論　文

ウェブを利用した専門用語の分野判定

木田　充洋[†]　外池　昌嗣[††]　宇津呂武仁[†††a)]　佐藤　理史[††††]

Domain Classification of Technical Terms Using the Web

Mitsuhiro KIDA[†], Masatsugu TONOIKE[††], Takehito UTSURO[†††a)], and Satoshi SATO[††††]

あらまし　本論文では，用語の出現する文書中における分野の割合に基づいて，用語の分野判定を行う手法を提案する．用語の分野判定は，専門用語集の自動生成を行う上で必要な要素技術の一つである．提案手法では，判定対象の用語と，対象分野の既知の専門用語サンプルのみを入力として，ウェブを利用することにより，自動で用語の分野判定を行う．具体的には，判定対象の用語が出現する文書を収集し，これらに対して分野判定を行うことで，用語が一般の文書に比べて専門分野の文書に偏って用いられる度合を測定し，専門用語であるかどうかを判定する．評価実験の結果，7 割以上の精度，9 割前後の再現率で，用語の分野判定を行うことができることを確認した．また，本論文では，既存の辞書に未登録の用語をウェブから収集し，これに対して提案手法を適用することにより，当該分野の新たな用語が獲得できることを示した．これにより，提案手法が専門用語集の自動生成において有用な技術であることを示した．

キーワード　専門用語，ウェブ，用語抽出，分野判定

1. まえがき

近年，自然言語処理の研究は盛んに行われており，様々な言語資源が利用されている．一般の自然言語処理でよく用いられている言語資源の例としては，新聞記事コーパスなどが挙げられる．一方，専門文書の言語処理のための言語資源の整備は，相対的に遅れているといえる．例えば機械翻訳では，専門分野の用語辞書がないために専門用語の翻訳に失敗するといった問題がある．最近の機械翻訳ソフトでは，いくつかの分野において，人手で用意された専門用語辞書を同梱している場合もあるが，これらの専門用語辞書を種々の分野について人手で作成するには，コストがかかる．またコーパスについては，新聞記事のような一般のコーパスに，文書分類などの研究目的で分野タグが付けられたものは存在するが，専門分野のコーパスとして研究用に整備されたものは少ない．このように，今後，専門分野における自然言語処理の研究においては，言語資源の収集が必要となると考えられる．

専門分野における自然言語処理に必要な言語資源としては，主に (1) 専門分野コーパス，(2) 専門用語辞書の二つが考えられる．このうち，専門分野コーパスについては，既に文書分類などを目的として研究が多く行われており [4], [14]，これらの手法を応用することで，作成が可能であると考えられる．しかし専門用語辞書の作成においては，有効な手法の研究はあまり行われていない．専門用語に関する研究としては, [1] や [2] などが挙げられる. [1] や [2] では，分野コーパスでの出現頻度と一般コーパスでの出現頻度の比を用いて，用語がその分野の専門用語かどうかを判定している．しかし，これらの手法では，判定する専門分野の分野コーパスをあらかじめ人手で用意しており，分野の言語資源が乏しい状況ではすぐに利用できない．し

[†] 任天堂株式会社，京都市
Nintendo Co., Ltd., 11–1 Hokotate-cho, Kamitoba, Minami-ku, Kyoto-shi, 601–8116 Japan
[††] 京都大学情報学研究科知能情報学専攻，京都市
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto-shi, 606–8501 Japan
[†††] 筑波大学大学院システム情報工学研究科知能機能システム専攻，つくば市
Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba, 1–1–1 Tennodai, Tsukuba-shi, 305–8573 Japan
[††††] 名古屋大学大学院工学研究科電子情報システム専攻，名古屋市
Department of Electrical Engineering and Computer Science, Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, 464–8603 Japan
a) E-mail: utsuro@iit.tsukuba.ac.jp

たがって，専門分野におけるコーパスが用意されていないような任意の分野において適用できる専門用語辞書作成手法を確立する必要がある．

専門用語辞書を作るというタスクは，(1) まず専門用語の候補を収集し，(2) その候補がその分野の専門用語かどうか判定する，といった二つの過程に分けられる．このうち，(1) については，大まかにその分野の語と考えられる語を収集する研究が既に行われている．例えば [9] では，複合名詞を構成する要素間で統計的な結び付きの強さを測定し，要素間の結び付きの強い複合名詞を専門用語として抽出している．また, [12] では，構成要素間の結び付きの文法パターンを利用して，専門用語の候補を抽出している．しかし，(2) についてはあまり研究が行われていない. [9] や [12] の手法では，統計情報や文法パターンを用いることにより，用語としてのまとまりの強い複合名詞が抽出できるが，その用語が当該分野に属しているかどうかという観点での厳密な判定は行われていない．そこで本論文では，用語の分野を判定する手法を確立することを目的とする．

用語の分野判定を行う場合，まず判定する分野の情報を与える必要がある．ここで，分野の情報としては，文書や用語などが考えられる．本論文では，その分野について既知の専門用語をサンプルとして与えることにする．本論文が提案する用語の分野判定手法では，判定対象の用語について，まず用語の出現する文書を収集する．次に，これらの文書それぞれについて，与えられた専門用語サンプルにより自動で収集した当該分野の文書との間で類似度を計算することにより，文書の属する分野を判定する．そして，文書の分野判定結果に基づいて用語の分野を判定する．用語の出現する文書及び当該分野の文書を自動で収集する情報源として，本論文ではウェブを利用する．ウェブ上には，様々な分野の多岐にわたる情報が含まれており，これらを収集することにより，様々な分野において文書の形で多くの情報を収集することができる．本論文の手法では，ウェブを利用することにより，判定対象の用語と判定したい分野の専門用語のサンプルを与えるだけで，用語の分野を判定することができる．

以下，本論文ではまず，**2.** で提案手法の基本的な考え方について述べる．次に，**3.** で，ウェブを利用した用語の分野判定手法について具体的に説明し，実験においてその性能を評価する．更に **4.** では，提案手法を応用した，辞書未登録語の収集，及び，分野判定方法について述べ，実験において，その性能を評価する．また，**5.** において，関連研究について紹介する．

2. 用語の分野判定

2.1 専門分野・一般コーパスにおける頻度分布に基づく方法

用語が出現する文書を利用して，用語の分野判定を行う研究としては, [1] がある(注1)．

[1] は，ある分野に対して，用語の「専門用語らしさ」を表す指標として，その分野の文書からなる専門分野コーパス及び，内容に分野の偏りがない一般コーパスの二つにおける，用語の出現頻度の分布を用いている．用語が当該分野の専門用語であるならば，一般コーパスでの出現頻度に比べて，専門分野コーパスでの出現頻度が十分に大きくなることが予想される．これに対し，当該分野の専門用語でないような用語では，一般コーパスでの出現頻度に比べて，専門分野コーパスでの出現頻度は大して大きくはならないと考えられる．したがって，専門分野・一般コーパスでの出現頻度の比の値によって，用語がどの程度，当該分野に属するかを表すことができる．この比の値に適当なしきい値を設けることで，用語の分野判定を実現できる. [1] は，以上の方法により，解剖学の分野において分野判定実験を行った．実験では，専門分野コーパスとして解剖学の教科書を，また一般コーパスとして Lancaster-Oslo/Bergen（LOB）コーパスと Wellington コーパスを用いた．これらのコーパスは人手であらかじめ作成されたものである．

