専門分野における用語の分野基礎性に関する研究

内山 清子
国立情報学研究所
〒東京都千代田区一ツ橋2－1－2

kiyoko@.nii.ac.jp

1　はじめに

特定の専門分野を新たに学ぼうとする人には，単なる知識として情報を得る，学部学生が専門分野を探す，異なる分野の研究者が他分野を学ぶ，企業の技術者がシステム実装に必要な手法を学ぶなど，多種多様な目的と対象者がいる．これらの目的のために読む文書には多くの専門用語が含まれ，その用語理解が重要な枠割を果たす．専門外の人がその専門用語を理解するために，専門用語辞書や解説書で調べることができるが，その用語が，どの程度の重要性を持っているのかが明確ではない．つまり，その専門用語が分野において，どれほどの重要性，基礎性，先端性を持っているのかがわからないため，どの専門用語に焦点を当てて学べばよいのか判断できない．また，概要を把握したい場合は，全ての用語を網羅する時間がないため，基礎的な要素だけを効率よく選択して学ぶことが望まれる．

本研究では，専門分野の分野基礎性について，様々な観点と指標を提案し，実験に基づき客観的に評価を行う．ここでの分野基礎性とは，その分野における最低限抑えておくべき，必須の用語であることを示す．分野基礎性が高い用語は，その用語を知らなければ，その専門分野のことを理解することができない，他の専門用語も理解することができない用語とする．一方，分野基礎性が低い用語は，基礎性の高い用語の理解を深めた上で，その知識を利用しなければ理解することができない専門性の高い用語であるとする．当然，分野基礎性が高く専門性が高い語もある．

以下，論文の構成は，分野基礎性に必要な観点と指標の整理，指標の妥当性を評価するために，具体的なデータを用いて様々な尺度を用いた評価を行い，提案尺度と比較を行う．その実験結果に基づいて分野基礎性の高い用語の抽出とレベル分けについて考察を述べる．

2　専門用語の分野基礎性

2.1　分野基礎性判定に必要な観点

専門用語の分野基礎性では，(1) 用語の適切性，(2) 用語のレベル分け，(3) レベル内における優先度（用語の理解難易度）を考える必要がある．まず (1) の用語の適切性は，専門用語らしさと対応するが，辞書の見出し語と成り得るような特定分野において広く長く認知されていることである．これは用語として適切なものは，分野基礎性も高いと判断されるが，適切な用語を抽出することは難しい．重要語抽出と類似した問題ではあるが，文書における重要語と分野全体における基礎的用語との間には違いがあると考え，本研究では，辞書や事典の索引語となっている用語が適切であると仮定した上で実験を進めていく．

次に (2) の用語のレベル分け問題として，がんの関連性による用語分類と選択基準について議論したものがあるが [4]，用語の選択基準をがんとの関連性の度合いによって4つの段階を設定している．本研究では，がん用語のように網羅性を追求するのではなく，初心者への学習手順を提示できる効率性を重視する．そこで関連性や理解の容易さを基準とするのではなく，学習者の知識レベルをレベル分けの基準とした．学習者には，個々の目的や背景に応じた必要な知識や情報が与えられるべきである．そこで暫定的に4段階に分類した．

1. 一般，大学学部生，他の研究分野の研究者（分野の概要を知りたい）
   Wikipedia レベルの情報　コアな情報のみ
2. 大学学部生（これからその分野を専門にする学生），企業の開発者（新しい技術の動向調査）
   分野の成り立ちも含めた詳細な概要
3. 大学院修士（修士論文テーマ探し），企業の技術者（システムの実装）
   最新動向も踏まえた，比較的広い範囲での詳細な情報

