単語の専門性に着目した気象学論文からの専門語抽出

宇井 敬一朗[*1]，天笠 俊之[*2,*3]，北川 博之[*2]

Term Extraction from Meteorological Papers Based on Degree of Specicity of Terms

Keiichirou UI[*1], Toshiyuki AMAGASA[*2,*3] and Hiroyuki KITAGAWA[*2]

Abstract

Recent remarkable development of ICT technologies has made us possible to process huge amount of data. In particular, we can get massive text information though the World-Wide Web. It should be noted that, among such information sources, there have been many Web sites, such as Wikipedia, where high quality information is provided. As a consequence, it has been an important issue how to extract useful information out of such heterogeneous information sources. In the meantime, in many scientific fields, extracting useful information out of massive information resources has been a major challenge. For example, in the biomedical field, PubMed, a well-known bibliographic database, has been used to extract useful knowledge. In this paper we try to extract technical terms out of heterogeneous information sources, such as papers, Wikipedia, and WWW in the meteorological field. To this end, we introduce the degree of specificity of a term to judge that whether the term is a technical term.

Keywords: Technical term extraction, Wikipedia, Paper, WWW.

概　要

近年のコンピュータ，インターネットの発達により，人々が扱う事のできるデータが膨大になっている．特に，テキストデータはWWWを通じて膨大な量をアクセスすることが可能である．また，Wikipediaのように，質の保証されたテキストデータも出現し，これら多様かつ膨大なテキスト情報から有用な情報を抽出することが重要である．一方，科学分野でも膨大な情報をいかに活用するかは重要な課題である．テキストデータの観点からは，バイオ医学分野のPubMedが著名な例であり，PubMedからの知識抽出などが試みられている．本研究では気象分野を対象に，論文データベース，Wikipedia, WWWを利用した専門語の抽出を行う．専門語の自動抽出は，特定分野のオントロジー構築に利用できるなど，さまざまな応用が期待される．本研究では，特に単語の専門性の指標を導入することにより，専門語の判別を行う．

キーワード：専門語抽出，Wikipedia，論文，WWW.

1　はじめに

近年のコンピュータ，インターネットの発達により，我々が扱う事のできるデータ量は爆発的に増加している．この膨大なデータを整理し，より高速に目的のデータを取得するための研究や，意味のある情報を抽出する研究が盛んに行われている．

[*1] 筑波大学大学院システム情報工学研究科（Graduate School of Systems and Information Engineering, University of Tsukuba）
[*2] 筑波大学システム情報系（Faculty of Engineering, Information and Systems, University of Tsukuba）
[*3] 宇宙航空研究開発機構宇宙科学研究所（Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency）

その中でも自然文によって記述されたテキストデータなど非構造データから構造化データを抽出する研究は，ウェブ等でアクセス可能な情報の多くがテキスト形式であるため，特に盛んになっている．例えば，テキストから単語を抽出し，その語の間の関係を推定することによって，類義語辞書が自動的に構築できる．そのために，大量のテキストデータから名詞や名詞句を抽出，利用される文脈が似た名詞（句）との関連付けといった処理が困難であり様々な研究がなされている．この類義語辞書は，そのまま類義語辞書として使われたり，同意語を用いた全文検索
に使われたり，文書の語彙統一に使われる．

こういった，分野に固有の語彙の自動抽出に関する研究は，これまで主にバイオ分野で盛んに行なわれてきた．しかしながら，同様のニーズは他の分野にも存在すると考えられる．例えば気象学を例にとると，気候学，気象力学，メソ気象学，航空気象力学など様々な関連分野が存在し，各分野において新しい用語が常に出現し続けている．このため，大量のテキストデータから専門語を自動的に抽出することができれば，類義語辞書の自動構築やその他の応用につながり大変有用である．

そこで本研究では，気象学論文などのテキストデータから専門用語を自動的に抽出する手法を提案する．専門語の自動抽出は，特定分野のオントロジー構築に利用できるなど，さまざまな応用が期待される．本研究では，特に語の専門性の指標を導入することにより，専門語の判別を行う．本提案で用いるテキストセットは，英語 Wikipedia [1], 英語気象学論文，AMS Glossary [8] である．それぞれの特徴を活かした専門語抽出を行う．また提案手法により抽出した専門語の評価を行った．Wikipedia，気象学論文からそれぞれ，20,200 語，7,200 語を抽出し，その精度を人手により評価した．

本論文の構成は，以下の通りである．第 2 節で関連研究について述べる．第 3 節で提案手法を説明し，第 4 節で提案の評価を行う．第 5 節でまとめと今後の課題について述べる．

2 関連研究

DIPRE [2] (Dual Iterative Pattern Relation Expansion) は，1998 年，Sergey Brin によって提案されたウェブ検索エンジンを用いた関係抽出システムである．DIPRE は，特定の関係の名詞句ペア集合を入力として与えると，ウェブ検索エンジンを活用することで，ウェブから入力ペア集合と同様の関係の名詞句ペアを抽出することが可能なシステムである．

Danushka Bollegala らは，Relational Duality を利用した関係抽出システム [3] を提案している．この手法は，DIPRE とは違い利用者からの入力ペアを要求しない手法で，文書セットから抽出可能な関係，ペア全てを抽出する．出力は，意味関係付きの名詞句ペアになる．意味関係付き名詞句ペアとは，その名詞句ペアがどう行った関係かといった情報が付加された名詞句ペアである．

対象分野の文書セットのみを用いた専門語抽出手法としては，Frantzi らが 1996 年に提案している手法 [4] と，中川らが 2001 年に提案している手法 [5] がある．Frantzi らが提案した手法は，例えば “cyclone” という語を専門語かどうか判断したい時に，“cyclone” を含むより長い名詞句（例えば，“tropical cyclone” など）の統計情報を利用することで，どれだけ専門語らしいかをスコアリングする．中川らが提案した手法は，評価対象の名詞に隣接する名詞の統計情報を利用した手法を提案している．

専門語抽出システムの先行研究として TermExtractor [6, 7] がある．TermExtractor はウェブアプリケーションとして提供されるシステムで，テキストデータを zip 形式で固めてアップロードすると，そのテキストデータで使われている専門語が

