

単語の専門性に着目した気象学論文 データベースからの専門語抽出

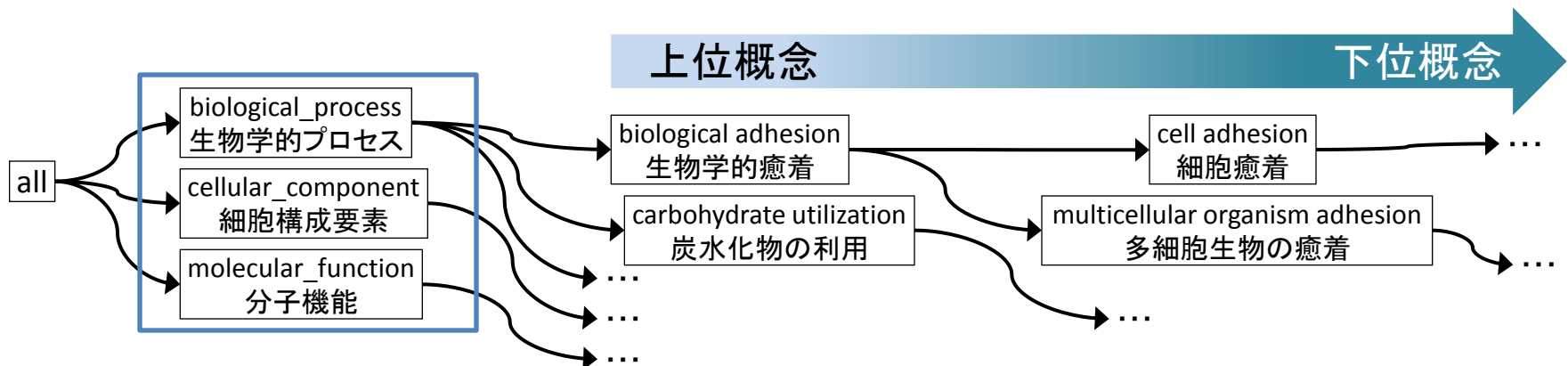
宇井敬一郎，天笠俊之，北川博之
(筑波大学)

科学におけるオントロジー

- 特定分野における概念・語彙体系
 - 共通の概念・語彙を形式的に定義
- 情報技術の活用のため, ますます重要に
 - 多様な情報源
 - データ, 論文, 書籍, Web, ...
 - 情報技術を利用した
 - 検索
 - 知識発見
 - 相互利用
 - オントロジーマッチング, マッピング, アラインメント

オントロジーの例

- 分子生物学オントロジー (GeneOntology)
 - 生物学における概念を記述
 - 各生物学用語は三つのカテゴリに分かれている
 - biological process (生物学的プロセス)
 - cellular component (細胞構成要素)
 - molecular function (分子機能)



オントロジー構築の問題

- 高コスト
 - 多数の専門家による合意形成
 - 数年に渡る議論
- 網羅性の担保が困難
 - 新しい概念, 語彙の出現
 - Web, ソーシャルメディアなどのあたらしいメディアの出現
- 完全に人手によるオントロジー構築はますます困難に

研究の目的

- 多様な情報源を利用したオントロジーの半自動構築
 - ターゲットは「気象学」
- 想定する情報源
 - 論文, Web, Wikipedia, WordNet, ...
- 手順
 - 語彙(専門語)の抽出
 - 語彙間の関連抽出
 - 同義語, 上位語, 下位語, ...
- 特徴
 - データ工学的アプローチを最大限利用

