

概要把握のための複数テキスト要約

福原 知宏, 武田 英明, 西田 豊明
奈良先端科学技術大学院大学
知能情報処理学講座
<http://ai-www.aist-nara.ac.jp/~tomohi-f/>

はじめに

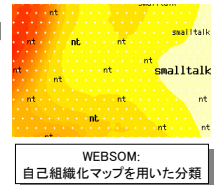
- 目的
 - 概要把握のための複数テキスト要約
- 提案手法
 - 統計情報を用いた話題特定
 - 結束性を考慮した文脈構築
- Topic Showcase
 - 複数テキスト要約システム
- 評価
 - 分類されたテキスト集合から特定した話題に有効
 - 話題・要約の閲覧により、テキスト集合の概要把握に有効

背景

- 情報の増加
 - WWW, メールングリスト, Netnews, etc.
- 情報収集における大量テキストの負荷
 - WWWからの検索結果
 - 閲覧に要する時間と労力
- 概要把握の必要性
 - 情報収集の効率化
 - テキスト集合を構成する話題の把握
 - それぞれの話題の内容の把握

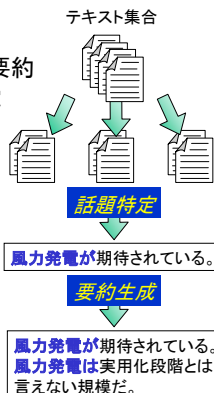
先行研究

- テキスト自動要約
 - 複数テキスト要約[McKeown]
 - 同一事件に関する複数新聞記事の要約
 - Netnewsの自動要約[佐藤]
 - 論文告知記事から会議名・日時・場所を抽出・一覧表示
- テキスト自動分類
 - 検索結果の可視化[Honkela]
 - 検索結果をキーワードで表示

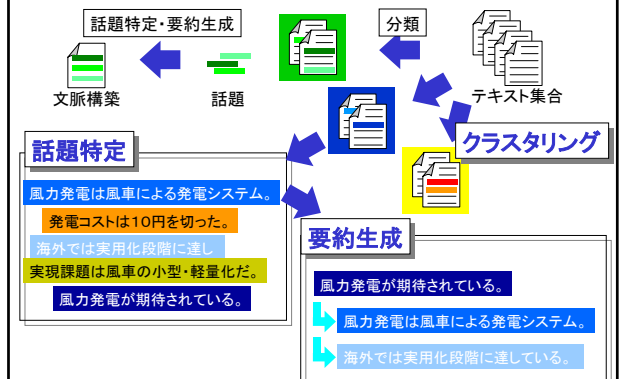


目的

- 概要把握のための複数テキスト要約
 - 対象テキストに依存しない話題特定
 - 文脈構築による話題補足
- 対象
 - テキスト全文検索結果
- 提案手法
 - 統計情報を用いた話題特定
 - 結束性を考慮した文脈構築

