TOSHIBA

DEWS 2006 ミニサーベイ

統計処理による日本語テキストからのキーワード抽出

2006 年 3 月 3 日

(株)東芝 研究開発センター
白井 智

TOSHIBA

発表概要

- 背景
- 形態素解析と課題
- 統計量を使った情報抽出
- まとめ

TOSHIBA

発表概要

- 背景
- 形態素解析と課題
- 統計量を使った情報抽出
- まとめ

TOSHIBA

日本語文書自動処理の必要性

大量に蓄積された文書資源
- WEBページ内の文章,メール
- 企業内の書類,特許情報 等々…

格納された情報をどう使うか？
- 整理分類，検索を通じた情報の取得
- 取得した情報を要約，解析

人間が関与せずに自動的に単語やキーワードを抽出したい！

TOSHIBA

文書自動処理の難しさ

コンピュータが扱いやすい文書
- 帳票、定型文
- XML文書

扱いたいのはそれだけではない
- 構造化されてない文書のほうが圧倒的に多い
- 人間向けに書かれた文書はほぼ構造化されてない

TOSHIBA

日本語自動処理の難しさ

文字の区切りを発見するのが困難(曖昧性)
- わかち書きされない
  - 例）「うらにわにはにわにわにはにわにわとりがいる」=>
    「裏庭には二羽，庭には二羽鶏がいる」
    「裏にワニ埴輪にワニは庭庭鳥がいる 」

組みあわせによる名詞
- 漢字、ひらがな、カタカナ
- 柔軟に組み合わせられる単語、キーワード
  - 例）日本データベース学会、問い合わせ処理
  - 例）モーニング娘。

TOSHIBA

ではどうするか？

辞書等の情報
- 単語、文法、品詞
- 人間が与える

文書に内在する統計量
- 出現頻度、生起確率
- データ群から自動的に抽出

を用い単語の区切り、品詞等々の情報を抽出

TOSHIBA

ではどうするか？

辞書等の情報
- 単語、文法、品詞
- 人間が与える

文書に内在する統計量
- 出現頻度、生起確率
- データ群から自動的に抽出

を用い単語の区切り、品詞等々の情報を抽出

本発表では、人手によらない統計量を情報の解析に用いる手法を紹介

TOSHIBA

発表概要

- 背景
- 形態素解析と課題
- 統計量を使った情報抽出
- まとめ

TOSHIBA

形態素解析([1])

日本語の文章を形態素に分割し品詞を付与する処理

- 形態素
  - 意味を持つ最小単位
- 良い精度（95%以上）

「日本語の文章を形態素に分割」の解析

日本語	の	文章	を	形態	素	に	分割
名詞	助詞	名詞	助詞	名詞	名詞	助詞	名詞

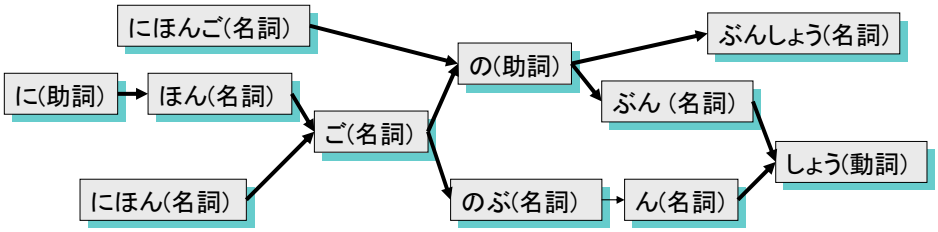
日本語処理といえば、まず形態素解析

TOSHIBA

形態素解析(統計的言語モデル)

品詞辞書を元に
全ての組みあわせから一番“ありそう”な物を選ぶ
- 統計的言語モデル（隠れマルコフモデル）
- 単語の出現頻度 と 品詞間の遷移確率

「にほんごのぶんしょう」の解析

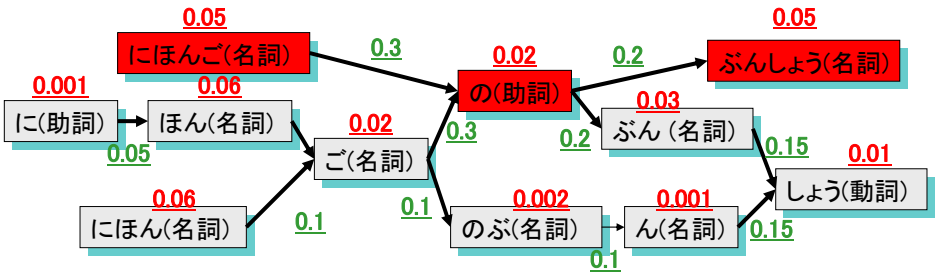


TOSHIBA

形態素解析(統計的言語モデル)

