Web 情報抽出のための専門用語獲得

池野 篤司† 濱口 佳孝† 山本 英子‡ 井佐原 均‡

†沖電気工業株式会社 ‡独立行政法人 情報通信研究機構

E-mail: †{ikeno546, hamaguti662}@oki.com, ‡{eiko, isahara}@nict.go.jp

1. はじめに

我々は情報検索と情報抽出を応用したアプリケーションである「産学連携支援ツールBluesilk®[1]」[1,2] への組み込みを目指した専門用語獲得技術の開発に取り組んでいる．Bluesilk®には，特定の属性を持つ関連語を表示する機能があり，人名・地名などの固有表現属性を持つ語に加えて，「磁性体」「活性汚泥」などの専門用語属性を持つ語を表示することを特徴の一つとしている．

そこで我々は，最新の Web ページを収集した大規模文書集合から用語を統計的に獲得した後，技術用語などの特定分野の専門用語であるかどうかを判別し、属性ラベルを付与するという手順で，Web からの専門用語獲得を試みている．図 1 に専門用語獲得の流れを示す．まず用語獲得のステップで Web 文書集合から専門用語候補リストを得る．この段階では，リストには様々な属性を持つ用語が含まれている．次に，そのリストから，属性ラベル付与のステップにより専門用語と確定した語のリストを得る，という流れになっている．

実験により我々の提案した手法が実用可能であることを確認できたので本稿で報告する．

2. 統計的用語獲得[3]

統計的用語獲得に関しては，形態素を単位とした n-gram の統計・表層情報を利用し，時間的・物理的コストを考慮した手法を用いた．主に「候補の選定」と「単位性の確認」の工程を経る．

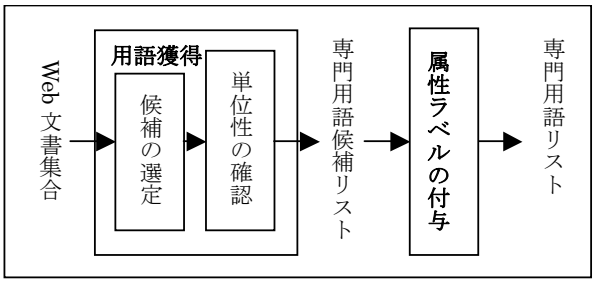

図 1 専門用語獲得の流れ

2.1 候補の選定

ここでは，2 以上で適当な長さまでの形態素 n-gram をすべて数え上げ，その中から文書集合中の特徴的な用語候補だけを選定する．

選定にあたっては，特定の文書に集中して出現する程度（集中度）と，文書集合全体での出現の散らばりの程度（分布度）との双方が極端に高低に寄らないことを条件とした．これは，文書集合中の特徴的な用語として，文書の内容を代表する用語と，分野を代表する用語との二通りの基準を考慮したためである．

文書の内容を代表する用語は，ある範囲内の df2/df を持つ傾向があると報告されている[4]．そこで，このような集中度を測る指標として df2/df を利用することとした．ここで，df は用語が出現する文書の数，df2 はその語が 2 回以上出現する文書の数，をそれぞれ表している．

一方で，分野を代表する用語は，ある特定の文書だけに極端に集中するのではなく，適度に散らばって出現する．ただし，あまり多くの文書に出現するものは機能語の類が含まれるので適当ではない．本研究では，このような分布度を測る指標として df/cf を利用することとした．ここで，cf は Web 文書集合全体でその語が出現する頻度である．

[1] Bluesilk®は（株）三菱総合研究所と沖電気工業（株）により共同開発されている．Bluesilk®は（株）三菱総合研究所の登録商標である．

これにより，文書集合中の特徴的な用語を選び出すために，集中度 df2/df と分布度 df/cf とが一定の値の範囲に収まっているという条件を課すことにした.

