日本語読解支援のための語義毎の用例抽出機能について

小林朋幸, 大山浩美, 坂田浩亮, 谷口雄作, 太田ふみ, Noah Evans, 浅原正幸, 松本裕治
奈良先端科学技術大学院大学 情報科学研究科
{tomoy-ko,hiromi-o,kosuke-s,yuusaku-t,fumi-o,noah-e,masayu-a,matsu}@is.naist.jp

1 はじめに

大規模かつ種々の言語データが利用可能になり，言語研究に用いることが容易になってきた．大規模なコーパスは，言語研究に限らず，言語学習のための有用な資源として用いることができる．辞書には十分な用例が示されていない場合が多いため，言語学習者が新しい語について学ぶ場合，大規模なコーパスや Web 検索によって多くの用例を取得し，提示することができれば，語の用法や意味の理解に大いに役立つと考えられる．

我々の研究グループでは，言語教育のためのコーパスの有効利用を考えており，特に，単語 (あるいは語義毎) の例文の提示，頻出用法の検索，類義語の用法の差異，などを表示できる学習支援システムの構築を計画している．このシステムは，学習者だけでなく，教師が例文や用法のバリエーションを検索したり，準備するためのツールとしての利用も考えている．本稿では，特に，学習者にとってよい例文を選択する基準とその実装について述べる．

2 背景

日本語学習者のための日本語読解学習支援システムとしてよく知られたシステムに「リーディングチュウ太」[1]がある．このシステムは，利用者が入力した文を形態素解析し，各単語に対して読みと EDR 辞書の定義文，および，その英訳を表示する．同様の機能を提供するページに Rikai.com[2] があり，利用者が入力した日本語文章に含まれる単語の辞書項目をポップアップにして表示する．日本語学習者にとっては，知らない漢字を含んだり読みのわからない単語を辞書で引くのは極めて困難であるので，これらは大変有用なシステムと言える．ただ，両システムは，各単語の辞書エントリを表示するだけであり，それらの語釈文があまりに簡潔な場合 (多くの場合はそうであるが)，あるいは，複数の語義が表示される場合，正しい意味を理解することは容易ではない．

これらのシステムが提供していないが有用と考えられる機能として，語義の違いに関する情報，および，それぞれの用例の提供がある．本稿で提案する日本語学習支援システムでは，このような機能を提供することを考えている．

[1] http://language.tiu.ac.jp/
[2] http://www.rikai.com/perl/Home.pl

3 システムの基本機能

上記システムと同様，我々のシステムも，日本語の単語または文章を受け取り，形態素解析を行なって単語に分かち書きし，各単語の辞書エントリの表示を行なうが，さらに，用例の検索を行なう．各単語に対し，次の処理を行なう．

1. 単語を含む文のコーパスからの検索
2. 検索された文を単語の語義ごとに分類
3. 各文の「よさ」の評価値計算およびランク付け

多義語についての語義分類は，容易な問題ではないが，語義の自動分類については多くの研究が行われており [1]，別モジュールとして実装することを想定しており，ここでは省略する．本稿では，3 のよい用例の評価値計算について説明する．「よい」用例は，利用者によって異なると考えられる．例えば，初学者に長い文や難しい語を含む文を提示しても益々理解を損ねるだけである．

4 よい用例の選択基準

前節の最後で述べたように，「よい」用例は学習者によって基準が異なるはずである．ここでは，現在想定している種々の用例選択基準について述べるが，これらの基準のどれをどのように重要視して用いるかは，システムを利用する学習者あるいは教師によって調整できることを考えている．以下に示す選択基準には，各文に関する有用性 (あるいは難易度) の度合いを数値として返すもの，上限値や平均値などの条件を満足するか，または，どの程度違反するかを返すものなど

種々の性質の基準がある．システムのインタフェースでは，それぞれの選択基準に対するパラメータ値 (上限値や平均値などの数値)，選択基準をどの程度重視するかを示す重みなどを利用者が設定できる．