品詞辞書を元に
全ての組みあわせから一番”ありそう”な物を選ぶ
- 統計的言語モデル（隠れマルコフモデル）
- 単語の出現頻度 と 品詞間の遷移確率

「にほんごのぶんしょう」の解析



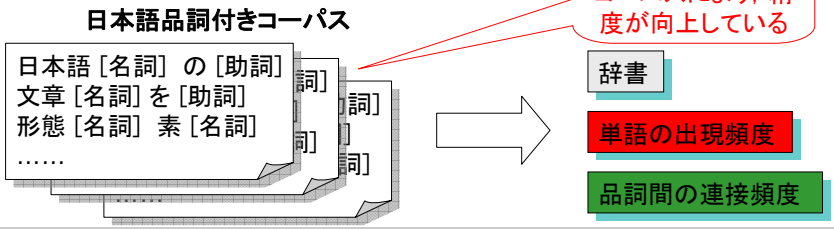
TOSHIBA

形態素解析(統計的言語モデル)

日本語品詞付きコーパスからの学習

- 品詞辞書
- 単語の出現頻度
- 品詞間の連接確率

例) 名詞が100回，名詞の直後に助詞が2回
名詞から助詞の遷移確率は 0.02



TOSHIBA

形態素解析の課題

未知語(辞書にない単語)
- 品詞がつかない(「未知語」)
- 異なる品詞に分割される場合もある
  「らうめん」=>「ら(接尾辞)，うめ(動詞)，ん(助動詞)」

複合名詞
- 複数の単語が複合してできる未登録な単語

例)「日本データベース学会」=>「日本」「データベース」「学会」

ルーズな文法への対応
- 話し言葉 等々

「未知なもの」に弱い

TOSHIBA

発表概要

- 背景
- 形態素解析の基本と課題
- 統計量を使った情報抽出
- まとめ

TOSHIBA

統計量を用いたキーワード抽出

統計量を用いた「未知の言葉」への対処

- 形態素解析を前提
  - 環境
  - 出現頻度と連接頻度
- 形態素解析を前提としない
  - 出現集中
  - 繰返しのパターンと連接確率

TOSHIBA

環境

「Nグラム統計によるコーパスからの未知語抽出」[2]

- 森 信介, 長尾 眞,
- 情報処理学会論文誌, Vol.39, No.7, pp.2093-2100 (1998).

- 大規模なコーパスを前提にした未知語自動抽出
- ある文字列の品詞情報を推測

TOSHIBA

環境とは？

ある文字列の前と後への連接確率の情報

- 各文字列、各品詞が出現することができるパターン

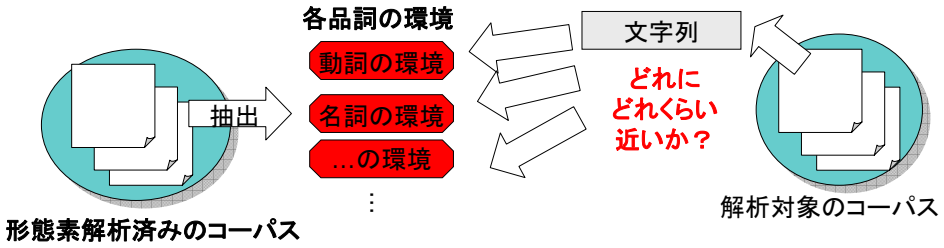
頻度	確率	文字		文字	頻度	確率
13	6．8%	が	楽し	い	16	8．3%
6	3．1%	も		か	2	1．0%
2	1．0%	う		く	3	1．6%

品詞の環境は，各品詞に含まれる文字列の総計とする

TOSHIBA

環境による抽出

- 各品詞の環境を学習
  - 形態素解析済みのコーパスから，品詞ごとの環境を計算．
- 解析対象の文書の任意のNグラムの環境と比較
- 近さに閾値を設定し、品詞及び単語かどうか推定
- 日経サイエンス（1年分　61万文字）から268個の未知語を抽出