この方法には，以下の欠点がある．

- コーパスに十分な頻度で出現しない語の判定が難しい．
- 人手でコーパスを用意する必要がある．

[1] では，解剖学の教科書を専門分野コーパスとして与えているが，この教科書にある程度以上の頻度で出現しない用語については，適切な分野判定を行うことは難しい．専門分野コーパスとしてどのような文書を使用するか，また専門分野コーパスをどのような方法で作成するかについては，いろいろな方法が考えられる．しかし，専門分野コーパスがどのような文書から構成されていても，その中に低頻度でしか出現しないような専門用語は存在する．これらの用語に対して，

(注1)：このほかの研究として，用語の構成要素のそれぞれに対して同様の方法を適用し，用語の分野判定を行っている [2] などがある．

専門分野・一般コーパス間の頻度分布の比較によって適切に分野判定をすることは，容易ではない．

2.2 文書の分野の割合に基づく方法

前節で述べたように，専門分野・一般コーパス間の頻度分布に基づく方法では，コーパスに十分な頻度で出現しない用語に対する分野判定が難しいという問題があった．この問題に対処するため，本論文では，用語が出現する文書の分野の割合に基づく分野判定手法を提案する．

用語がどのような専門分野の語であるのかは，その用語が出現する文書の分野から考えることができる [5]．用語がある専門分野の語であるとき，その用語は主にその専門分野の文書において使用される．逆に，用語が一般語である場合，その用語は様々な分野の文書に使用されると考えられる．

図 1 に，用語とそれが出現する文書の専門分野についての例を示す．「インピーダンス特性」という用語は，電気分野の専門用語である．この用語が出現する文書は，そのほとんどが電気分野のトピックに関する文書であり，このことからこの用語は電気分野でしか用いられない専門用語であることが分かる．また，「電磁気現象」という用語が出現する文書は，電気分野のトピックに関する文書がある程度の割合で含まれており，この用語もまた電気の専門用語であると考えられる．しかし，ここには電気分野のトピック以外の文書もある程度の割合で含まれている．このような用語は，電気分野以外の分野においても使用される可能性のある用語であることが分かる．実際のところ，「電磁気現象」は地震予知のトピックにおいて重要な用語であり，文書集合中にも地震のトピックの文書がある程度

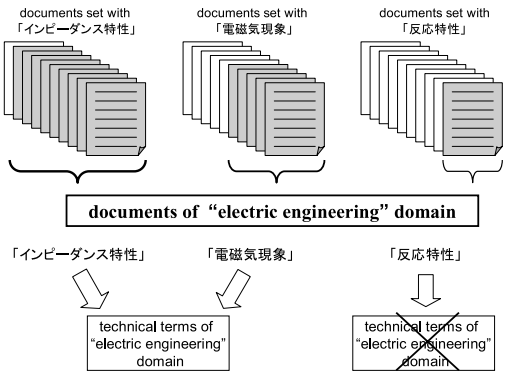

図 1 文書の分野に基づく用語の専門性
Fig. 1 Degree of specificity of a term based on the domain of the documents.

含まれている．一方，「反応特性」という用語は，電気分野に限らず広く使用される一般語である．この用語を含む文書集合中には，電気分野の文書はほとんど含まれておらず，その他のトピックの文書が大半を占める．

このような，用語とそれが出現する文書の分野に関する関係を利用して，本論文では以下のような用語の分野判定手法を提案する．

まず，判定対象の用語に対して，それが出現する文書のサンプルを用意する．次に，サンプル文書のそれぞれについて，専門分野コーパスとの間で内容の近さを測ることで，文書の分野を判定する．そして，サンプル文書中の，当該分野の内容を含む文書の割合に基づいて，用語の分野を判定する．

以上の方法で用語の分野判定を行うためには，(1) 判定対象の用語が出現する文書のサンプルと，(2) 専門分野コーパスの二つが必要である．本論文では，多様な分野において，広範囲にわたる用語について分野判定を行うことを目標としている．ここで，任意の分野，用語に対して，これらの文書資源を人手で用意するのは，非常にコストがかかる．そこで，本論文では，これらを自動で収集するための情報源として，ウェブを利用する．ウェブ上には，多種多様な専門分野の情報が存在する．また，ウェブ上の文書は日々更新されており，既存の辞書に載っていないような新語や，最新のトピックに関する情報を含んでいる．そこで，ウェブを利用することにより，多くの用語，分野に対して，これらの文書資源を自動で収集することができる．ウェブを利用して用語の分野判定を行う具体的な方法については，次章で述べる．

3. ウェブ文書を用いた用語の分野判定

3.1 概　　要

本論文においては，用語の分野判定の問題を，ある分野 C において，判定対象の用語 t が，C の専門用語であるかどうかを判定するタスクとして設定する．本論文では，用語 t の分野 C に対する専門性の度合 $g(t, C)$ を，以下の 3 段階で定義する．そして，用語 t が出現する文書中の分野の割合の値に基づいて，用語 t の分野をこの 3 段階で判定する．

$$g(t,C)=\begin{cases} + & (t\text{は，}C\text{に属する文書にのみ出現する}) \\ \pm & (t\text{は，}C\text{に属する文書，属さない文書の双方に出現する}) \\ - & (t\text{は，}C\text{に属する文書には出現しない}) \end{cases}$$

ここで，本論文では，用語の専門性が‘+’または‘±’となる用語を分野 C の専門用語であるとし，専門性が‘−’となる用語は分野 C の専門用語ではないとする．例えば図 1 において，「インピーダンス特性」のような，その分野でのみ用いられる用語は，専門性が‘+’の用語である．また，「電磁気現象」のような用語は，専門性が‘±’の用語であり，これらの‘+’，‘±’に属する用語は電気分野の専門用語であるとする．一方，「反応特性」のような用語は，専門性が‘−’の用語であり，電気分野の専門用語ではないとする．専門用語のうち，専門性が‘+’である語と‘±’である語の違いは，その用語が当該専門分野のみで用いられるのか，それとも，他の分野でも用いられるのかの違いである．これらの語の区別は，分野判定技術を用いた応用を考える場合に有用となる．例えば，当該分野の文書を集める場合に，専門性が‘+’である用語を含む文書を集めればよい．

本論文の手法では，情報源としてウェブを利用する．本手法では，入力として判定対象の用語 t と，分野 C の既知の専門用語集合 T_C の二つのみを与えることとする．そして，用語 t の分野判定において必要な，(1) 用語 t が出現する文書，(2) 分野 C のコーパス，の二つをウェブから自動で収集する．

本論文における用語の分野判定タスクの入出力を以下に示す．また，分野判定の流れを図 2 に示す．

入力	判定対象の用語 t
	分野 C の既知の専門用語集合 T_C
出力	用語 t の分野 C に対する専門性 $g(t,C)$

用語の分野判定は (a) 分野 C のコーパス D_C を作成するプロセス，(b) D_C を用いて用語 t の分野判定を行うプロセス，の二つに分けられる．以下に，それぞれのプロセスについて詳しく説明する．

