記述式小テストの解答自動分類のための特徴抽出に関する一考察

大井健太郎　高瀬治彦　森田直樹*　北英彦　林照峯
三重大学大学院 工学研究科　*愛媛女子短期大学 生命科学研究所
ooi@ce.elec.mie-u.ac.jp

1. はじめに

講師が, 講義の内容をどれだけ学生が理解しているのかを把握することは, わかりやすい講義を行うために必要である. 講義中に学生の理解度を知るための手段にはさまざまなものがある. そのひとつに小テストがある. 小テストの結果, 学生の理解が不十分である箇所を知ることができ, 追加説明などのフィードバックをすることで, 学生の理解度が向上する. しかし, 小テストは解答の回収, 集計, 分析に手間がかかるため, 講義時間中に効果的なフィードバックを行うことは難しい.

近年, e-ラーニングが注目されている. 計算機を使用することで, 講義にさまざまな効果がもたらされる. これを記述式小テストに用いて, 解答の回収, 集計, 分析を素早く行うことを, 本研究の目的とする. これが実現できれば, 学生へ適切なフィードバックをすみやかに行えるようになる.

記述式小テストでは, 解答中にさまざまな表現が現れるため, 全体の傾向を即時につかむことは困難である. そこで, 本研究では, 記述式小テストの解答の傾向を講師が短い時間で把握できるように, 解答を分類する手法について検討する.

2. 記述式小テストの解答の自動分類

2.1. 解答の自動分類における注意点

多数の文書をその内容にしたがって分類するための手法が数多く提案されている[1]. これらの手法が対象としている文書の集合に対して, 本研究で対象とする小テストの解答の集合は, 以下の 2 つの特徴を持つ.

(1) 分類に時間をかけすぎてはいけない.

小テストの解答を自動的に分類することにより, 講師がすみやかに学生にフィードバックできるようにすることを目的としているので, 分類に時間がかかっては意味がない. 本研究では, 分類の開始から完了までにかかる時間が 10 秒以内であることを目標とする.

(2) 分類対象は小規模なデータである.

一般的な文書の分類では, 件数の多いデータを対象とすることが多い. また, 文章の長さも小テストの解答に比べると長いものが多い. 本研究では, 学生の人数は 100 名程度, 学生の解答の長さは 2, 3 文までを想定している.

本研究では, これらの点を考慮に入れた新しい分類手法を提案する.

2.2. 文書の自動分類

文書の自動分類にはベクトル空間モデルがよく用いられている[1]. ベクトル空間モデルは, 出現単語にもとづいて文書を 1 つのベクトルで表現する. ベクトル空間モデルを利用した一般的な分類の手順を以下に示す.

1. 名詞を抽出する.

   形態素解析により, 普通名詞, 固有名詞を分類のためのキーワードとして抜き出す.

2. 文書をベクトル化する.

   出現したキーワードとその出現頻度に応じて文書の情報を数値化し, ベクトル化する. これを特徴ベクトルと呼ぶ.

3. 特徴ベクトルをもとに分類する.

   ベクトルの余弦あるいは距離からベクトルの類似度を測り, それをもとに分類する.

本研究では, 手順 2 の文章をベクトル化する部分について検討する. これは, 文書数が少なくかつ短いため, 文書内の情報をできるだけ詳細に数値化しないと, 分類が困難になるためである.

2.3. 従来法による文書のベクトル化

文書のベクトル化手法として, tf･idf 法を利用した手法がある[1]. tf･idf 法は, 文書中にある単語が出現する回数に応じてその重みを重く (tf 値), 多くの文書に使われている単語ほど重みを軽くする (idf 値) ことで, 文書中の各単語を数値化し, それをベクトル化したものを特徴ベクトルとして用いる.

