

記載必要項目の抽出と 文章作成支援

鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
教授 村田 真樹

2022年3月8日

自己紹介(テキスト処理を研究)

- 1992年度～1997年度 京都大学
 - 故長尾真教授(元国会図書館長、元京都大学総長、元情報通信研究機構理事長)の指導の下で研究
 - 文章の解析の研究
- 1998年度～2009年度 情報通信研究機構
 - 教師あり機械学習を利用したテキスト処理の研究
 - 情報アクセス技術の研究
 - その他の種々の研究
- 2010年度～現在 鳥取大学
 - 教師あり機械学習を利用したテキスト処理の研究、情報アクセス技術の研究など

研究業績

テキスト処理の分野において

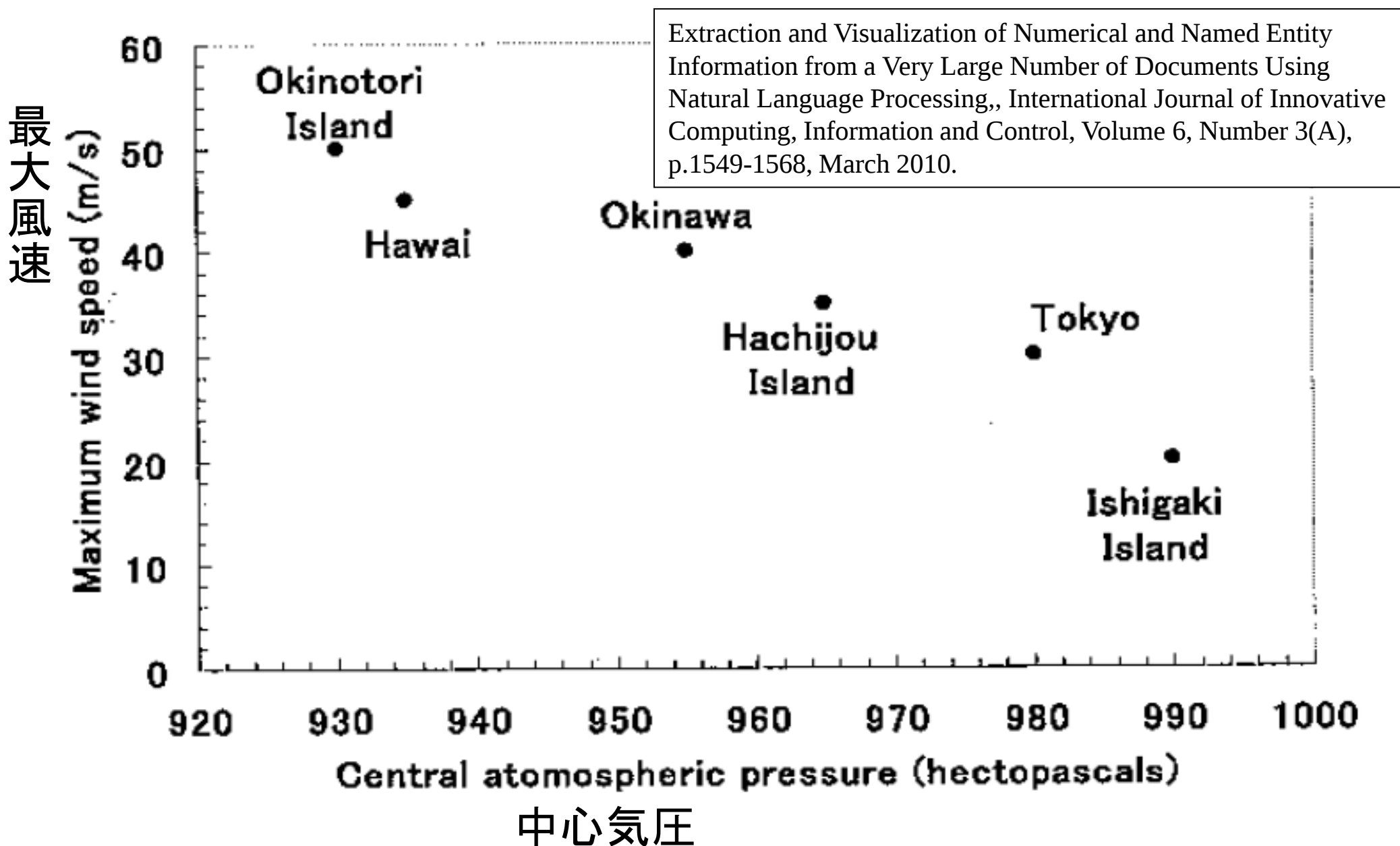
- 著書、4件
- 査読付論文、80件以上
- 査読付国際会議、100件以上
- 特許登録、40件以上、特許出願、80件以上
- 受賞3件（FIT2005論文賞（FIT2005 第4回情報科学技術フォーラム、電子情報通信学会情報処理学会共催）など）
- 招待講演6件