2.2 単位性の確認

上記の方法によって選定された用語候補は，語の境界の妥当性に関する判断が行われていないため, ひとかたまりの用語とみなすには不適なものも含まれている. そこで, 用語全体の「単位性(unithood)」[5]（ひとかたまりの単語とみなせる度合い）が強いことを条件としてさらに候補を絞り込んだ.

単位性を調べる関数の例として C-value[6]を変形した田中ら[7]の関数(1)がある.

$$f(Z)=\log(ml(Z)+1) * \log(cf(Z)) * (1-1/cd(Z)) \quad (1)$$

ここで，Z は任意の n-gram, ml(Z)は Z を構成する形態素数，cf(Z)は文書集合全体での Z の出現頻度，cd(Z)は Z に連接する形態素の異なり数である.

(1)の関数において，第 1 項 log(ml(Z)+1)が長さの項，第 2 項 log(cf(Z))が出現頻度の項，第 3 項(1-1/cd(Z))が連接する形態素の異なり数の項である.

(1) の関数から派生するいくつかの関数の効果を実験で調べた結果，本研究では，(1) の関数から長さの項を削除し，さらに，異なり数の項を出現頻度の影響を考慮して補正した項 cd (Z)/(cf(Z)+cd(Z))に入れ替えた関数(2)を用いることとした.

$$f(Z)=\log(cf(Z)) * cd(Z)/(cf(Z)+cd(Z)) \quad (2)$$

用語候補として残しておくための基準は，「ある形態素列が自分より一形態素だけ短い部分形態素列よりも単位性が強いこと」とした．この条件は，用語の候補として挙げられた形態素 n-gram よりも，右（後方）または左（前方）を短くした部分形態素列の方が単位性が強いときに，当該の形態素 n-gram を候補から外す役割を果たす．これにより，単位性の強い形態素列を包含する長い形態素 n-gram は，用語としての境界が妥当でないとして候補から除去される.

2.3 実験結果

実験は，大学の工学系 Web ページ約 200ＭＢを形態素解析したものを対象として実施した. 候補の選定の工程では 10-gram までの形態素列を数え上げた.

単位性の確認においては, すべての部分形態素列について条件をチェックしなくても, 先頭と末尾の一形態素を短くした場合についてのみチェックするだけでほぼ同じ効果が得られることがわかった.

また，(2)の関数を用いることにより，専門用語性を反映すると考えられる df2 の値が高い用語を多く残すことができた．よって，この関数を用いた手法が本研究の目的に合致すると判断して, 以降の実験ではその結果を利用することにした.

3. 属性ラベルの付与[8,9,10]

ここでは, 上記手段により獲得された用語候補（リスト）を処理対象とする．獲得された用語は複数の要素語（形態素）が結合された形をしている．用語全体の属性ラベル[2]が専門用語であるかどうかを判定するにあたって, 用語を構成する各要素の属性ラベルが専門用語であるかどうかの情報を利用することにした.

提案手法は，用語のいずれかの構成要素の属性ラベル[3]を全体に反映させる単要素属性適用をまず行う．次に，それでも属性ラベルが付与できなかった用語を対象として既に属性ラベルが付与されている用語の構成要素から選択した属性影響語を用いて，再度適用を試みる.

3.1 単要素属性適用

構成要素の属性ラベルに着目した属性ラベル決定ルールにより，単要素の属性を用語全体に適用する.

- 属性ラベル決定ルール

  各構成要素の持つ属性ラベルの組み合わせを考慮して全体の属性ラベルを判断するルールを用意しておく. たとえば末尾の構成要素の属性ラベルを重視するといったルールである.

  ただし，今回の実験データにおいては，構成要素がとり得る属性ラベルは専門用語または固有表現のみであり，かつ，事前に全体が固有表現のラベルを持つ用

[2]固有表現か，専門用語か，無ラベル（それ以外の一般的な語）かを判別した情報を属性ラベルと呼ぶことにする.
[3]各構成要素には固有表現抽出や専門用語辞書とのマッチングにより属性ラベルを付与しておくものとする.

