N-020

e-Learning システムにおけるコンテンツ自動分類手法の検討
Examination of Automatic Content Classification Method in e-Learning System

宮川　裕介†　　瀬沼　航太郎†　　泉　隆†
Yusuke　Miyakawa　　Kotaro　Senuma　　Takashi Izumi

1. はじめに

e-Learning システムが教育機関や企業で利用されている．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を行ってきた．[1]しかし，e-Learning システムにはコンテンツの作成･管理に要する作業量が多いため，コンテンツの拡充と共に管理者の負担を軽減させる仕組みが必要とされている．そこで，コンテンツの自動分類手法を考案することで管理者の負担軽減をはかる．

本研究では，基本情報技術者試験を対象とし，IPA が公表するシラバス[2]に沿って過去に出題された問題を管理しやすいように分類することを目的とする．

本報告では形態素解析エンジン「MeCab」[3]を利用して特徴候補となる単語の抽出を行った．また，シラバスに記載されている用語をもとに，特徴となる単語か識別するための辞書を作成し，ラベル付けを行う．そして，これらからコンテンツの分野別分類が可能か検討した．

2. 対象試験の特徴

基本情報技術者試験の問題は大分類，中分類，小分類の 3 項目によって分野ごとに振り分けられる．そして，午前の試験は問題番号によって出題される問題の系列（テクノロジ系，マネジメント系，ストラテジ系）が定められている．小分類は計 99 項目あり，午前の試験はこの小分類順に出題されることが多い．よって，これらの特徴から午前に出題される試験問題を分類する．

3. 自動分類手法の概要

本節ではコンテンツ自動分類手法の概要を説明する．なお，本論におけるコンテンツとは基本情報技術者試験に出題された問題のことを指す．また，以下に言葉の定義を記す．

特徴語　：分類の際に特徴量となる単語
識別辞書：特徴語とそれに対応するクラスをまとめたもの

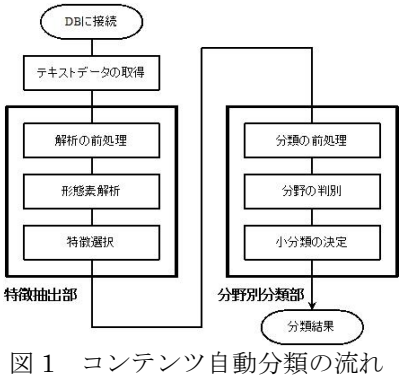

図 1　コンテンツ自動分類の流れ

コンテンツ自動分類における処理の流れを図 1 に示す．

- **テキストデータの取得**
  データベースに接続し，問題文と正解選択肢のテキストデータを取得する．
- **解析の前処理**
  特徴抽出部にて解析に不要なテキストを削除する．
- **形態素解析**
  MeCab を用いて形態素解析を行い，特徴語の候補となる単語を抽出する．
- **特徴選択**
  抽出した特徴語候補と識別辞書の情報を比較し，一致するものを特徴語としてラベル付けする．
- **分類の前処理**
  対象としている試験は問題番号によって系列が定められているので，問題とラベル付けした特徴語の系列名が一致しているもののみ特徴量として用いる．
- **分野の判別**
  特徴量として用いる特徴語のうち，特徴語間で大分類名から順に一致しているものをカウントする．そして，数が最も大きいものをその問題の分野とする．
- **小分類の決定**
  もし分野の判別ができなければ，特徴語の分類番号がもっとも若いものをその問題の分野とする．なお，特徴語が一つも抽出されていなければどの分野にも振り分けない．

4. 形態素解析

4.1 概要

形態素解析とは自然言語の文字列を意味のある最小の文字列（形態素）まで分割し，それぞれの品詞を判別する作業のことである．

本報告では条件付き確率場(CRF)[4][5]に基づく高い解析精度を誇るオープンソースの形態素解析エンジン「MeCab」を用いて，特徴語候補の抽出とラベル付けを行う．

4.2 条件付き確率場（CRF）

CRF は主に“系列ラベリング”問題を解く際に使われている．系列ラベリング問題とは入力されたデータ列に対し，出力として個々のデータにラベル付けを行うものである．以下に例を示す．

入力：私 / は / 夕飯 / を / 食べ / た
出力：名詞 / 助詞 / 名詞 / 助詞 / 動詞 / 助動詞

† 日本大学　Nihon University

上記のような連続する形態素に対して，それぞれに当てはまる品詞に分類する．このように系列ラベリング問題は分類問題として解くことができる．

4.3 識別モデル

CRF は個々のデータにラベル付けを行うことができると記したが，どのようにラベル付けを行っているのか．機械学習の代表的な手法として「識別関数」と「識別モデル」がある．

