2ZF-4

形態素解析を用いた特徴語抽出手法の検討

宮川　裕介†　　瀬沼　航太郎†　　泉　　隆†

日本大学†

1. はじめに

ICT 技術の発展により，インターネットを用いて学習を行う e-Learning システムが教育機関や企業で利用されている．そこで，我々は基本情報技術者試験を対象としたシステムの構築と学習評価方法の検討[1]を行ってきた． e-Learning システムを使用する利点として，時間や場所の制約がない，学習や教育の効率化を図れるといった点が挙げられる．しかし，その一方で管理者の負担が大きいことや利用頻度は学習者の学習意欲に依存するといった欠点も挙げられる．特に学習意欲の低下は目標の達成度に大きな影響を及ぼすため，学習意欲を向上・維持させるようなシステムの構築が求められる．

本研究では，システムを通じて学習者から得られる解答情報に加え，問題文が持つ属性情報を学習評価に用いることを目的としている．この学習評価により，各学習者が苦手とする属性の把握とその対策をフィードバックすることが可能になり，学習意欲の維持・向上につながると考える．

問題文の属性分析手法としてデータマイニングで用いられている知識発見プロセス[2]を用いる．これを本研究の対象である基本情報技術者試験[3]に応用し，問題文から有用で意味のある文字列である特徴語を発見する．本報告では，試験問題に特徴語抽出手法を適用した実験結果を示し，その有効性を考察する．

2. 対象試験の特徴

本研究で対象とするのは基本情報技術者試験である．この試験で出題される問題は，経験則より以下の問題形式に分類することができる．

① 専門用語の意味や説明を問う問題
② ある説明に最も適した用語を選ぶ問題
③ 数値や思考による計算結果を求める問題

各問題にはその問題を特徴付ける用語が含まれており，本報告ではこのような用語を問題の属性とすべく特徴語と定義する．なお，本報告における特徴語は索引に記載されるような重要度の高い用語とする．また，特徴的であっても一般的に使用されるような重要度の低い用語（「ネットワーク」など）は非特徴語とする．

“Examination of The Feature Word Extraction Method using Morphological Analysis”
†Yusuke Miyakawa, †Kotaro Senuma, †Takashi Izumi
†Nihon University

3. 知識発見プロセスの概要

知識発見プロセスとは巨大なデータの中から特徴的なパターンを見つけるためのアルゴリズムである．これは Fayyad らによって定式化され，データマイニングに使われている．その概念図を図 1 に示す．

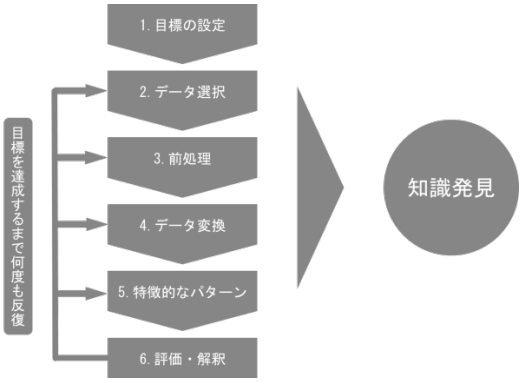

図 1　知識発見プロセスの概念図

4. 対象試験における特徴語発見プロセス

3.に示した知識発見プロセスを対象試験に対応させる．ここでは，本研究に適用するようアレンジしたプロセスを述べる．

4.1 目標の設定

本報告では 2.で定義した特徴語のみを対象試験の問題文から抽出し，尚且つ誤抽出を最小限にとどめるプロセスの構築を目標とする．

4.2 データ選択

対象試験から得られるデータは問題文，選択肢，正解選択肢である．本報告では問題文を分析し，特徴語を抽出するので分析に必要な問題文のデータのみを選択する．

4.3 前処理

選択した問題文のデータには出題年や問題番号，全角半角が統一されていない英数字などが含まれている．そこで，これらを正規化し，意図する分析手法が適応できる文字列に変換する．