語を実験対象から外したので，今回は，専門用語という属性ラベルを持つ構成要素が一つでも存在する場合に，用語全体に専門用語という属性ラベルを割り当てるというルールのみを用いる．

例えば「磁場／配向」という用語に対して，「配向」が専門用語という属性ラベルを持っていた場合，単要素属性適用により「磁場／配向」全体に専門用語という属性ラベルを与える．ここで「磁場」は本来属性ラベルを持っていなかったとすると，「磁場／配向」全体が専門用語という属性ラベルを付与されたことにより，「磁場」を属性影響語と見なして「磁場」にも専門用語という属性ラベルを仮設定する．この操作の結果，例えば「交流／磁場」という用語に手がかりとなる属性ラベルが本来は存在しなかったとしても，属性影響語「磁場」に仮設定された専門用語という属性ラベルの適用により「交流／磁場」全体に専門用語という属性ラベルを与えることができることになる．

3.2 属性影響語による適用

3.1 節の単要素属性適用を行っても属性ラベルを割り当てることができなかった用語に対して以下の処理を行う．

ここで処理対象となった用語は属性に関する決定的な情報を持たないため，既知の情報を用いて，未適用用語に関連する情報を補間する必要がある．そのため，既判別の専門用語の構成要素から「属性影響語」を選択する方法を提案する．

(1) 「属性影響語」の抽出

専門用語という属性ラベルを付与された用語の集合から頻出する構成要素をリストアップして，それらを「属性に影響を与える語（属性影響語）」であるとする．

(2) 属性の仮設定

属性影響語のうち，現在は無ラベルである語（固有表現でもなく専門用語でもない語）に対して，専門用語という属性ラベルを一時的に設定する．

(3) ルール適用

属性影響語の属性ラベルを専門用語に設定した状態で，3.1 節のルールを再度適用する．

3.3 実験結果

実験は，用語獲得のステップにおいて獲得された 6000 語強のうち，固有表現抽出器により固有表現と判定された語を除く 4896 語を対象として実施した．この 4896 語に対して人手で属性ラベルを付与して正解データを作成したところ，2787 語に専門用語というラベルが付与された．