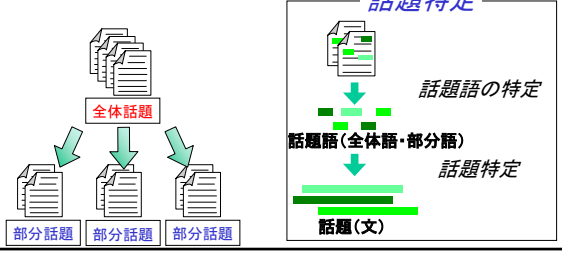


システム全体像



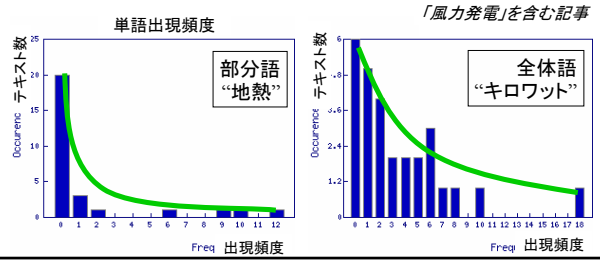
統計情報を用いた話題特定

- 目的
 - テキスト集合を代表する文の特定
- 手法
 - クラスタリングと歪度・尖度の利用



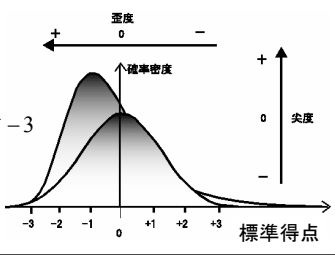
話題の特徴

- 単語の分布形状に相違
 - 全体話題を示す語: ならかな分布形状
 - 部分話題を示す語: 急な分布形状

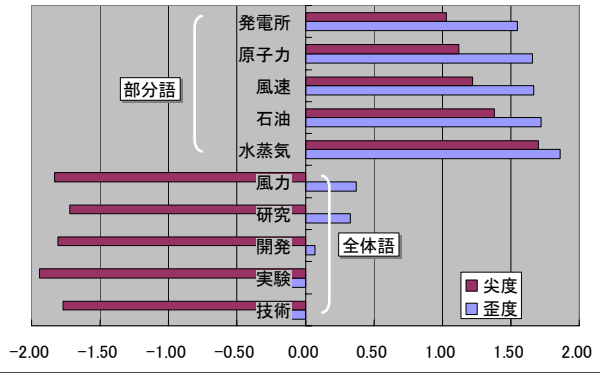


歪度・尖度

- 歪度
 - 分布の歪み
$$\alpha_3 = 1/n \sum_{i=1}^n (x_i - \bar{x})^3 / s^3$$
- 尖度
 - 分布の尖り
$$\alpha_4 = 1/n \sum_{i=1}^n (x_i - \bar{x})^4 / s^4 - 3$$



単語の部分性



話題語の特定

- 評価関数
 - 歪度と尖度の線形和
$$P(term_i) = w_1 \alpha_3 + w_2 \alpha_4$$

α_3 : 歪度, α_4 : 尖度
 w_1, w_2 : 重み
- 部分語・全体語の判定
 - 歪度と尖度に閾値
$$P(term_i) \leq \tau_a: \text{全体語}$$

$$P(term_i) \geq \tau_b: \text{部分語}$$

話題特定

- 話題の評価値
 - 文中の部分語/全体語の評価値の総和を計算
$$score(S_i) = \sum_{term_j \in S_i} |P(term_j)|$$

S_i : 文 i
 $term_j$: S_i 中の話題語

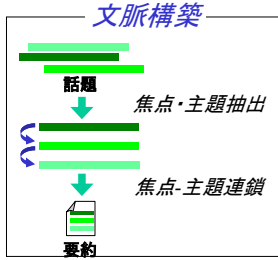
0.47 0.34 2.57 3.44
S = 風力/発電の一キロワット時当たりの単価は十二円を切った。

	歪度	尖度	評価値
風力	0.32	0.15	0.47
発電	0.12	0.22	0.34
キロワット	1.24	1.33	2.57
単価	1.67	1.77	3.44
得点			3.38

$w_1 = w_2 = 1$
 $score(S) = (0.32+0.15)+...$
 $+ (1.67+1.77)$
 $= 3.38$

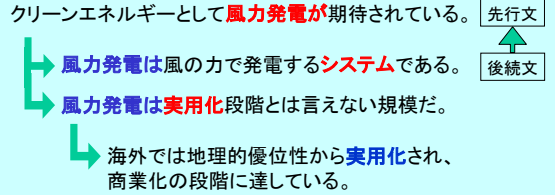
結束性を考慮した文脈構築

- 目的
 - 話題の内容を補足する要約の生成
- 要約とは
 - 話題と話題に関連する文の集合
- アプローチ
 - 先行文に関連する文の展開による文脈構築
- 手法
 - 焦点-主題連鎖