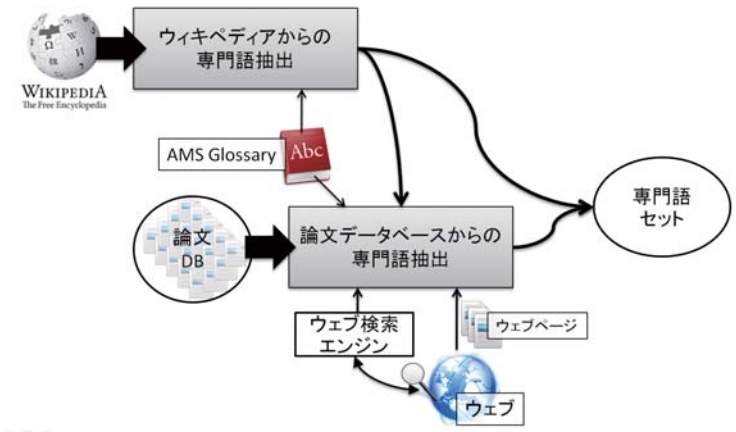

図 1 提案手法の概要.

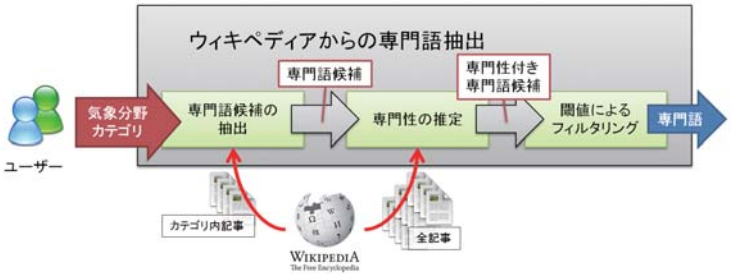

図 2 Wikipedia からの専門語抽出.

抽出され，XML 形式で表現されたデータをダウンロード出来る．本研究では，TermExtractor の手法を応用し，ウェブ，Wikipedia を利用した新たな手法を提案する．

3　提案手法

3.1　概要

本研究における専門語とは，その分野固有の概念に強い関連性を持つ語を指す．この関連性の強さを専門性とし，本手法ではこの専門性を推定するために，「専門語は対象とする分野の文書において高頻度で出現し，他分野の文書では出現しないもしくは頻度が低い」という考え方を採用する．

本研究は，専門語を抽出するために次の情報源を利用する：1) 気象分野の用語集である AMS Glossary [8]，2) オンライン百科事典 Wikipedia [1]，3) 気象分野の英語論文，4) ウェブ文書．

提案手法の概要を図 1 に示す．

Wikipedia からの専門語抽出

気象分野に関するカテゴリを指定することで，気象分野に関わる Wikipedia 記事を選定し，名詞句を抽出する．Wikipedia は様々な分野に関する記事が存在するため，気象学以外の分野におけるその名詞句の出現頻度も取得可能である．このため，Wikipedia 内の情報だけで専門性の推定が可能である．抽出した名詞句の専門性を推定によるスコアを算出・ソートした後に，AMS Glossary を指標として，上位何件を専門語として抽出するか決定する．

論文からの専門語抽出

気象学論文からテキストを抽出し名詞句の抽出をする．抽出した名詞句の専門性の推定には，ウェブを用いる．しかし，ウェブを用いると，インターネットアクセスに時間がかかる．このため，気象学論文を用いて予め専門性の高そうな単語を推定する．ウェブを用いて専門性を推定によるスコア算出・ソートした後に，Wikipedia のとき同様，AMS Glossary を指標に上位何件を専門語として抽出するか決定する．

3.2　Wikipedia からの専門語抽出

この節では，提案提案手法の最初の手順である Wikipedia からの専門語抽出について説明する．図 2 に処理の流れを示す．

専門語候補の抽出

Wikipedia の気象学のカテゴリに属する記事から名詞句を全て抽出する．この名詞句を専門語候補と呼ぶ．

専門性の推定

専門語候補それぞれに対し，専門性というある種のスコアを推定する．この専門性は，気象学の記事において高い頻度で出現し，気象学以外の記事ではあまり出現しない語に高いスコアが与えられる．

評価と専門語の決定

推定した専門性を，AMS Glossary を用いて評価し，専門語を決定する．

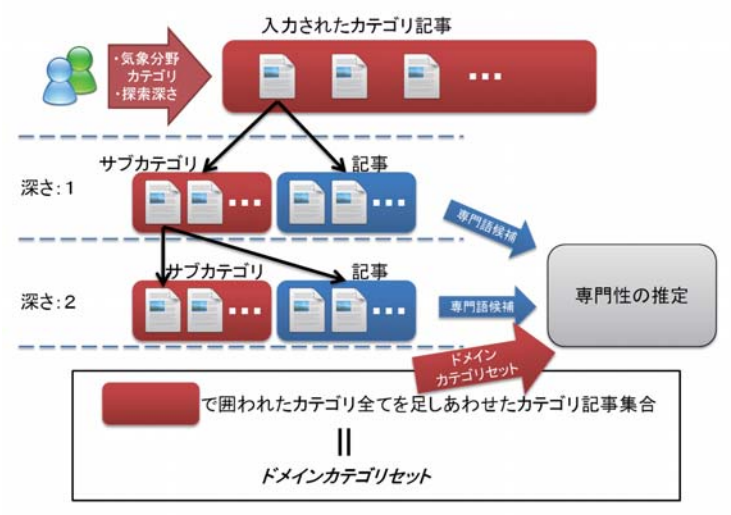

図 3 Wikipedia 記事からの専門語候補抽出（深さ 2 の例）.

