IDF を用いた単語レベル判定システムの構築と検証

Development and Evaluation of a Word Level Rating System
Based on Inverse Document Frequency

†北村達也・†冨岡洋介・‡川村よし子
KITAMURA Tatsuya・TOMIOKA Yosuke・KAWAMURA Yoshiko
†甲南大学・‡東京国際大学
†Konan University・‡Tokyo International University
†〒658-8501 神戸市東灘区岡本 8-9-1
†E-mail: t-kitamu@konan-u.ac.jp

Abstract : The aim of this study is to establish a reliable criterion for rating the level of Japanese words, and to develop an automatic word level rating system based on the criterion. We have already developed word level rating systems based on the coverage for the Japanese language proficiency test (Kawamura, 1999), the word familiarity, and the word frequency (Kawamura, 2008), and released them on our web page (http://language.tiu.ac.jp/) as parts of Reading Tutor. In the present study, the inverse document frequency (IDF) calculated from a newspaper database is employed and evaluated.

キーワード：単語難易度判定，IDF，単語親密度，単語出現頻度，日本語能力試験出題基準

１． はじめに

言語教育においては，学習者の学習段階に応じた語彙を含む教材を用いることが極めて重要である．しかし，この「学習段階に応じた語彙」をいかに決定するかは難しい問題である．本研究では，単語の重要度の指標の１つである IDF (inverse document frequency)を単語レベルの基準として用いることを提案する．IDF は，単語の出現頻度を出現度数ではなく，出現の偏りの有無によって判定する指標である．

これまで単語のレベル判定システムとしては，日本語能力試験の出題基準にもとづく語彙チェッカー (川村,1999)，単語親密度と単語出現頻度にもとづく親密度チェッカーと頻度チェッカー(川村，2008)がある．本研究では，新たに IDF を基準として用いて文中の単語レベルを自動判定するシステムの開発をするとともに，このシステムの検証を行った．

２． IDF を用いた単語レベル判定システム

2.1 IDF

IDF とは，文書集合において，少数の文書に偏って出現する単語に大きな重みを与える尺度である(尾内，2008)．$df(w_i,D)$ を文書集合 $D=(d_1,d_2,\ldots,d_N)$ において単語 w_i が出現する文書数とすると，その語の IDF は次式で与えられる．

$$idf(w_i,D)=\log\frac{N}{df(w_i,D)}$$

対数をとるのは N が大きく変化した場合でも値の変化を少なくするためである．単語 w_i が文書集合 D の一部の文書にしか現れない場合には $idf(w_i,D)$ は大きくなり，それが多くの文書に現れる場合には $idf(w_i,D)$ は小さくなる．

言うまでもなく，文書集合が異なれば同じ単語の IDF であってもその値は異なる．従って，利用目的に応じた文書集合から IDF を求めることが重要である．

2.2 単語レベル判定システム

本研究では，IDF の値が小さい単語ほど難易度が低い単語であると仮定し，入力文章中の単語のレベルを判定するシステム(以下「IDF チェッカー」)を開発した．

IDF は「CD-毎日新聞 2006 データ集」から求めた．まず，形態素解析システム MeCab(IPA 辞書使用)(工藤，2006)により上記データベースの全ての文章を形態素解析した．そして，各単語に関してそれが含まれる記事数を計算し，IDF を求めた．本研究では 95,762 文書から 128,592 語の IDF を求めた．

IDF と単語レベルとの関係は，表 1 のように設定した．このレベル分けにおいては，頻度チェッカー(川村，2008)における各レベルに含まれる単語数と一致させている．なお，IDF が 11 以上のものは「その他(レベル 0)」として分類した．ただし，このレベル分けは暫定的なものであり，必要に応じて今後変更していく．

IDF チェッカーの仕組みは，語彙チェッカーや他の川村(2008)の難易度判定システムと同一である．すなわち，ユーザーはシステムの web ページに web ブラウザーでアクセスし，分析し

表 1: IDF と単語レベルの関係

単語レベル	IDF	累計単語数
5(初級)	4 未満	1,053 語
4	4 以上 6 未満	6,110 語
3	6 以上 7 未満	12,097 語
2	7 以上 7.5 未満	16,406 語
1	7.5 以上 8 未満	21,496 語