気象オントロジー構築

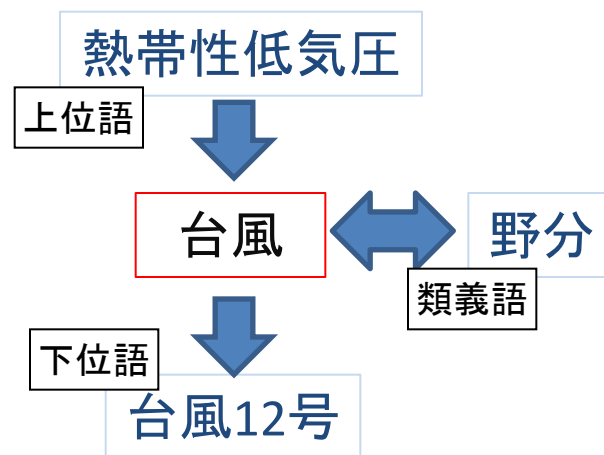
気象オントロジー構築のステップ

- 論文データベース構築
 - 日本気象学会が発行している論文を利用

- 専門語の抽出
 - 頭字語-展開語のペア抽出

JMA - Japan Meteorological Agency

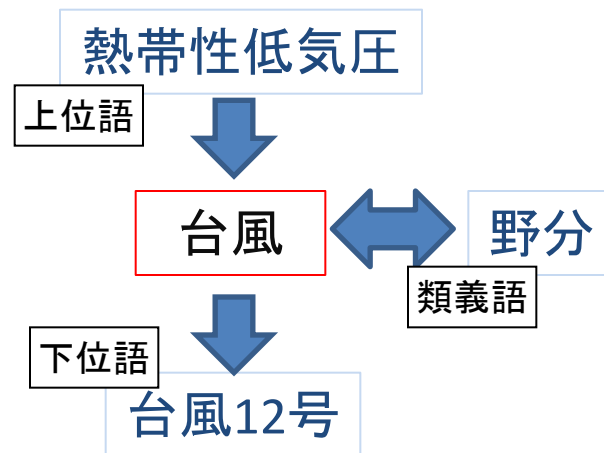
- 語彙間の関連抽出
 - 類義語
 - 上位語, 下位語



気象オントロジー構築のステップ

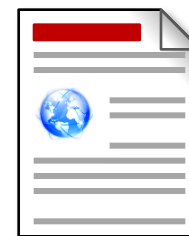
- 論文データベース構築
 - 日本気象学会が発行している論文を利用
- 専門語の抽出
 - 頭字語-展開語のペア抽出
- 語彙間の関連抽出
 - 類義語
 - 上位語, 下位語

JMA - Japan Meteorological Agency



論文データベース構築

- 日本気象学会が発行している 2 誌を利用
 - 気象集誌
 - 隔月発行
 - 1965 年から最新号 (45 年間)
 - 計 3733 の論文
 - SOLA
 - 毎年発行
 - 2005 年から最新号 (5 年間)
 - 計 494 の論文
 - 2 誌共にPDFで発行, 日本語・英語が混在
- PDF からテキストデータのみを抽出
- ここから英語論文のみを選択
 - UTF-8に変換
 - 日本語の文字数が閾値以下の論文 → 英語論文と判断
- 計 **3067** の英語論文による気象学論文データベース



テキストデータ抽出



英語論文のみ



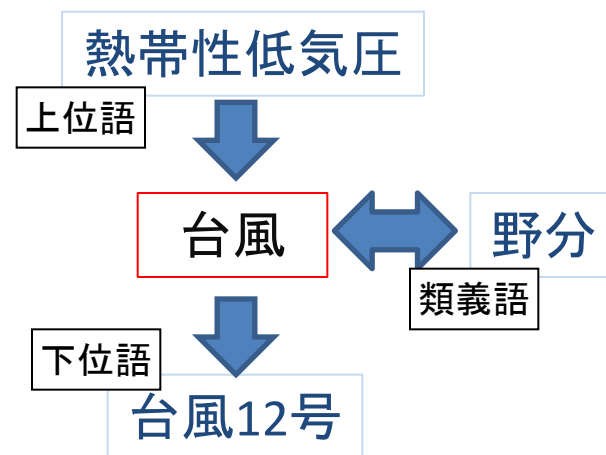
気象オントロジー構築のステップ

- 論文データベース構築
 - 日本気象学会が発行している論文を利用

- 専門語の抽出
 - 頭字語-展開語のペア抽出

JMA - Japan Meteorological Agency

- 語彙間の関連抽出
 - 類義語
 - 上位語, 下位語



頭字語-展開語の抽出

- 括弧に注目したパターンマッチ
 - 文書を走査し, 「(」 - 括弧開き」を検出, 前後を走査
 - 括弧開きの前後が以下の 2 パターン
 - 頭字語(展開語
 - 例: JMA (Japan Meteorological Agency
 - 展開語(頭字語
 - 例: Japan Meteorological Agency (JMA
 - 括弧閉じは無視
 - JMA (Japan Meteorological Agency; JMA is the Japanese ...
 - このように補足情報が付加されることがあるため
 - 頭字語の n 文字目 = 展開語の n 単語目の頭文字
 - 例: JMA (Japan Meteorological Agency