4. 大学院博士（博士論文テーマ探し），研究者（新しいテーマ探し）
   過去の研究成果も含めた，狭い範囲での詳細な情報

最後に (3) のレベル内における優先度（用語の理解難易度）は，同じレベルでも優先的に学ぶべき用語と，このレベルで知っておいた方が良いが，必須性が高くない語の差が出てくる．たとえば，言語処理に直結している用語（形態素解析のアルゴリズム等）はレベル 2 や 3 で学ぶべきであるが，研究テーマによっては深い言語学の知識が必要な場合もある．今回はこのレベル内での差は特に考えず，大枠のレベル分けを第一段階としている．以下では，分野基礎性判定に関連する指標を整理する．

2.2 基礎性判定の指標

これまで基礎性判定の指標として，以下で説明する指標に加えて優先度（学ぶ順序，優先度）と定義明確度（定義文で導入される度合い）を考えてきた．しかし，これらの指標は実際のテキストから抽出でき，データとして有意でなければならない．そのため予備調査の結果，優先度では，必要となる講義資料と教科書のデータが不足していたこと，定義明確度は定義を明確に表現した文「〜とは」があまりなかったことから今回の指標からは除外した．

2.2.1 頻度

重要な語は文中で繰り返し用いられることから，様々な指標に頻度情報が含まれている．分野基礎性においても，分野において繰り返し出現する用語は重要であり，その分野の研究者が親しみ，馴染みのある語として頻繁に出現する．今回は，頻度情報として，用語の出現回数だけでなく，用語が出現する年数（経年推移度）も含めた．用語がどれだけ長く使われてきたかという時系列情報も基礎性を示す重要な指標である．

2.2.2 網羅度

専門分野内でいくつかの更に詳細な専門分野（サブカテゴリ）に分かれることがある．自然言語処理の場合は，言語学，人工知能（機械学習），認知科学,，応用システムとして情報検索，質問応答などに分類できる．そのため学ぶ人がすでに持っている知識によって用語の理解が変わってくる．しかし，分野基礎性が高い語は，どのサブカテゴリにおいても関連があり，どんな背景知識を持っている人にとっても必要な用語である．複数のサブカテゴリに共通して出現する語を網羅度が高い用語として指標の一つに定めた．

2.2.3 語構成度

分野基礎性が高い用語は，単独でも多く出現するが，前や後ろに様々な語が結合する傾向がある．たとえば，「形態素解析」という用語の場合，後ろに「システム」が結合して「形態素解析システム」，前に「統計的」が結合して「統計的形態素解析」など，様々な派生の専門用語，新しい複合語（いずれは専門用語として認識されるものも含む），臨時一語（専門用語や辞書の見出し語にはならない語）を生成することができる．この基準は重要度計算の時でも利用されているが，どれだけの語と接続する可能性があるのかで，その基となる用語の重要性が計算できる．ある用語が新規の用語を構成している数が多ければその用語の概念は重要であると考えられる．

今回は，上記の複合語を生成する数だけでなく，用語の長さも考慮した．用語の長さは重要な指標で，短い単語は専門用語になりにくい．専門性が高まるほど語の構成は複雑化すると考えられる．そこで，今回は単純に文字の長さも語構成度の指標の中に含めた．

3 実験

従来行われてきた尺度を用いて，分野基礎性の高い用語候補の抽出を実験した．その実験結果をベースとして，提案手法として上記の指標を数値化し，ランキングしたものと比較を行う．

3.1 対象データ

用語抽出のために，以下のように一般分野と専門分野のコーパスと正解セットを用意した．

1. 一般分野のコーパス
   毎日新聞 1995 年版 [2] 一年分の新聞記事データ（MAI）述べ語数 6,419,610 語，異なり語数 651,050 語
2. 専門分野のコーパス
   情報処理学会自然言語処理研究会 14 年分 1993 年から 2006 年の論文アブストラクト（IPSJ）述べ語数 61,518 語，異なり語数 10,755 語

