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[†]　　佐藤　理史[††]

[†] 京都大学 情報学研究科　　[††] 名古屋大学大学院 工学研究科

1. はじめに

専門用語集は，その分野に従事する人にとっては欠かせない資源のひとつであり，これを整備することが必要とされている．しかし，専門用語は様々な分野において日々増え続けており，これらを逐一人手によって用語集に収録するのは大変なコストがかかる．そこで，専門用語集を自動で生成，更新する技術が必要になる．

専門用語辞書を生成のタスクは，(1) 専門用語の候補を収集し，(2) 候補が当該分野の専門用語かどうか判定する，といった 2 つの過程に分けられる．(1) に関しては，重要語の抽出手法 [1] などが知られている．(2) については自動で行うことを前提にした研究はあまりない．(2) に関する研究としては，Chung[2] や Drouin[3] があげられる．これらの研究では，専門用語が一般の文書で出現することは少なく，もっぱ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．具体的には，人手で作成した専門分野コーパス・一般コーパスの間での用語の出現頻度の比を用いて用語の分野判定を行っている．しかしこれらの手法では，コーパスに十分な頻度で出現しない用語の分野判定が困難である，コーパスを人手で用意する必要がある，などの問題がある．

そこで本論文では，次のような手法を提案する．まず，判定対象の用語に対して，用語が出現する文書のサンプルを収集し，それぞれの文書に対して，専門分野コーパスとの間で内容の近さを測ることで，文書の分野判定を行う．そして，文書サンプル中での，当該分野の文書の割合を利用して，用語の分野判定を行う．ここで，用語が出現する文書のサンプルおよび，専門分野コーパスを自動で収集するため，本研究ではウェブを情報源として利用する．ウェブ上には多種多様な分野の情報があふれており，これを利用することで，専門分野コーパスが整備されていないような多くの分野に対して，提案手法による用語の分野判定を行うことが可能になる．提案手法では，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を入力として与え，上記の文書をウェブから収集することにより，自動で用語の分野判定を行う．

このような用語の分野判定技術の適用例として，本論文では，大規模汎用日英対訳辞書「英辞郎☆」の日本語エントリについて分野の判定を行うというタスクを取り上げる．「英辞郎」は，約 129 万語 (ver79) を収録している非常に大規模な辞書であり，そのエントリには様々な専門用語も含まれている．実験では「英辞郎」の日本語エントリについて実際に分野判定を行い，その結果を評価した．

☆ http://www.alc.co.jp/

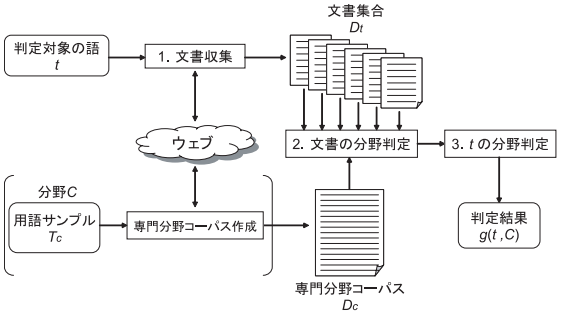

図 1　ウェブ文書を用いた用語の分野判定の流れ

2. ウェブ文書を用いた用語の分野判定

2.1 概　　要

本研究で扱う分野判定は，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問題である．本研究では，用語 t の分野 C に対する専門性の度合い $g(t,C)$ 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もとづいて，用語 t 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，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．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'+' である用語を用いることで，分野の文書を容易に集めることができる．

提案手法による，分野 C における用語 t の分野判定の入出力を以下に示す．また，分野判定の流れを図 1 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 2 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．

2.2 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 種類のクエリをサーチエンジンに入力し，t を含む文書の集合 D_t を収集する．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，500 文書を選び，これを専門分野コーパス D_C とする．

サーチエンジンは，goo☆を用いている．また，クエリとして「t」以外に付属語を付加して検索するのは，t に関して詳しく記述されているページを得られやすいという理由による．また，(2) で T_C 内の語を多く含む 500 文書を選ぶ理由は，分野 C の文書以外の文書を用いないようにするためである．