2 単語目

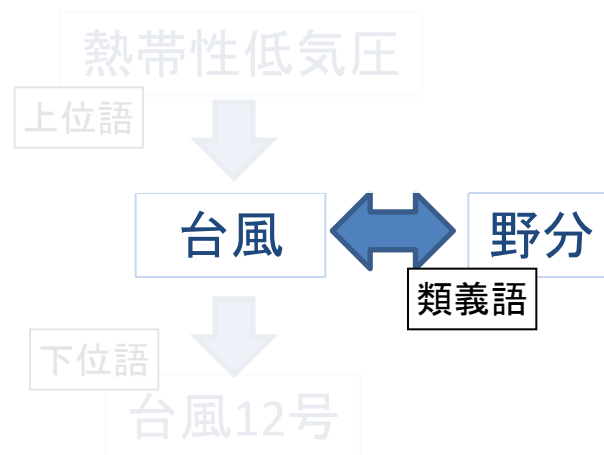
気象オントロジー構築のステップ

- 論文データベース構築
 - 日本気象協会が毎月発行している2誌(気象集誌, SOLA)の論文を利用

JMA - Japan Meteorological Agency

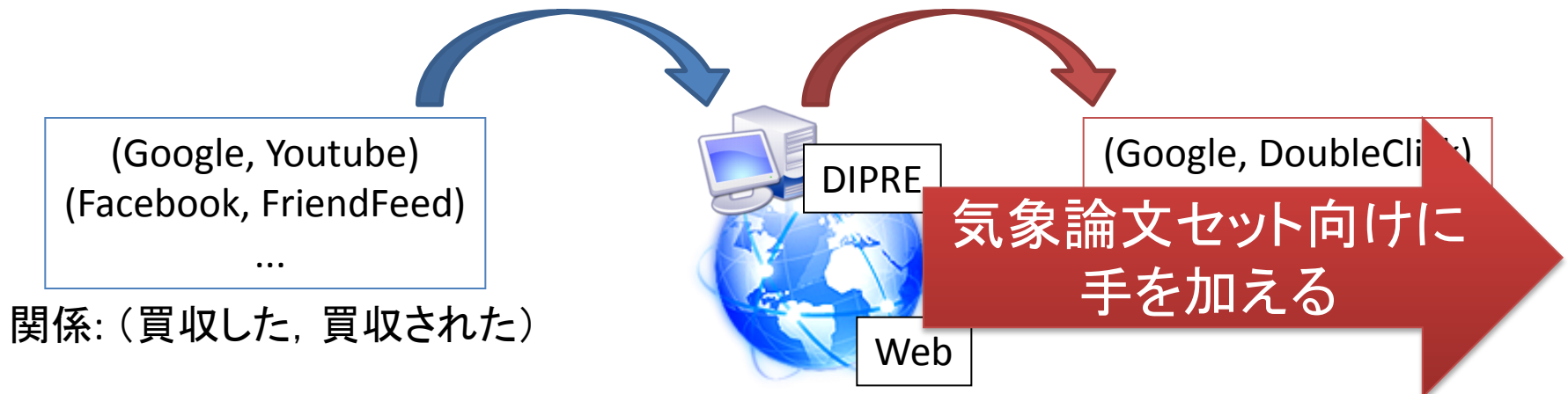
- 専門語の抽出
 - 頭字語-展開語のペア抽出

- 語彙間の関連抽出
 - 類義語
 - 上位語, 下位語



類義語抽出

- DIPRE
 - 特定の関係を抽出する手法の一種
 - セルゲイ・ブリンによって提案
 - Google の創業者のひとり
 - 入力: 特定の関係のペア集合
 - 出力: 入力と同じ関係のペア集合
 - 特徴
 - Web 検索エンジンを活用 → 文書の走査, 情報の取得を効率化