デジタル言語処理学事典 [3] の本文データ（NL）述べ語数 89,662 語，異なり語数 17,485 語

3. 正解セット
デジタル言語処理学事典の索引語（正解リスト）2577 語

3.2 判定尺度と評価基準

本研究における実験では，英語教育のための分野特徴単語の選定尺度 [1] を利用した．内山ら [1] の研究では，特定分野の英語学習の効率化のために特徴的な語彙を抽出する方法を提案している．この特徴語抽出は，尺度毎に用語の特徴（初級，中級，上級など）が報告されていることから，本研究における分野基礎性用語の抽出およびレベル分けに共通している部分があると考えられる．

尺度は，対数尤度比 (LLR)，カイ二乗値 (Chi2)，イエーツ補正カイ二乗値 (Yates)，自己相互情報量 (PMI)，コサイン (Cosine)，Dice 係数 (Dice)，補完類似度 (CSM)，頻度 (Freq) の 8 種類である[1]．

特徴語は，各コーパスのテキストを，形態素解析 Mecab によって分割し，名詞を抽出した．複合名詞の抽出は連続する名詞をまとめて一単語とした．正解セットはデジタル言語処理学事典 [3] の索引語を利用したが，句になっている索引は助詞を除いて複合名詞や単独名詞にした．各尺度に従って値を求め，値の降順にソートして特徴語リストを作成した．

3.3 各尺度の平均精度

尺度の精度評価を評価するため，平均精度 (Average Precision, AP) を利用した．平均精度は特徴語リストの上位から順番に正解セットの単語と比較し，正解であった時にその時の順位を r とし，それまでの順位における正解数を l とし，l/r を求め，特徴語リスト中の全正解数での平均精度を算出した．表 1 に毎日新聞 (MAI) とデジタル言語処理学事典 (NL)，毎日新聞 (MAI) と情報処理学会 NL 研抄録 (IPSJ) の特徴語リストの平均精度を示す．

表 1 の結果では，MAI と NL の方が，MAI と IPSJ よりも平均精度が高かった．正解セットが NL にある索引語を利用していることもあり，コーパスとして正解セットとの親和性が高かったことが理由である．一方，IPSJ は精度がかなり低く，NL との出現単語の種類に違いがあることがわかった．分野基礎性が高い語は，事典のように分野全体について万遍なく説明し，用語も漏れがないようなテキスト集合に特徴的であると考えられる．

一方で，事典はあらかじめ用意しておかなければならないことや，新しい用語をフォローすることが難しい．論文であれば，テキストを蓄積でき，用語の変化についても把握できるはずである．そこで，前述の 3 つの指標（頻度，網羅度，語構成度）をベースにして論文からの情報による精度の向上を検討した．

[1] 各尺度で共通するパラメタは，a=専門分野コーパスの単語 α の頻度,b=一般分野コーパスの単語 α の頻度，c=専門分野コーパスでの単語の総頻度-a，d=一般分野コーパスの単語総頻度-b によって計算．詳細は文献 [1] を参照されたい．

表 1: 各尺度とコーパス別平均精度

尺度	MAI-NL 平均精度	MAI-IPSJ 平均精度
Yates	0.37167	0.22994
Cosine	0.37089	0.22741
Chi2	0.36975	0.22529
LLR	0.35263	0.19558
Dice	0.34012	0.16316
CSM	0.30733	0.23446
Freq	0.30733	0.23446
PMI	0.91532	0.042585

3.4 提案尺度

各指標に対応して，IPSJ の異なり語 10,755 語を対象として以下のような値を取得した．

1. 頻度
単語 α について IPSJ の抄録に出現する数 a，IPSJ のタイトル，著者キーワードの出現頻度，出現年度数，NL の本文中に出現する数 n

2. 網羅度
単語 α が NL 中に出現する章の数

3. 語構成度
単語 α に複数の語が接続して複合語（臨時一語も含む）を生成している場合は，その複合語の数が NL の語構成数，IPSJ の語構成数，単語 α の文字数 l

語構成度は，2 文字から生成される単語の場合，非常に多くの複合語を生成するため一般的な語が高い値を取ってしまう．しかし，分野基礎性が高い用語は複

合語が多いため，2 文字の単語より長い単語が他の語と接続する数を数えた．また，頻度ベースでは一般的に用いられている尺度と同様の数値となる可能性が高いため，できるだけ頻度の影響が出ない重みづけを検討した．