2.3 用語の分野判定

前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 3 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 中の各文書と専門分野コーパス D_C の類似度を計算し，分野 C に属する文書からなる集合 $D_t(C, L)$ を構成する．

ステップ 3 $D_t(C, L)$ と，D_t の文書数の割合を計算し，これにもとづいて，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．

2.3.1 判定対象の用語を含む文書収集

判定プロセスでは，まず判定対象の用語 t を含む文書をウェブ上から収集する．具体的には，専門分野コーパス作成時と同様に，「t」「t は」など 5 種類をクエリとして，t を含む文書を収集し，最大 100 文書からなる文書集合 D_t を構成する．

2.3.2 文書の分野判定

ここでは，収集された文書集合 D_t 中の各文書に対して，専門分野コーパス D_C との間の類似度を用いて分野判定を行い，分野 C の文書のみからなる文書集合 $D_t(C, L)$ 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) を利用した方法を用いる．

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．

次に，これらの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，分野 C に属する文書の集合 $D_t(C, L)$ に含める．

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

D_C と d_t の類似度がどの程度ならば d_t を分野 C の文書といえるのかは，D_C を構成する文書や分野自体の特性に依存する．本研究では，パラメーター調整用用語セットを用意して，類似度下限値 L を経験的に定めている．

2.3.3 用語の分野判定

最後に，用語 t の分野判定を行う．まず，2.3.1 節で収集した文書集合 D_t と，2.3.2 節で得た分野 C に属する文書の集合 $D_t(C, L)$ 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 2 つの判定境界 $a(\pm)$ および $a(+)$ をもうけ，用語 t の専門性 $g(t, C)$ を 3 段階で判定する．

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

ここで用いる判定境界 $a(+)$ および $a(\pm)$ についても，本研究では文書の類似度下限値 L と同様に，パラメーター調整用用語セットによって決定している．

3. 汎用対訳辞書エントリの分野判定

3.1 概　　要

汎用対訳辞書エントリの分野判定を，以下の入出力からなるタスクであると定義する．また，入力から出力までの流れを，図 2 に示す．

入力	汎用対訳辞書の日本語エントリ集合 T_{dic} 分野 C の既知の専門用語集合 T_C
出力	分野 C の専門用語であると判定されたエントリ集合 $T_{dic,C}$

まず，汎用対訳辞書の日本語エントリ集合 T_{dic} に対して，(1) コーパスフィルタ，(2) 構成要素フィルタ，の 2 つのフィルタリングを行い，分野 C の専門用語である可能性の高い語だけを残す．そして，フィルタを通過したエントリに対して，(3) 分野判定，を行うことで，分野 C の専門用語エントリの集合 $T_{dic,C}$ を出力として得る．以下にそれぞれの手順について詳しく述べる．

3.2 コーパスフィルタ

用語の分野判定では，既知の専門用語集合 T_C から，2.2 節の手順で専門分野コーパス D_C が作成される．このと

☆ http://www.goo.ne.jp/

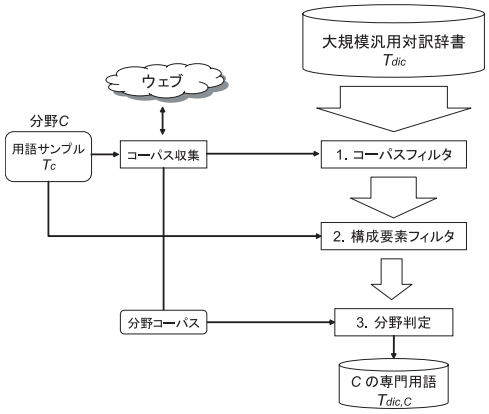

図 2 汎用対訳辞書エントリの分野判定の流れ

き，専門用語集合 T_C により収集した文書集合 $D(T_C)$ が得られている．そこで，これを用いてエントリのフィルタリングを行う．具体的には，判定対象のエントリ集合 T_{dic} 中の各エントリ t について，$D(T_C)$ における出現頻度 $f(t)$ が下限値 f_0☆以上となる t のみを残す．

フィルタリングで D_C ではなく，$D(T_C)$ を用いる理由は，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分野 C の文書は存在し，その中に一定の割合で分野 C の専門用語が含まれるからである．

