

Web上のアニメ関連ページを使ったクイズ生成について

小出恭平 廣川佐千男

日本への留学生の多くは、来日前にも日本語検定のための学習をしている。しかし、授業で学ぶより多くの情報が Web にあり、漫画やアニメなどの情報から日本への興味を持つ人も多い。通常の授業では困難な日常的な話し言葉の学習を、アニメ関連 Web 文書を用いたクイズとして提示することで、日本語の学習を支援する方法を検討した。

Quiz Generation based on Animation Related Web Pages

Kyohei Koide, Sachio Hirokawa

Many of foreign students to Japan are learning Japanese before they really come to Japan. Many interesting information, including comics and anime, are available on the Web. They are different type of material that cannot covered by lesson at schools. The present paper propose a system that generates quizzes by which students can learn everyday talk languages.

1. はじめに

近年、世界のグローバル化によって、国を超えた人の出入りが多くなった。その中で、日本に来る留学生は日本学生支援機構によると、約14万人であり、多数の留学生が在籍している。彼らの大多数は日本語学校などで日本語を勉強してきた。ところが、それだけでは日常会話などで使われる日本独特の表現を全て学んだとは限らない。

日本の漫画・アニメは世界でも人気がある。その漫画・アニメ中には、日常的な会話が多いため、漫画・アニメを利用して、日本語を勉強している外国人が多く存在する。ICT 技術や Web の技術の発展により、漫画・アニメに関する Web ページが多く出てきている。特に、google や yahoo のように Web ページの検索が多くあり、さまざまなサイトで使用されている。しかし、膨大な情報量が自分に合う漫画・アニメの検索を困難にしている。

そこで、本研究では留学生の日本語学習を支援する為に、新しい学習手段として漫画・アニメに関する Web ページを集め、検索エンジンを構築し、クイズ形式によって日本語の勉強に役立つシステムを作成した。

外国人が学ばなければならない日常的な会話表現は多くある。たとえば、「あんた」や「お前」のように、英語では「you」で示される二人称の多様性や「～わ」や「～よ」などの接尾語などがある。今回は学習の素材として日本のマンガ・アニメに関する Web ページを使用する。選択した理由として、日本のマンガ・アニメは世界的にも高い評価を受け、多くの言語に翻訳され、出版されていることがある。これらの内容は主に会話で成り立っているため、教科書や辞書では学ぶことのできない表現が多く、日常的な話し言葉による日本語の学習に役立つと考えられる。インターネットの普及により膨大な Web ページの中には、アニメや漫画に関するページも含まれている。それらのページの中には、アニメを見ている人達のブログも含まれているので、そこからページ内の文章を抜き出すことで、日本語特有の表現を得ることが期待できる。

学習者の能動的な学習には、高い学習効果がある。そこで一方的に内容を提示するのではなく、本研究では、クイズ形式とし、クイズに答えるゲームとすることで利用者は楽しむことができ、それにより動機づけが行え、学習効果が上がると考えた。

2. 関連研究

クイズの生成についての関連研究として[1,2,3]がある。本研究ではアニメに対する検索エンジンを作ることで、アニメに関するクイズを生成し、日本語を学ぶ人がこれを回答することで、日本語の勉強に役立つと考えられる。ここではクイズ生成に関する関連研究を挙げる。

2.1 XMLドキュメントからの練習問題自動生成システム[1]

教育機関へのコンピューターやLANが導入されたことによって、これらに対応した教育方法が必要とされている。学生は、使用する教材に沿った練習問題をブラウザ上で回答し、練習問題を解くことで理解を深めることが可能となる。学生の理解度に対応した問題を出題し、これを学生が解答することで、さらに理解度を深めることが可能となる。しかし、学生の理解度によって、難易度別に練習問題を作成し出題するのは、教師に対しての負担が大きくなる。

この論文は、XML文書から練習問題を生成するシステムAEGIS(Automatic Exercise Generator based on the Intelligence of Student)について述べている。このシステムは、学生の理解度に応じて、問題の自動生成を行い、解答に対して自動採点を行う。生成する問題は選択形式、穴埋め形式、誤り訂正形式の出題が可能である。