4.1 文の長さ

用例として利用者に提示すべき文はは，短すぎても長すぎてもよい文とは言えない．利用者が，自分がほしい文の平均文長 (単語数あるいは文節数)，上限，および，下限を指定することで，検索された文を順序つける．指定された平均文長に等しい文を最高値 (1) とし，それから離れるにつれて一定の比率で低い数値を出力する．上限および下限を逸脱する文に対しては，大きな負数を評価値として与える．

4.2 文中の単語の難易度

文の難易度を測るため，2 種類の評価値を考えている．一つは，コーパスにおける単語の出現頻度 (確率) に基づく評価値，もう一つは，日本語能力試験 (JLPT) の分類 (1 級語から 4 級語) [3] に応じた評価値を定義する．前者では，出現頻度の対数値を元に，最大値が１になるように正規化した数値を評価値とする．後者では，4 級語の評価値を１，1 級語の評価値を 0.5 として評価値を決定する．なお，JLPT 漢字表に現れない単語は，難解語として 0 級と想定する．(なお，固有名詞はこのクラスに分類されてしまうため，扱いを別途考える必要がある)

4.3 文中の漢字の難易度

文に含まれる単語を構成する漢字が，日本語能力試験のどの級に対応するか[4] を元に，単語の評価を行なう．文に含まれる漢字の平均の級，および，最も難しい漢字の級の値，全漢字数，文の全文字数などの数値を返す．

なお，日本語能力試験 (JLPT) の各級の単語および漢字の種類の数はほぼ次の通りである．

級	単語	漢字
1 級	10000 語	2000 字
2 級	6000 語	1000 字
3 級	1500 語	300 字
4 級	800 語	100 字

4.4 文の構造の難易度

文の構造上の難しさを測ることは容易ではないが，現在は，いくつかの擬似的な方法で文の難易度を測ることを想定している．一つは，n-gram 確率に基づく評価値であり，大規模なコーパスから計算した単語 n-gram(現在は tri-gram) によって文の単語あたりの確率を求め，それに応じた評価値を計算する．これは，単語や品詞の連接だけによる文の難易度尺度である．

もう一つは，文の構造的な難しさを測る評価値であり，日本語文の係り受け構造の複雑さ (あるいは難しさ) に基づいた尺度である．日本語係り受け解析システム南瓜 [2] は，2 つの文節間の係り受けの有無を Support Vector Machines(SVM) を利用して判断しており，各係り受け関係には SVM が出力した分離平面からの距離が数値によって与えられている．この数値が低いほど SVM にとって係り受けの有無の判断が難しかったことを示している．南瓜の係り受け解析は 100%正しいわけではないが，これらの数値により，個々の係り受け関係がどの程度容易なものであるかを推定することができる．この数値は上限，下限は理論的にはないので，この値をシグモイド関数によって 0 から 1 の間 (正数の場合は，0.5 から 1 の間) の数値に写像し，その平均値を文の統語構造の難易度として用いる．

4.5 総合的な文評価

以上の評価尺度は，様々な数値の組み合わせを返すので，それらを元に重み付き平均値や足切り値などを設定するインタフェースを準備する予定である．利用者が種々の設定を trial-and-error によって調整することにより，好みの基準で用例をランク付け，各自にとって「よい」用例を検索例文から選択することが可能なシステムを構築する予定である．

5 おわりに

我々が構築中の日本語学習者支援システムの用例選択基準とその実装法について概略を述べた．ここで述べた多くの機能は現在実装中である．これらの実装を完了し，実際に学習者あるいは教師に利用してもらうことにより，有用なシステムとして実現していきたい．

参考文献

[1]Eneko Agirre and Philip Edmonds: *Word Sense Disambiguation: Algorithms And Applications*, Text, Speech and Language Technology Vol.33, Springer, 2006.

[2]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.

[3] http://www.thbz.org/kanjimots/jlpt.php3

[4] http://www.coscom.co.jp/kanji/j-index.html