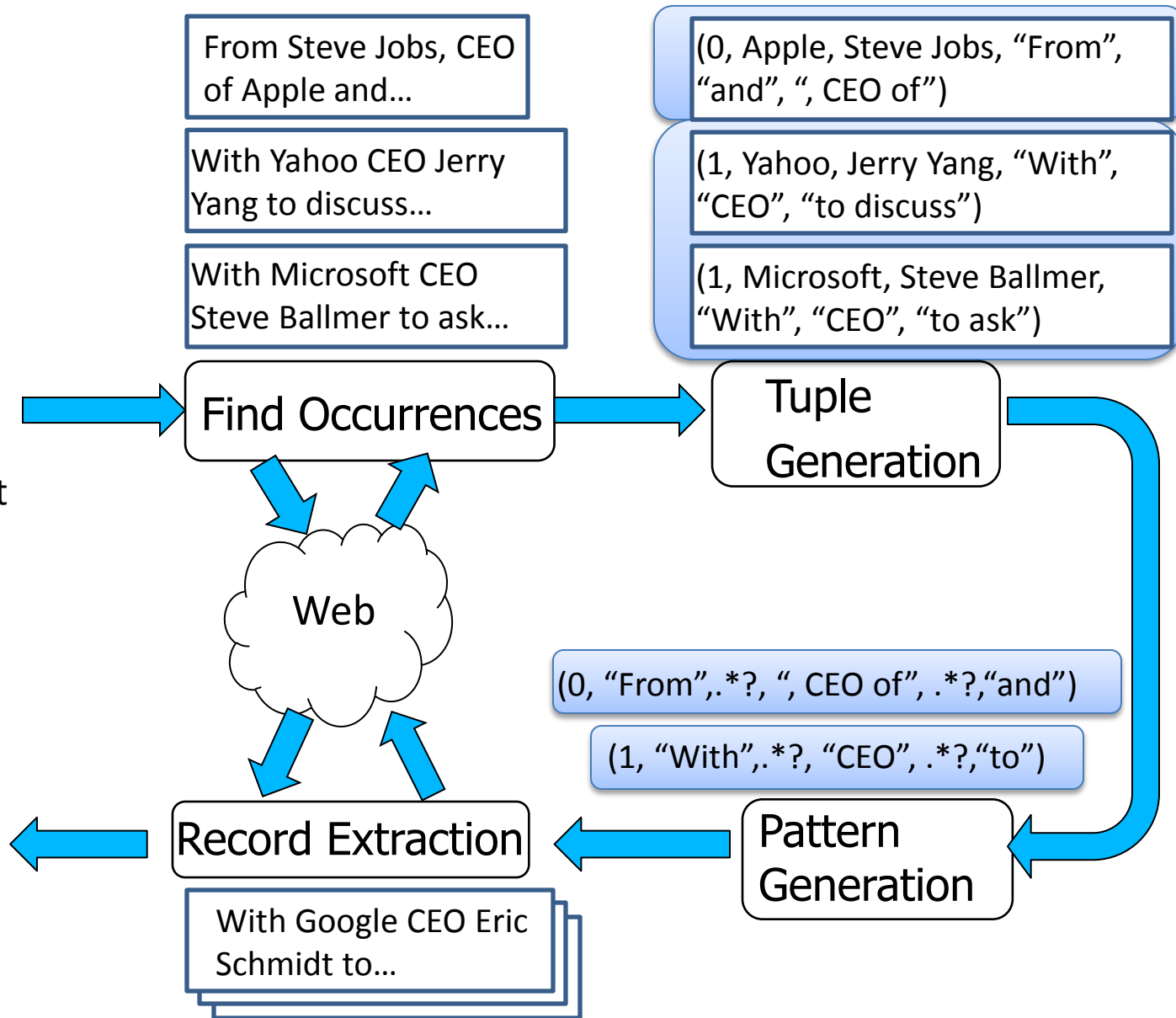
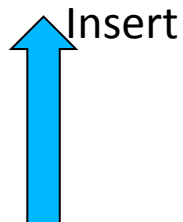
DIPRE [WebDB 1998]

Seed Records

Company	CEO
Microsoft	Steve Ballmer
Yahoo	Jerry Yang
Apple	SteveJobs

New Records

Company	CEO
Google	Eric Schmidt
IBM	Samuel J. Palmisano
...	...



実験

実験

- 頭字語-展開語 ペア抽出
- DIPREを用いた実験
 1. 頭字語-展開語ペアを使用
 2. 手作業で抽出したペアを使用

頭字語-展開語の抽出

- 気象学論文データベースに対して,
先の抽出手法を用いた

- 論文数: 3067

- 頭字語-展開語 ペア数: 2435

- 頭字語数: 1675 ← 少ない

- 展開語数: 2435

- 結果例

頭字語	展開語
JMA	Japan Meteorological Agency
NGSST	New Generation Sea Surface Temperature
MC	Mesoscale Convection
MC	Mindanao Current

} 同じ頭字語

DIPRE を用いた実験①

- 入力ペア集合
 - 頭字語-展開語ペア数:
2,435
 - 気象論文データベースから括弧を用いて抽出
- 結果
 - 出力(新たなペア): 17

新たに抽出された 17 ペア

cloud system	cs-lat and cs
bay of bengal	bo
arabian sea	ar
indian ocean	ion for the e
so	enso cycle
values	rt
pri	rt
atmospheric delay	rt
g system	rs
threat scores	tss
sst	sst
average ne	or prmse
ne	prmse
hic methods of pro	radio occultation
ro	and
expansion of letkf	letkf

DIPRE を用いた実験②

- 入力ペア集合
 - “also known as”, “a kind of” などを含むセンテンスから類義語らしき 8 ペアを手作業で抽出

- 結果
 - 新たなペア: 0

入力ペア

western boundary current	Kuroshio
OPeNDAP	DODS
moisture influx	congo airmass
flowdependent localization	hierarchical ensemble filter
mesoscale model	MM5
Western Ghats	semi-arid region
Baiu front	quasi-stationary front
Nighttime drainage winds	gravity current