また, tf･idf 法に加えて, 簡単な構文情報に着目することで, 各単語の重みを決定する手法が提案されている. 例えば, 黒橋らは助詞に着目してキーワードの重み付けをする手法を提案している[2]. この手法は, 助詞に着目して tf･idf 法で作った特徴ベクトルにさらなる重み付けをする. 例えば, 「は」が付属する単語の重みを一定倍することで主語にあたる単語を中心に分類することができ, 「を」が付属する単語の重みを一定倍することで目的語にあたる単語を中心に分類することができる. このように, 特定の助詞に付属する単語の重みを重くすることで重要視したい情報をもとに分類できる.

しかし, これらを小テストの分類に用いた場合, 以下の問題点がある. tf･idf 法では多くの解答に含まれている語の重みは軽くなるため, 正解に含まれる単語の idf 値が小さくなる. その結果, そのような単語が分類に際して重要視されなくなり, 正答と誤答が混在した分類結果を得る.

また, 助詞の用い方が解答者によって異なるため, 黒橋らの手法では, 必ずしも重要視したい単語の重みが重くならない.

3. 提案する解答群のベクトル化手法

2.1 で述べたように, 小テストの解答が小規模なので, その解答に含まれる名詞の使用回数のみに着目するだけでは情報が足りないことが考えられる. そこで本研究では, 次の二つの考え方に基づき解答を数値化する.

1. 数値化対象の品詞の拡大
2. 係り受けに基づく重み付け

以下でそれぞれについて詳しく説明する.

3.1. 数値化対象の品詞の拡大

小テストの解答にはそれほど多くの単語が含まれていない. その中から名詞のみを取り出して分類に用いるのでは, 情報が足りないことが考えられる. そこで, 名詞だけではなく動詞も分類に用いることにする. これは, 文章の主な構成要素は主語 (名詞) と述語 (動詞) であるためである. ただし, 動詞の活用形を考慮に入れると複雑になりすぎるので, 語幹のみを扱うものとする.

3.2. 係り受けに基づく重み付け

これは, 単語の出現回数以外に, 解答中での文節の係り受けの関係に着目する方法である. 具体的には, より多くの文節からの係を受けている文節ほど, その解答で中心的な働きをしていると考え, それに含まれる単語を重要視する. その結果, ささいな修飾語の違いに惑わされず, その文で表したかった内容に沿ったベクトル化が行われる.

具体的には, 各解答文を係り受け解析器[3]で解析し, 各文節について, 修飾されている回数を数え, その回数をその文節に含まれる単語の重み (tf値×idf値) に掛け合わせる.

2.1 で述べたように, 分類にあまり時間をかけることはできない. しかし, 小テスト解答は文書が短いため, 係り受けを解析しても分類に要する時間はそれほど長くならないと考える.

4. 実験

提案手法を, 実際に行った小テスト解答に適用した結果をこの章で示す.

2002 年度前期に, 三重大学工学部電気電子工学科 1 年生を対象に開講された講義「計算機基礎」で行われた記述式小テスト「電子メールを使うときの注意事項は何でしょうか。特に重要と思うものをひとつだけ書いて下さい」に対する, 85 名分の解答を分類する. 講義では, これに関して以下に 5 点を説明した.

1. ウイルスに注意する.
2. 添付ファイルの大きさに注意する.
3. 意味のないメールを送らない.
4. 相手を中傷したり不快にさせたりする内容を書かない.
5. 宛先に気をつける.

つまり, 分類結果はこれら 5 種類およびその他に分かれることが望ましい.

