逆文献頻度と文字の難易度に基づく単語レベル判定システムの構築

北村達也（甲南大学知能情報学部，t-kitamu@konan-u.ac.jp）
川村よし子（東京国際大学言語コミュニケーション学部，kawamura@tiu.ac.jp）

研究の背景と目的

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重要
→ いかに決めるべきか？

我々が過去に用いた単語レベル判定基準：
- 日本語能力試験の級別の出題基準（川村, 1999）
- 単語親密度（川村, 2008）
- 単語出現頻度（川村, 2008）

→ よりよい判定基準はないか？
- 客観的な
- 時代の変化を反映可能な

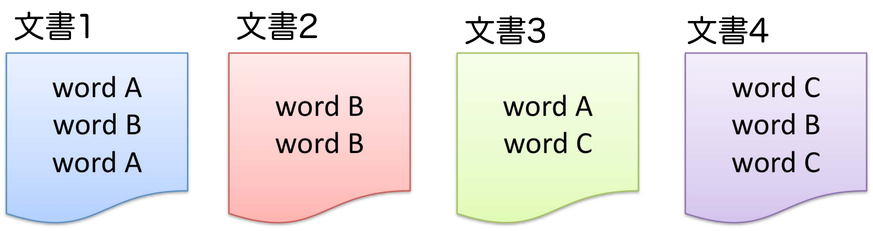
→ 本発表は2つの基準について考察
- 逆文献頻度（単語の重要度を評価）
- 単語の総画数（文字面の難易度を評価）

逆文献頻度（IDF）とは

Inversed Document Frequency: IDF (Jones, 1972)

単語の重要度に関する指標の1つ．文書集合において当該単語を含む文書の数に対応する指標

具体例



単語Aは2つの文書に現れるのでIDFは4/2
単語Bは3つの文書に現れるのでIDFは4/3
単語Cは2つの文書に現れるのでIDFは4/2　（実際には対数をとる）

- 多数の文書に現れる単語 → IDF小
- 少数の文書にのみ現れる単語 → IDF大
- 1つの文書に当該単語が複数回現れてもIDFには影響しない．

重要：同じ単語であってもIDFを計算する文書集合が異なればIDFの値が異なる！

式による表現

$df(w_i, D)$ を文書集合 $D=(d_1, d_2, \ldots, d_N)$ において単語 w_i が現れる文書数とすると，単語 w_i のIDFは次の式で与えられる

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

単語の総画数による難易度評価の試み

仮定
1. 難しい漢字は画数が多い
2. 単語の文字面のレベルが単語に含まれる文字数と難易度で規定できる

→ 単語中の文字の総画数が多い単語ほどレベルが高い

例
科学技術振興機構 > 文書集合 > 語彙 > 日本 > 式

平仮名・片仮名の取り扱い
0もしくは1に設定にしてはどうか？

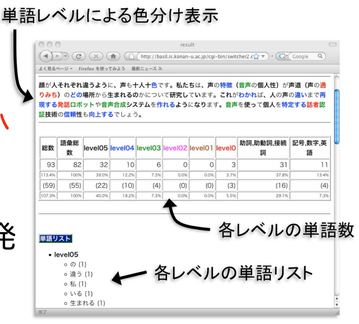
考察
- 単語のレベルとある程度対応しているのではないか？
- 単語の重要度等を全く考慮していないため，この基準だけで単語レベルを判定するのは困難．他の基準と組み合わせられないか？

http://basil.is.konan-u.ac.jp/chuta2/にて公開中!!

IDFチェッカー（IDFにもとづく単語レベル判定システム）

以下のように考える．
- 多数の文書に現れる単語 → IDF小 → 重要度大
- 少数の文書にのみ現れる単語 → IDF大 → 重要度小

IDFは「CD-毎日新聞2006データ集」から計算
入力文章中の単語のレベルを判定するシステムを開発



図：IDFチェッカーの出力画面．文章入力と結果出力はwebブラウザで行う

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

運用実験（北村, 冨岡, 川村, 2009）

データ
- 新聞記事：09年1月13～19日(1週間分)の朝日，読売，日経の1面トップ記事各3編（32,375語）
- 社説：同3紙の社説（21,102語）
- 中上級教材：チュウ太の読解教材バンクの「日本を読む」23課（20,237語）
- 初級教材：チュウ太の読解教材バンクの「日常生活に見る日本の文化 」49課（27,007語）

方法
- 単語出現頻度にもとづくシステムとの比較（出題基準でも同様の調査）
- IDF順および出現頻度順に並べた単語リストで各々の上位12,000語のカバー率を調査

結果

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	83.3	85.9	93.9	94.4

← IDFのカバー率が高い！

判定精度の評価
- 初級（4級）語彙の23.8%がレベル2以下と判定された
- IDFを新聞記事から求めているため，初級の語彙の出現頻度が低いことに起因する

→ 話し言葉のコーパスや初級教科書からIDFを用いることによって解決可能

まとめと課題

まとめ
- 単語レベル評価のための基準について検討した
  - 逆文献頻度（IDF）
  - 単語の総画数

課題
- 単語レベル評価のための基準について検討した
  - 逆文献頻度（IDF）
  - 単語の総画数

謝辞　本研究の一部は，（独）科学技術振興機構 平成21年度シーズ発掘試験（11-143）及び（独）日本学術振興会 平成21年度科研費 基盤研究（B）（21320095）により実施された