

IDFを用いた単語レベル判定 システムの構築と検証

北村達也(甲南大学知能情報学部)

富岡洋介(甲南大学理工学部)

川村よし子(東京国際大学言語コミュニケーション学部)

<http://basil.is.konan-u.ac.jp/>

研究の背景

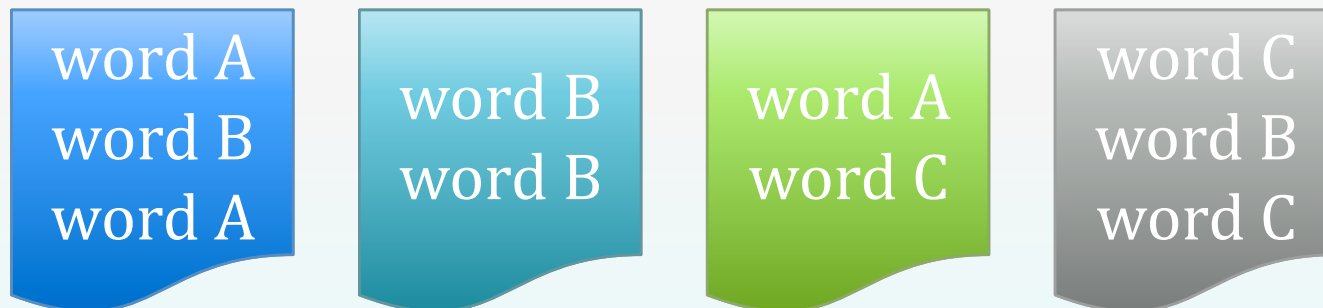
- ◆ 言語教育においては、学習者の学習段階に応じた語彙を含む教材を用いることが重要
- ◆ 「学習段階に応じた語彙」の決め方は？
- ◆ 先行研究：単語レベル判定の基準の例
 - ◆ 日本語能力試験の出題基準(川村, 1999)
 - ◆ 単語親密度(川村, 2008)
 - ◆ 単語出現頻度(川村, 2008)
- ◆ 本研究：IDF (Inverse Document Frequency) の評価

IDFとは

- ◆ 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
 - ◆ 多数の文書に現れる単語 → IDF小
 - ◆ 少数の文書のみにも現れる単語 → 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

IDFの具体例

- ◆ 以下の4文書から成る文書集合において,
 - ◆ 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 $4/3$



- ◆ 重要: 同じ単語であっても文書集合が異なればIDFの値は異なる!

IDFチェッカー

- ◆ 本研究では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
 - ◆ 少数の文書に現れる単語 → IDF大 → 重要度小
- ◆ IDFを「CD-毎日新聞2006データ集」から計算
 - ◆ 95,762文書から128,592語のIDF
 - ◆ IDFの値を5段階の単語レベルと対応づけ(表1)
- ◆ 入力文章中の単語のレベルをIDFにもとづいて判定するシステムを開発

IDFチェッカーの処理例

チュウ太のレベルチェッカー

http://basil.is.konan-u.ac.jp/chuta2/

よく見るページ Firefox を使ってみよう 最新ニュース

甲南大学
KONAN UNIVERSITY


チュウ太のレベルチェッカー

頻度&親密度&IDFチェッカーβ版

入力した文章を単語ごとにレベル分け・色分け・カウントをします。

↓文章を入力して下さい。

顔が人それぞれ違うように、声も十人十色です。私たちは、声の特徴（音声の個人性）が声道（声の通りみち）のどの場所から生まれるのかについて研究しています。これがわかれば、人の声の違いまで再現する発話ロボットや音声合成システムを作れるようになります。音声を使って個人を特定する話者認証技術の信頼性も向上するでしょう。

freq fami IDF リセット 

参考文献

- 川村よし子, 北村達也, 文章の難易度判定のための単語親密度研究会誌, 15(2), 24-25 (2008) [スライド](#)

お知らせ

- 2009年10月 「リーディング・チュウ太ワークショップ&シンポジウム」を開催しました。ひお越してください。(2009/02/06)

完了

result

http://basil.is.konan-u.ac.jp/cgi-bin/switcher2.c

よく見るページ Firefox を使ってみよう 最新ニュース

顔が人それぞれ違うように、声も十人十色です。私たちは、声の特徴（音声の個人性）が声道（声の通りみち）のどの場所から生まれるのかについて研究しています。これがわかれば、人の声の違いまで再現する発話ロボットや音声合成システムを作れるようになります。音声を使って個人を特定する話者認証技術の信頼性も向上するでしょう。

総数	語彙総数	level05	level04	level03	level02	level01	level0	助詞,助動詞,接続詞	記号,数字,英語
93	82	32	10	6	0	0	3	31	11
113.4%	100%	39.0%	12.2%	7.3%	0.0%	0.0%	3.7%	37.8%	13.4%
(59)	(55)	(22)	(10)	(4)	(0)	(0)	(3)	(16)	(4)
107.3%	100%	40.0%	18.2%	7.3%	0.0%	0.0%	5.5%	29.1%	7.3%

単語リスト

- level05
 - の (1)
 - 違う (1)
 - 私 (1)
 - いる (1)
 - 生まれる (1)

完了

IDFチェッカーの性質

- ◆ 先行研究で用いられた基準(日本語能力試験の出題基準, 単語出現頻度)との単語カバー率の比較
 - ◆ 新聞記事, 社説, 中上級教材ではIDFが他の基準を上回るカバー率を示した.
 - ◆ 初級教材では日本語能力試験の出題基準のカバー率が他の基準を上回った.
- ◆ 新聞記事からIDFを計算したため, 易しい単語のレベルが高く評価される場合がある.



ポスターにて

- ◆ IDFチェッカーのデモ
- ◆ IDFチェッカーの運用実験結果の詳細
- ◆ チュウ太バーのデモ
- ◆ リーディング・チュウ太ワークショップ&シンポジウム2009のご案内