G-19

SBT(Social Based Testing)における自動カテゴリ分類手法の精度向上へ向けた一検討

A study on improving the method for automatically classifying the category in social based testing

○中西洋介[1]，村上和希[1]，綿名一樹[2]，五味悠一郎[3]
*Yosuke Nakanishi[1], Kazuki Murakami[1], Kazuki Watana[2], Yuichiro Gomi[3]

Abstract: We are developing an education information system that users can post questions and answers. We call this system social based testing (SBT). When the representation of question categories in SBT is biased toward certain categories, it can be difficult for users to learn evenly. I propose an information system that classifies the questions posted by user into categories automatically.

１．まえがき

SBT(Social Based Testing)とは，筆者らが開発している，学習者が問題と解答を投稿し合える機能を持つ教育情報システムである．

SBT に投稿された問題がカテゴリ別に分類されないと，問題のカテゴリに偏りがあった場合に，万遍なく学習することが難しい．また，投稿された問題数が少ないカテゴリが分からないと，投稿者はどのカテゴリの問題を補えばよいのか分からない．そのため，SBT の投稿問題をカテゴリに分類する必要がある．

本報告では，SBT に投稿される問題を対象に，自動でカテゴリ分類する機能を持ったシステムを提案した．

２．提案手法

２．１．概要

本報告では医療情報技師能力検定試験を対象に，2015 年時点の医療情報技師育成における新到達目標に沿ったカテゴリに分類することを目標とした．提案手法の概要を Figure1 に示す．提案手法では，カテゴリ別の過去問題から専門用語を自動抽出し，教師データとなるカテゴリ別の専門用語群を構築した．投稿問題に含まれる名詞と構築した教師データを比較し，最も一致したカテゴリに分類した．

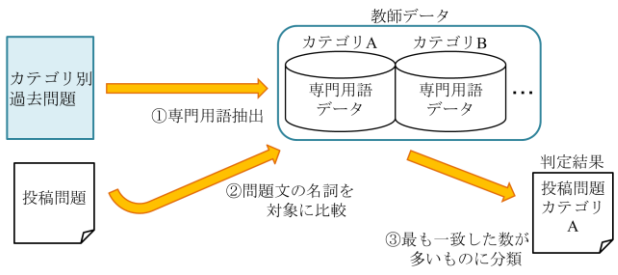

Figure1. Overview of the proposed method

２．２．専門用語の抽出手法

２．２．１．形態素解析

カテゴリ分類の対象となるテキストを，Mecab を用いて形態素解析した．Mecab は文章を形態素(語)ごとに区切り，品詞，活用形，原形，読みを付与するプログラムである．専門用語は複合語や複合名詞である場合が大半を占めているため，名詞(単名詞と複合名詞)を専門用語候補として抽出した．

２．２．２．専門用語の特定

中川らが発表した FLR 法を用いて，抽出された各候補語から専門用語を特定し，教師データを構築した[1]．本報告では，専門用語の定義を「その用語を見ただけで意味，内容がわからない用語，一般的に使われていないと判断した用語」とした．FLR 法で用いられる用語の定義を Table1 に示す．

Table1. Definition of FLR method

用語	定義
FLR法	連接頻度と出現頻度を利用したスコア付け手法
連接頻度	ある単名詞が複合名詞を形成するために連接する頻度の平均 「複合名詞の中でよく使われる単語」
連接する単名詞	前後への単語の繋がり
出現頻度	ある名詞が単独で出現する頻度
単独で出現する名詞	他の複合名詞に包含されることなく出現する名詞
スコア	複合名詞の用語としての重要度

FLR 法を専門用語「医療情報技師」に用いた例を Figure2 に示す．2.2.1 項の形態素解析で抽出された各候補語に重要度をスコア付けし，専門用語を特定した．

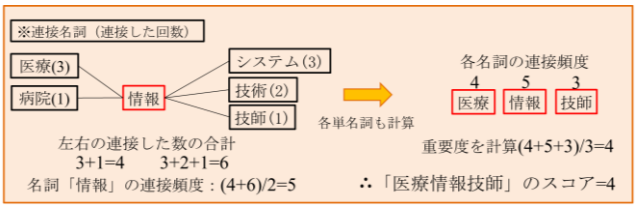

Figure2. Scoring example

２．３．投稿問題のカテゴリ分類

2.2.2 項で述べた手法を基に，カテゴリ別の過去問題から専門用語を特定し，教師データとなるカテゴリ別の専門用語群を構築した．投稿問題に含まれる名詞と構築した教師データを比較し，最も一致した数が多いカテゴリに分類した．投稿問題においても 2.2.1 項で述べた方法で名詞を抽出した．