焦点-主題連鎖

- 先行文の記述を後続文で説明する談話構造



文脈構築

- 主題・焦点の決定
 - 格構造の利用
 - 主題: 格、助詞ハ
 - 焦点: 主題以外の格
- 関連文の選択
 - 話題の得点

格	文節
ノ格	最近の
提題、テ格	ソフトウェア開発では
ガ格	エージェント利用が
連格	新たな
ト格	流れと
提題受	なっている

文	得点
クリーンエネルギーとして風力発電が期待されている。	1.23 0.46 0.73
風力発電は風力によって発電するシステムである。	2.46
風力発電は実用化段階とは言えない規模だ。	1.23 0.46 1.45 1.22
自然エネルギーとして実用化が進んでいるのが風力発電だ。	4.36 1.23 0.46 0.22 1.45
	4.58



Topic Showcase

- 主な機能
 - キーワード検索
 - クラスタリング
 - 話題特定
 - 要約生成
- テキストデータ
 - 日本経済新聞95年度版

話題

風が発電システムのタービンを回す。

要約

風が発電システムのタービンを回す。発電効率の高いタービンの開発が発電コストの低下につながる。

要約について詳しく知りたい

要約元のテキスト集合

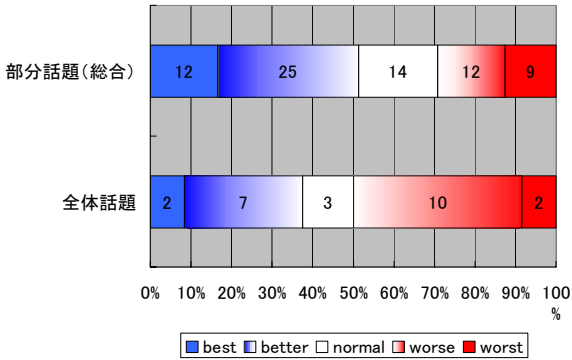
デモビデオ



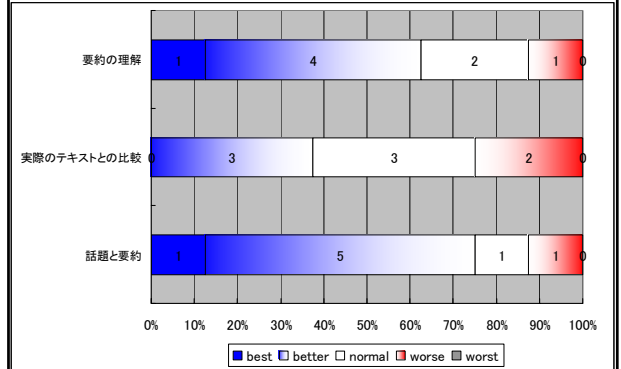
評価

- 評価方法
 - アンケート調査(情報系大学院生8名)
- 評価項目
 - 話題がテキスト集合の内容を反映しているか?
 - 要約の内容を理解できるか?
 - 要約は実際のテキスト集合の内容を反映しているか?
 - 話題と要約からテキスト集合の内容を判断できるか?
- 評価データ
 - 話題
 - “風力発電”のキーワード検索結果
 - 要約
 - ユーザの選択による任意話題

話題の評価



要約の評価



考察

- 話題特定
 - 部分話題に有効
 - 統計情報に加え、単語の意味知識により精度改善
- 文脈構築
 - 内容の理解が可能
 - 焦点・主題連鎖以外の文脈構造の適用
 - 焦点の選択戦略により、より結束性のある文脈構築
- 話題と要約による概要把握
 - 提案手法の有効性

議論

- 良い要約とは何か？
 - 少ない文字数＝良い要約？
 - ┆ よく説明された長い要約でも良い評価となる
 - 読者の理解度
 - ┆ 対話的に情報を引き出せる要約
- 要約コーパスの準備・公開・共有
 - 複数の関連する文書の要約
 - ┆ システム設計者の考える要約だけでは偏りが生じる
 - メーリングリスト(自然言語処理・心理学・知識処理・etc.)

まとめ

- 概要把握のための複数テキスト要約
 - 統計情報を用いた話題特定
 - 結束性を考慮した文脈構築
- Topic Showcase
 - 話題と文脈による複数テキスト要約システム
- 評価
 - テキスト集合の概要把握への有効性
- 今後の課題
 - 観点に応じた要約
 - ┆ 対話的な要約