3.2 専門分野コーパスの作成

ここでは，入力された既知の専門用語集合 T_C を用いて，以下の手順でウェブから専門分野コーパス D_C を作成する．

(1) T_C 中の各語 t に対して，「t」「t は」「t の」「t とは」「t という」の 5 種類のクエリ $q_i(t)$ $(i=1,\ldots,5)$ をサーチエンジンに入力し，収集される文書集合 $D(q_i(t))$ の和集合を求めることにより t を含む文書の集合 D_t を得る．

$$D_t=\bigcup_{i=1,\ldots,5} D(q_i(t))$$

そして，それらの和集合を $D(T_C)$ とする．

$$D(T_C)=\bigcup_{t\in T_C} D_t$$

(2) $D(T_C)$ 中の文書から，T_C 中の語を多く含む順に，指定した文書数だけを選んで文書集合を作成し，これを専門分野コーパス D_C とする．

サーチエンジンとしては，goo(注2)を用いている．クエリとして「t」以外に付属語を付加して検索することにより，t に関して詳しく記述されているページが収集しやすくなる．また，(2) において，T_C 内の語を多く含む文書を選ぶ理由は，t を含む文書が必ずしも分野 C の文書であるとは限らないので，その中でもその分野について書かれている可能性の高い文書だけを用いるためである．

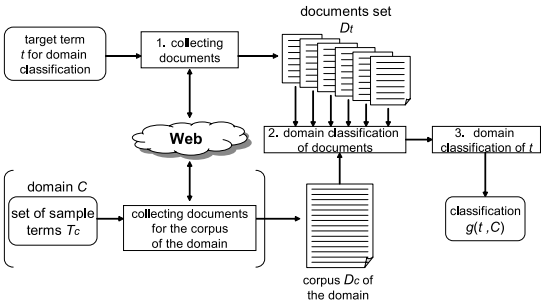

図 2 ウェブ文書を用いた用語の分野判定
Fig. 2 Domain classification of terms based on Web documents.

3.3 用語の分野判定

前節のプロセスにおいて収集した専門分野コーパス D_C を用いて，用語 t の分野判定を行う．用語 t の分野判定は，次の三つのステップで行う．

ステップ 1 用語 t を含む文書をウェブから収集し，文書集合 D_t を構成する．

ステップ 2 D_t 中の各文書 d_t と専門分野コーパス D_C の類似度を計算し，類似度が下限値 L 以上となる文書の集合 $D_t(C,L)$ を構成する．

(注2)：http://www.goo.ne.jp/

ステップ 3　二つの文書集合 $D_t(C,L)$ と，及び，D_t の文書数の割合を計算し，これに基づいて，用語 t の専門性 $g(t,C)$ を 3 段階で判定する．

以下に，それぞれのステップについて詳しく説明する．

3.3.1　判定対象の用語を含む文書の収集

専門分野コーパス作成時と同様に，「t」「t は」など 5 種類のクエリを用いて，t を含む文書をウェブから収集し，最大 n 文書からなる文書集合 D_t を構成する．ただし，サーチエンジンにより得られるページの中には，テキストとしての情報をあまりもたないようなサイズの小さいページや，画像ばかりのページなどが含まれる．そこで，収集したページのテキスト部分の文字数が 500 未満のページは文書として使用しないようにしている．

3.3.2　文書の分野判定

次に，文書集合 D_t 中の各文書 d_t と専門分野コーパス D_C の間の類似度を測定し，類似度が下限値 L 以上となる文書の集合 $D_t(C,L)$ を構成する．文書間の類似度を計算する手法としては，文書の単語の頻度ベクトル（以下，文書ベクトル）の余弦を利用した方法を用いる(注3)．

具体的な文書の分野判定手順を以下に述べる．まず，専門分野コーパス D_C を一つの文書 d_C とみなして，文書ベクトル $dv(d_C)$ を作成する．また，t を用いて収集された文書集合 D_t 中の各文書 d_t についても，文書ベクトル $dv(d_t)$ を作成する．

これらの文書ベクトルは，以下の条件をすべて満たす形態素列を次元とし，その頻度を値とすることで構成している．形態素解析には Juman（ver4.0）[8] を用いている．

- 形態素数が 5 以下．
- 各形態素の品詞が接頭辞，名詞，動詞，カタカナ(注4)のいずれか．ただし，各形態素のうちの少なくとも一つの品詞は接頭辞でない．
- 形態素列を連結した語がストップワードリスト(注5)に含まれない．

次に，文書ベクトル間の余弦 $\cos(dv(d_t),dv(d_C))$ を計算し，これを文書 d_t と専門分野コーパス D_C の間の類似度 $sim(d_t,D_C)$ とする．

$$sim(d_t,D_C)=sim(d_t,d_C)=\cos(dv(d_t),dv(d_C))=\frac{dv(d_t)\cdot dv(d_C)}{|dv(d_t)||dv(d_C)|}$$

そして，$sim(d_t,D_C)$ の値が下限値 L 以上となるような d_t を分野 C に属する文書であると判定し，文書集合 $D_t(C,L)$ に含める．

$$D_t(C,L)=\{d_t|sim(d_t,D_C)\geq L\}$$

専門分野コーパス D_C と文書 d_t の類似度がどの程度ならばその文書が分野 C に属するのかという境界は，使用した専門分野コーパス D_C や分野 C 自体の特性によって異なる．したがって，本論文では，パラメータ調整用用語集合を用意して，類似度下限値 L を経験的に定めている．

3.3.3　用語の分野判定

最後に，用語 t の分野判定を行う．まず，**3.3.1** で収集した文書集合 D_t と，**3.3.2** で得た文書集合 $D_t(C,L)$ の文書数の割合 r_L を求める．

$$r_L=\frac{|D_t(C,L)|}{|D_t|}$$

そして，r_L の値に二つの判定境界 $a(\pm)$ 及び $a(+)$ を設け，分野 C に対する用語 t の専門性 $g(t,C)$ を 3 段階で判定する．

$$g(t,C)=\begin{cases}+ & (a(+)\leq r_L)\\ \pm & (a(\pm)\leq r_L<a(+))\\ - & (r_L<a(\pm))\end{cases}\qquad(1)$$

ここで用いる判定境界 $a(+)$ 及び $a(\pm)$ についても，文書の類似度下限値 L と同様に，パラメータ調整用用語集合によって決定している．

3.4 評　　価

本節では，用語の分野判定実験を行い，提案手法の性能を評価する．

3.4.1　対 象 分 野

「電気工学」「光学」「航空宇宙工学」「核工学」「天文学」の 5 分野を対象として評価実験を行った．

(注3)：文書の分野判定手法に関する研究では，機械学習を用いる方法 [4] が盛んに研究され，文書分類タスクなどにおいて利用されている．機械学習を用いる手法によって，高精度で文書の分野判定を行うことが可能となるが，正例及び負例の 2 種類の文書集合を用意して学習を行わなくてはならないため，訓練データを用意するためのコストが高い．本論文の目的は「用語」の分野判定であるため，ここでは手法にかかるコストを優先し，文書ベクトルの余弦を利用したより簡便な方法を採用することにする．

(注4)：Juman（ver4.0）の解析結果では，カタカナは品詞分類「未定義語」の中に含まれる細分類として扱われている．

(注5)：1 文字の平仮名や形式名詞などの頻出語の中から事前に人手で定めた 153 語からなる．

3.4.2 専門分野コーパスの作成

各々の分野について，既存の専門用語辞書(注6)のエントリから，用語 t_C を無作為に 100 語選び，既知の専門用語集合 T_C を作成した(注7)．そして，T_C 中の用語 t_C を入力として，各用語につき最大 500 文書を収集し，この中から，T_C 内の語を異なりで多く含む上位 500 文書だけを選んで，分野コーパス D_C を作成した．これは，上位 500 文書よりも下位の文書を人手で調べたところ，当該分野の内容を含む文書は 3〜4 割程度で，当該分野以外の文書が大半を占めていたためである．今回使用した，上位 500 文書からなる専門分野コーパス D_C における，当該分野の内容を含む文書の割合は，8〜9 割程度であった(注8)(注9)．

