e-Learningシステムにおける学習者の質問への自動応答

吉田　賢史　　井上　温子　　中山　弘隆
情報・システム科学専攻博士課程　情報システム工学科　情報システム工学科

（受理日 2005年4月12日）

1　はじめに

既存のe-Learning管理システム（LMS：Learning Management System）は、学習の進捗状況を管理する機能および学習者のポータルサイト的な役割を持たせたものが主流である。

しかし、通常e-Learningは、学習者個人が時間や場所にとらわれず独りで学習することが多く、受講を続けることが難しいとされる。そのため、学習者の注意を画面に引き戻す効果を付けたコンテンツや、孤独な学習を回避するため、掲示板などを利用してディスカッションを行わせるなどの機能を組み込んだLMSも実験的に利用されている。

しかし、掲示板の活用は時間の拘束を受ける。つまり、教員に対して質問が発せられたとしても、その掲示板を見るまで学習者の質問に対して返答することはない。そのような時間差が学習者のモチベーションを下げてしまうことになる。

そこで我々は、学習者の質問に対しリアルタイムに応答できないかと考え、入力された質問内容をテキストマイニングの手法を用いて、予め想定した質問群、あるいは、過去に質問された質問群から適切な応答を学習者に自動的に表示する機能を提案する。

第2節では、学習者の質問文を解析する形態素解析につて述べ、第3節で学習者の質問があらかじめ想定した質問のそのカテゴリーに属するか判定するための重み付けと類似度について述べる。また、第4節で自動応答とその精度について述べる。

2　形態素解析

2.1　形態素解析とは

英語やフランス語などの多くのヨーロッパ系の言語の場合には、単語が空白で区切られているために、単語を抽出するのは容易であるが、日本語では、単語間に空白がないため単語を抽出することが難しい。

そこで、日本語の文書から単語を正確に抽出するためには、形態素解析（morphological analysis）と呼ばれる技術を用いる。形態素解析とは、与えられた文章を、形態素（語のなかで変化しない最小単位）に分解し、各単語に品詞や語形変化などの情報を与える処理であり、自然言語処理の分野で活発に研究開発が行われている。文章を形態素に自動的に分類する形態素解析器には、茶筌、Mecab、senなどがある。

例えば、茶筌を用いて「2次関数について教えてください。」を品詞別に分類すると、Table 1のようになる。

Table 1: 形態素解析

2次関数について教えてください。		
2	2	名詞-数
次	次	名詞-接尾-助数詞
関数	関数	名詞-一般
について	について	助詞-格助詞-連語
教え	教える	動詞-自立
て	て	助詞-接続助詞
ください	くださる	動詞-非自立
。	。	記号-句点
EOS		

さて、学習者の質問に対し自動応答する場合、日本語の助詞（「は」、「が」、など）は、極めて頻繁に使われる単語であり、これらの単語を用いても、ほとんどすべての質問に含まれる語であるため、特定の質問内容に絞り込むことができない。

このように高い頻度で出現する単語は、質問文を特定する能力が低く、質問を決定する語として適当ではない。このような単語のことを不要語（stop word）と呼ぶ。質問語群から不要語を除去することにより、質問語群の総数を減らすことができるため、処理の効率化や高速化、あるいはデータベース容量の削減することが可能である。

自然言語における単語は、大きく内容語（content word）と機能語（function word）の2つに分けられる。内容語は、それ自体意味を持った単語で、名詞や動詞などがこれに相当する。機能語は、単語と単語の間の関係を表している単語で、日本語の場合には助詞や助動詞などがこれに含まれる。

一般に、機能語は質問内容を特徴付けるうえで役に立たないため、不要語として扱う。また、内容語であっても重要度が低く、不要語にした方がよい代名詞なども除去するべきである。形態素解析を利用し質問語の抽出をおこなうと、各単語に対する品詞が得られ、品詞情報から機能語や代名詞などを判別することができる。その品詞情報に基づき不要語を除去することが可能である。

2.2 想定質問と学習者質問

中学3年生の学習者を対象に「2次関数のグラフとその平行移動」を単元とするe-Learningによる学習を体験させた後、単元に関する質問を書かせた。これを学習者質問$S_i(i=1,2,\cdots,N)$とする（Table 3）。これに対し、教員が予想した学習者質問を想定質問$T_j(j=1,2,\cdots,n)$とする（Table 2）。

Table 2: 想定質問

T_1	aの値を変化させるとグラフはどのように変化しますか
T_2	bの値を変化させるとグラフはどのように変化しますか
T_3	cの値を変化させるとグラフはどのように変化しますか
T_4	切片とa,b,cの関係はありますか
T_5	頂点とa,b,cの関係はどのようになっているのですか
T_6	頂点の座標はどうしてわかるのですか？
T_7	2次関数のグラフはいつも左右対称になるのですか
T_8	グラフを横に動かすにはどのように(a,b,c)のどの値を変化させるのですか？
T_9	グラフを縦に動かすにはどのように(a,b,c)のどの値を変化させるのですか？
T_{10}	放物線を横向きに描く方法はありますか？