3.3 構成要素フィルタ

まず，既知の専門用語集合 T_C 中のすべての専門用語を形態素に分割し，形態素集合 $U(T_C)$ を作成しておく．そして，コーパスフィルタを通過した各エントリ t のうち，$U(T_C)$ に含まれる形態素を構成要素として持つ t のみを残す．

3.4 用語の分野判定

以上のフィルタを通過した各エントリ t に対して，2 章で述べた手法により，用語の分野判定を行う．そして，分野判定の結果，専門性が'+' または'±' と判定された語を分野 C の専門用語であるとし，そのような t を持つエントリからなる集合 $T_{dic,C}$ を構成し，これを出力する．

4. 実　　験

本節では,「英辞郎 ver79」の日本語エントリに対して，分野判定実験を行う．評価は (1) フィルタの性能，(2)2 章で提案した分野判定手法の性能の 2 点について行う．

4.1 実験の概要

対象とするエントリ集合 T_{dic} は,「英辞郎 ver79」(約 129 万エントリ) の日本語エントリのうち，2 形態素以上からなるもので，既存の専門用語辞書エントリ☆☆に含まれていないものとした．このようなエントリの総数は $|T_{dic}| = 274,784$ であった．

判定を行う分野は,「電気工学」「光学」「航空宇宙工学」「核工学」「天文学」の 5 分野とし，入力として与える既知の専門用語 T_C は，既存の専門用語辞書エントリから，各分野 100 語ずつ選んで使用した．

☆ 本論文の実験では $f_0 = 5$ としている．

☆☆ 本研究では 2 種類の専門用語辞書 (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) を用いた．

表 1 フィルタリングに使用するコーパスの大きさ

	コーパスサイズ (Mbyte)	文書数 $\|D(T_C)\|$
電気工学	62.4	8,179
光学	82.0	10,263
航空宇宙工学	114.9	12,090
核工学	106.6	10,939
天文学	95.2	10,448

表 2 フィルタリングに使用する構成要素

	構成要素数 $\|U(T_C)\|$	構成要素の例
電気工学	164	抵抗，荷，極
光学	143	偏光，フィルター，放射
航空宇宙工学	161	飛行，機，推進
核工学	146	線，吸収，熱
天文学	138	座，平均，星雲

表 3 フィルタによるエントリ数の変化

	コーパスフィルタ後		構成要素フィルタ後	
	エントリ数	推定専門語数	エントリ数	推定専門語数
電気工学	8,503	969(11.4)	1,776	604(34.0)
光工学	9,546	783(8.2)	1,579	477(30.2)
航空宇宙工学	13,542	704(5.2)	1,869	602(32.2)
核工学	13,568	706(5.2)	2,934	534(18.2)
天文学	8,471	424(5.0)	990	206(20.8)

カッコ内の数値はその時点での総エントリ数に対する推定専門用語数の割合 (%)
推定専門語数は，無作為に選んだ 500 語について人手で調査した結果から推定．

4.2 フィルタの性能評価

判定対象のエントリ T_{dic} に対し，コーパスおよび構成要素によるフィルタリングを行った．ここで用いたコーパス $D(T_C)$ の大きさを表 1 に示す．また，各分野 100 語の T_C から得られた構成要素集合 $U(T_C)$ 中の構成要素数とその例を表 2 に示す．

これらのフィルタリングにおける，エントリ数の変化と，それぞれの時点での推定専門用語数を，表 3 に示す．表 3 から，総エントリ数は，入力エントリ数 $|T_{dic}|$ に対して，コーパスフィルタによって 20〜30 分の 1，さらに構成要素フィルタによって 5〜15 分の 1 に減少していることがわかる．このことから，フィルタによって判定対象エントリ数を大幅に減少させることができていることがわかる．また，構成要素フィルタの前後において，総エントリ数における推定専門用語数の割合が約 1 割から 4 割へと変化している．このことから，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よいことが確認できる．

4.3 分野判定の性能評価

ここでは，フィルタリングを通過したエントリに対して用語の分野判定を行うことで，提案手法の性能を評価した．判定対象の用語は，フィルタリングを通過したエ

表 4 「英辞郎」の日本語エントリに対する分野判定性能

	L	a(±)	精度	再現率
電気工学	0.2	0.4	0.711(150/211)	0.882(150/170)
光学	0.2	0.4	0.878(130/148)	0.861(130/151)
航空宇宙工学	0.2	0.4	0.739(136/184)	0.845(136/161)
核工学	0.25	0.2	0.704(88/125)	0.967(88/91)
天文学	0.15	0.4	0.687(101/147)	0.971(101/104)

精度のカッコ内の数値は，(出力中の正解数/出力数) を示す.
再現率のカッコ内の数値は，(出力中の正解数/対象中の正解数) を示す.