3.4.3 判定対象とする用語・文書

実験用の用語集合として，それぞれ 100 語からなるパラメータ調整用用語集合 T_{dev}，評価用用語集合 T_{eval} の二つを，分野ごとに作成した．ここで用語集合に用いた用語は，以下の条件を満たす語を候補として，専門用語とそうでない語の双方が十分な割合で含まれるように選んだ．

- 汎用対訳辞書「英辞郎 ver79(注10)」のエントリに含まれる．
- 既存の専門用語辞書のエントリに含まれない．
- 文書集合 $D(T_C)$ での総出現頻度が 5 以上．
- goo におけるヒット数が 100 以上 10,000 未満．

「英辞郎 ver79」は，約 129 万語を収録している非常に大規模な対訳辞書であり，そのエントリには一般語だけでなく様々な分野の専門用語が含まれていることが確認されている．このような大規模な汎用辞書のエントリに専門分野の情報を付与することは，専門用語集を生成する上では重要なタスクである．本項の評価実験においては，本論文で提案する分野判定手法を，汎用辞書エントリの分野判定タスクに適用することを想定して，上記の基準に基づいて用語集合を選んだ．なお，実際に汎用対訳辞書「英辞郎」エントリの分野判定タスクに提案手法を適用した結果については，[6] にて詳細を述べている．

分野判定対象の用語 t が出現する文書のサンプル D_t の収集においては，文書数 $n = 100$ として文書を収集した．文書数については，予備実験において，100 文書よりも少ない文書を用いた場合には，分野判定の性能が安定しなかった．したがって，本実験では，性能が安定していることが確認されている 100 文書を用いた．

表 1　実験用用語集合の用語数
Table 1　Number of terms for experimental evaluation.

	それぞれの専門性に属する用語の数					
	調整用集合 T_{dev}			評価用集合 T_{eval}		
	＋	±	−	＋	±	−
電気工学	43	14	43	48	20	32
光学	35	15	50	40	24	36
航空宇宙工学	39	10	51	36	24	40
核工学	22	24	54	34	28	38
天文学	41	12	47	35	15	50

3.4.4 評価尺度

作成した用語集合 T_{dev}，T_{eval} には，当該分野について人手により専門性 '＋'，'±'，'−' のいずれかを付与した．これらの用語の語数を表 1 に示す．T_{dev} あるいは T_{eval} について，それぞれの専門性を付与された用語集合を $T_{ref}(+)$，$T_{ref}(\pm)$，$T_{ref}(-)$ と表す．また，$T_{ref}(+)$ と $T_{ref}(\pm)$ の和集合により構成される用語集合を，$T_{ref}(+ \cup \pm)$ と表記することにする．これらの用語集合は，T_{dev}，T_{eval} のそれぞれに対して作成される．

更に，**3.3.3** の式 (1) に従って，実験用用語集合 T_{dev} 及び T_{eval} のそれぞれに対して，専門性 '＋'，'±'，'−' が自動判定された用語の集合をそれぞれ $T_{sys}(+)$，$T_{sys}(\pm)$，$T_{sys}(-)$ と表す．また，$T_{sys}(+)$ と $T_{sys}(\pm)$ の和集合により構成される用語集合を，$T_{sys}(+ \cup \pm)$ と表記することにする．提案手法の評価は，実験用用語集合 T_{dev}，T_{eval} のそれぞれについて，以下の式で計算される精度と再現率によって行う．

$$\text{精度} \qquad P_+ = \frac{|T_{sys}(+) \cap T_{ref}(+)|}{|T_{sys}(+)|}$$

(注6)：本論文では，既存の専門用語辞書として 2 種類の辞書（106 分野，約 12 万 6 千語収録のものと 23 分野，約 19 万語収録のもの）を用いている．
(注7)：既知専門用語を選ぶ際には，goo での検索ヒット数を基準として，十分な文書数を集められ，かつ一般語としての用法が大きな割合を占めないと考えられる用語を対象とした．また，既知専門用語の数を 100 語よりも少なくした場合（10〜70 語）でも，用語の分野判定の性能が急激に劣化するということはなかった．ただし，少数の既知専門用語を効果的に選定する方法については，今後の研究の課題である．
(注8)：使用する文書を上位 500 文書よりも少なくして専門分野コーパスを作成した場合でも，コーパス中の当該分野の文書の割合は，500 文書の場合とあまり変わらなかった．また，これらの専門分野コーパスを使用して用語の分野判定実験を行ったところ，500 文書よりも少ない文書数での性能は，上がる場合や下がる場合などがあり，一貫した結果は得られなかった．
(注9)：自動収集した文書集合から，人手で当該分野の文書を選ぶことで作成した高品質な専門分野コーパスを用いて，用語の分野判定性能を評価することも行ったが，自動的に作成した専門分野コーパスを用いた場合と比較して，性能の差はほとんど見られなかった．
(注10)：http://www.alc.co.jp/

表 2　判定境界 $a(\pm)$ における分野判定の精度と再現率
Table 2　Precision/recall of domain classification with threshold $a(\pm)$.

	L	$a(\pm)$	調整用集合 T_{dev}		評価用集合 T_{eval}	
			精度 $P_{+\cup\pm}$	再現率 $R_{+\cup\pm}$	精度 $P_{+\cup\pm}$	再現率 $R_{+\cup\pm}$
電気工学	0.2	0.4	0.96 (54/56)	0.95 (54/57)	0.95 (59/62)	0.87 (59/68)
光学	0.2	0.4	0.94 (49/52)	0.98 (49/50)	1.00 (60/60)	0.94 (60/64)
航空宇宙工学	0.2	0.4	0.94 (42/44)	0.86 (42/49)	0.79 (54/68)	0.90 (54/60)
核工学	0.25	0.2	0.92 (36/39)	0.78 (36/46)	0.95 (60/63)	0.97 (60/62)
天文学	0.15	0.4	0.96 (51/53)	0.96 (51/53)	0.86 (48/56)	0.96 (48/50)

$$P_{+\cup\pm} = \frac{|T_{sys}(+\cup\pm) \cap T_{ref}(+\cup\pm)|}{|T_{sys}(+\cup\pm)|}$$

再現率　$R_{+} = \dfrac{|T_{sys}(+) \cap T_{ref}(+)|}{|T_{ref}(+)|}$

$$R_{+\cup\pm} = \frac{|T_{sys}(+\cup\pm) \cap T_{ref}(+\cup\pm)|}{|T_{ref}(+\cup\pm)|}$$

3. 4. 5　文書類似度しきい値及び文書割合しきい値の決定

文書類似度下限値 L 及び，当該分野に属する文書数の割合のしきい値 $a(+)$，$a(\pm)$ の値は，パラメータ調整用用語集合 T_{dev} を用いて決定する．T_{dev} に対して，文書類似度下限値を $L = 0.05, 0.1, \ldots, 0.3$，文書割合しきい値を $a(+)$，$a(\pm) = 0.1, \ldots, 0.9$ と変化させて用語の分野判定を行い，以下の式で計算される重み付き F 値の値が最も大きくなるときのしきい値の組合せを選択した．