3.2.1 専門語候補の抽出

この節では，Wikipedia からの専門語候補の抽出について記述する．専門語候補の抽出には，初めに抽出する Wikipedia 記事を選定し，OpenNLP[9] を用いて専門語候補の抽出を行う．

Wikipedia 記事の選定

図 3 に記事の選定から専門語抽出の過程を示す．専門語候補の抽出をする際，予め抽出対象の記事を選定する．選定の基準は，気象学に関連した記事であれば抽出対象，そうでなければ対象外になる．

記事の選定を行うに当たって，本システムは初期入力として，気象学に関わる複数の Category 名前空間記事と，探索するサブカテゴリの深さ受け取る．与えられた Category 名前空間記事のサブカテゴリを深さ分探索する．

探索した全ての Category 名前空間記事がリンクしている記事が，専門語候補抽出対象として選定される．深すぎる探索を行うと，より多くの Category 名前空間記事を探索することになる．探索した Category 名前空間記事は，のちの専門性の推定にも利用するため，多すぎる Category 名前空間記事は，推定の精度を下げてしまうため，探索するサブカテゴリの深さは限定している．

探索を行った Category 名前空間記事は，一つの集合にまとめて，*DomainCategorySet* と呼ぶ．*DomainCategorySet* はのちの専門性の推定に利用する．

選定した記事からの専門語候補の抽出

選定した気象分野記事に対し，既存の名詞句抽出システムを利用して専門語候補の抽出を行った．利用した名詞句抽出システムは，OpenNLP[9] である．

気象分野記事から OpenNLP によって抽出した名詞句を，専門語候補と呼ぶ．この専門語候補を，*DomainCategorySet* を用いて専門性の推定を行う．

3.2.2 専門性の推定

抽出した専門語候補の専門性推定について述べる．専門性とは，その気象分野における専門語らしさを数値化したものである．専門性は，その語が気象学に関する文書により多く出現すると高なり，他分野の文書により多く，もしくは偏りなく出現すると，専門性は低くなる．

専門性の推定には，二つの手法を提案している．Lucene[1] の検索エンジンを用いた手法と，ベクトル空間モデルによる手法である．

[1] http://lucene.apache.org/

Lucene の全文検索エンジンによる専門性推定

Lucene とは，オープンソースな Apache ライセンスプロジェクトで，全文検索システムを実現するための Java クラスライブラリ及びそのクラスライブラリを用いて実装されているシステムの総称である．Lucene の標準の全文検索エンジンは，TF-IDF [10] によるスコアリングとベクトル空間モデル [11] を用いたブーリアン検索 [12] である．ベクトル空間モデルでは，ユーザから投入されたれたクエリから生成したベクトルと，文書ごとに得られるベクトルがより近い文書がより高いスコアになる．

この Lucene 標準の全文検索エンジンを用いて，専門語候補 *term* の専門性を次のように定義した．

$$score_N(term) = \sum_{k=1}^{N} slope^k \cdot f(d_k)$$

$$f(d) \quad = \quad \begin{cases} 1 & d \text{ が } DomainCategorySet \text{ に属する．} \\ 0 & \text{上記以外．} \end{cases}$$

ここで，*slope* ($0 < slope < 1$) は利用者が与える定数．*DomainCategorySet* とは，専門語候補の抽出の際に探索を行った Category 名前空間記事の集合である．また，*dk* は，専門語候補 *term* をクエリとした検索結果の第 k 位の記事を指している．

気象分野の記事に出現する専門語候補は，Lucene 検索エンジンで検索すると，検索上位に *DomainCategorySet* が占めるため，その専門語候補のスコアは高い値を示す．

ベクトル空間モデルによる専門性推定

ベクトル空間モデルによる専門性推定は，*DomainCategorySet* から作った V_{Cat} と，専門語候補 *term* から作るベクトル V_{term} の内積をとり，*df* で割ることで，推定値を求める．

$df_{term} =$ *term* を含んでいる記事の数　　(*DF)

$df_{term} = \log$（*term* を含んでいる記事の数）
(*LogDF)

$v_{term,d} =$ 記事 d に含まれる *term* の数

$$v_{cat,d} = \begin{cases} 1 & d \text{ が } DomainCategorySet \text{ に属する．} \\ 0 & \text{上記以外．} \end{cases}$$

$$V'_{term} = \{v_{term,d}\}_{d \in D}$$

$$V_{term} = \mid V'_{term} \mid \cdot V'_{term}$$

$$V_{Cat} = \{v_{cat,d}\}_{d \in D}$$

$$score(term) = \frac{V_{Cat} \cdot V_{term}}{df_{term}}$$

ここで，d は記事，D は全 Wikipedia 記事集合を表している．ベクトル V_{Cat} と，ベクトル V_{term} の内積を取ることで，ベクトル V_{term} から，*DomainCategorySet* に属する記事に該当する要素のみが非ゼロ要素なる．このため，*DomainCategorySet* に属する記事にあまり出現しない専門語候補は低い *score* (*term*) になる．また，ベクトル V_{term} が正規化されたベクトルであることと，df_{term} の除算をすることで，どの分野の記事にも高い頻度で出現する専門語候補にも低い *score* (*term*) が与えられる．

3.2.3　評価と専門語の決定

これまでの処理によって，専門語候補と，その専門性スコアによる順位付けが得られた．この中から，専門語候補の上位 N 語を専門語として抽出する．そのために，専門語彙である AMS Glossary を利用する．これを正解集合として，専門語候補のうち，AMS Glossary に収録された専門語がどの程度含まれているかを指標として，N を決定する．その指標には，*Recall*, *Precision*, *Fmeasure* 利用する．

Recall, *Precision*, *Fmeasure* の算出方法は以下の通りである．ここで T_{AMS} とは，AMS Glossary に属する語の集合，T_N は本手法によってソートされた語の上位 N 語の集合である．

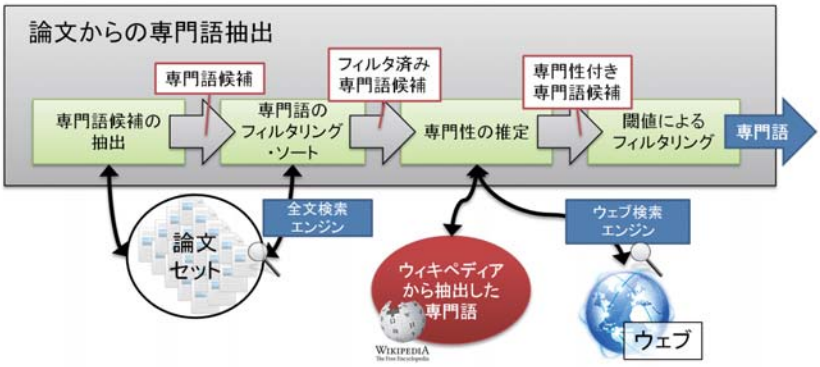

図 4 論文からの専門語抽出.

$$Recall = \frac{|T_{AMS} \cap T_N|}{|T_{AMS}|}$$

$$Precision = \frac{|T_{AMS} \cap T_N|}{|T_N|}$$

$$Fmeasure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

3.3 論文からの専門語抽出

この節では，論文からの専門語候補の抽出とその専門性の推定について記述する．手順は大まかに 4 つのフェーズに分かれている（図 4）.