出題するに当たり、出題箇所・出題文・選択肢の情報が必要となる。これらを自動で抽出するのは困難である。そのため、出題者は予め教材に問題作成のための情報をタグとして埋め込み、これを元に、問題作成を行っている。この論文は出題方法に出題者の意向が関わり、事前に設定を行うが、本論文では出題方法・選択肢を全て自動で行った。

学習者の理解度の定義として、回答した練習問題の内、正解した問題の難易度を合計し、回答した練習問題数で割った値となる。正解数が多いほど、理解度の値が大きくなり、少ないほど値は0に近くなる。

現在AEGISを使用する為に必要なXMLタグは、出題者によって手動での入力となっており、作業に大きな負荷がかかる。また、このシステムは曖昧さがないことが前提となっているので曖昧さがある場合の対応が課題となっている。

2.2 理解状況に適応した多肢選択式問題の自動生成に関する構想[2]

近年、資格取得のニーズが高まり、検定試験の対策講座などが多くなった。資格試験や検定試験は迅速かつ主観的な採点の少ない多肢選択肢が最も多く使われている形式となっている。一般に、検定試験・資格試験の対策学習として過去に試験として出題された問題を用いての学習が行われる。実際の試験はこれと同じか類似した問題が多く出題されるので、過去の問題を解くことは、学習として有効であると考えられる。

また、短時間で効果的な]対策学習を行う方法の一例として、学習者の不得意分野の繰り返し学習などの個別の理解度に対しての問題演習が必要となる。しかし必要となる問題を事前に用意し、学習者の理解度の状況に応じて適切な問題を提供することは困難である。そこで学習者の理解度によって選択問題の自動生成を行っている。

選択肢となる単語は事前に知識ベースに格納しこれを使用しての語彙に関する万大演習システムの構築を行っている。問題の生成方法として、解答となる単語のデータの情報を取り出し、テンプレートに埋め込むことで問題の生成を行っている。

選択肢の生成方法として、単語間の関係を表す語彙空間を元にして、解答となる単語との概念距離によって、難易度の調節をしている。語彙空間の概念図を図3に示す。

学習者の理解度の設定方法は、0~1の数字とし、全ての単語に語彙についての理解度を与えている。解答者が正解の場合は、正解となった単語のみ、不正解の場合は、選択肢となった単語全てに理解度の計算を行い設定している。具体的な計算式は、今後の課題となっている。この論文では単語を樹形図的に概念の関連度によって選択肢の設定を行っている。本研究ではGETAIによって得られた単語間との類似度によって選択肢の生成を行う。

2.3 4択クイズを連想問題として解く[3]

自然言語で与えられた質問の答えを探す技術は質問応答と呼ばれ、近年注目されている。このシステムは、大きく2つのプログラムに分けられる。一つは解答の候補を集めるプログラムで、もうひとつは最適な解答の選択である。この論文では後者の解答の選択に注目している。今までにも解答の選択方法として、問題文と解答候補の論

理的変換を行い、推論の操作によって2つが対応しているかを調べる手法がある。しかし、この手法は、語彙的知識と世界的知識という膨大な情報を必要とする。当論文では、これとは別の手法によって解答の選択を行っている。その手法は、問題文に含まれる重要な語句と、解答の候補となった単語との語彙的な関係性の強さを調べることによって、解答を選択する。

この手法は2つの手順によって構成される。まず、問題文から適切なキーワードの抽出を行う。次に、このキーワードと解答の候補との関係の強さを調べる。キーワードと解答の候補の関係の強さは、ウェブを知識源として、サーチエンジンのヒット数を利用したいくつかの尺度で測定し、それらの尺度によって連想の強さを測定する。この論文は、クイズ生成ではなく解答の選択を問題としているが、単語同士の関連度を使用しているので関連研究として紹介した。