テキスト処理において豊富な経験と実績があります。
テキスト処理関係でなにかあればご相談ください。

技術移転の実績あり

- 前職(情報通信研究機構)での成果
- テキストから数値固有名詞情報を抽出可視化
 - テキストマイニングシステムに実装
 - 商用システムで業界初
- 企業(株式会社NTTデータ数理システム(当時
は株式会社数理システム))に**技術移転**
- 情報通信研究機構の**報道発表**(2008.9.30)
 - <http://www.nict.go.jp/press/2008/press-20080930-5.pdf>

例：台風の圧力と風速の関係



原理

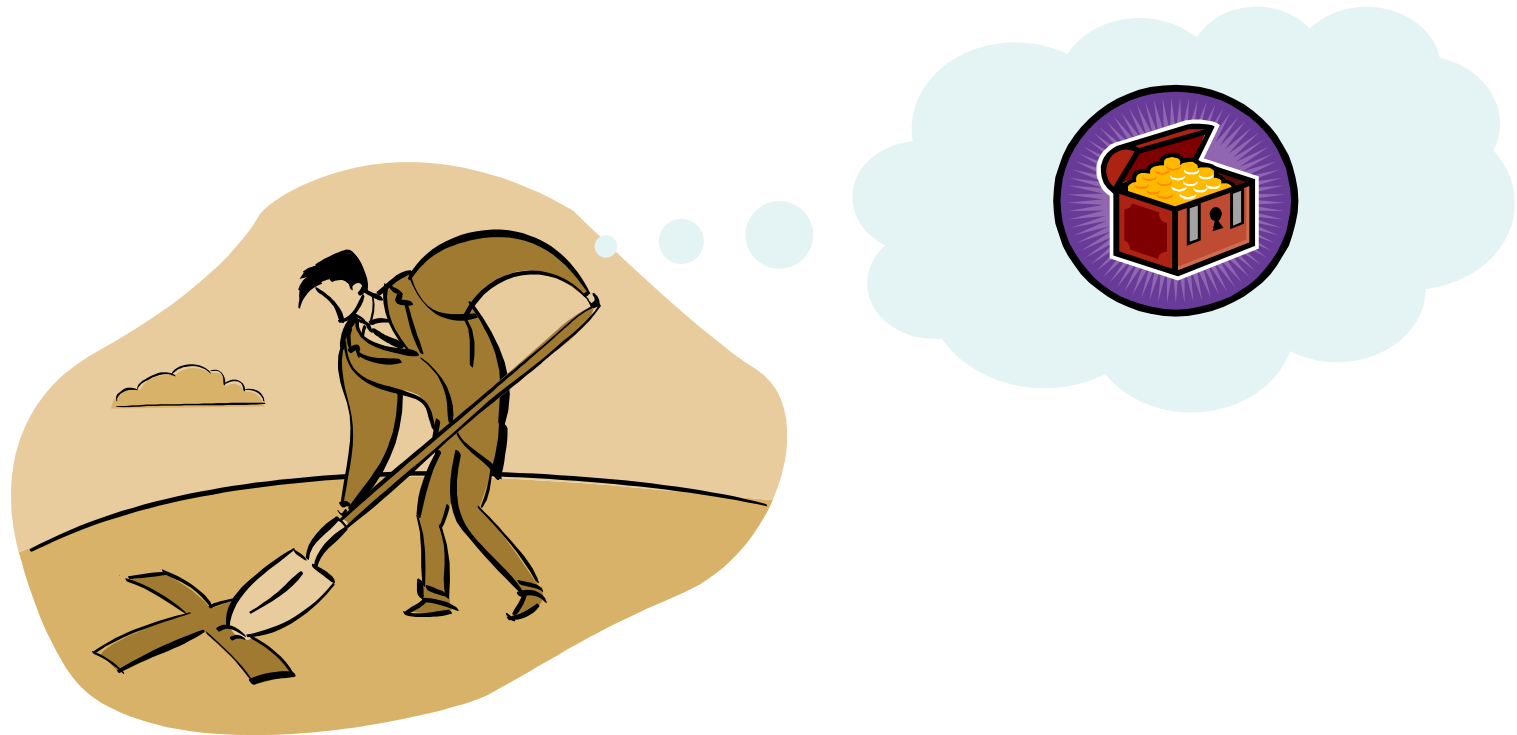
- 文書例
 - X月Y日に**沖ノ鳥島**近海で中心気圧**930ヘクトパスカル**、最大風速**50メートル**の非常に強い台風が発生した。
- 文書中から、数値固有名詞対を取得
 - 930ヘクトパスカル、50m/s、沖ノ鳥島
 - 935ヘクトパスカル、45m/s、ハワイ
 - ……
- それをグラフ化