識別関数はパーセプトロンや SVM など，教師データとの距離をもとにデータをクラスに分類する手法である．一方，識別モデルはデータがあるクラスに分類される確率をもとに分類する手法である．CRF ではこの「識別モデル」として，基本的に対数線型モデルを使用している．

4.4 構造学習

また、CRF は「構造学習」という手法を用いている．構造学習とは個別にクラス推定を行わず，データを全て渡してからまとめてデータごとにラベル付けを行う手法である．

4.2 節の例より，通常の分類問題では個別のデータに対して品詞を分類するため，個々の品詞としては最適であっても全体からすると最適でない分類結果となることがある．そこで，個別にクラス推定を行わず，前後の素性から最適な結果を得る．たとえば，助詞の後ろにある単語の品詞を推定するとき，考えられるのは名詞や動詞である．しかし，さらにその後ろの品詞に助動詞が推定されていれば動詞である確率がかなり高いのでこの助詞の後ろは動詞であることが推定される．このように全体として最適な分類結果を得るために，データを全て渡して分類問題を解くのが「構造学習」である．

5. 実験

5.1 実験条件

4 節で説明した手法を使用した形態素解析エンジン「MeCab」と 3 節の分野別分類部のアルゴリズムより試験問題の分類実験を行った．表 1 に実験概要を示す．

表 1　実験概要

対象	平成 21 年度春期 基本情報技術者試験　午前問題　80 問
測定項目	大分類，中分類，小分類における 特徴語の正抽出率・分類正解率

本報告における正抽出率と分類正解率の定義を以下に示す．なお，識別辞書には IPA が公表するシラバスの小分類ごとに記載されている用語例を使用し，作成した．

正抽出率　：分野を正しく判別できる特徴語が抽出できた割合
分類正解率：試験問題のうち正しく分類できた割合

5.2 実験結果

上記の実験条件より以下のような結果を得た．

表 2　実験結果

	正抽出率 [%]	分類正解率 [%]
大分類	53.75	40.00
中分類	46.25	33.75
小分類	32.50	23.75

実験結果より正抽出率および分類正解率は全体的に低い値となった．正抽出率が低くなった原因として，本実験で使用した識別辞書の用語数が不十分であったことがあげられる．シラバスには，全ての用語ではなく，各小分類から抜粋した用語のみを記載しているためであり，試験問題に出現する特徴語をすべて網羅するには至らない．このことから，識別辞書の作成方法について再検討する必要がある．

また，各正抽出率と分類正解率を比較すると正抽出率より分類正解率の方が約 10 [%] 低くなっていることが分かる．このことから図 1 における分野別分類部が分類精度を悪くしている原因の一つとしてあげられる．また，特徴抽出部の特徴選択において特徴語候補から特徴語を選別する際，識別辞書にある用語と完全に一致しているもののみを特徴語としたことも原因の一つである．これらのことから，本来なら特徴語となり得る単語が非特徴語としてはじかれ，正抽出よりも低い分類正解率となったのではないかと考える．このことから，特徴選択の選択基準の見直しや分類精度の高い分類器を構築する必要があると言える．

6. まとめ

本報告では基本情報技術者試験を対象とし，特徴となる単語の抽出と問題の分類について検討した．その結果，正抽出率と分類正解率はともに低い値となった．これは識別辞書を構成するデータの数が不十分であったことや分類に用いたアルゴリズムの汎化能力が低かったことが主な原因であると推測される．これらのことから識別辞書の作成方法の再検討や汎化能力の高い分類器を構築する必要があると考える．

今後は，識別辞書の作成方法の検討および汎化能力高い分類器を比較し，再度実験を行うことで分類精度の向上を行う．

参考文献

[1] 金子勇太:「利用者の学習意欲を維持する e-Learning システムの開発 —利用者評価に関する検討—」，情報処理学会大会，N-021 (2013-03)

[2] 「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 (2013-6)

[3] MeCab：Yet Another Part-of-Speech and Morphological Analyzer:http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html (2013-6)

[4] 工藤拓，山本薫，松本裕治：「Conditional Random Fields を用いた日本語形態素解析」，情報処理学会研究報告,2004(47), 89-96, (2004-5-13)

[5] 奥村学，高村大也：「自然言語処理のための機械学習入門」，コロナ社 (2010)