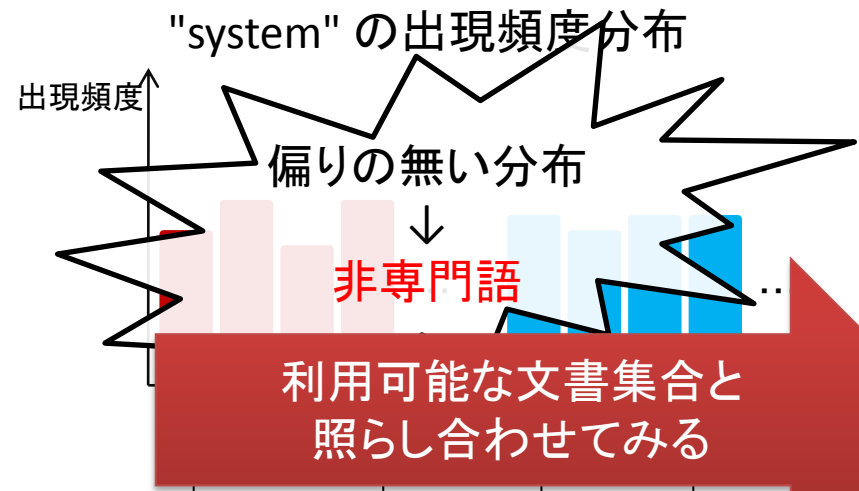
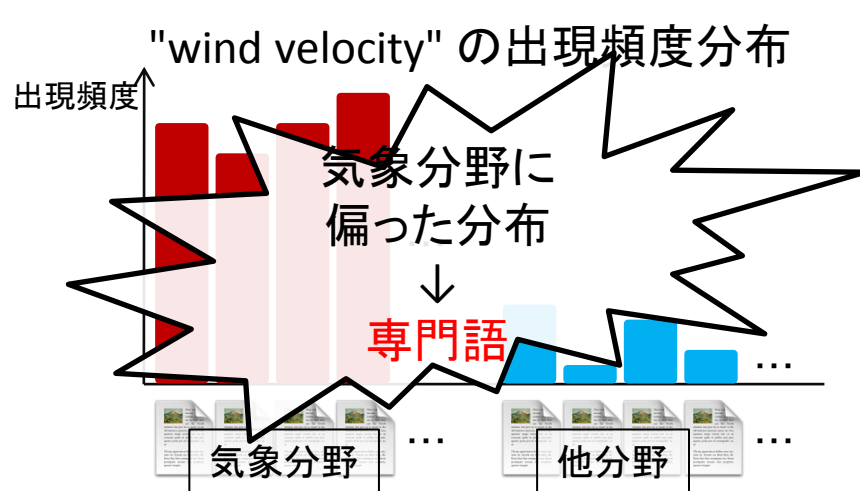
実験のまとめと考察

- DIPRE による類義語抽出結果まとめ
 - 頭字語-展開語ペアを使ったとき
 - 抽出したペアは妥当だとは言えない
 - そもそも単語として意味を成していない
 - 手作業で作成したペアを使ったとき
 - 一つもペアが抽出できなかった
- 考えられること
 - 論文の数が足りない
 - DIPRE は HTML のような半構造化データ向き
 - Brin の論文の例でも HTML が挙げられている
 - 専門語を予め抽出することにより、より良い結果の可能性
 - (単語として意味を成していないと、抽出の意味が無い)

專門語抽出

専門語の抽出

- TermExtractor (Sclano and Velardi, 2007)
 - 単語の出現頻度分布からその単語の専門性を導出
 - 専門性: 対象分野での専門語らしさ
 - 例: 対象分野 → 気象学
 - "wind velocity" と "system" の専門性を導出
 - 専門性が閾値以上するとき → 専門語



利用可能な文書集合

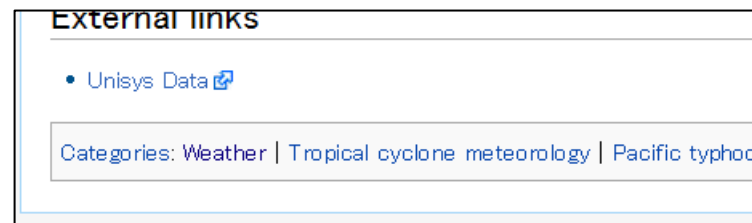
- 気象学論文データベース
 - 気象学分野のみの文書集合
 - 論文の数: 3067
 - ✓ 他分野の文書が必要



- Wikipedia (英語)
 - 様々な分野の文書集合
 - Category:Weather など Category 表記により、どの分野に関する文書か明示
 - ✓ 記事の数: 1000万件