１：日大理工・学部・子情　２：日大理工・院(前)・情報　３：日大理工・教員・情報

２．４．キーワードを用いた比較方法

専門用語は，教師データ内の複数のカテゴリに存在する場合があるため，対象問題が複数のカテゴリと一致してしまう可能性がある．教師データの構成要素として，専門用語だけではなく，カテゴリの内容をよく表すキーワードを特定し，利用することにした．キーワード特定の概要を Figure3 に示す．

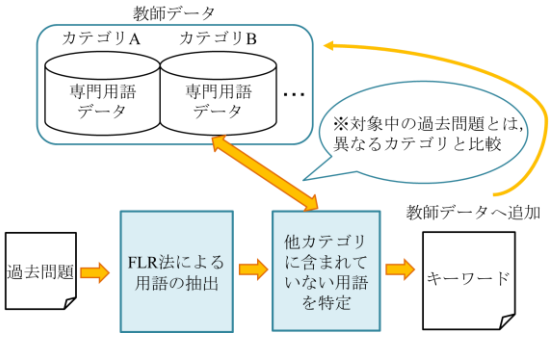

Figure3. Overview of Keyword specific

３．評価実験

３．１．専門用語抽出実験

2.2 節で述べた専門用語抽出手法の評価実験を，医学・医療系，情報処理技術系，医療情報システム系の三つの系で構成された第 3 回医療情報技師能力検定試験問題(総単語数 1,766，総専門用語数 638)を対象に行なった．本報告では，提案手法を再現率(網羅性を表す)，精度(正確性を表す)，F 値(再現率と精度の調和平均)で評価した．再現率，精度は(1)，(2)の式で求める．専門用語の定義を基に，正解データとなる専門用語を決定した．実験結果を Table2 に示す．なお，カッコ内はそれぞれ(1)式，(2)式の数値を表す．

$$\text{再現率} = \frac{\text{正解した出力結果の数}}{\text{正解データの専門用語数}} \quad (1)$$

$$\text{精度} = \frac{\text{正解した出力結果の数}}{\text{出力結果の総数}} \quad (2)$$

Table2. experimental result

	再現率	精度	F値
医学・医療系	0.95(189/198)	0.49(189/387)	0.65
情報処理技術系	0.94(191/204)	0.48(191/396)	0.64
医療情報システム系	0.94(222/236)	0.31(222/714)	0.47

すべての系において 0.94 以上の再現率が得られたが，精度に関しては 0.31 以上であったため，F 値は 0.47 以上という結果が得られた．正解した出力結果の数は多かったが，不正解の出力結果も多くあったため，精度が低くなった．

３．２．自動カテゴリ分類実験

提案手法の評価実験を行なった．2.3 節で述べた手法を純粋に評価するため，2 種類の教師データを用意した．一方は，2.2.2 項で述べた専門用語の定義を基に，専門用語を手作業で抽出した，正解データのみの教師データである．もう一方は，過去問題を対象に，2.2 節で述べた手法で構築した教師データである．教師データには第 3,4 回医療情報技師能力検定試験問題を使用した．分類対象として第 5 回医療情報技師能力検定試験問題を使用し，3.1 節と同様の評価値を用いて評価した．また，カテゴリ分類の正解は到達目標と医療情報の教材を基に上級医療技師が手作業で分類したものを用いた．実験結果を Table3 に示す．なお，全データに対する再現率と精度は一致するため，F 値のみを記載した．カッコ内の数値は，正解数/問題数を表す．

Table3. experimental result

	正解教師データ	提案手法での教師データ
医学・医療系	0.50(30/60)	0.48(29/60)
情報処理技術系	0.70(42/60)	0.77(46/60)
医療情報システム系	0.63(44/70)	0.59(41/70)

最小で 0.48，最大で 0.77 の F 値が得られた．教師データの語数が多いとカテゴリ分類の性能が向上すると考えられ，多くの過去問題が必要と分かった．また，2 種類の教師データで，最大 0.07 の数値の差(情報処理技術系の結果より，0.77－0.70)が生じたが，情報処理技術系においては，提案手法の方が上回る数値であった．

３．３．キーワードを追加した自動カテゴリ分類実験

2.4 節で述べた比較方法の評価実験を，3.2 節と同様の方法で行なった．しかし，評価値は 3.2 節の結果よりも増加しなかった．教師データが少なく，複数のカテゴリに存在する専門用語も少なかったことが，評価値の増加につながらなかった要因として考えられる．教師データを増やし，提案手法の有用性を確かめ，比較方法を改善する必要がある．

４．まとめ

本報告では，SBT の投稿問題を自動でカテゴリ分類する手法の検討を行なった．教師データの語数が多いほど，分類が正しく行われた割合が大きい結果であった．今後は，比較方法の検討や教師データを増やして検証する予定である．

５．謝辞

本研究は JSPS 科研費 26870619 の助成を受けたものである．

６．参考文献

[1] 中川裕志，森辰則，湯本紘彰：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，Vol10，No1，pp.27-45，2003.