3. Webデータの収集

本研究で作ったのは一般のWebページを使用しているクイズシステムである。そのためにGoogleを使用しているアニメ・漫画に関するWebページの収集が必要である。本章では、それらWebページの収集方法と検索のためのインデックス化の前処理について述べる。

3.1 Googleによる検索結果の保存

Googleにアニメタイトルを入力し、その検索結果上位1000件のurlを一覧に残す。

使用するアニメのタイトルを、検索語とする。次に、検索語をGoogleの検索結果に反映できるurlエンコードを行う。その後、エンコード処理を行った検索用urlをMozillaに送り、検索結果上位100件を表示したページを取得する。この処理を上位100件から順に10回、合計1000件の検索結果の保存を行う。次に得られた検索結果より、該当するWebページのリンクとなるurlの一覧を取得する。この処理をperlプログラムによって行った。得られたurlの一覧について、wgetを利用し、Webページを取得する。

今回、wgetで辿るリンクの階層は3回とし、urlエンコードされたページの取得も行った。この結果、19600件のhtmlファイルを得ることができた。

3.2 Webページのテキスト化

3.1節で保存したWebページはディレクトリ構造になっているので、データを保存したディレクトリ内を対象にし、再帰的にhtmlファイルを探す。htmlファイルを取得した場合、そのファイルをeuc-jpにエンコード後、txt形式で開き、タグ消しを行う。エンコードされなかったものはインデックス対象より除外した。タグ消し処理を行い、残った部分はWebページの本文であるので、これを新たにアニメ・漫画を追加可能となる形式である“(英語タイトル)-(ID).txt ”というファイル名で保存し、同時にIDと

htmlファイル場所を表す対応表を出力した。こうして得られた、14297件のテキストファイルに対して検索システムを構築した。

4. 検索エンジンの作成

本研究では、使用者によって入力された単語に関連する単語や文を使用して、クイズの自動生成を行うことを目的としている。そのため、本システムでは、汎用連想計算エンジン(GETA)を使用している検索システムの製作を行った。

また、GETAを使用するに当たり、検索対象となる文章の形態素解析を行わなければならない。本章では、これらの紹介と実行結果を述べる。

4.1 形態素解析

形態素解析を行うに当たり、MeCabという京都大学等で開発された形態素解析エンジンを使用する。MeCabは、辞書、コーパスに依存しない汎用的な設計であり条件付き確立場(CRF)に基づく高い解析精度を持つ。

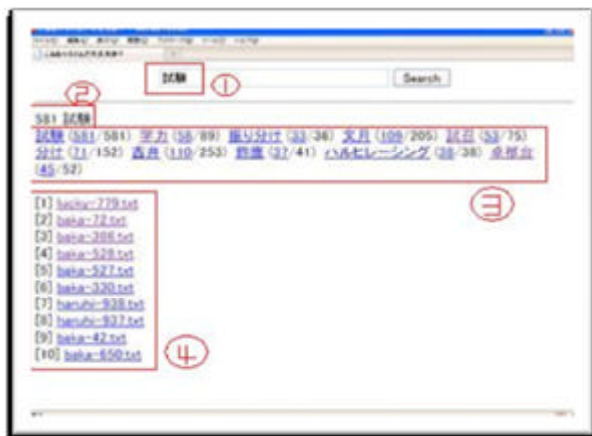
3.2節で生成したテキストファイルの一つずつ、Mecabを使用して各文に対して、形態素解析を行い、結果をfreqfileとして保存した。今回、MeCabには、名詞のみを、解析対象とした。解析を行った結果は付録に示す。

4.2 GETAによる検索システムの製作

汎用連想計算エンジンGETAは文書における頻度付き索引データを典型とする大規模かつ疎な行列を対象として文書間および単語間の類似度を内積型メジャーで高速計算するツールである。

4.1節で生成したfreqfileを使用して検索システムの製作を行った。図1に検索ワードを「試験」とした場合の結果を表示する。

図1. 「試験」と関連の深い単語・ファイルの検索結果



に検索したい言葉を入力する。

に検索した言葉が含まれているテキストファイルの数を表示している。

に検索した言葉と同時に出現した単語を多い順に 10 件表示する。単語をクリックすることで、その単語での検索結果が表示され、右にある数字をクリックすると、その単語との AND 検索の結果が表示される。