専門語候補の抽出

論文から名詞句を全て抽出する．この名詞句を専門語候補と呼ぶ.

専門語候補のフィルタリングとソート

専門性推定を行う前に，予め抽出に用いた論文セットを用いてフィルタリング，スコアリング，ソートを行う.

専門性の推定

専門語候補それぞれに対し，Wikipedia の専門語抽出と同様に専門性の推定を行なう．ただし，Wikipedia の時と違い，他分野の文書セットが用意できない，そのページがどの分野に関して記述しているのか判別できないなどの違いがある．このため，ウェブ検索エンジンと Wikipedia から抽出した専門語を用いてそれぞれの推定を行う.

評価と専門語の決定

専門性を推定した専門語候補をソートし，AMS Glossary を用いて評価，推定した専門語の閾値を決定する．この閾値を元に専門語候補をフィルタリングし，残った候補を専門語として出力する.

専門語候補の抽出は前節と同様なので，ここでは手順 2 の専門語候補のフィルタリングとソートから説明する.

3.3.1 専門候補のフィルタリングとソート

専門性の推定にはウェブ検索エンジンを利用するため，大量の専門語候補を全て処理することは難しい．そのため，system, paper などの一般語を除く必要がある．ここでは，一般語に対して低いスコアが与えられるようなスコアリングを行ない，閾値によるフィルタリングによって専門語候補の絞り込みを行なう.

専門語候補のフィルタリングは図 5 手順で行なう.

DF の小さい語をフィルタリング

抽出した専門語候補から，DF（その語が出現する論文の数）が特に低い候補を除く．予め Lucene によって作成した完全転置インデックスを用いて，DF を求める.

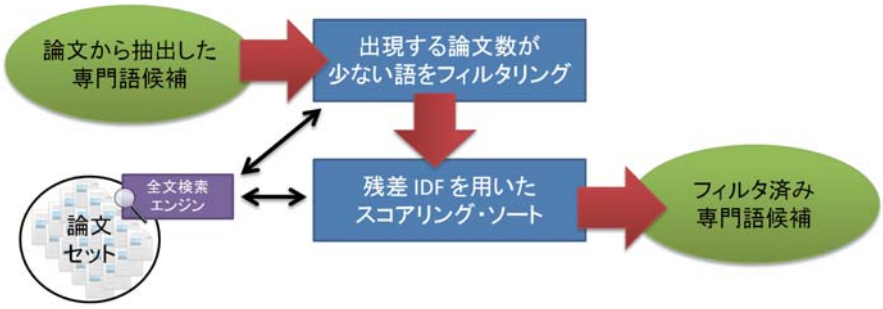

図 5 論文から抽出した専門語候補のフィルタリング

残差 IDF を用いたスコアリングとソート

各専門語候補に対し残差 IDF を用いたスコアを算出し，降順にソートする．スコアに必要な，単語の出現に関する数字は，前の処理と同じく Lucene の完全転置インデックスを用いる．

***DF* の小さい語のフィルタリング**

DF 値（その語が出現する論文の数）が特に低い候補を除く．これにより，その論文・著者が独自に定義している語や，偶然他分野から引用されてくるような語を除去する．

通常 *DF* を求めるには，文書セット全体を走査する必要がある．しかし，本研究では，文書セット全体の走査を避ける為に，予め Lucene によって作成した完全転置インデックスによる全文検索エンジンを用いることとした．

残差 IDF を用いたスコアリングとソート

専門語候補のソートは，残差 IDF [18] (Residual IDF，RIDF) というスコアリング手法の一つを利用する．残差 IDF は，文書セット全域の単語分布の偏りを利用したスコアリング手法である．計算式は以下の通りである．

$$idf_{term} = \log\left(\frac{NumDocs}{DocFreq_{term}}\right)$$

$$estimatedIdf_{term} = -\log\left(1 - e^{-\frac{GrobalTermFreq_{term}}{NumDocs}}\right)$$

$DocFreq_{term}$ = 語 *term* が出現する論文の数

$GrobalTermFreq_{term}$ = 語 *term* が文書セット内に出現する頻度

$$\begin{aligned} ridf_{term} &= idf_{term} - estimatedIdf_{term} \\ &= \log\left(NumDocs \cdot \frac{1-e^{-\frac{GrobalTermFreq_{term}}{NumDocs}}}{DocFreq_{term}}\right) \end{aligned}$$

idf_{term} は IDF と呼ばれ，どの文書にも出現するような一般語に対して，低いスコア，出現する文書が限られるような語に対して高いスコアが与えられる．$estimatedIdf_{term}$ は，語の出現を，それぞれ独立であると仮定したときに導かれる推定 IDF である．その語の出現確率が独立であると仮定すると，ポアソン分布に従った出現確率分布になる．

ただし，実際にソートに用いたスコア手法は以下の式である．

$$score1_{term} = \frac{1 - e^{-\frac{GrobalTermFreq_{term}}{NumDocs}}}{DocFreq_{term}} \quad (RIDF)$$

$$score2_{term} = ridf_{term} \cdot \log DocFreq_{term} \quad (RIDFLogDF)$$

これは，$DocF\,req_{term}$ がある程度大きく，残差 IDF も高い値を持つ語により大きなスコアが算出されるよう調整されている．

3.3.2　専門性の推定

専門性の推定は，Wikipedia からの専門語抽出同様，推定対象の語が気象分野の文書により多く出現，他分野の文書にあまり出現しない語であるときに高いスコアが与えられるような，推定手法を用いる．