Table 3: 学習者質問

S_1	aの値を変化させるとグラフにどのような影響がありますか
S_2	aの値を変化させるとグラフにどのように影響しますか
S_3	aはグラフとどのような関係がありますか。
S_4	aの値はグラフにどのような影響がありますか。
S_5	bの値を変化させるとグラフにどのような影響がありますか
S_6	bの値を変化させるとグラフにどのように影響しますか
S_7	bはグラフとどのような関係がありますか。
S_8	bの値はグラフにどのような影響がありますか。
S_9	cの値を変化させるとグラフにどのような影響がありますか
S_{10}	cの値を変化させるとグラフにどのように影響しますか
S_{11}	cはグラフとどのような関係がありますか。
S_{12}	cの値はグラフにどのような影響がありますか。
S_{13}	切片の傾きは何の値が関係しているのですか
S_{14}	頂点と各値a,b,cの関係はどのようになっているのですか
S_{15}	頂点の座標はどのようにもとめるのですか？
S_{16}	2次関数のグラフはいつも線対称になるのですか
S_{17}	グラフを横に平行移動させたいのですがどの値を変化させればいいのですか
S_{18}	グラフを横に動かしたいのですが、どうすればいいですか？
S_{19}	グラフをx軸方向に平行移動させるためにはどうすればよろしいでしょうか？
S_{20}	グラフを縦に平行移動させたいのですがどの値を変化させればいいのですか
S_{21}	グラフを縦に動かしたいのですが、どうすればいいですか？
S_{22}	グラフをy軸方向に平行移動させるためにはどうすればよろしいでしょうか？
S_{23}	グラフを横向きに回転させたいのですが値をどのように変化させればいいですか
S_{24}	グラフを90°回転させる方法はありますか？
S_{25}	グラフを下に開くように描きたいのですがどうすればいいですか？
S_{26}	aはグラフの開き方を変えるだけですか？
S_{27}	グラフの開き方とaとの関係を教えてください

これらの質問文から

- 未知語
- 名詞-サ変接続
- 名詞-一般
- 名詞-代名詞-一般
- 名詞-接尾-一般
- 名詞-接尾-助数詞
- 名詞-接尾-特殊
- 名詞-数
- 名詞-形容動詞語幹
- 動詞-自立

の10品詞を内容語として抜き出し、学習者の質問がどの想定質問に該当するかを決定する。

3 重み付けと類似度

3.1 質問の数値ベクトル化

想定質問と学習者質問の類似度を計算するために質問文を数値ベクトル化する。例えば、T_1の「aの値を変化させるとグラフはどのように変化しますか」の場合、形態素解析器により品詞分解した後、不要語を除去し、$WT_1=\{\text{a, 値, 変化, する, グラフ}\}$を得る。$T_2$についても同様に、$WT_2=\{\text{b, 値, 変化, する, グラフ}\}$が得られる。このようにして得られた想定質問語集合を$WT_i(i=1,2,3,\cdots,n)$とする。学習者質問も同様に、学習者質問語集合を$WS_j(j=1,2,3,\cdots,N)$とする。この想定質問語集合と学習者質問語集合の全体集合をWとする。

$$W=\left\{\bigcup_{i=1}^{n} WT_i\right\}\bigcup\left\{\bigcup_{j=1}^{N} WS_i\right\}$$

また、質問語の総数を$|W|$で表し、$m=|W|$とする。
このWの語$w_i\in W$が想定質問T_jに含まれる数をf_{ij}とし、次のようなl_{ij}を定義する。

$$l_{ij}=\begin{cases}1 & \left(f_{ij}>0\right)\\ 0 & \left(f_{ij}=0\right)\end{cases} \tag{1}$$

このl_{ij}を用いて、想定質問T_iの質問ベクトルを

$$\boldsymbol{t}_i=\left(l_{1i}, l_{2i}, l_{3i}, \cdots,, l_{mi}\right)$$

Table 4: 記号の説明

w_i	質問の全体集合Wに含まれる質問語
m	質問集合全体にわたる質問語の総数
n	想定質問数
f_{ij}	想定質問WT_jに含まれる質問語w_iの個数
F_i	想定質問全体をに含まれる質問語w_iの個数
n_i	質問語w_iを含む想定質問数