技術分野の動向

- 電子化テキストの増加
 - 電子テキストから有用な情報を取り出す技術の重要性が高まる
- テキストマイニング
 - テキストデータ中から有用な情報を取り出すこと
- 関連概念：データマイニング
 - データ中から有用な情報を取り出すこと
 - テキストマイニングはデータマイニングの一種

データマイニングとは

- データ(data)を発掘(mining)して、宝物を見つける手法・プロセス



テキストマイニングとは

- **書籍の紹介**
- 事例で学ぶテキストマイニング、上田太一郎監修, 村田真樹ほか著、共立出版, 2008.1
- テキストマイニングとは、大量のテキストデータから新しい事実や傾向を発見する支援技術
 - 人間の分析作業の効率化(人間の作業時間の短縮)
- 種々のテキストマイニング事例を詳述
- 実際のマイニングソフトの利用事例も紹介



今回の技術の概要

- 情報抽出と頻度情報を利用して、文書群において記載必要項目を好適に検出できる技術を構築した。
- 辞書に基づく方法で、数種類の記載必要項目の欠如を0.6から0.8の精度で検出できることを示した。
- 本技術は情報の整理にも利用できる。

記載不足の指摘

機械学習、情報抽出、辞書を利用して、記載必要項目が欠如している文章を自動検出して、文章作成支援をする。

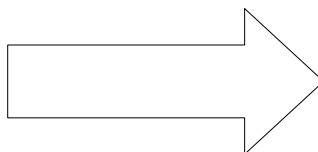
記載必要項目：文書で記載の必要な項目

例：論文であれば、「目的」「例」「問題点」など
(例が欠けている論文であればわかりにくい。)

記載必要項目が欠落している

文章

自動判別



文章作成支援

従来技術とその問題点

- 誤字の修正・適切な語の選択、語順の修正、語と語の係り受けの誤りおよび複雑性の修正に関する従来技術は多数あるが、必要項目の記載が欠如した文章の改善に資する研究はあまりなされていない。
- 項目を事前に指定せずに情報を表に整理する研究はあまりなされていない。

新技術の特徴・従来技術との比較

- 新技術では従来技術で利用していない、記載必要項目の情報等を利用
- これにより、性能を向上させた
- 記載必要項目の検出性能
 - 従来技術のF値(正解率): 0.1~0.5
 - 新技術のF値(正解率): 0.6~0.8

記載不足の指摘

機械学習、情報抽出、辞書を利用して、記載必要項目が欠如している文章を自動検出して、文章作成支援をする。

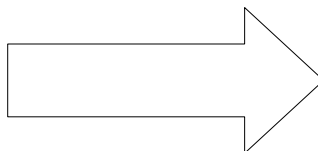
記載必要項目：文書で記載の必要な項目

例：論文であれば、「目的」「例」「問題点」など
(例が欠けている論文であればわかりにくい。)

記載必要項目が欠落している

文章

自動判別



文章作成支援

- 文章の種類ごとに書くべき項目がある程度決まっている。
 - 新聞—— 5W1H
 - 論文—— 研究対象・研究成果・必要性・有効性
 - 面接用原稿—— 自分の能力とその根拠など
- 書くべき項目が書かれていなければ文章の説得力や可読性が低下する

論文での記載必要項目

記載必要項目	内容	検出に役立つ単語
目的	研究の目的	目的 目標 目指す
問題点	先行研究の問題点、 背景	問題
比較	先行研究との比較	比較 比べる
例	具体例	例えば 例 具体

• ある種の機械的处理

- 「目的」「目指す」などの語がなければ、目的が記載されていないと判断する。
- 執筆者にそれを伝えて、目的を書き足すように促す。

文章作成支援の結果(F値)

- ベースライン = すべての論文を記載必要項目が欠落している論文であると判別

項目名	ベースライン	ルールベース (提案手法)
比較	0.32	0.59
問題点	0.54	0.69
目的	0.47	0.56
例	0.06	0.86

- ルールベース(提案手法)のF値 : 0.6~0.8
- F値 : ベースライン < ルールベース(提案手法)

F値

- 再現率 システムの網羅率

$$\text{再現率} = \frac{\text{システムの正解数}}{\text{テストデータ内の正解数}}$$

- 適合率 システム精度

$$\text{適合率} = \frac{\text{システムの正解数}}{\text{システムの出力数}}$$

- F値 再現率・適合率の調和平均

$$\text{F値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

調和平均＝逆数の平均の逆数

まとめ

- 記載必要項目「比較」「問題点」「目的」「例」の検出ができた
 - ルールベース(提案手法)のF値 : 0.6~0.8
 - F値 : ベースライン < ルールベース(提案手法)
- 論文からの記載必要項目の抽出と文章作成支援
 - 岡田拓真, 村田真樹, 徳久雅人, 馬青
 - 言語処理学会第21回年次大会, 2015.

