じ文中で出現すると仮定し，本研究ではその関係と類似する関係を表すことができる文節係り受け構造で表すことができると考えた．文節係り受け構造とは各文節間の修飾関係を表す構造であり，その構造から得られる次の特徴量を提案する．

修飾文節数，被修飾文節数： 後方修飾の特徴を持つ係り受けを再帰的に辿ることで重要度の算出対象である候補語を含む文節が修飾する文節数とその候補語を修飾する文節数を算出する．ある文を係り受け解析した図 1 の文節 C に着目すると，修飾文節数は文節 C が文節 D を修飾し，その文節 D が文節 E を修飾するため二つになる．文節 C の被修飾文節数は文節 A と文節 B の修飾対象になっているため二つになる．

修飾経路数，被修飾経路数： 修飾文節数，被修飾文節数を算出するために辿った対象の候補語までの経路の数とその候補語からの経路の数を算出する． 図 1 の文節 C に着目すると，修飾経路数は文節 C から文節 E までの 1 経路，被修飾経路数は文節 A と文節 B からの 2 経路がある．

修飾，被修飾経路上の品詞別の出現頻度： 修飾経路，被修飾経路上の各形態素の品詞の合計を算出する．図 1 の文節 C に着目すると，修飾経路上の名詞や助詞，動詞は 1 で，他の品詞は 0 である． 被修飾経路

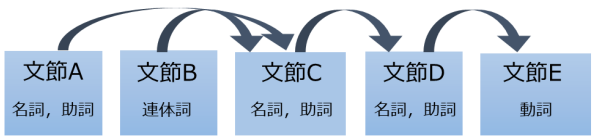

図 1　係り受け解析を行った一例
上の名詞や助詞，連体詞は 1 で他の品詞は 0 である．

3　評価実験

本研究で提案した手法によってドメイン固有のキーワードがどれだけ正確に抽出されたかを評価するため，候補語を重要度で降順に並び替え，候補語を n 語抽出した際の適合率 P で評価を行う．

$$P = \frac{TP}{TP + FP} \tag{1}$$

ただし，TP は正しく抽出できたドメイン固有のキーワードの数を表し，FP は誤って抽出したドメイン固有のキーワードではない候補語の数を表す．

評価のために人工知能分野の論文抄録からあらかじめ人手でドメイン固有のキーワードが抽出された NTCIR-1 用語抽出研究用テストコレクション[1]を用いる．

本研究ではキーワードではない語が誤って判定されてしまうことを防ぐために，ドメイン固有のキーワードのうち，使用するデータセットで最も多く出現する品詞である名詞を候補語とする．

4　おわりに

本研究ではドメイン固有のキーワードに関するあらゆる特徴量を用いた教師あり分類学習手法を提案した．今後の課題として，本研究の提案手法はドメインが特定されたテキストにしか適用できないため，未知のドメインに属する文書からドメイン固有のキーワードを抽出できる手法が必要になる．

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] 佐藤理史, 佐々木靖弘. 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.

[1]国立情報学研究所（2015）「NTCIR Project テストコレクション利用手続き・覚書（研究目的用）」 http://research.nii.ac.jp/ntcir/permission/perm-ja.html（閲覧日：2020 年 2 月 9 日）