TOSHIBA

出現頻度と連接頻度

「出現頻度と連接頻度に基づく専門用語抽出」[3]

- 中川裕志,森辰則,湯本紘彰
- 自然言語処理,Vol.10 No.1,pp.27-45,2003

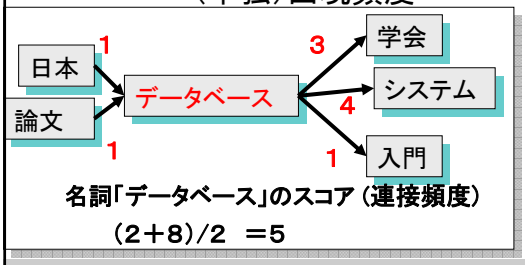
- 形態素解析で抽出された名詞の列からキーワード（専門用語）を選択（ランク付け）する
- 分割されすぎた複合名詞を復元
  - 例）「日本」「データベース」「学会」 => 「日本データベース学会」

TOSHIBA

出現頻度と連接頻度とは？

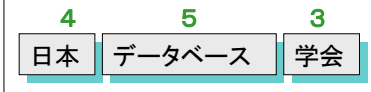
連接頻度と出現頻度を用いたスコアリング

- スコア： 連接頻度×出現頻度
- 連接頻度
  - 前後への単語の繋がり頻度　の平均
  - 「複合名詞の中でよく使われる単語である」ということを示す
- （単独）出現頻度



名詞「データベース」のスコア（連接頻度）

（2＋8）/2　＝5



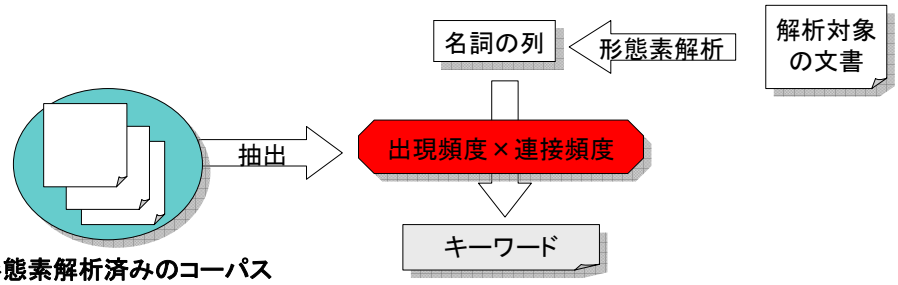
「日本データベース学会」のスコア
（単名詞のスコアの平均）

（4＋5＋3）/3　＝4

TOSHIBA

出現頻度と連接頻度による抽出

- 解析対象の文書を形態素解析し名詞を抽出
- 連続する名詞の組み合わせ全てをスコアリング
- 妙なキーワードは抽出されにくい



形態素解析済みのコーパス

TOSHIBA

出現集中

「未踏テキスト情報中のキーワードの抽出システム開発」[4]

- 梅村恭司
- 未踏ソフトウェア創造事業,2000

- 文章に自動的にキーワードを付与したい
- 辞書を使わず「出現集中」と呼ばれる統計量のみを利用
- 出現集中
  - 「ある単語が一つのドキュメントに現われたという条件で，」
    同じ単語がもう一度出現する確率」
  - キーワードは，出現集中の値が大きくなるという特徴

TOSHIBA

出現集中とは？

ある文字列の出現集中の推定 =

その文字列が二回出現した文書数 / その文字列が一回出現した文書数

一回出現	二回出現	出現集中	文字列
124696	79894	0．64	ロ
3672	2413	0．65	ロボ
3320	2237	0．67	ロボッ
3319	2237	0．67	ロボット
577	96	0．16	ロボットに

キーワードの区切りを越えると減少

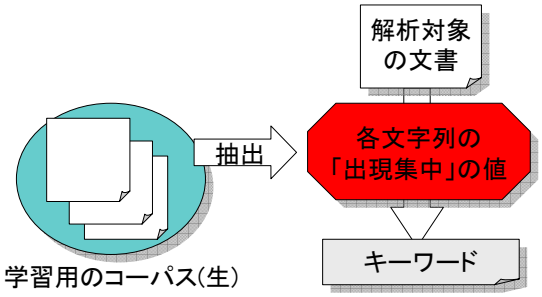
キーワードの区切りを越えると出現集中が減少する

出現集中の積が最大になるように文字列を分割

TOSHIBA

出現集中による抽出

- 辞書データを用いず，生のコーパスだけを学習データに
  - 形態素解析や単語分割等の事前処理が必要ない
- 単語境界精度85%、学習文書数が多いほど高精度、2回以上出現しない単語に弱い