重み付き F 値　$F_{+} = \dfrac{1}{\alpha\dfrac{1}{P_{+}} + (1-\alpha)\dfrac{1}{R_{+}}}$

$$F_{+\cup\pm} = \frac{1}{\alpha\dfrac{1}{P_{+\cup\pm}} + (1-\alpha)\dfrac{1}{R_{+\cup\pm}}}$$

$$(0 \le \alpha \le 1)$$

本論文では，当該分野の用語を高精度で自動収集するという用途を想定し，用語分野判定の再現率よりも精度を重視して，$\alpha = 0.75$ とした．

3. 4. 6　評 価 結 果

本項では，パラメータ評価用用語集合 T_{dev} を用いてしきい値を決定し，評価用集合 T_{eval} を用いて用語の分野判定を行い，その性能を評価した．

(1)　用語集合 $T_{ref}(+\cup\pm)$ における評価

ここでは，用語が当該分野の専門用語であるかどうかの判定における，システムの性能を評価する．ここでの判定境界は $a(\pm)$ である．

$a(\pm)$ を判定境界として分野判定を行った場合の，2 種類の用語集合における精度と再現率を表 2 に示す．結果を見ると，精度については，どちらの用語集合においてもほぼすべての分野において，9 割以上の高い性能を示している．また，再現率についても，ほぼ 8 割から 9 割以上の値を得られている．これらの結果から，本手法による分野判定によって，当該分野の専門用語を高い性能で判定できることが分かる．

なお，誤って専門用語と判定されてしまった語としては，以下のような原因によるものがあった．

［専門分野コーパスの純度の問題］

作成した専門分野コーパス中での当該分野の文書の割合は 8〜9 割であった．しかし，1〜2 割含まれる，当該分野以外の文書が原因となって誤判定をしてしまう場合があった．例として，電気分野の専門用語と誤判定された「未利用エネルギー」がある．この用語は，電気に限らず広くエネルギー関係の文書に出現する語である．この用語が誤判定された理由として，既知の専門用語集合 T_C に，「原子力発電」などの発電関係の用語が含まれていることが挙げられる．これらの語により収集されるエネルギー関係の文書が，専門分野コーパスに含まれたことにより，エネルギー関係の文書を，電気分野の文書であると誤判定したと考えられる．

［トピックが関連する分野の語］

判定対象の語が，判定する分野と関連する分野の用語であるような場合に，誤判定となることがあった．例として，航空宇宙工学の専門用語と判定された「軍事大国化」などが挙げられる．「軍事大国化」を含む文書での，分野に属する文書の割合は $r_L = 0.47$ であった．この用語が出現する文書は，トピックとしては軍事関係の文書が大半を占めるが，軍事関係の文書の中には，戦闘機やロケットなど，航空宇宙分野に関係のある用語が多く出現する．これにより，軍事関係の文書は，航空宇宙工学の分野に属すると誤判定されたことが原因と考えられる．なお，評価用用語集合 T_{eval} において，航空宇宙工学での精度が 8 割弱と低い性能を示しているのは，このような軍事関係の用語が T_{eval}

表 3　判定境界 $a(+)$ における分野判定の精度と再現率
Table 3　Precision/recall of domain classification with threshold $a(+)$.

	L	$a(+)$	調整用集合 T_{dev}		評価用集合 T_{eval}	
			精度 P_+	再現率 R_+	精度 P_+	再現率 R_+
電気工学	0.2	0.7	0.97 (32/33)	0.74 (32/43)	0.92 (24/26)	0.50 (24/48)
光学	0.2	0.7	0.83 (20/24)	0.57 (20/35)	0.82 (23/28)	0.58 (23/40)
航空宇宙工学	0.2	0.5	0.90 (28/31)	0.72 (28/39)	0.53 (27/51)	0.75 (27/36)
核工学	0.25	0.3	0.55 (18/33)	0.82 (18/22)	0.57 (32/56)	0.94 (32/34)
天文学	0.15	0.7	0.89 (34/38)	0.83 (34/41)	0.87 (33/38)	0.94 (33/35)

に多く含まれていたことが原因である．

［特定の分野によく出現する一般語］

一般語でありながら，ある分野においてよく用いられるため，その分野の専門用語であると誤判定されてしまう場合があった．例として，核工学の専門用語と判定された「異常事象」が挙げられる．「異常事象」を含む文書の集合において，核工学分野に属すると判定された文書の割合は $r_L = 0.78$ であった．この用語は，核工学に限らず広く用いられる用語であるが，近年よく報道される原発事故関連のトピックの文書ではこの語が頻繁に用いられている．そのため，この用語が出現する文書をウェブから収集すると，原発事故関連の文書が大きな割合を占めてしまい，この用語を核工学の用語であると判定してしまったことが誤判定の原因である．

(2)　用語集合 $T_{ref}(+)$ における評価

ここでは，専門性が ‘+’ となる用語，すなわち当該分野でのみ用いられるような専門性の高い用語の判定について，その性能を評価する．ここでの判定境界は $a(+)$ である．

$a(+)$ によって用語の分野判定を行った場合の，2 種類の用語集合における精度と再現率を表 3 に示す．

結果を見ると，両方の用語集合での核工学，及び，評価用集合での航空宇宙工学を除く場合において，8 割からほぼ 10 割近くの精度で専門性 ‘+’ の用語を判定することができている．このように，判定境界 $a(+)$ の場合は，精度においては十分に高い性能を示している．一方，再現率では，判定境界 $a(\pm)$ の場合に比べて，多くの分野で 2〜3 割低い値を示している．このことから，専門性が ‘+’ となるような専門性の高い用語を高精度で選ぶことはできているが，その一方で本来は専門性 ‘+’ であるが，選びそこねて専門性 ‘±’ と判定してしまっている用語が多いことが分かる．

既に述べたように，専門用語の中でも専門性が ‘+’ である用語を区別する利点は，専門性 ‘+’ の用語を用いることで，分野の情報を容易に収集できることである．この場合は，専門性が ‘+’ の専門用語を，十分高い精度で一定数収集できればよく，網羅する必要はない．この目的においては，提案手法は十分有用であるといえる．

また，精度が低くなった分野については，以下の理由が考えられる．

核工学分野では，人手で専門性 ‘±’ と判定した用語の中に，原爆等の放射能被曝関係の用語や，放射線医療関係の用語が多く含まれる．これらの用語は，原爆関連のトピックの文書や，医療関係のトピックの文書にも出現する．しかしこれらのトピックの文書の中には，自動判定により核工学分野の文書と判定されたものが多い．その結果，自動判定によって，それらの用語を専門性 ‘+’ であると判定したことが原因と考えられる．

評価用集合での航空宇宙工学分野の精度が低い理由は，調整用集合と比べて軍事関係の用語数が多く，これらの用語の専門性の度合の判定に失敗したことが原因と考えられる．

4. ウェブから収集した辞書未登録語の分野判定

本章では，これまでに述べた用語の分野判定手法を用いて，辞書未登録語の分野判定を行う．

専門用語集作成のタスクの中で，用語集に含めるべき専門用語は，(1) 既存の汎用辞書に登録されている用語，(2) 既存の汎用辞書に登録されていない用語，の 2 種類に分けて考えることができる．このうち，(1) の用語に対する分野判定については, [6] で詳しく述べている．ここでは，(2) の用語を収集し，分野判定を行うことで，既存の辞書には登録されていない専門用語を獲得することを試みる．このような用語を収集し，専門用語として認定する技術の確立は，まだ辞書が作成されていない最新のトピックなどに対して専門用語集を作成できるだけでなく，年々進歩する様々な分野に対して，辞書を更新するコストを大きく削減させる