たい文章を入力する(http://basil.is.konan-u.ac.jp/chuta2/). 入力された文章はサーバーに送られ，まず MeCab により形態素解析される．次にサーバーに用意されている IDF のリストと照合して，IDF の値に応じて単語レベルを決定する．最後に，単語レベルに応じて単語を色分けした入力文章，各単語レベルに含まれる単語数(総数，異なり語数)，各単語レベルの単語リストを表示させる．

３． IDF チェッカーの運用実験

IDF チェッカーが日本語教育において有用なツールとなりうるかどうかを検証するため，新聞データおよび日本語教育用教材を用いた運用実験を行った．

3.1 データ

新聞データとしては，2009 年 1 月 13 日から 19 日(1 週間分)の朝日新聞，読売新聞，日本経済新聞の 3 紙の 1 面トップ記事各 3 編(以下「新聞記事」計 32,375 語)と同 3 紙の社説各 2 編(以下「社説」計 21,102 語)を使用した．

日本語教材としては，『チュウ太の読解教材バンク』に収められた教材のうち，中級・上級教材として「日本を読む」の全 23 課(以下「中上級教材」計 20,237 語)および初級教材として「日常生活に見る日本の文化」の全 49 課(以下「初級教材」計 27,007 語)を用いた．

3.2 方法

それぞれの文章を IDF チェッカーおよび頻度チェッカーに入力し，IDF 順および頻度順に並べた単語リストで各々の上位 12,000 語(以下「IDF リスト」「頻度リスト」)のカバー率を調査した．さらに，比較のために，日本語能力試験の 1 級から 4 級までの出題基準の単語リストの語(以下「出題基準」)のカバー率も同様に調査した．

3.3 結果と考察

調査の結果を表 2 に示す．IDF リストのカバー率は，新聞記事，社説，中上級教材のいずれにおいても他と比べて高く，IDF リストで 95 %以上の語彙がカバーできることがわかった．

表 2: 各種リストのカバー率(%)

	新聞記事	社説	中上級教材	初級教材
IDF	95.9	96.5	95.2	94.5
頻度	93.2	94.1	92.6	90.5
出題基準	82.6	87.4	93.4	95.7

ちなみに個々の記事および社説を調査した場合も同様の結果が得られた．ただし，初級教材では出題基準のカバー率のほうが高かった．これは，現行の IDF チェッカーが新聞データという書き言葉から求めた IDF を用いていることによると考えられる．

さらに，出題基準を用いて IDF チェッカーのレベル判定精度を検証したところ，初級(4 級)語彙の 23.8 %の語がレベル 2 以下と判定されることが明らかになった．例えば，「涼しい」「たぶん(多分)」「水曜」「眼鏡」「コップ」などがレベル 2，「辛い」「どなた」がレベル 1，「おとどい」「あさって」「そちら」「あちら」等がレベル 0 に判定されてしまう．こうした問題については，話し言葉のコーパス等を利用して IDF を求めることによって解決できるはずである．

４． おわりに

本研究では，単語の重要度の指標の 1 つである IDF を用いて単語レベルを判定するシステムを構築した．本システムでは，IDF を新聞記事データベースから求めた．IDF の値は元となった文書集合に依存するため，本システムは新聞記事の単語レベル判定に最も適している．初級レベルの教育で使われる平易な文章や話し言葉の分析のためには，話し言葉のコーパス等他の文書集合から求めた IDF を用いてシステムを開発する必要がある．

謝辞 本研究の一部は文部科学省 ORC 整備事業(平成 16-20 年度)による助成および東京国際大学平成 20 年度特別研究助成を得て行われた．IDF に関してご助言いただいた甲南大学の永田亮先生に感謝いたします．

参考文献

川村よし子(1999)「語彙チェッカーを用いた読解テキストの分析」『講座日本語教育』34, 1-22.

川村よし子(2008)「単語親密度と頻度情報を活用した難易度判定システムの開発」『ヨーロッパ日本語教育』13.

尾内理紀夫(2008)『マルチメディアコンピューティング』コロナ社.

工藤拓(2006)「MeCab: Yet Another Part-of-Speech and Morphological Analyzer」http://mecab.sourceforge.net/.