TOSHIBA

文字列の繰返しパターンと連接確率

「反復文字列階層グラフによる文書からのキーワード自動抽出」[5]

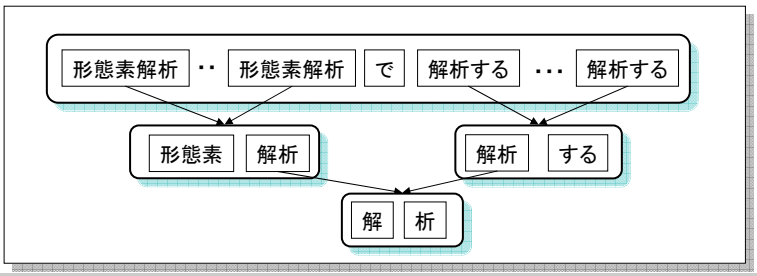
- 白井 智 鳥井 修 金井 達徳
- DEWS2005にて発表

- 辞書を用いず，解析文書のみの情報でどこまでキーワードを抽出できるか？
- 解析対象の文書 のみ を利用する
  - 学習データなし，辞書なし
- 繰返しパターンを用いてキーワード候補生成
- グラフのトポロジと連接頻度によるフィルタリング

TOSHIBA

文字列の繰返しパターンと連接確率

- 反復文字列階層グラフ
  - 文字列の繰返しをノードとして括りだす
  - 二回以上繰返された文字列がノードに
  - ノードがキーワード候補

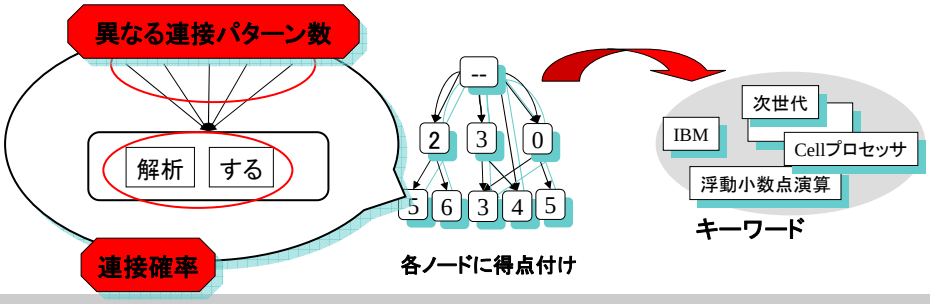


TOSHIBA

文字列の繰返しパターンと連接確率

各ノードに点数を付け上位をキーワードとして抽出

- 文字列の利用のされ方（異なる連接パターン数）と文字列の結びつき（連接確率）でスコアリング
- 複合名詞も上位に抽出。妙なノイズも多い。一回しか出現しない単語は抽出できない



TOSHIBA

発表概要

- 背景
- 形態素解析と課題
- 統計量を使った情報抽出
- まとめ

TOSHIBA

まとめ

- 日本語自動処理の重要性
- 形態素解析の課題
  - 未知語の扱い
  - 複合名詞の細分化
- 統計量を用いたキーワード，未知語の抽出手法
- 今後
  - さらなる統計量(？)
  - 統合されたモデル/システム

TOSHIBA

参考文献

[1] 永田昌明: “形態素解析”, 言語と心理の統計, pp.62-73, 岩波書店, 2003.

[2] 森信介,長尾眞: “n グラム統計によるコーパスからの未知語抽出”, 情報処理学会論文誌, Vol.39, No.7, pp.2093-2100, 1998.

[3] 中川裕志, 森辰則, 湯本紘彰: “出現頻度と連接頻度に基づく専門用語抽出”, 自然言語処理, Vol.10, No.1, pp.27-45, 2003.

[4] 梅村恭司: “未踏テキスト情報中のキーワードの抽出システム開発”, 未踏ソフトウェア創造事業, 2000.

[5] 白井 智, 鳥井 修, 金井 達徳: “反復階層グラフによる文書からのキーワード自動抽出”, 日本 データベース学会 Letters, Vol.4, No.1, pp. 77-80, 2005.