ことができるという点で，非常に有用である．

4.1 概　　要

ここでは，分野 C について，ウェブから辞書未登録語を収集し，分野判定を行うことで，分野 C における辞書未登録の専門用語を獲得する．本節では，このタスクの入出力を以下のように定義する．また，入力から出力までの流れを図 3 に示す．

入力	分野 C の既知の専門用語集合 T_C
出力	C の専門用語と判定された用語集合 $T_{web,C}$

まず，分野 C の既知の専門用語集合 T_C を入力としてウェブから文書を収集し，辞書に登録されていない専門用語の候補語を文書中から抽出する．こうして収集された候補語集合に対して，構成要素を利用したフィルタリングを行い，当該分野の専門用語である可能性の高い語を残す．ここでは，ウェブを用いた分野判定処理には時間コストがかかることを考慮して，低コストで適用可能なフィルタリングにより，あらかじめ対象用語数を絞っている．そして，フィルタを通過した語のそれぞれに対して，前節で述べた用語の分野判定手法を適用し，分野 C の用語であると判定された語の集合 $T_{web,C}$ を得る．

以下に，それぞれの手順について詳しく説明する．

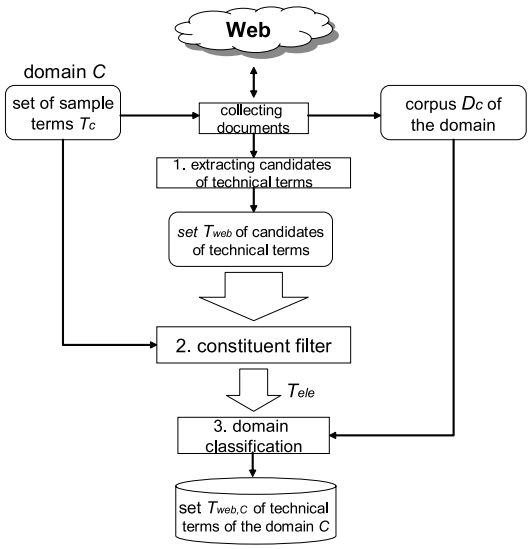

図 3　ウェブを利用した辞書未登録語の分野判定
Fig. 3　Web based domain classification of terms not included in existing lexicons.

4.2 手順の詳細

4.2.1 専門用語候補抽出

ここでは，入力された分野 C の既知の専門用語集合 T_C を用いて，専門用語候補を収集する方法を説明する．

用語の分野判定では，まず専門用語集合 T_C から，**3.2** の手順で専門分野コーパス D_C を作成する．このとき，専門用語集合 T_C から収集した文書集合 $D(T_C)$ が得られている．これらの文書集合には，分野 C における多岐にわたるトピックの文書が含まれており，その中では辞書未登録の専門用語が使用されていることが期待できる．

そこで，$D(T_C)$ 中から以下の条件をすべて満たす用語を，辞書未登録の専門用語候補として抽出し，専門用語候補語集合 T_{web} を得る．

- 2 形態素以上から構成される複合名詞．
- $D(T_C)$ での総頻度が 5 以上．
- 既存の専門用語辞書及び汎用辞書のエントリ(注11)に含まれない．

専門用語候補を D_C からではなく，$D(T_C)$ から抽出する理由は，専門分野コーパス D_C には選ばれなかった文書の中にも，低い割合ではあるが専門分野の文書は存在し，その中に専門用語が含まれている可能性があるため，そのような用語を獲得できるようにするためである．

4.2.2 構成要素フィルタ

ここでは，構成要素フィルタについて説明する．

まず，入力として与えた既知の専門用語集合 T_C 中の用語を形態素に分割し，T_C 中の用語の構成要素の形態素集合 $U(T_C)$ を作成しておく．次に，候補語集合 T_{web} 中の各用語 t について，t を形態素に分割し，t の構成要素である形態素のうち少なくとも一つが $U(T_C)$ に含まれる用語だけを残して，候補語集合 T_{ele} を構成する．

フィルタリングの方法としては，汎用コーパスを用いる方法など，ほかにもいくつかの方法が考えられる．その中でも構成要素フィルタは，入力の用語以外に言語資源を必要としないので，コストのかからない非常に簡便なフィルタである．したがって，本論文では構成要素フィルタを用いることにしている．

4.2.3 用語の分野判定

構成要素フィルタを通過した候補語集合 T_{ele} 中の各

(注11)：**3.4** で用いたものと同じものを用いている．

用語 t の分野判定を行う．そして，専門性が '+' あるいは '±' と判定された語を分野 C の専門用語であるとして，そのような t からなる集合 $T_{web,C}$ を出力する．

4.3 評　　価

本節では，**4.2** で述べた，(1) 構成要素フィルタ，(2) 分野判定，の二つの項目に対して評価を行った．対象とする分野及び入力として与える既知の専門用語集合 T_C は，**3.4** で用いた 5 分野，100 語とした．また，分野判定において，用語 t が出現する文書のサンプル D_t としては，100 文書を収集して用いた．分野判定時に必要な，文書類似度下限値 L 及び，判定境界 $a(\pm)$ の値についても，**3.4** でパラメータ調整用用語集合 T_{dev} を用いて決定した値を用いた．

4.3.1 フィルタの性能

ここでは，構成要素フィルタの性能評価を行う．評価は，構成要素フィルタ前後での候補語数及び，その中に含まれる専門用語数の推定値によって行った．ここで，専門用語数の推定値は，候補語集合から 500 語を無作為に選び，人手で調査した結果を用いて推定している．

構成要素フィルタ前後での専門用語候補の数，及び，その中に含まれる専門用語数の推定値を表 4 に示す．表 4 から，構成要素フィルタによって，専門用語候補の数はおよそ 4〜7 分の 1 に減少していることが分かる．一方，フィルタ後の専門用語数は，フィルタ前の 3 分の 2 程度である．このことから，構成要素フィルタによって，専門用語の多くを残しつつ，そうではない用語を大幅に候補語集合から除外できていることが分かる．したがって，構成要素フィルタは専門用語を残すフィルタとして効率が良いことが確認できる．

4.3.2 分野判定の性能

ここでは，分野判定の性能評価を行う．構成要素フィルタによって得られた候補語集合 T_{ele} の中から，文書集合 $D(T_C)$ の頻度が高頻度であった 1,000 語について，分野判定を行った(注12)．これらの 1,000 語に対して，あらかじめ，人手で当該分野の用語かどうかを判定しておいた．これを用いて，**3.4.3** で述べたように，判定境界 $a(\pm)$ における精度 $P_{+\cup\pm}$ 及び再現率 $R_{+\cup\pm}$，また判定境界 $a(+)$ における精度 P_+ 及び再現率 $R_\pm$ を計算し，これによって分野判定の性能を評価した．

各分野 1,000 語に対する分野判定の精度と再現率を表 5 に示す．また，本手法により，当該分野の用語であるとして出力された辞書未登録語の例（電気分野）を表 6 に示す．