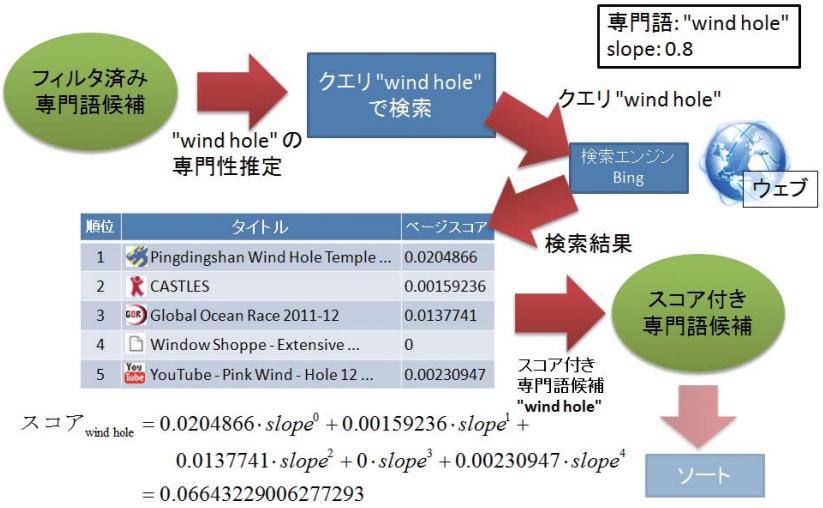

図 6 専門性の推定.

ただし，様々な分野に関する文書が存在する Wikipedia とは違い，論文セットだけでは他分野の文書を用意することは難しい．そこで，様々な分野のテキストセットが存在するウェブを用いて，論文から抽出した専門語候補の専門性の推定を行う．ただし，ウェブページは，そのページがどの分野に関する記述をしているかといった情報を持っていないケースが殆んどである．そのため，そのページが気象分野にどの程度関係のあるページなのかといった推定も行う必要がある．

専門語候補の専門性推定の大まかな流れを，図 6 に示す．この図では，専門語候補 “wind hole” の専門性推定を行なっている．ページスコアとは，そのページがどの程度気象分野に関係のある語であるかを数値化したものである．スコア$_{wind\ hole}$ は “wind hole” の専門性推定スコアである．

専門語候補 “wind hole” でウェブ検索　専門性の推定を行う専門語候補をクエリとしてウェブ検索エンジンに掛け，検索結果の上位 N 件のウェブページを取得する．

ウェブページのスコアリング　検索結果とし返ってきた上位 N 件のウェブページが気象分野にどれだけ関係があるかスコアリングをする．スコアリングには，Wikipedia から抽出した専門語を利用する．

専門語候補 “wind hole” のスコアリング

スコア付きの N 件のウェブページを使い，専門語候補 “wind hole” のスコアリングをする．

ウェブページのスコアリング

専門語候補をクエリとしてウェブ検索し，検索結果として返ってきたウェブページがどれだけ気象分野に関係性をページスコアとして数値化する．ウェブページ w のページスコアを以下に示す．*KnownTermSet* とは，Wikipedia から抽出した専門語セットである．

$$
\begin{aligned}
WebPageScore_w &= \frac{KnownTermFreq_w}{NumWords_w} \\
KnownTermFreq_w &= (\text{ページ } w \text{ における } KnownTermSet \text{ に属する語の出現頻度}) \\
NumWords_w &= (\text{ウェブページ } w \text{ の単語数})
\end{aligned}
$$

ウェブページのテキストと *KnownTermSet* 内の全専門語の照らし合わせにおいて，*KnownTermSet* 内の語が名詞句であり，2 万語以上あるため，総当り的な実装をすると処理に時間がかかってしまう．そこで，Aho-Corasick 法 [18, 19] に手を加えたものを利用し，高速化を計った．

検索結果のウェブページをスコアリングした後，専門語候補のスコアリングを行う．

専門語候補のスコアリング　専門語候補でのウェブ検索，ウェブページのスコアリングの後，専門語候補のスコアリングを

行う．専門語候補 *term* に対する専門性推定スコアは以下のように定義した．

$$score_N(t) = \sum_{N}^{k=1} slope^k \cdot WebPageScore_{w_k} \quad (1)$$

wk = クエリ t での検索結果 k 位のウェブページ

ここで，*slope* ($0 < slope < 1$) は，利用者が与える定数．$WebPageScore_{w_k}$ は，ウェブページ w_k に対するのページスコアである．

例えば，気象学に多く出現する専門語候補 t は，検索結果上位のウェブページが，気象学分野に深く関わるウェブページ（ページスコアの高いウェブページ）であることが多くなるため，高いスコアを示す．逆にどの文書にも出現するもしくは気象学以外の分野に多く出現するような専門語候補 t だった場合，検索結果の上位ウェブページは，気象分野にかかわりの薄いウェブページ（ページスコアの低いウェブページ）が多くなるため，$score_N(t)$ は低い値を示す．

3.3.3　評価と専門語の決定

Wikipedia の時同様，*Recall*, *Precision*, *Fmeasure* を算出し，上位何語を専門語として抽出するか決定する．ただし，ウェブページを用いたスコアリングを行うと *Fmeasure* が明確に最大値を取る上位 N 語の判定が難しいため，代わりに *Precision* を用いて専門語を決定した．

4　評価実験

提案手法を実装して実際に専門語の抽出を行ない，結果の評価を行なった．

4.1　Wikipedia からの専門語抽出

実験に用いた Wikipedia データは，2010 年 9 月 16 日からダンプされた，英語 Wikipedia の全記事データ 24GB（1,000 万記事）と，Category 名前空間記事によるリンクデータを MySQL[13] に格納して利用した．

入力した Category 名前空間記事は，Atmospheric sciences, Meteorology, Climate, Weather，Mesoscale meteorology の計 5 記事，探索する深さは 1，である．この入力により，抽出した専門語候補の数 : 96,927, 専門語候補の抽出を行った記事数 : 2,199, *DomainCategorySet* のサイズ（要素数）: 248 という結果になった．