に検索した単語の出現回数が多い順に 10 件表示される。ファイル名をクリックすることで該当するページの本文が表示される。

5. クイズ生成システム

本研究では、4 章で述べた検索エンジンを元に、クイズのネタを入力し、クイズの文と正解・誤答を生成するシステムを作った。このシステムは、通常の実験エンジンと同様に、各文書についてのインデックスを持っている。まず、ユーザから与えられた検索条件にもとづき検索を行う。次に、検索結果の特徴語を抽出する。文書の重要度に応じてランキングし、抽出した特徴語を関連語として表示するようなシステムは他にもある。本手法の特徴は、文書単位のインデックスの他に、文単位のインデックスを利用するところにある。提案手法では、検索結果として得られた文書を構成するそれぞれの文について、指定された個数の特徴語がどのように出現しているかを文単位の DB により求め、それに基づき、文書の重要度を評価する。

5.1 クイズ自動生成システムの構築

このシステムは使用者によって入力された単語から正解となる単語、これと関連のある文章、そして不正解となる単語を出力する。正解となる単語が入力された単語と

一致することを避ける為、検索は 2 段階で行い、文章を出力する。

予め、収集した Web ページを問題文として、使用できる形式に成形する。3.2 節で生成したテキストファイルを開き、1 文ごとに区切り保存を行った。今回文末の判断として「。」「!」「?」を使用した。これらのいずれかを基準として、開いたテキストファイルを区切り、配列へ格納する。この配列を順に、ファイルに書き出し、保存を行う。

まず、入力された単語(以後、一次入力)を GETA に入力し、関連のある単語 10 個のリストを返す。そのリストより正解となる単語(以後、解答)を無作為に設定する。次に、得られた回答を GETA に入力し解答と関連のある単語、文章のリストを返す。

誤答となる単語のリスト(以後、誤答リスト)を生成する為に、解答と関連のある単語のリストから、ランダムに要素を 1 つ抽出する。その単語が一次入力、及び解答と重複していなければ、誤答リストに格納する。誤答リストに重複している要素があれば、その要素を 1 つにまとめる。以上の処理を繰り返し、誤答リストが予め指定した要素の数になれば処理を終了する。

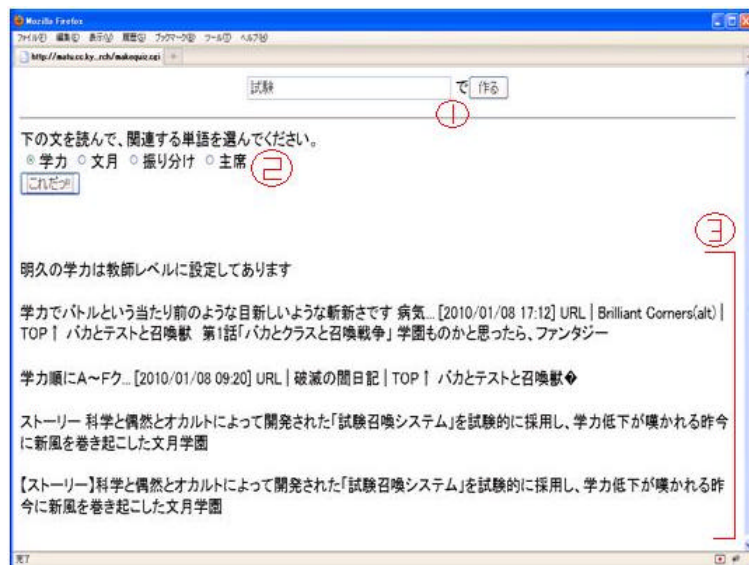
問題文となる文章(以後、問題リスト)の抽出を行う。解答を GETA に入力し、得られた解答と関連の高いテキストファイル 10 個のリストを受け取る。そのリストから無作為にテキストファイルを選択する。これを開き、1 行毎に配列へ格納し、そこから要素を 1 つ選択する。問題文に指向性を持たせるため、一次入力とこれによって GETA から得られる一次入力と関連のある単語上位 5 個のいずれかを含んでいる文を選び、問題文リストへ格納する。選択した文の重複を避ける為、格納した問題文の内、同一のものを 1 つにまとめる。以上の処理を、指定した問題文の数となるまで繰り返す。