今回，分野判定の対象としている用語の中には，当該分野ではよく用いられる語であるが，専門用語の単位としては適切ではないと考えられる語が含まれている．表 6 において，「△」で示した語がこれにあたる．例えば，「予備変圧器」という語は，「予備」として用意されている「変圧器」という意味であり，この語は専門用語の単位としては適切ではないと考えられる．表 5 における精度と再現率は，このような語も含めて，当該分野でよく用いられる用語をすべて正解として計算した値である．また，表 5 における「用語の単位が適切な割合」は，上記の基準で正解とした出力中の用語における，専門用語の単位という点からも適切な語の占める割合を表している．

表 5 から，判定境界 $a(\pm)$ においては，7 割弱から 8 割弱の精度で分野判定を行うことができていることが分かる．また再現率は，7 割強以上の値であり，最も高い分野では約 95%という高い値となっていることが分かる．誤判定されている語のほとんどは，**3.4.6** で述べた原因によるものであり，本手法では原理的に判定が難しい．したがって，このような語に対しては，フィルタリング等，分野判定以外の技術によって除外する必要がある．

判定境界 $a(+)$ では，ほぼ 6 割から 7 割 5 分の精度で分野判定を行うことができている．また再現率は，約 85%から 95%以上という高い値となっている．表 3 における精度と比較して，それほど高い精度は得られていないが，表 5 での判定境界 $a(\pm)$ での精度と比較すると，妥当な値であると考えられる．なお，航空宇宙工学と核工学では，他分野に比べて低い精度を示しているが，これは **3.4.6** で述べた原因と同様の原因

表 4　構成要素フィルタによる専門用語候補語数の変化
Table 4　Changes in number of technical term candidates with constituent element filter.

	構成要素フィルタ前 T_{web}		構成要素フィルタ後 T_{ele}	
	候補語数	専門用語数（推定）（%）	候補語数	専門用語数（推定）（%）
電気工学	24,460	1,272 (5.2)	6,623	848 (12.8)
光工学	29,090	1,047 (3.6)	6,985	866 (12.4)
航空宇宙工学	41,279	660 (1.6)	6,364	458 (7.2)
核工学	40,439	890 (2.2)	10,834	650 (6.0)
天文学	29,240	1,170 (4.0)	5,491	659 (12.0)

(注12)：予備実験により，分野判定の対象となる候補語集合 T_{ele} 中の語では，$D(T_C)$ での頻度が高い方が，当該分野の専門用語である可能性が高いことが確認されている．

表 5　辞書未登録語の分野判定における精度と再現率
Table 5　Precision/recall of domain classification of terms not included in existing lexicons.

	判定境界 $a(\pm)$			判定境界 $a(+)$		
	精度 $P_{+\cup\pm}$	再現率 $R_{+\cup\pm}$	用語の単位が適切な割合	精度 P_{+}	再現率 R_{+}	用語の単位が適切な割合
電気工学	0.754 (399/529)	0.828 (399/482)	0.393 (157/399)	0.697 (168/241)	0.853 (168/197)	0.494 (83/168)
光学	0.766 (454/593)	0.875 (454/519)	0.368 (167/454)	0.743 (234/315)	0.932 (234/251)	0.453 (106/234)
航空宇宙工学	0.797 (408/512)	0.739 (408/552)	0.402 (164/408)	0.666 (277/416)	0.936 (277/296)	0.502 (139/277)
核工学	0.685 (470/686)	0.953 (470/493)	0.377 (177/470)	0.580 (362/624)	0.981 (362/369)	0.406 (147/362)
天文学	0.747 (480/643)	0.945 (480/508)	0.475 (228/480)	0.763 (350/459)	0.888 (350/394)	0.520 (182/350)

によるものである．

また，表 5 において，「用語の単位が適切な割合」は約 4 割であった．このことから，今回分野の判定が正しく行われた用語の中には，「予備変圧器」のように，専門用語の単位としては適切ではないと考えられる語が多数含まれていることが分かる．このような語を除外するには，用語としての単位の認定を厳密に行う必要がある．今回の方法では，単に複合名詞と考えられる品詞列を抽出しているだけで，用語としての単位の認定を厳密には行っていないため，このような語を誤認定してしまう．そこで，このような用語を専門用語と区別し，除外する技術が必要である．

用語としての単位を認定する方法としては，用語を構成する各形態素がどのような語と連接し得るかを考慮して，用語の単位を認定し，収集する [9] や [12] などが知られている．今後の課題は，そのような用語の単位を厳密に認定する手法を適用することで，専門用語の単位として適切な辞書未登録語を収集し，専門用語集の自動生成を実現することである．

5. 関 連 研 究

本論文では，ウェブを用いて自動で用語の分野判定を行う手法を提案した．用語の分野判定を行っている研究の中でも，特に本論文と関連の深い研究として, [1] や [2] が挙げられる．これらの研究では，人手で作成した専門分野コーパス・一般コーパス間における用語の出現頻度の比を利用することで，用語の分野判定を行っている．これらの研究は，専門用語が一般の文書で出現することは少なく，専ら専門分野の内容の文書において使用されるという特性を利用している点で，本研究と共通している．しかし，これらの手法では，判定する専門分野の分野コーパスをあらかじめ人手で用意しており，分野の言語資源が乏しい状況における適用可能性が低い．

自動で専門用語集の生成を行っている研究としては，ウェブから関連用語を収集する手法を用いる [10] がある．この手法では，一つの用語を中心としてその語の周りに用語集を構築する．これに対して本論文では，複数の語により規定される分野のモデルを構築し，そのモデルが対応できる範囲で用語集を生成しており，異なったアプローチをとっている. [10] では，ウェブ上のヒット数を利用して語と語の関連度を測る手法を用いている．この考え方を本論文のタスクに取り入れる方法の一つとして，既知専門用語との間の関連度を併用した用語分野判定尺度を導入することが挙げられる．また，既知専門用語と判定対象の語が共起して出現する文書を収集し，この文書の分野判定を行うことにより，用語の分野判定の性能を向上させることなどが挙げられる．

分野のモデルを構築し，分野判定に利用している研究としては, [15] や [3] が挙げられる. [15] では，多言語のウェブディレクトリから構築した多言語分野モデルを利用して，言語横断情報検索の検索質問の分野判定を行い，言語横断情報検索の性能向上を実現している. [3] では，複数分野の専門用語辞書中の見出し語を構成する用語の分布を用いて，分野のモデルを構築し，用語の定義を解説する用語説明の分野判定を行っている．これらの研究は，いずれも，検索質問あるいは用語説明の構成語を用いて，これらの構成語の組と分野モデルとの間での分野判定を行うことにより，分野判定対象の分野を直接判定している．これに対して，本論文では，判定対象の用語が出現する文書を複数収集し，まず，それぞれの文書と分野モデルとの間の関連性を測ることで，文書の分野判定を行う．そして，収集された文書群における分野の分布を用いることで，間接的に「用語」の分野判定を行っており, [15] や [3] とは分野判定の枠組みが異なっている．

専門分野の中には，言語資源の作成や，それを利用した言語処理技術の研究が盛んに行われている分野もある．例として，生命科学分野が挙げられる．この分

表 6 「電気工学」における辞書未登録語の分野判定出力例
Table 6 Examples of domain classification of terms not included in existing lexicons. (“Electrical Engineering” domain)