WIKIPEDIA
The Free Encyclopedia



Wikipedia 記事 "Typhoon" のページ下部

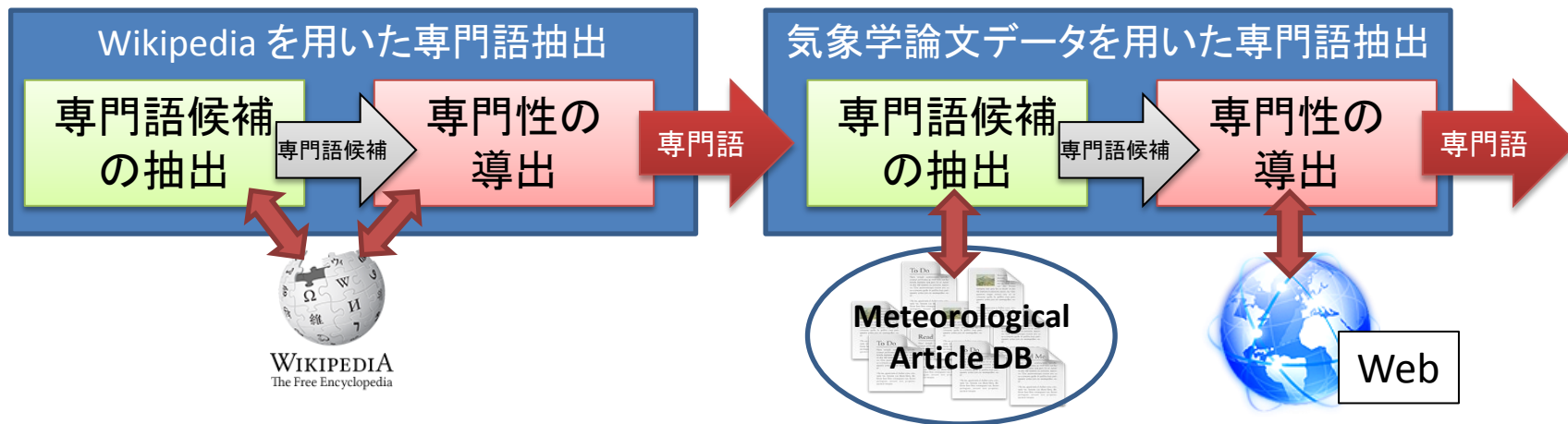
- ウェブ
 - ✓ 文書の数: 無数
 - ✓ その文書が気象学に関するものか否か

TermExtractor を使うのは難しい

それぞれの文書集合の
特徴を生かした手法を提案

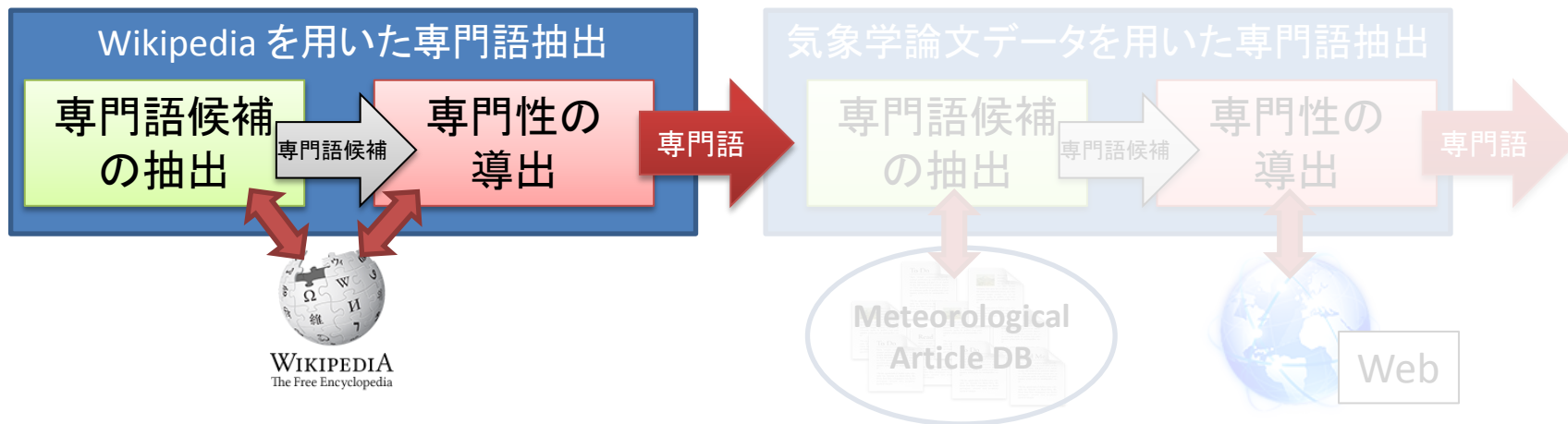
提案手法

- 2つのステップ
 - Wikipedia を用いた専門語抽出
 - 専門語候補の抽出
 - 専門性の導出
 - Wikipedia 記事中の Category 表記を利用
 - 気象学論文データを用いた専門語抽出
 - 専門語候補の抽出
 - 専門性の導出
 - Wikipedia を用いて抽出した専門語と, ウェブを利用



提案手法

- 2つのステップ
 - Wikipedia を用いた専門語抽出
 - 専門語候補の抽出
 - 専門性の導出
 - Wikipedia 記事中の Category 表記を利用
 - 気象学論文データを用いた専門語抽出
 - 専門語候補の抽出
 - 専門性の導出
 - Wikipedia を用いて抽出した専門語と, ウェブを利用



Wikipedia を用いた専門語抽出

Wikipedia 記事から気象学に関わる Category を選出

Category 例: Climate, Weather, Meteorology, Wind, etc.