と定義する。

一方、学習者質問のベクトル化も同様に、Wの語$w_i \in W$が学習者質問S_jに含まれる数をf'_{ij}とし、l'_{ij}を次のように定義する。

$$l'_{ij} = \begin{cases} 1 & \left(f'_{ij} > 0\right) \\ 0 & \left(f'_{ij} = 0\right) \end{cases} \tag{2}$$

このl'_{ij}を用いて、学習者質問S_iの質問ベクトルを

$$\boldsymbol{s}_i = \left(l'_{1i}, l'_{2i}, l'_{3i}, \cdots,, l'_{mi}\right)$$

と定義する。

3.2 TF-IDF

前節では、l_{ij}を2値化してベクトル化した。これを2進重みという。このような重み付けは、局所的重み付けといわれ、質問文中に頻繁に出現する質問語に対して大きな値が与えられる。つまり、質問文を特徴づけている語に対して大きな値が与えられる。

また、質問集合全体Wにわたる質問語w_iの分布を考慮して決定される重み（g_iとおく）を大域的重みという。特定の想定質問に集中して現れる質問語に対して大きな値が与えられる。

このl_{ij}によって想定質問の特徴となる語が得られ、g_iによって想定質問内の一般的な語の影響を抑えることが可能となる。

3.2.1 TF(Term Frequency)

局所的重みについては、2進重み以外にもさまざまなものが提案されている。ここでは、索引語頻度（TF）と対数化索引語頻度（Logarithmic TF）について説明する。

2進重みでは、質問に含まれる語数の多い語と少ない語の差別化が不可能である。その問題を解消するため、索引語頻度（TF）は、質問内の質問語の個数を用いる。

$$l_{ij} = f_{ij} \tag{3}$$

しかし、質問語の個数が多い場合、その語に対して過大な重みを与える傾向がある。このような個数の多い語の影響力を押さえるため、対数を用いて、

$$l_{ij} = \log\left(1 + f_{ij}\right) \tag{4}$$

とする。これを対数化索引語頻度という。

3.2.2 IDF(Inverse Document Frequency)

質問全体にわたる質問語w_iの個数の偏りを考慮した重みg_iについて説明する。よく知られている文書頻度の逆数（IDF）は、

$$g_i = \frac{n}{n_i} \tag{5}$$

で与えられる。これは、多くの想定質問に含まれている語は、小さな値となり、特定の想定質問に含まれている語は大きな値となる。対数化しているのは、IDFの値の変化を小さくするためである。

また、確率的IDFは、

$$g_i = \frac{n - n_i}{n_i} \tag{6}$$

で与えられ、想定質問の半数以上に含まれる語は負の値となる。半数以上の想定質問に含まれる語の判定に用いることも可能である。

想定質問文に含まれる語数を用いる重み付け以外に、情報理論におけるエントロピーを用いた大域的重み付けがある。事象$E_1, E_2, \cdots, E_n$の起きる確率をそれぞれ$p_1, p_2, \cdots, p_n$とすると、1事象あたりのエントロピーHは次の式で定義される。

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{7}$$

エントロピーHの取り得る範囲は$0 \le H \le \log n$であり、各事象E_iが等確率で起こるとき最大値$\log n$をとる。また、各事象の起こる確率に偏りがあるに従いエントロピーは小さくなる。エントロピーが最小となるのは、1つの事象E_iの確率が1で、他の事象$E_j (i \neq j)$の確率がすべて0となるときである。このとき、エントロピーの最小値は0となる。そこで、質問語w_iに対してw_iが想定質問WT_jに含まれる事象E_jの確率は、$\frac{f_{ij}}{F_i}$であるからこのときのエントロピーは、

$$H_i = -\sum_{i=1}^{n} \frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i} \tag{8}$$

となる。ここで、$\frac{H_i}{\log n}$を考えると、$0 \le \frac{H_i}{\log n} \le 1$である。この値は、質問語が各想定質問に含まれる個数が近いほど1に近づき、少数の想定質問にしか含まれない場合には0に近づく。

よって、質問語w_iの大域的重み付けは、

$$g_i = 1 - \frac{1}{\log n} H_i = 1 + \frac{1}{\log n} \sum_{i=1}^{n} \frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i} \tag{9}$$

で与えられる。

3.3 類似度

入力された学習者の質問がどの想定質問に該当するかを判断するために、類似度を以下のように定める。学習者の質問ベクトルを

$$\boldsymbol{s}_i = \left(l'_{1i}, l'_{2i}, l'_{3i}, \cdots,, l'_{mi}\right)$$

とし、想定質問に重み付けをしたベクトルを

$$\boldsymbol{q}_j = \left(l_{1j}g_1, l_{2j}g_2, l_{3j}g_3, \cdots,, l_{mj}g_m\right)$$