我々の他の技術 単語単位で表を作成

- 単語のクラスタリングにより、同じ種類の単語群を収集
- 同じ種類の単語群を列にまとめる

	構築年	別名	構築者	...
大阪城	1583年	錦城	豊臣秀吉	...
姫路城	1346年	白鷺城	赤松貞範	...
熊本城	1600年	銀杏城	加藤清正	...
名古屋城	1612年	金鯱城	徳川家康	...

検索エンジンを用いた記載欠落箇所の補完

- 文章に重要情報の書き漏れ

→ 読みにくい文章

- 文書から重要情報を抽出

- 表にまとめて出力

→ 表の空白箇所が文章の書き漏れ部分
空白箇所を指摘すれば修正可能

- 検索エンジンを用いて 空白箇所を埋める情報を取得 検索エンジンで文書の情報補完

→ 文章を書き足しやすくなる

→ 書き足す内容を教えられる
文章修正支援

キーワード:
城名

重要情報:
時代 所在地

	時代	所在地
大阪城	安土桃山	大阪
道本城	(空白)	長野

道本城

検索

“道本城は、根井氏の城郭...
根井氏は平安時代末期の武将“

	時代	所在地
大阪城	安土桃山	大阪
道本城	(平安時代)	長野

提案手法(重要情報の抽出)

● Wikipediaの文書

“道本城(どうほんじょう)は、長野県佐久市にあった中世の日本の城(城郭)。湯川の蛇行が洗う断崖上にあり、郭や空堀が残っていたが、現在は耕作地化によって消滅している。”

● 文書から情報抽出

- ✓ 固有表現抽出により「地名」「人名」「組織名」を抽出
- ✓ 表にまとめる

	地名	人名	組織名
道本城	<u>長野</u>		



空白箇所を検索エンジンで補完する

提案手法(検索エンジンを利用)

- 検索エンジンで文書を取得
 - ✓ 「地名」を検索キーワードとする
 - ✓ 検索結果50件を文書とする

(例) 道本城

検索

“木曾義仲軍団の四天王で...
道本(どうほん)城(長野 県佐久市大字根々井字道本城)は、...
木曾義仲に与し、四天王・根井幸親は...”

- ✓ 文書を「地名」「人名」「組織名」に分類して抽出

- 表にまとめる

- ✓ 正解候補が複数→候補の出現した文書数の多さでソート
→上位5位までで抽出の性能評価

	地名	人名	組織名
道本城	<u>長野</u>	<u>根井幸親</u>	<u>木曾義仲軍団</u>

Wikipediaでは取得できなかった情報を取得できる

- 7割前後の性能で、表作成
- 5位正解率(5個の候補のうちに一つでも正解があればよい)で、5～7割前後の性能で、ウェブを利用して空白箇所を補完
- 検索エンジンを用いた記載欠落箇所の補完
 - 野浪尚哉 村田真樹 馬青
 - 言語処理学会 第23回年次大会 発表論文集 (2017年3月)

想定される用途

- 文書群の情報を表に整理できるシステム
- 記載必要項目の検出、記載必要項目の記載の欠如した文章を検出するシステム
- 文書群において記載必要項目の内容・欠如を一覧できる表を示す文書作成支援システム

実用化に向けた課題

- 性能は100%にはできない。性能が100%でなくてもよいような実用化の検討が必要
- 支援システムならば、性能100%でなくてもよいのではないか。

企業への期待

- 技術を利用したシステムの開発と販売
- 実用システムの社会への普及
- テキストデータを持つ、企業との共同研究を希望。
- テキストマイニング技術を開発中の企業、自然言語処理分野への展開を考えている企業には、本技術の導入が有効と思われる。

本技術に関する知的財産権

- 発明の名称 : 文書解析装置、プログラム
- 出願番号 : 特願2015 - 93710
- 特許番号 : 特許第6535858号
- 出願人 : 鳥取大学
- 発明者 : 村田真樹、岡田拓真、
藤原隆太、徳久雅人

お問い合わせ先

**国立大学法人鳥取大学
研究推進機構**

T E L 0857 - 31 - 5546

F A X 0857 - 31 - 5571

e-mail sangakucd@ml.cjrd.tottori-u.ac.jp