最初に，ベクトル化に使用した単語数の比較を行う．提案法では，分類に使用できる（2 人以上が使用した）87 個の単語（名詞，動詞）について，各解答での平均使用単語数は 7.6（最低 4，最高 15）であった．黒橋らの手法では，名詞のみを対象とするため，総単語数 55，各解答での平均使用単語数は 4.8（最低 2，最高 12）となり特徴ベクトルの情報量を増やすことができた．なお，黒橋らの手法では，「に」と「を」が付属する単語の重みを 5 倍している．これはこの小テストが目的語にあたるものを問う内容になっていると考えたためである．

次に，得られた特徴ベクトルを，一般的な分類手法であるウォード法により分類した結果を比較する．図 1 の (a)，(b) は，それぞれ黒橋らの手法，提案法により得られた分類結果をデンドログラムで表示したものである．なお，デンドログラムを作成するまでに要した時間は，Pentium4 2.4HHz の計算機で約 5 秒であった．より詳細な分類に必要な時間を考慮に入れても，許容できる範囲である．

以下では予想される 5 種類の解答のうち「ウイルスに注意する」という内容を含む解答の分類状況を比較する．そのために，それに該当する解答を筆者が抽出し図 1 において着色した．そのような解答は 30 個あった．

デンドログラムでは，文書間の距離に応じて木構造を作成し，必要なクラスタ数に分かれる高さで木を切断することで分類結果を得る．例えば，(a) では高さ 35，(b) では高さ 2 で木を切ると，それぞれのクラスタ数は 12 と 11 個になる．

ウイルスに注意するという旨を記した解答を多く含むクラスタを挙げると，(a) は①と②のクラスタ，(b) は③と④のクラスタが上位の二つである．それぞれ，クラスタ内のウイルスに関する解答数は①で全 20 個中 12 個，②で全 5 個中 5 個，③で全 9 個中 9 個，④で全 8 個中 6 個である．この結果より，約半数のウイルスに関する解答が，これらのクラスタに含まれていることが分かる．これらのクラスタに含まれる「ウイルスに注意する」記述のない解答の数は提案法によるもののほうが少ないので，提案法のほうが余分なものを含まない分類ができたといえる．

また，誤った分類をしたクラスタ①と④について，そこに含まれていたいくつかの解答を表 1 に示す．黒橋らの手法では，誤って分類された解答の内容が多岐にわたっているが，提案法では，添付ファイルに関するものだけである．講義では，「添付ファイルにウイルスが含まれているかもしれないので，気をつけよ」と教えたので，比較的類似した内容の解答を誤判断したといえる．

以上より，期待した分類結果により近い結果を提案法により得ることができた．

表 1．各クラスタに含まれる解答の例

(a) ①に含まれる解答の例

内容	生徒の解答
ウイルス	送られてきた添付ファイルのウイルス相手に対する言葉遣い
ウイルス	不愉快な電子メールを受け取ったらウイルスかもしれないので削除する。
ウイルス	知らない人からのメールは開かない
無意味	意味のないメールは送らない。
宛先	送り間違えをしない。
添付	添付ファイルは必要な大きさだけ貼る。

(b) ④に含まれる解答の例

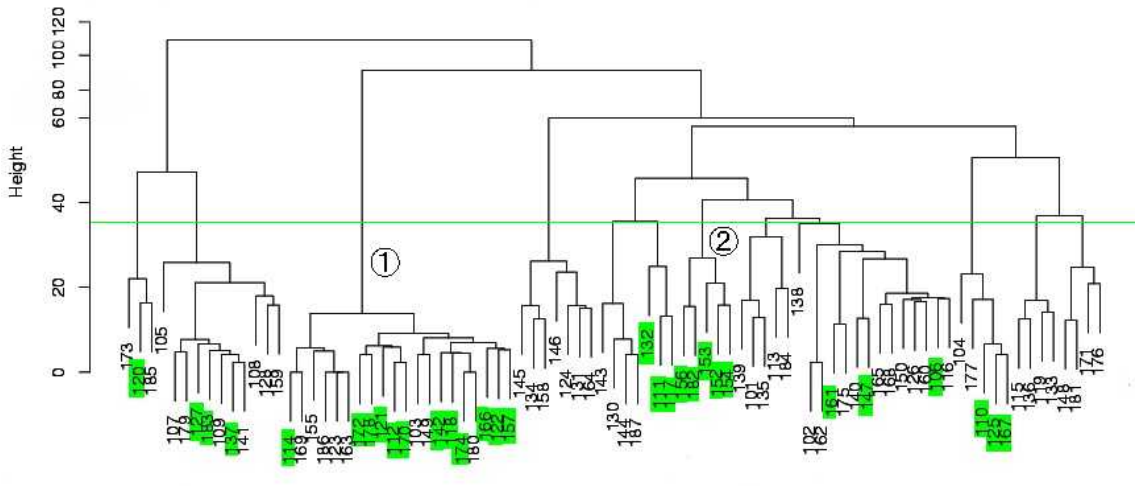
内容	生徒の解答
ウイルス	送られてきた添付ファイルのウイルス相手に対する言葉遣い
ウイルス	知らない人からのファイルが添付されたメールをむやみに開かない。知っている人でも注意する。
添付	容量の大きい添付ファイルを確認なしに送らない。
添付	添付ファイルは必要な大きさだけ貼る。