とする。この$\boldsymbol{s}_i, \boldsymbol{q}_j$の単位ベクトルの内積を考える。2つのベクトルのなす角をθとすると、

$$\cos\theta = \frac{\boldsymbol{s}_i}{\|\boldsymbol{s}_i\|} \cdot \frac{\boldsymbol{q}_j}{\|\boldsymbol{q}_j\|} \tag{10}$$

と表すことができる。

$-1 \le \cos\theta \le 1$であり、$\cos\theta$の値が1に近い（θの値が小さい）ほど、2つの質問は似ていると判断する。

4 自動応答とその精度

教員があらかじめ用意した想定質問（Table 2）と学習者の質問（Table 3）を用いて、教員の意図する応答ができるかどうかを実験した。質問の判定には類似度が最大のものを選択する方法とり、すべての想定質問に対する類似度が0.1未満のものは、判定不能とした。

実験の結果、局所重み付けについては、索引語頻度、または、対数化索引語頻度を用いるのが適当であることがわかる（Table 5、Table 6）。対数を使う場合と使わない場合の差が出なかったのは、質問が単文であるため、質問語の出現頻度が少なかったからと考えられる。

また、大域的重み付けについては、確率的IDFと情報エントロピーを用いた場合の正答率に差はみられなかった。

しかし、誤判定した質問に関する類似度が低いものを調べると、確率的IDFでは類似度が0.1未満のものが2質問含まれていることがわかった。類似度の定義からも0.1未満は自動応答の候補として判断することは適切ではない。

Table 5: 正答率

平均類似度	2進重み	索引語頻度	対数化索引語頻度
確率的IDF	88.5	88.5	88.5
情報エントロピー	80.8	88.5	88.5

Table 6: 平均類似度

平均類似度	2進重み	索引語頻度	対数化索引語頻度
確率的IDF	0.43	0.45	0.45
情報エントロピー	0.64	0.66	0.66

Table 7: 正答数と判定不能数

確率的 IDF	正答 最大値 と一致	誤答 類似度 0.1以上	判定不能 類似度 0.1未満
索引語頻度	23	1	2
対数化索引語頻度	23	1	2

エントロピー	正答 最大値 と一致	誤答 類似度 0.1以上	判定不能 類似度 0.1未満
索引語頻度	23	3	0
対数化索引語頻度	23	3	0

そこで、大域的重み付けでは、情報エントロピーを用いた重み付けが適当であると考えられる。

5 今後の課題

局所重み付けとして、索引語頻度あるいは対数化索引語頻度を用い、大域的重み付けとして、情報エントロピーを用いることで、ある程度正確に応答することが可能であることがわかった。今後は、学習者の質問とそれに対する応答をシステム自体が学習し、より正確な応答が可能なシステムにする必要がある。

謝辞

本研究の一部は日本文部科学省オープン・リサーチ・センター整備事業（平成16年～平成20年）による私学助成を得て行われた。

参考文献

[1] 高志修, 富永浩之, 林敏浩, 山崎俊範. 対話的な授業支援のための一問一答クイズAQuAs. 信学技報 ET2004-79, 2004, pp. 37-42.

[2] 風間淳一, 光石豊, 牧野貴樹, 鳥澤健太郎, 松田晃一, 辻井潤一. チャットのための日本語形態素解析. 言語処理学会第5回年次大会発表論文集, 1999, pp. 509-512.

[3] 堤豊, 牛島和夫; 電子メールを用いた日本語文による質問応答システムにおける類似質問の抽出について, 自然言語処理, 117-2,1997, pp. 161-166.

[4] 加藤尚吾,赤堀侃司;電子掲示板を用いたコミュニケーションにおける参加者の感情表出の容易性の分析, 教育情報研究, 第20巻第2号 VOL.20 NO.2, 2004, pp. 3-13.

[5] 松河秀哉, 中原淳, 西森年寿, 望月俊男, 山内祐平; 電子掲示板上での学習者の活動を把握する指標の検討, 日本教育工学会論文誌 28(1), 57-68, 2004, pp. 57-68.

Auto-answering to learners' questions in an e-Learning system

Kenji YOSHIDA and Hirotaka NAKAYAMA

Department of Information Science and Systems Engineering,
Faculty of Science and Engineering, Konan University
Okamoto 8-9-1, Higashinada, Kobe 658-8501, JAPAN

(Received April 12, 2005)

Abstract:

Many LMS (Learning Management System) aim to manage learning through internet, and have a function of portal site mainly. It is difficult for many students to continue learning by such an e-Learning system, because it provides lecture contents only. There is no communication between students and teachers. In order to overcome this difficulty, bulletin board systems can be utilized. In many practical cases, however, teachers can not answer students' questions immediately due to the time limitation. Learner's motivation deteriorates unless they can receive answers from teachers in real time.

This paper reports our trials to constitute an e-Learning system which can reply automatically learners' questions.