タイトル，著者キーワード，出現年度数，出現する章数，語構成数を文字の長さで乗算したものを，テキスト中の出現頻度に加算する方法を考えた．各項目の組み合わせにより数値の調整をし，降順にランキングし，実験と同様に平均精度を算出した．表 2 に提案手法の平均精度を示す．指標をテキスト別に分け，IPSJmeta は IPSJ のタイトル，著者キーワード，出現年度数，語彙構成数を加算，NL は NL における網羅度，語構成度を加算，ALL は IPSJmeta と NL を加算，Simple-Freq は単純に全ての値を加算したものである．

表 2: 提案手法と平均精度

提案尺度	平均精度
NL*l+n	0.48121
ALL*l+a+n	0.46757
Simple-Freq	0.38801
IPSJmeta*l+a	0.33066

4 結果と考察

提案尺度の平均精度は，実験で使用した尺度よりも精度が高かった．この理由として，従来の尺度は頻度ベースで，コーパスの種類を分けてもどうしても頻度が高い語が上位にランキングされる傾向がある．提案尺度では頻度の影響よりも，語の構成に着目した指標となっているために比較的良い結果が出たと考えられる．しかし，論文情報からの値を基にすると精度が下がる傾向がある．正解セットに含まれる用語の適切性にも関連してくるが，今後，どの単位の用語を学習者に提示するかが課題となる．

では実際のレベル分けには提案手法は有効であるのかを調べるために，暫定的に作成した分野基礎性スコア単語リスト 587 語を使って順位を調べてみた．このリストはデジタル言語処理学事典の索引語の中から抜粋したものを専門家によってスコアを付与してもらったものである．このスコアに従って，提案手法の順位付けの特徴を簡単に調べた．分野基礎性が高い語が 1，低いものを 4 とした 4 段階にレベル分けを行った．スコア 1 の平均出現順位は 209 位，2 は 779 位，3 は 1137 位，4 が 595 位という順番となった．スコア 1 から 3 までは基礎性の高い順にランキングされていることがわかった．但し，レベル分けの基準が学習者を基準としてもまだ揺れるため，更に検討が必要である．

5 おわりに

本研究では，用語の分野基礎性について，基礎性判定に必要な観点，様々な指標を整理し，実際のテキストデータに当てはめて指標の数値化実験を行った．提案手法と比較するために，従来手法の尺度を用いてランキングした語の平均順位精度の算出を行った．一般分野コーパスとして毎日新聞，専門分野コーパスとして論文抄録と事典における用語の出現頻度に基づいて評価した．結果は毎日新聞と事典を利用して計算したものが，論文抄録と毎日新聞との平均精度よりも高かった．提案手法では分野基礎性に必要な指標として頻度，網羅度，語構成度を実際のテキストデータに適用して数値化し，独自の重みづけを行って順位付けを行った．提案指標を含めた場合，従来尺度よりも平均精度が高く，有効性を確認した．

今後の課題は，用語の適切性を再度検討することと，レベル分けの基準を学習者の知識レベルだけでなく，より詳細な細分化された観点として検討する必要があると考える．

謝辞

情報処理学会刊行誌掲載論文データに関して，研究利用することを許諾していただいた，社団法人情報処理学会に感謝いたします．

参考文献

[1] 内山将夫，中條清美，山本英子，井佐原均 英語教育のための分野特徴単語の選定尺度の比較　自然言語処理, 11(3)，pp.165-197, 2004.

[2] 毎日新聞社 CD-毎日新聞 95 データ集 1995 年版

[3] 共立出版株式会社 デジタル言語処理学事典 自然言語処理学会編，2010.

[4] 中川晋一，内山将夫，三角真，島津明，酒井善則 コーパスに基づくがん用語集合の作成と評価 自然言語処理, 16(2), pp.3-44. 2009.