4.4 データ変換

形態素解析により得られる品詞情報と形態素を用いて特徴語の抽出を行う．なお，形態素解析器は MeCab[4]を使用した．

4.3 で前処理を行った問題文に対し，形態素解析を行い，品詞情報と形態素に変換する．次にあらかじめ用意した特徴語に対して同様の処理を行い，品

詞情報を特徴語の識別パターンとして用いる．なお，特徴語の品詞情報は対象試験のシラバス[3]に記載されている用語 2809 語から取得した．

4.5　特徴的なパターン

4.4 で得た品詞情報を使用し，問題文の中から特徴語と思われるパターンを発見する．そして，それに対応する複数の形態素を一つの文字列として抽出する．

抽出した文字列より文字数を調べる．特徴語は 3 文字以上で構成されている場合が多いので抽出した文字列のうち，2 文字以下のものは抽出しない．また，3 文字以上であっても先頭に「実－」や「何－」などの接頭辞を含む文字列，末尾に「－中」や「－的」などの接尾辞を含む文字列は一般的に使われる用語であることが多いので，これも抽出しない．

4.6　評価・解釈

以上のプロセスにより特徴語の正解率を求め，この手法の有効性を評価する．また，非特徴語がどのような特徴を有しているのか評価し，誤抽出数を抑えるための処理を 4.3 の前処理や 4.5 のパターン発見部分に組み込む．

5.　特徴語抽出実験

5.1　実験条件

4.で示した提案手法を用いて以下の実験条件より特徴語抽出実験を行った．

表 1　実験条件

対象	平成 22 年　秋期 基本情報技術者試験　午前問題　80 問
測定項目	・抽出した特徴語の正解率 ・特徴語の抽出率
特徴語数	143 語

表 1 より本実験では，2.に示した特徴語の定義に当てはまる用語を対象から抽出し，これを特徴語とした．そして，提案手法を用いて抽出した用語のうち特徴語を正抽出，非特徴語を誤抽出とし，抽出できなかった特徴語を未抽出とする．

抽出した用語に含まれる正抽出の割合を正解率と定義し，式(1)から求める．また，特徴語における正抽出の割合を抽出率と定義し，式(2)から求める．

$$(\text{正解率}) = \frac{(\text{正抽出数})}{(\text{正抽出数}+\text{誤抽出数})} \times 100[\%] \qquad (1)$$

$$(\text{抽出率}) = \frac{(\text{正抽出数})}{(\text{特徴語数})} \times 100[\%] \qquad (2)$$

5.2　実験結果

実験結果を表 2 に示す．

表 2　特徴語抽出実験結果

	全体	問題形式 ①	問題形式 ②	問題形式 ③
特徴語数［語］	143	52	44	47
正抽出数［語］	141	51	44	46
誤抽出数［語］	49	6	13	30
未抽出数［語］	2	1	0	1
正解率［%］	74.2	89.5	77.2	60.5
抽出率［%］	98.6	98.1	100	97.9

表 2 より全体の抽出率は 98.6[%]となった．これは形態素解析により得た品詞情報を識別パターンとして用いたことで，未知の特徴語に対応できたことを示している．しかし，全体の正解率は 74.2[%]と抽出率より低下している．この原因として，特徴的なパターンを持つ非特徴語が抽出されたためと考える．誤抽出となった用語は「タスク」，「使用時間」など特徴的であるが重要度の低いものであった．

各問題形式の正解率に着目する．3 つの問題形式の中で①が最も高い値となった．これは他の形式より問題文の長さが短い傾向にあり，非特徴語が抽出されることが少ないためと考える．これに対して，正解率が最も低い③は問題文が長い傾向にある．このことから正解率の精度は問題文の長さに影響されることがわかる．

これらより，本手法は未知の特徴語に対して高い抽出率となるが，特徴的なパターンを持つ非特徴語の誤抽出を抑制するのは困難である．また，それに伴い，正解率の精度は問題文の長さに影響される．

6.　まとめ・今後の課題

本報告では知識発見プロセスに基づき，形態素解析を用いた特徴語抽出手法の検討を行った．その結果，未知の特徴語に対して正解率 74.2[%]となった．

今後は抽出した用語を用いて問題文のクラスタリングを行い，クラス内での各用語の出現頻度から誤抽出を抑制できないか検討する．

参考文献

[1] 久津間啓右：「インターネットを利用した情報技術学習支援システム ―S-P 表を用いた学習状況の評価の検討―」，第 9 回情報科学技術フォーラム，pp.(4-461)-(4-462)，N-006，(2010).

[2] U.Fayyad, G.Piatetsky-Shapiro, and P.Smyth: The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, Vol. 39, No. 11, pp. 27–34 (1996)

[3] 「基本情報技術者試験（レベル 2）」シラバス（Ver 3.0）：http://www.jitec.ipa.go.jp/1_13download/syllabus_fe_ver3_0.pdf (2014-1)

[4] MeCab : Yet Another Part-of-Speech and Morphological Analyzer: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html (2014-1)