ントリから選んだ各分野 1,000 語の用語とした．これらの用語に人手で正解を付与し，提案手法の判定結果の精度と再現率によって，提案手法を評価した．なお，提案手法は，用語を'+'，'±'，'−' の 3 段階で判定するが，ここでは'+' と'±' を区別せず，専門用語であるか否かという観点での判定性能のみを評価した．

評価セットにおける分野判定の精度と再現率を表 4 に示す． 表 4 から，提案した用語の分野判定手法は 7 割から 9 割弱の精度を示していることが分かる．また，再現率は 8 割台半ばから，ほぼ 10 割に近い値を示している．これらのことから，専門用語については，ほぼすべての語について，正しく分野判定を行うことができており，それ以外の語について，一部誤って専門用語であると判定してしまっていることがわかる．

誤判定の原因としては，主に以下のようなものがあった．

(1) 専門分野コーパスに他分野の文書が含まれる場合
(2) 共通のトピックを持つ分野の用語
(3) 特定の分野によく出現する一般語

(1) の例としては「未利用エネルギー」があげられる．この語は，電気分野のコーパスにエネルギーに関する文書が含まれたことにより，電気用語であると誤判定された．(2) の例としては「精密誘導爆弾」があげられる．航空宇宙工学と軍事工学は航空機やロケット技術において内容や用語が共通するために，本手法では，2 つの分野を同じ分野であると判定し，この語を航空宇宙用語だと誤判定した．(3) の例としては「異常事象」があげられる．この語は専門用語ではないが，原発事故の話題で頻繁に用いられたために，収集した文書の大半が原発関連の文書であり，核工学用語であると誤判定された．

(1) の原因による場合は，コーパス作成方法の改良により改善できる余地がある．(2) と (3) の場合は，本手法では原理的に判定が難しい．したがって，フィルタリング等，分野判定以外の技術によってこれらの語を除外する必要がある☆.

5. おわりに

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もとづいて，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，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．本研究では専門用語集の自動生成を目指しており，その枠組において用語の分野判定は重要な要素技術の一つである．本論文では，提案手法によってこの技術を確立した．

評価実験では，提案手法が 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，この実験により，専門用語集の自動生成のタスクのひとつである，汎用辞書エントリの分野判定について，用語の分野判定技術を利用して実現できることを示した．

専門用語集の自動生成を行っている関連研究として，ウェブから関連用語を収集する手法 [4] がある．この手法では，一つの用語を中心としてその語のまわりに用語集を構築する．これに対して本研究では，複数の語により規定される分野のモデルを構築し，そのモデルに対応する範囲で用語集を生成するという，異なったアプローチをとっている．

今回扱った，汎用辞書エントリの分野判定の他に，提案手法を応用した例として，辞書未登録の専門用語の獲得があげられる．既存の辞書に未登録の用語を収集し，提案手法により分野判定を行うことによって，新規の専門用語として獲得することができる．我々が行った実験では，自動で収集した文書から抽出した各分野 1,000 語の評価セットに対して提案手法を適用することにより，150〜200 語程度の辞書未登録語を専門用語として獲得することができることを確認している．

☆ 英辞郎エントリの分野判定では，英語訳の情報を用いてコーパス，構成要素フィルタを両言語で行うことにより，より効果的に専門用語以外の語を除外できることを確認している．

参 考 文 献

1) 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).
2) Chung, T. M.: A corpus comparison approach for terminology extraction, *Terminology*, Vol. 9, No. 2, pp. 221–246 (2004).
3) Drouin, P.: Term extraction using non-technical corpora as a point of leverage, *Terminology*, Vol. 9, No. 1, pp. 99–117 (2003).
4) 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).