様式 C-19

科学研究費補助金研究成果報告書

平成　22年　5月　1日現在

研究種目：若手研究(B)
研究期間：2008 ～ 2009
課題番号：20700654
研究課題名（和文）　Web検索結果から学習問題文を自動生成するシステム

研究課題名（英文）　Automatic question generation using knowledge capture from Web

研究代表者
越智　徹（OCHI TORU）
広島国際大学・工学部・助教
研究者番号：10352048

研究成果の概要（和文）：
現在、様々なWBTシステムが開発・運用されているが、コンテンツの作成の費用やコストが大きいという問題がある。我々はWeb上に存在する様々な用語解説サイトに着目し、Web上の知識情報を抽出し、ネットワーク用語についての日本語によるWBTの問題を自動作成する試みを行った。
この研究によって構築されたシステムは、1）検索エンジンを用いてWeb上からある用語について記載されているWebページのデータを取得、2）そのデータへテンプレートを適応して用語の説明文とキーワードを抽出する、3）抽出されたキーワードから適切なものを選択する、4）選択されたキーワードから作問を行う、というものである。

研究成果の概要（英文）：
The advent of various WBT systems appear attractive to teachers, however, making learning material is still a time consuming job. This paper presents a method to automatically generate questions for studying computer network management in Japanese. For this purpose, we have been developing an automatic generation system that extracts descriptive sentences from the Internet. The system consists four steps; acquisition, extraction, selection and question generation. As a result, our system generates questions in Japanese.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008年度	900,000	270,000	1170,000
2009年度	500,000	150,000	650,000
年度			
年度			
年度			
総　計	1,400,000	420,000	1,820,000

研究分野：総合領域
科研費の分科・細目：科学教育・教育工学
キーワード：知識獲得、e-Learning、自然言語処理

様式 C-19

科学研究費補助金研究成果報告書

1．研究開始当初の背景

近年、WebCT、Moodle 等の WBT(Web Based Training)システムが多く利用されている。インターネットを利用して自宅からでも学べる WBT は学習者にとって大変便利であるが、その反面、教師にとっては学習問題・コンテンツを作成するコストが大きい。

2．研究の目的

WBT のためのコンテンツ作成に関する問題を解決するため、近年様々な研究が試みられてきた。[1-2]

これらの研究では、数式問題を対象とし、人手によって作成ルールを定め、それに従って問題生成を行っている。本研究では、問題生成ルールの半自動生成も含めて、最終的に図 1 に示すような問題文を自動生成することをめざしている。

このために、Web 上に存在する様々な用語解説サイトに着目し、Web 上の知識情報を抽出し、ネットワーク管理の基礎知識を学ぶための教材として、ネットワーク用語について WBT 用の問題を自動生成する試みを行った。

> 問題：MAC アドレスはどのような番号ですか。次の選択肢から選びなさい。
>
> ・固有の番号
> ・特定の番号
> ・自由な番号
> ・10 進数の番号

図 1　目標結果の例

3．研究の方法

(1) 本研究では、図 2 に示すシステムを構築・使用して研究を行う。提案するシステムは、取得、抽出、選択、生成の 4 ステップから構成される。図 2 にシステム概要を示す。各ステップについて以降の節で述べる。

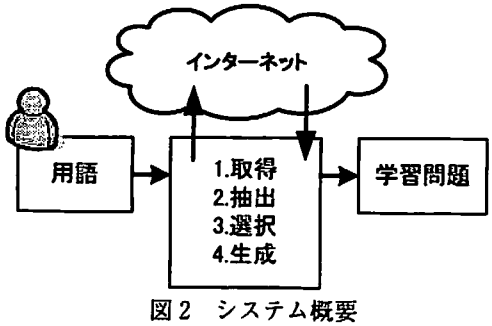

図 2　システム概要

(2) 取得

まずシステムは、問題作成者から与えられた用語 X についての説明文を Web から取得する。まず、システムは Yahoo! Japan の検索エンジンを使用し、用語 X について 500 件のヒット結果の URL を得る。さらに、その URL が示した Web ページをすべて取得し、テキスト部分だけを蓄積する。日本語に対応した検索エンジンは Yahoo!以外にも存在するが、Yahoo! Japan は、WebAPI を使用した検索エンジンの外部利用手順を公開しており[3]、この WebAPI を使用することで、検索結果の自動取得が可能なため、本研究に使用するには最適であると判断した。なお、WebAPI の利用により最大 1000 件のヒット結果を得ることが出来るが、1000 件目までとなると経験的に用語との乖離が大きいため、500 件までを取得している。