5. まとめ

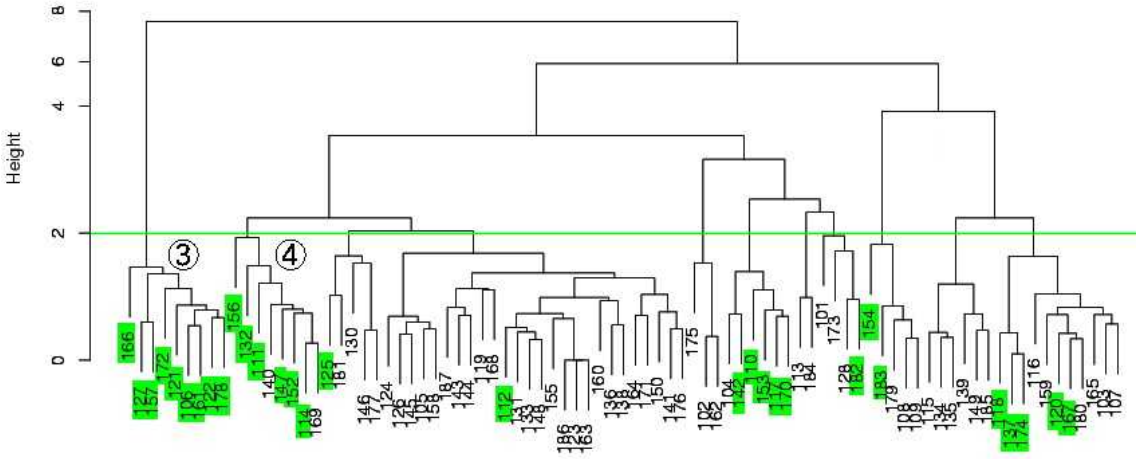
記述式の小テストを行う際に，学生の解答を講師にわかりやすく分類して提示することは，講師が学生にフィードバックを行うことへの手助けとなる．本論文では，その分類法について検討した．

解答を分類するために，各解答から，名詞および動詞を抽出し，その係り受けの構造を解析して得られた情報をもとに分類を行うことで，よりわかりやすい分類結果が得られる可能性を，実験により示した．今後の課題と

して，分類方法や分類の結果をわかりやすく提供する方法を検討することがあげられる．



(a) 黒橋らの手法



(b) 提案法

図1. 分類結果のデンドログラム

参 考 文 献

[1] 徳永健伸：情報検索と言語処理，東京大学出版会 (1999)

[2] 黒橋禎夫，中村順一，長尾真:構文情報を利用した電子ニュース記事のクラスタリングシステムの作成と評価，電子情報通信学会技術研究報告，言語理解とコミュニケーション，Vol.98, pp.15-22 (1998)

[3] 工藤拓，松本裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌，Vol.43，No.6，pp.1834-1842 (2002)

[4] 神嶌敏弘:データマイニング分野のクラスタリング手法(1) クラスタリングを使ってみよう！，人工知能学会誌，Vol.18，No.1，pp.59-65 (2003)