出力された専門用語候補	文書の分野の割合 r_L（$L = 0.2$）	人手での判定	
陰極箔	1.00	±	
カソード電圧	0.96	+	
誘導性負荷	0.96	+	
一定電流	0.95	+	△
相互誘導回路	0.94	+	
両端電圧	0.94	+	△
電圧帰還管	0.93	+	
フェーザ表示	0.93	+	
負帰還増幅回路	0.92	+	
系統側	0.91	−	
電流値	0.91	+	△
電流定格	0.90	+	△
純抵抗	0.89	+	
差動増幅回路	0.89	+	
整流後	0.89	+	△
出力電力	0.88	+	△
進相コンデンサ	0.87	+	
最大出力電流	0.86	+	△
交流波形	0.86	+	△
低力率	0.85	−	
回路例	0.85	+	△
誘電体中	0.84	+	△
誘電体損失	0.83	+	
基準電圧源	0.83	+	
圧電体	0.82	+	
直流分巻電動機	0.82	+	
アナログ電話用設備	0.82	−	
高電圧化	0.81	+	△
一次側コイル	0.81	+	△
電気二重層キャパシタ	0.81	+	
直角座標表示	0.80	−	
光電面	0.79	±	
電気工作物	0.79	+	△
電気影像法	0.77	+	
電力用半導体	0.77	+	△
測温抵抗体	0.76	−	
消光係数	0.76	−	
電子機器用固定抵抗器	0.75	±	△
電荷減衰測定	0.75	+	
オン抵抗	0.74	+	
圧電特性	0.73	±	
回路計算	0.73	+	△
受電設備	0.73	+	△
非直線性	0.73	−	
圧電セラミックス	0.72	+	
分極処理	0.72	+	
本質安全回路	0.72	+	△
固定子巻線	0.71	+	
電荷移動遷移	0.71	±	
分極反転	0.71	±	
はんだ接続部	0.71	±	△
組合せ論理回路	0.71	−	
契約負荷設備	0.70	±	△
予備変圧器	0.70	+	△
半固定抵抗	0.70	+	

「△」が付与されている語は，当該分野で用いられるが，辞書に登録すべき用語の単位としては不適な語．

野では，人手，自動を問わず広く言語資源を作成する研究が行われており，多くの言語資源が整備されている [7], [13]．また，これらの言語資源を利用して，生命科学分野に特化した多くの言語処理技術の研究が行われている [11]．

6. むすび

本論文では，用語の出現する文書中における分野の割合に基づいて，用語の分野判定を行う方法を提案した．提案手法では，ウェブ上から文書を収集することにより，判定対象の用語と，当該分野の既知の専門用語のサンプルのみを入力として，用語の分野判定を自動で行うことができる．評価実験では，7 割以上の精度，9 割前後の再現率で，用語の分野判定を行うことができることを確認した．また，本論文では，既存の辞書に未登録の用語をウェブから収集し，分野判定を行った．これにより，専門用語集の自動生成において提案手法が有用な技術であることを示した．

提案手法は，ある一つの分野に対して，用語の分野判定を行う手法である．一方，複数分野に属する専門用語については，各分野における出現の分布をとらえる必要があると考えられる．この問題については，提案手法により，各分野に対して独立に分野判定を行い，複数分野における専門性の度合の分布を測定することで，対処できると考えている．ただし，判定対象の用語が出現する文書のうちの多数が当該分野以外に属す場合は，本論文の手法をそのまま適用しても適切な分野判定を行えない可能性がある．そのような場合には，前章で，文献 [10] の手法との関連において述べたように，既知専門用語をアンカーとして用いて，既知専門用語と判定対象の語が共起して出現する文書を収集し，この文書の分野判定を行うことにより，用語の分野判定の性能を向上させる手法が有効であると考えられる．

文　献

[1] T.M. Chung, “A corpus comparison approach for terminology extraction,” Terminology, vol.9, no.2, pp.221–246, 2004.

[2] P. Drouin, “Term extraction using non-technical corpora as a point of leverage,” Terminology, vol.9, no.1, pp.99–117, 2003.

[3] 藤井　敦，石川徹也，“World Wide Web を用いた事典知識情報の抽出と組織化，” 信学論（D-II），vol.J85-D-II, no.2, pp.300–307, Feb. 2002.

[4] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms, Springer-Verlag, 2002.

[5] 影浦　峡，“「専門用語の理論」に関する一考察，”情報知識学会誌，vol.12, no.1, pp.3–12, 2002.
[6] 木田充洋，外池昌嗣，宇津呂武仁，佐藤理史，“ウェブを利用した専門用語の分野判定，”言語処理学会第 12 回年次大会発表論文集，pp.388–391, 2006.
[7] 小池麻子，高木利久，“生命科学文献からの知識抽出と辞書構築，”情報処理，vol.46, no.2, pp.123–129, 2005.
[8] 黒橋禎夫，河原大輔，日本語形態素解析システム JUMAN version 4.0，東京大学大学院情報数理工学系研究科，2003.
[9] 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.
[10] 佐々木靖弘，佐藤理史，宇津呂武仁，“ウェブを利用した専門用語集の自動編集，”言語処理学会第 11 回年次大会発表論文集，pp.895–898, 2005.
[11] 平　博順，前田英作，“バイオ自然言語処理のための機械学習技術，”情報処理，vol.46, no.2, pp.130–136, 2005.
[12] 竹内孔一，影浦　峡，ダイユ　ベアトリス，小山照夫，“多言語専門用語抽出モデルの構築，”言語処理学会第 11 回年次大会発表論文集，pp.887–890, 2005.
[13] 建石由佳，大田朋子，辻井潤一，“バイオ NLP のためのコーパスと各種リソースの現状，”情報処理，vol.46, no.2, pp.130–136, 2005.
[14] 上田修功，“テキストモデリングの新展開，”言語処理学会第 9 回年次大会チュートリアル資料，2003.
[15] 木村文則，前田　亮，宮崎　純，吉川正俊，植村俊亮，“Web ディレクトリを言語資源として利用した言語横断情報検索，”情報処理学会論文誌：データベース，vol.45, no.SIG7 (TOD22), pp.208–217, 2004.

（平成 18 年 3 月 10 日受付）

木田　充洋

2004 京大・工・電気電子卒．2006 同大大学院情報学研究科修士課程知能情報学専攻了．現在，任天堂（株）勤務．在学中は，自然言語処理の研究に従事．

外池　昌嗣

2001 京大・工・情報卒．2003 同大大学院情報学研究科修士課程知能情報学専攻了．現在，同大学院情報学研究科博士後期課程在学中．自然言語処理の研究に従事．

宇津呂武仁　（正員）

1989 京大・工・電気工学第二学科卒．1994 同大大学院工学研究科博士課程電気工学第二専攻了．京都大学博士（工学）．奈良先端科学技術大学院大学情報科学研究科助手，豊橋技術科学大学工学部情報工学系講師，京都大学情報学研究科知能情報学専攻講師を経て，2006 より筑波大学大学院システム情報工学研究科知能機能システム専攻助教授．自然言語処理の研究に従事．

佐藤　理史

1983 京大・工・電気工学第二学科卒．1988 同大大学院博士課程研究指導認定退学．京都大学工学部助手，北陸先端科学技術大学院大学情報科学研究科助教授，京都大学情報学研究科助教授を経て，2005 より名古屋大学大学院工学研究科教授．工博．自然言語処理，情報の自動編集等の研究に従事．