この実験では，専門性推定によって得られたスコアを用いて降順にソートを行った語の配列の評価を行う．結果を，図 7 から図 11 に示す．これらの図は，それぞれの専門性推定によるスコア順にソートした上位 N 件の専門語候補を T_N としたときの，*Recall*, *Precision*, *Fmeasure* の値をプロットしたものである．それぞれの専門性推定手法同士の比較をするために，三つの *Fmeasure* をプロットしたものが，10 になる．三つの手法の横軸の数が違うのは，それぞれのスコアを導出したときに，スコアが 0 になった専門語候補を除外しているためである．

三つの手法に共通することは，上位 N 語を増やしていくにつれ *Recall* は単調増加を続け，*Precision* ははじめのうちは振動するがやがて振動自体は収束し，緩やかに単調減少を続けている点である．Recall が単調増加する理由は，T_N の数が増えるので自然とその中に T_{AMS} にも含まれる語が増えるのに対し，分母の $|T_{AMS}|$ は変化しないためである．逆に Precision が落ちて行く理由は，分母の $|T_N|$ の増加量に比べ，その中に正解する語の数があまり増えていかないことを示している．*Precision* が序盤において振動しているのは，$|T_N|$ は一定ずつ増加していくのに比べ，$|T_{AMS} \cap T_N|$ の変化量が安定しないためである．

次に *Fmeasure* に注目する．Lucene 標準検索エンジンを用いた手法，ベクトル空間モデル（DF）では *Fmeasure* が一定値で頭打ちになっているのに比べ，ベクトル空間モデル（LogDF）を用いたベクトル空間モデルは，*Fmeasure* が上位 20000 語を最大値に，それ以降落ちていることがわかる．

図 11 は，*Recall* − *Precision* 曲線と呼ばれるグラフの一種で，横軸が *Recall*，縦軸が *Precision* として，三つの手法の結果をプロットしている．今までのグラフをみてもわかるとおり，*Recall* と *Precision* にはトレードオフの関係がある．この関係において最も効率のよい *Recall*−*Precision* のトレードオフの関係を持つ手法はどれかを比較するためのグラフである．このグラフにおいてもやはり，Lucene 検索エンジンを用いた手法が振動しているときによりよい性能を示すことがあるものの，Recall 0.06 以降は，ベクトル空間モデル（LogDF）のほうがより良い *Precision* を示していることがわかる．

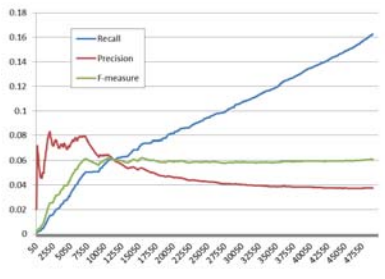

図 7 Lucene 検索エンジン.

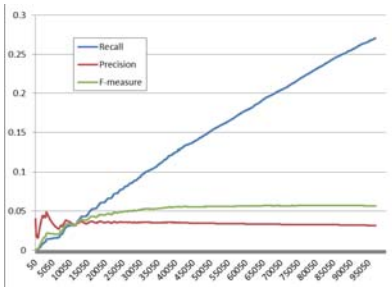

図 8 ベクトル空間モデル 1.

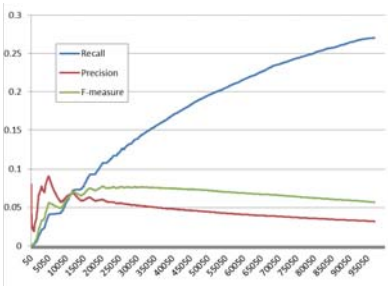

図 9 ベクトル空間モデル 2.

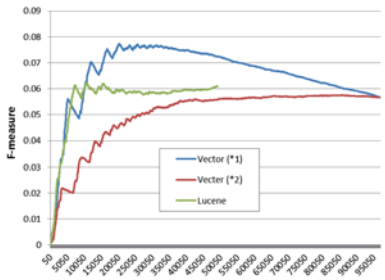

図 10 *Fmeasure* の比較.

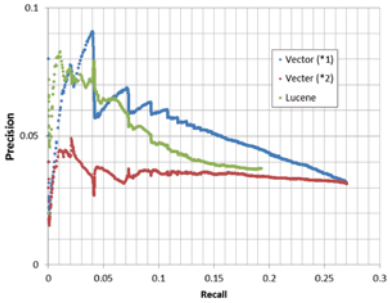

図 11 *Recall-Precision* 曲線の比較.

このことから，三つの手法のうち，ベクトル空間モデル（LogDF）が今回の手法に最適であることがわかる．この結果から，前述したとおり *Fmeasure* が最大になる上位 20,200 語の専門語候補を専門語として抽出した．また，これらの *Recall*, *Precision*, *Fmeasure* の値が 0.1 を満たない程低い理由は，正解セットに用いた AMS Glossary が，気象学専門語のサブセットであるためである．

4.2 気象学論文からの専門語抽出

次に，気象学論文からの専門語抽出の実験結果を説明する．

専門語候補の抽出に用いたデータセットは，AMS が 1996 年から 2009 年に発行した 12 誌 [15] と，日本気象学会が 1965 年から発行している気象集誌 [16]，2005 年から発行している Scientific Online letter of the Atmosphere [17] のそれぞれ最新刊の英語論文である．具体的には，AMS の論文数は 21,396，日本気象学会の論文数は 3,067 であり，計 24,463 の論文を対象としている．

英論文テキストセットからの抽出結果は，名詞句数 1,273,873，処理時間は 25 時間であった．この名詞句を専門語候補とし，残差 IDF を用いたソートを行う．

スコアリングとソートの結果を，図 12 から図 15 に示す．ソートした候補の配列の上位 N 語を横軸，縦軸を *Recall*, *Precision*, *Fmeasure* としている．

RIDF と RIDFLogIDF の，AMS Glossary を正解セットとしたときの *Recall*, *Precision*, *Fmeasure* の結果に共通しているは，RIDFLogIDF の方が高い数値を出しているということである．このことから，専門性の推定を行う専門語候補は，RIDFLogIDF によってソートされた候補配列を使う．