- 指定した Category に含まれる記事を
気象学分野の記事とする



専門語候補を抽出

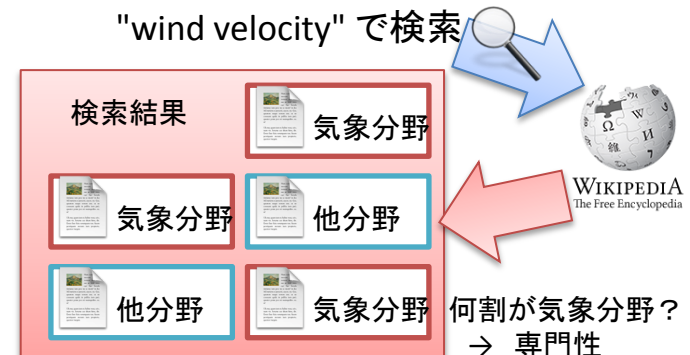
- OpenNLP を使い, 名詞句 (= 専門語候補) の抽出
 - OpenNLP: Java クラスライブラリ, POSタグ付加も可能
 - 最大エントロピー法による教師あり学習を用いた手法



専門性の導出

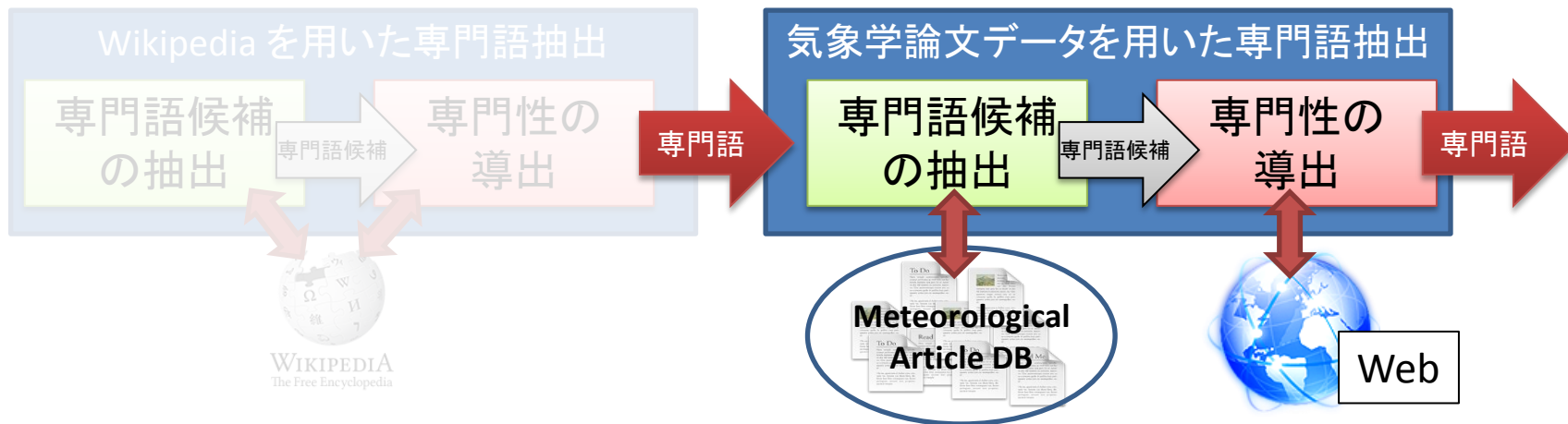
- 検索エンジン Lucene を利用
- 専門語候補 "wind velocity" の専門性を求める
 - "wind velocity" で Wikipedia を検索
 - 検索結果から専門性を導出
 - 検索結果記事の内, 何割が気象分野か
 - 閾値以上するとき, "wind velocity" は専門語

各候補の専門性導出



提案手法

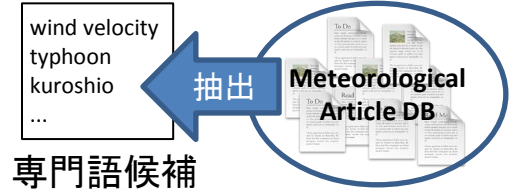
- Wikipedia を用いた専門語抽出
 - 専門語候補の抽出
 - 専門性の導出
 - Wikipedia 記事中の Category 表記を利用
- 気象学論文データを用いた専門語抽出
 - 専門語候補の抽出
 - 専門性の導出
 - Wikipedia を用いて抽出した専門語と、ウェブを利用