このプログラムの結果をブラウザで見えるように HTML タグを付け、cgi として構築する。

検索語を入力する場所・解答選択し、答える場所の追加をする。解答の一覧を生成する時に、解答となる番号をタグとして保存し、使用者によって選択された番号がこれと一致した場合、正解したメッセージを表示する。一致しなければ生成した問題と同一の html 文と不正解であるメッセージを表示する。

図 2 に検索語を「試験」として自動生成を行った場合の動作を示す

図 2.「試験」に関するクイズの自動生成



に、単語を入力し、「作る」を押して、問題文・解答の生成を行う。

に解答の候補が表示され選択をする。正解の場合正解のメッセージが表示される。不正解の場合はその旨のメッセージと選択した番号を表示する。この場合、「学力」が解答となっている。

に問題文の生成を行った。

5.2 評価

形態素解析を行う際に、人名や熟語が認識されずに分けられている場合があるので形態素解析の方法も考慮する必要がある。

例として、声優の「加藤英美里」を検索語とする。入力を「加藤 英美里」とした場合、検索結果は0件となった。一方、検索語を「加藤 英美 里」の場合、検索結果は504件となった。これは、形態素解析の時点で名前の「英美里」の部分が「英美」と「里」に分けられたことが原因となっている。

そして、図3の第3文、第5文のように、問題文が重複している場合が見られた。この原因として、収集したWebページの中に特定のサイトから、文章をコピーした記事であると考えられた。そのため、使用するWebページを1つずつ取捨選択する必要があると考えられた。

図3.問題文の重複



文の終了の判定として「。」「！」「？」を使用した、タイトルの中などに「！」「？」が使用されているために、図3の第1文のように文の区切りが不適切であったので文の区切りとなる条件の見直しが必要となった。

6. おわりに

本研究では留学生の日本語学習を支援する為に、新しい学習手段として漫画・アニメに関するWebページを集め、検索エンジンを構築し、クイズ形式によって日本語の勉強に役立つシステムを作成した。その結果、単語抽出が粗いことや、提案手法の有効性の評価ができていないことなど課題はあるが、文章を読みそれらを理解してから解答するシステムが完成したので、本研究の目標は達成できた。

今回使用したデータは2011年11月時点のものであり、若干情報に新鮮さが無い。アニメは約3ヶ月毎に入れ替わりがおこり、それらの情報を対象としているのでこの傾向が顕著にみられた。本研究ではデータの収集、インデックス作成の自動化まではできなかったが、最新データへの対応は技術的には実現可能である。これができれば3ヶ月毎に新しいデータについてのクイズ生成も可能と考える。

謝辞

本論文の最後にあたり、日頃から熱心に御指導いただいた九州大学情報基盤研究開発センターの廣川佐千男先生に深く感謝いたします。また、同センターの伊東先生、

中藤先生および殷先生からは有益な助言を戴きました。日頃の研究活動において様々な助言をいただいた研究室の皆様に感謝します。

参考文献

- [1] 菅沼明, 峯恒憲, 正代隆義: XML ドキュメントからの練習問題自動生成システム; AEGIS: マルチメディア時代の新しい教育研究会, pp.24-31, 2001
- [2] 津森伸一, 海尻賢二: 理解状況に適応した多肢選択式問題の自動生成に関する構想: 教育システム情報学会研究報告, 21(4), pp.3-8, 2006
- [3] 外池昌嗣, 佐藤理史, 宇津呂武仁: 4 択クイズを連想問題として解く: 情報処理学会研究報告. 自然言語処理研究会報告 2004(47), pp.53-60, 2004