(3) 抽出

前節で取得したテキストデータに対し、テンプレートとキーワードを使用した抽出処理を行う。

①テンプレートマッチング

用語 X に対し、「X は」「X とは」「「X」」「【X】」の 4 種類のテンプレートを用意し、このテンプレートを含む文章を Perl による正規表現によって、取得ステップで蓄積したテキストデータから抽出した。

Fujii らは[4]、14 種類のテンプレートを作成して百科事典データから抽出しているのに対して[4]、我々は上記の 4 種類のテンプレートを用いて取得データから抽出した。また、このテンプレートのマッチした文に続く 2 つの文までをテンプレートマッチングの出力とした。2 文を付加したのは、これは経験的に良いと判断したためである設定した。図 3 に示した例では、太字文字がテンプレートにマッチした部分、波下線文字がさらに続いて抽出される 2 つの文を示している。

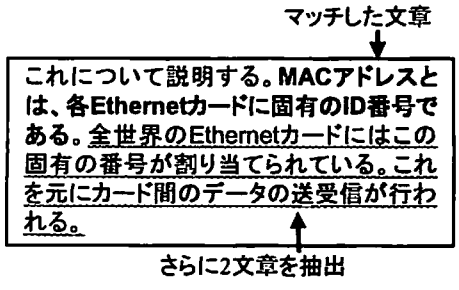

図 3　テンプレートマッチングの出力

②キーワードの抽出

テンプレートによって抽出された出力は、用語Xについての説明文とみなしてよい。この説明文から、用語Xを説明する重要なキーワードを抽出する。説明文には、「32ビット」や「16進数」、「ネットワーク機器」など、複合語がキーワードとして多く用いられているため、これらを単にMeCabで形態素単位で抽出すると、32、ビット、16、進数、ネットワーク、機器、と複数の形態素に分かれてしまうので、連続する名詞はつなげるように追加処理する。この結果、抽出されたキーワードを出現回数の多い順に出力する。

この抽出処理で出力されるキーワードは、1）用語そのものを表し用語と意味が近いとみなせるものと、2）用語の説明のために必要なものに二分される。例えば、「MACアドレス」という用語に対して、1）に該当するものは「物理アドレス」「番号」などで、2）に該当するものは「固有」「12桁」「16進数」などである。

(3)選択

ここでは、用語Xを説明するために、また問題生成のために必要かつ重要だと思われる文章群が、システムによって自動選択される。この処理では、抽出ステップで抽出された文章群のうち、ユーザに提供する上で重要と思われる文章を取り出すために、文章群の各キーワード出現回数と各文章のキーワード数を利用したスコアの計算の結果により値が高い文章、上位から100個を取り出す。このスコアは、文章群の各キーワードの出現回数と一文章に含まれるキーワード数とを加算平均、乗算平均した結果をスコアとする。

また、抽出された文章群は動的計画法を用いて類似されたものをグループとしてまとめる。

(4)生成

ここまでの処理によって、学習問題文が生成される。

4．研究成果

(1) 本研究では、MACアドレス、IPアドレスをはじめ、様々なネットワーク用語について一連のシステムを試行した。その結果の一部をここに示す。表1は「MACアドレス」で抽出されたキーワード例である。キーワード数は非常に膨大になるため、少なくとも2以上カウントされたキーワードのみをシステムは保存するようにしたが、それでも500以上キーワードが抽出された。表1は、上位10個を抜粋したものである。また、表2はIPアドレスについて同様に上位10個を抜粋したものである。