気象学論文データを用いた 専門語抽出

専門語候補を抽出

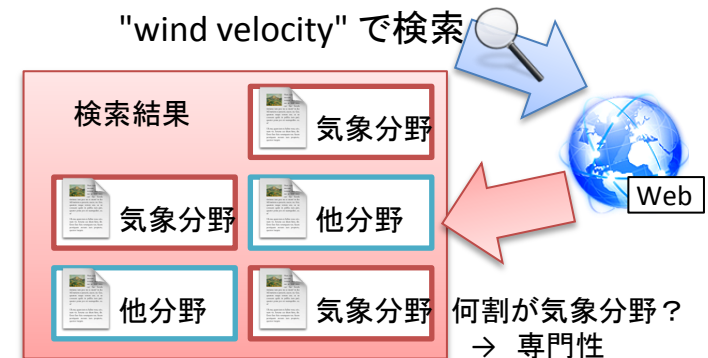
- OpenNLP を使い、名詞句 (= 専門語候補) の抽出



各候補の専門性導出

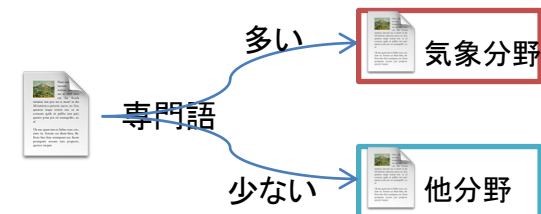
専門性の導出

- ウェブ検索エンジンを利用
- 専門語候補 "wind velocity" の専門性を求める
 1. "wind velocity" でウェブ検索
 2. 検索結果から専門性を導出
 - 検索結果ページの内、何割が気象分野か
 3. 閾値以上のとき, "wind velocity" は専門語



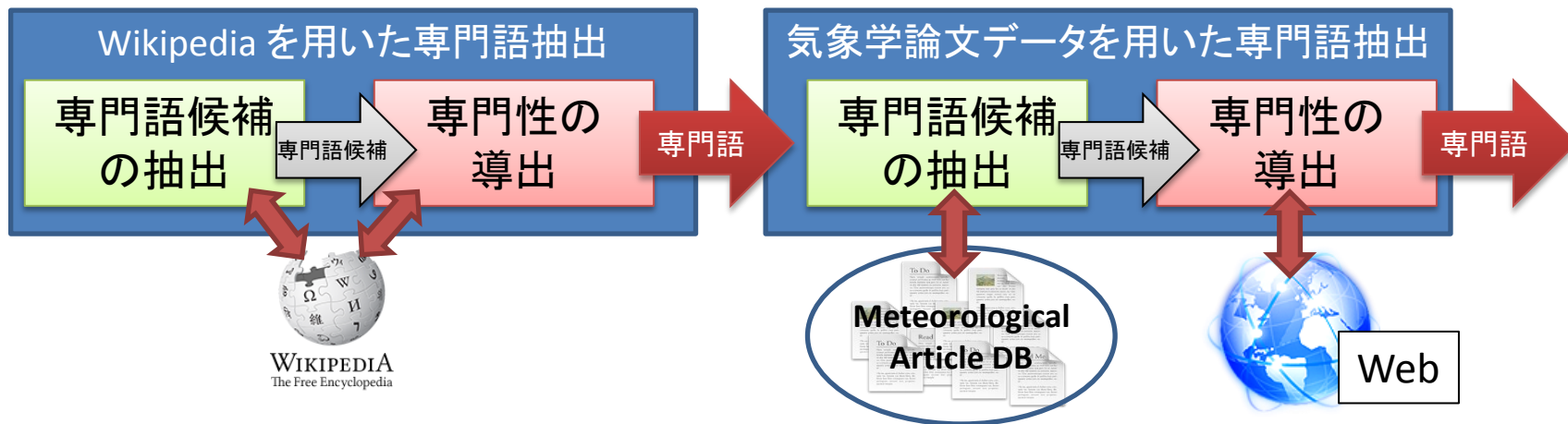
気象分野に関わるページかどうかの判定

- ページ内にどれだけ専門語が含まれるか
 - 閾値以上なら気象分野に関わるページ
 - 使う専門語は, Wikipedia から抽出したもの



提案手法

- Wikipedia を用いた専門語抽出
 - 専門語候補の抽出
 - 専門性の導出
 - Wikipedia 記事中の Category 表記を利用
- 気象学論文データを用いた専門語抽出
 - 専門語候補の抽出
 - 専門性の導出
 - Wikipedia を用いて抽出した専門語と、ウェブを利用



まとめと今後の課題

まとめ

- 気象学論文データからオントロジー構築を試みた
 - 論文データベースの構築
 - 3067 の論文によるデータベース
 - 頭字語-展開語ペアの抽出
 - 2435 ペアを抽出
 - DIPREを用いた類義語の抽出
 - 17ペアの抽出
- 新たな専門語抽出手法の提案
 - 気象学論文, Wikipedia, ウェブをデータソースに使用

今後の課題

- 専門語抽出
 - DIPREを用いた手法
 - データの量を増やして再度実験
 - WikipediaとWeb検索エンジンをもちいた専門性判定
 - 実装と実験
 - 人手による評価プロセス
- オントロジー構築

謝辞

- 堀之内先生(北大)から貴重なコメントとご議論をいただきました。お礼申し上げます。