RIDFLogIDF によってソートされた専門語候補配列から，上位 20,000 語を専門性推定する．この 20,000 という数字は，*Fmeasure* が最大なる数字以上で，かつ，少数語で専門語推定を何度かテストすることで，現実的な処理時間で専門語推定できると判断した結果である．

4.3 専門性の推定と専門語の決定

前述の 20,000 の専門語候補の専門性推定を行った．検索結果ウェブページは上位 N 件，*slope* は 0.9，ページスコアの算出に用いる *KnownTermSet* は 4.1 節で取得した結果を用いた．20,000 語の専門性推定には，14 時間かかった．

図 16 に，20,000 語の専門性推定，スコア降順にソートし，その上位 N 語を候補セット，AMS Glossary を正解セットとした結果を示す．

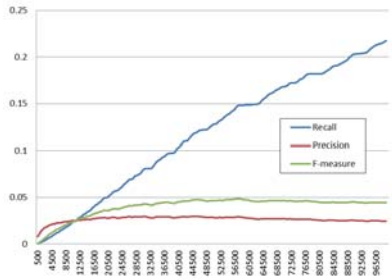

図 12 RIDF によってソートした専門語候補.

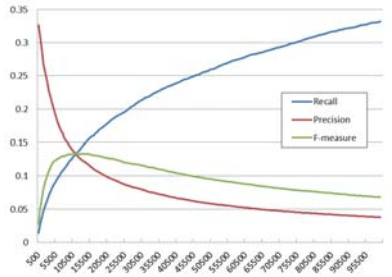

図 13 RIDFLogDF によってソートした専門語候補.

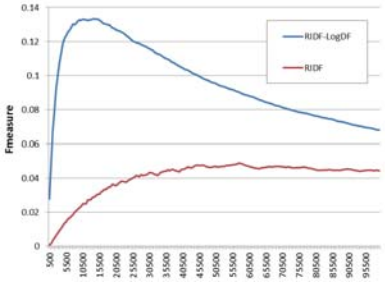

図 14 RIDF と RIDFLogDF の比較（*Fmeasure*）.

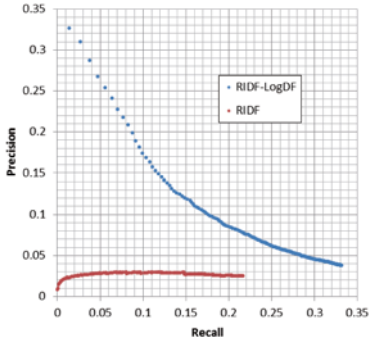

図 15 RIDF と RIDFLogDF の比較（***Recall-Precision*** 曲線）.

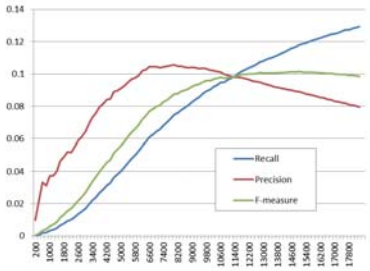

図 16 ウェブによる専門性推定の上位 N 語の ***Recalll, Precision, Fmeasure***

図 16 が示すとおり，*Precision* は上位 6,000 語から上位 8,000 語を候補セットとした時をピークに，それ以降は緩やかに減少している．*Fmeasure* は，上位 12,000 語を候補セットとした時をピークに，以降は横ばいになっている．

この結果から，*Precision* が最大となる上位 7,200 語を専門語として抽出した．

4.4 専門家による評価

専門家による判定によって，抽出した専門語の精度を評価した．評価は，筑波大学生命環境科学研究科の博士後期課程，前期課程の学生 7 名である．

評価の方法は，Wikipedia から抽出した専門語 20,200 語と，気象学論文から抽出した専門語 7,200 語から，AMS Glossary に登録済みの単語を除去，それぞれ 100 語を無作為抽出し，計 200 語を用意する．この 200 語の抽出を，学生の人数分行い，それぞれ 200 語が本当に専門語かどうか判断をした．

専門語か否か判断は大まかな基準として「気象学辞書を作るとしたらその語を辞書に加えるか否か」を挙げ，基本的には添削者の主観に任せた．しかし，どうしても判断に迷い，質問で挙がりそうな場合は具体的に指示を行った．具体的には下記である．

- “cloud and wind” のような，専門語の組み合わせになっている語は専門語とする
- 人名，組織名は気象学に関わる人物，組織であれば専門語とする
- “bright bands” のように，専門語の複数形は専門語とする
- “of tropical” や “forecast and” のように，名詞句としての形を成していない語は，非専門語とする

無作為抽出された 200 語のうち，専門語として判断された語の数を，添削者，抽出元別に表 1 に示す．

この結果から分かることは，まず添削者によって専門語として判断される数が大きく違うことである．これには二つの理由が考えられる．一つは，添削者毎に無作為抽出をしたためである．このため，添削者に評価してもらう専門語らしい語に偏りが出ている可能性がある．二つ目は，今回の評価方法は，採点者の主観が強く出てしまったためである．これは上記で

列挙した判断基準の具体例をより多くカバーすることにより，採点者間での結果の違いの差を埋められる可能性がある．

次に，専門語・非専門語として判断されたそれぞれの語が，それぞれの抽出手法でソートされていたときに上位何位にいたのかといった視点から，評価をする．表 2 は，Wikipedia，論文から抽出しそれぞれの専門性推定によりソートされた時，専門語・非専門語として判断された語は平均何位なのかをまとめたものである．

表 1 抽出した 200 の専門語中で専門語として判断された語数.

抽出元＼添削者	博士 A	博士 B	博士 C	修士 D	修士 E	修士 F	修士 G	平均
Wikipedia	70	45	34	39	59	35	51	47.57
気象学論文	87	54	34	35	77	37	56	54.29
合計	157	99	68	74	136	72	107	101.86

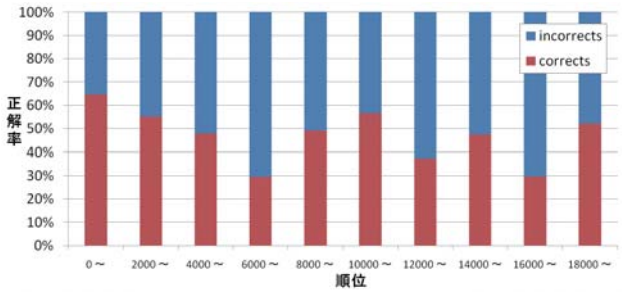

図 17 Wikipedia を用いた専門性推定によるランキング時の正解率分布.