表1 「MACアドレス」のキーワード例

管理	18
NIC	21
サービス	21
機器	21
変更	23
ルータ	24
番号	33
確認	33
接続	34
IPアドレス	61

表2 「IPアドレス」のキーワード例

32ビット	35
IPv4	35
番号	42
アクセス	46
ホスト	46
固定IPアドレス	50
数字	56
取得	56
ネットワーク	103
インターネット	126

(2) 選択処理においてスコア計算の上抽出された文章群の一部を以下に示す。

「MACアドレス」の例：

- MACアドレスとは、Ethernetネットワークカード自身につけられた48ビットの番号で、カードごとに固有の番号がつけられています。
- MACアドレスとは、ネットワーク上で各ノードを識別するために設定される、Ethernetカード毎に固有で割り当てられる固有のID番号のことである。
- MACアドレスとIPアドレスの相互変換には、ARPやRARPというプロトコルを用いる。
- Ethernetカードには1枚1枚固有の番号が割り当てられており、これを元にカード間のデータの送受信が行われている。

「IPアドレス」の例

- IPアドレスは、インターネット上で利用できる固有の「グローバルアドレス」と、LAN内で利用できる「プライベートアドレス」がある。
- IPアドレスは32ビットで表現されているので，コンピュータが世界で約40億台接続できることになる。
- IPアドレスは、ネットワークを識別する

ためのネットワーク部とネットワーク内のコンピュータを識別するためのホスト部から構成されています。

- インターネット上の機器を識別するための番号で、インターネットに接続中の機器それぞれに異なるIPアドレスが割り振られています。

(3) 以上の結果を用いて、システムは学習問題文の作成を行う。「MACアドレス」および「IPアドレス」について例を示す。

「MACアドレス」
問題文：Ethernet[2]には1枚1枚[1]の[4]が割り当てられており、これを元に[2]間の[3]の送受信が行われている。

選択肢：番号、固有、カード、データ、ARP、パケット

「IPアドレス」
問題文：[2]上の機器を[3]するための[1]で、[2]に [4]中の機器それぞれに異なるIPアドレスが割り振られています。

選択肢：番号、インターネット、識別、接続、LAN、通信

(4) 本研究では、第2節の図1で示したような完全な学習問題文と選択肢の提示までは至らなかったが、自動作成という点において、処理手法を確立させた。基本的にキーワード抽出、説明文抽出処理においてシンタックス手法によって処理しているが、自然言語を処理にするには限界がある。そのため、一部ヒューリスティック手法を取り入れた。

引用参考文献

[1] 金西計英, 林賢太郎, 光原弘幸, 矢野米雄: 「教材知識に基づきWBT上で演習問題を生成する機能の実現」, 教育システム情報学会誌, Vol.20, No.2, pp.71-82, 2003

[2] T.Kojiri, S.Hosono, T.Watanabe: "Automatic Generation of Answer for Complex Mathematical Exercise", ICCE 2006 Workshop Proceedings of Problem-Authoring, pp.25-32,2006

[3] http://developer.yahoo.co.jp/

[4] A.Fujii and T.Ishikawa, "Organizing Encyclopedic Knowledge based on the Web and its Application to Question Answering", Proc. of ACL-EACL 2001, pp.196-203, 2001.

5．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計1件）

① "Automatic question generation using knowledge capture from Web", Toru Ochi, Michio Nakanishi, E-Learn 2009 Proceedings., pp.3589-3592, 2009年（査読有）

〔学会発表〕（計2件）

① "Automatic question generation using knowledge capture from Web", Toru Ochi, Michio Nakanishi, E-LEARN 2009 - World Conference on E-Learning in Corporate, 2009年10月29日, Sheraton Vancouver Wall Centre Hotel, Canada.

② 「Webからの知識獲得による学習問題生成の試み」, 越智徹, 中西通雄, 電子情報通信学会Webインテリジェンスとインタラクション研究会, 2008年12月12日, 神奈川近代文学館

〔図書〕（計0件）
〔産業財産権〕
○出願状況（計0件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計0件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

6．研究組織

(1)研究代表者

越智 徹　　(OCHI TORU)
広島国際大学・工学部情報通信学科・助教
研究者番号：10352048

(2)研究分担者

なし

(3)連携研究者

なし