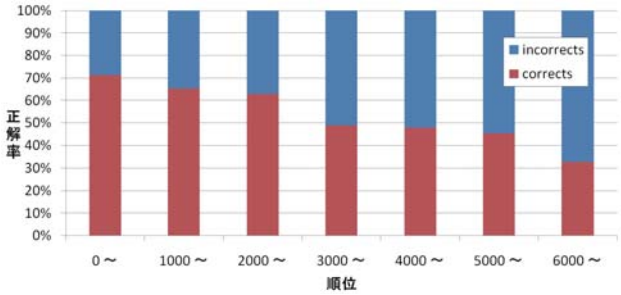

図 18 気象学論文を用いた専門性推定によるランキング時の正解率分布.

表 2 専門語，非専門語のそれぞれの平均ランキング.

抽出元	専門語	非専門語	抽出語数
Wikipedia	9542.27	10961.35	20,200
気象学論文	2994.20	3930.21	7,200

このことから，専門語として判断された語の方が，上位に位置していたことがわかる．

次に，専門語，非専門語として判断された語は，どの順位に分布するのかといった分析をする．図 17 と図 18 は，それぞれの抽出手法における，正解と判断された割合の分布を示している．この割合を正解率とする．

Wikipedia から抽出した専門語の順位と正解率の相関係数は -0.4143，論文から抽出した専門語の順位と正解率の相関係数は -0.9747 であった．このことから上位の語ほど，専門家に正解であると判断されていることがわかる．

以上から，Wikipedia を用いての専門性推定によるソートと，論文から抽出した名詞句に対し，ウェブを用いて専門性推定・ソートを行うことが有効であることが言える．

5 まとめ

本論文では，TermExtractor の手法をベースに，Wikipedia, 気象学論文，ウェブ，既存の専門語集を利用した，気象学用語の抽出手法の提案と実験を行った．1,000 万の Wikipedia 記事から，2,190 の気象学に関する記事を選定し，専門語候補として，96,927 語を抽出した．そこから，専門語候補の推定を行い，20,200 語を専門語として抽出した．気象学論文から英語論文のみを抽出し，24,463 の論文のテキストデータを用いた．このテキストデータから，1,273,873 の専門語候補を抽出し，残差 IDF を用いてソート，上位 20,000 語に対し専門性の推定を行った．結果，7,200 語を専門語として抽出した．気象学専攻の博士学生に，それぞれのテキストデータから抽出した専門語の精度を評価した．専門語非専門語が専門性推定によってソートされていたときに，どの順位にいたのかといった分布を見ることで，本手法の有効性を確認した．

今後の課題としては，ウェブを使った専門性推定をより多くの専門語候補に対して行う事，Wikipedia 記事や気象学論文

からの専門語抽出に用いるスコアの改良，図 17 において，正解率が上がっている順位，下がっている順位に存在する語について専門家からの意見を聞くなどが挙げられる．

謝辞

本研究を進めるにあたり，北海道大学堀之内武准教授，筑波大学田中博教授，日下博幸講師，海洋研究開発機構遠藤伸彦博士に貴重なご助言を頂いた．また，評価実験では，筑波大学秋元優子氏，池田亮作氏，高根雄也氏，石川真奈美氏，古橋奈々氏，田中翔太氏，加藤隆之氏にご協力頂いた．また，本研究の一部は科学研究費補助金（#21650017）の支援によるものである．ここに記して謝意を表す．

参考文献

1) Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Main Page
2) Sergey Brin. Extracting patterns and relations from the World Wide Web. In Proceedings of the 1998 International Workshop on the Web and Databases (WebDB '98), March 1998
3) Bollegala, Danushka Tarupathi and Matsuo Yutaka and Ishizuka Mitsuru. Relational duality: unsupervised extraction of semantic relations between entities on the web. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 151-160.
4) FRANTZI Katerina T, ANANIADOU Sophia, 辻井潤一 . 専門用語の自動抽出 . 情報処理学会研究報告 . 自然言語処理研究会報告 96(27), 83-88, 1996-03-14.
5) 中川裕志，湯本紘彰，森辰則 . 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.
6) P. Velardi, R. Navigli, P. D'Amadio. Mining the Web to Create Specialized Glossaries. IEEE Intelligent Systems, 23(5), IEEE Press, 2008, pp. 18-25.
7) Francesco Sclano, Paola Velard. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communitie. Conference TIA-2007, Sophia Antipolis, 8-9 octobre 2007.
8) American Meteorological Society. AMS Glossary. http://amsglossary.allenpress.com/glossary
9) OpenNLP. http://incubator.apache.org/opennlp/
10) Gerard Salton, Christopher Buckley. Term-weighting approaches in automatic text retrieval. Information Processing & Management, Volume 24, Issue 5, 1988, Pages 513-523.
11) G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing Communications of the ACM, vol. 18, nr. 11, pages 613-620.
12) LANCASTER, F.W. and FAYEN, E.G. Information Retrieval On-Line. Melville Publishing Company, Los Angeles, 1973.
13) MySQL. http://www.mysql.com/
14) 社団法人日本気象学会. http://wwwsoc.nii.ac.jp/msj/
15) American Meteorological Society. AMS Journals. http://journals.ametsoc.org/
16) 社団法人日本気象学会. 気象集誌. http://www.jstage.jst.go.jp/browse/jmsj
17) 社団法人日本気象学会. Scientific Online letter of the Atmosphere. http://www.jstage.jst.go.jp/browse/sola
18) 北研二，津田和彦，獅々堀正幹 . 情報検索アルゴリズム . 共立出版，2002 年 .
19) Aho, Alfred V., Margaret J. Corasick. Efficient string matching: An aid to bibliographic search. Communications of the ACM 18.