図 2 に実験結果を示す．

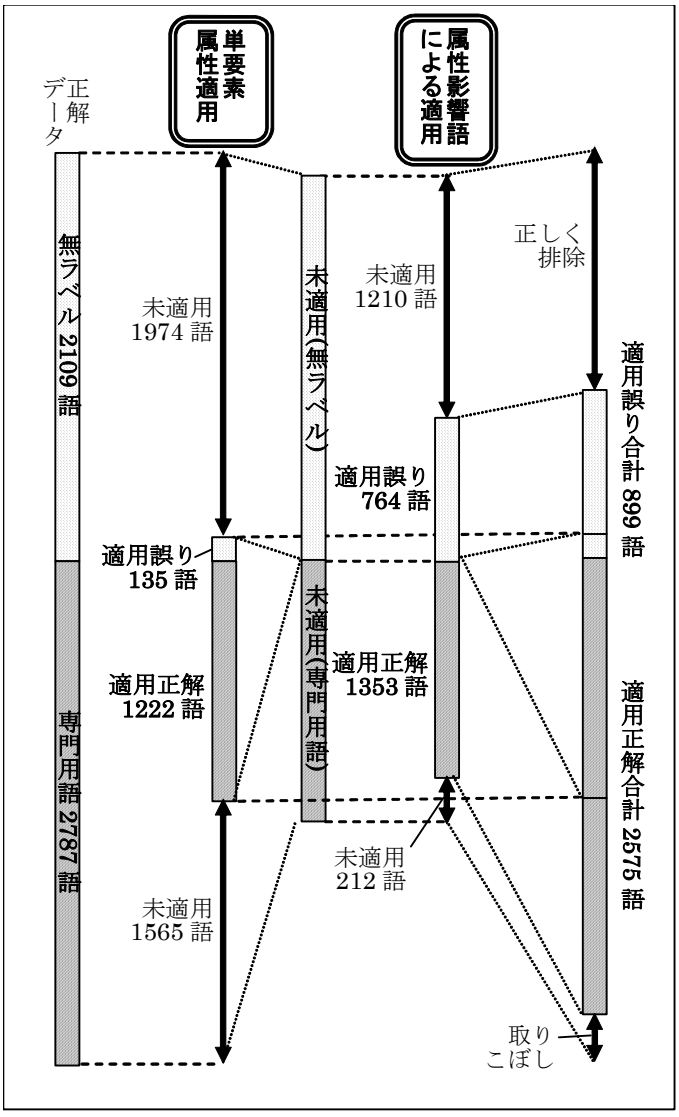

図 2　属性ラベルの付与の実験結果

単要素属性適用においては 1357 語が専門用語であると判定された．適合率は 90%であり，このルールは実用に耐えうる有効なルールであると言える．

一方で，どちらのルールも適用できなかったものが半数以上残り，その中に専門用語と判別されるべき語が多く

含まれていることから, このルールだけでは十分ではないことがわかる.

そこで，ルール適用できずに残った 3539 語に対して，属性影響語による適用を取り入れたところ，2117 語が専門用語と判定された．このルール適用の適合率は 63%であったが，残った専門用語のカバー率（再現率）は 86%に達した.

この二つの属性適用手法を通じての再現率は 92%，適合率は 74%となり，2787 語の専門用語のうち 212 語は取りこぼしたが，無ラベルの語のうち 1210 語を確実に外すことができた. 属性影響語による適用の導入により適合率は少し低下したが再現率を大幅に改善することができた.

4. 考察

実験の結果，本手法により，応用アプリケーションに供給するための専門用語辞書データを Web 文書集合から獲得できることがわかった. それぞれの手法について改善が必要な点はあるが，良好な結果を得ている.

属性ラベルの付与のステップにおいて，属性影響語による適用の性能にはまだ改善の余地がある. 適合率の低下を避けるためには, 属性影響語を何らかの手段で取捨選択する必要がある. 一般名詞として使われることが多い語を排除するには，idf 値の高いものを選択するという方法が考えられる. 用語獲得時に利用された統計情報から属性影響語の idf 値を求めて傾向を把握することを試みたところ, 確かに高頻度のものは idf 値が小さくなることが多いことが確認できた．しかし，idf が高い値を取るものの中にも一般名詞が混在しており，idf 値だけで選択するのは困難であると思われる.

5. まとめ

本研究では，Web 文書集合から専門用語獲得を行った. まず統計的に用語を獲得し，その後，専門用語属性を判別するというステップで処理した.

統計的用語獲得に関しては，形態素 n-gram の統計・表層情報を利用し，時間的・物理的コストを考慮した．実験では, 専門用語と判断できる複合名詞や名詞句などを獲得できた.

専門用語の判別では, 与えられた用語の構成要素の属性ラベルに着目した単純な属性適用ルールを用いた. さらに, 専門用語と判定されたものの構成要素を「属性影響語」と見なし, 属性影響語に専門用語という属性ラベルを仮設定して，先の処理で判定できなかった用語について，再度ルールを適用した. その結果, 全体を通じての適合率 70%, 再現率 90%を実現した.

実験の結果，本手法により，応用アプリケーションに供給するための専門用語辞書データを Web 文書集合から獲得できることがわかった.

今後の課題として，専門用語判別における属性影響語の選択基準について継続して検討したいと考えている.

一方で，属性影響語の導入がうまく働いたとしても，誤ったラベルを付与してしまう用語や, 判定の手がかりを持たない用語がまだ存在することがわかってきた. これらについても今後検討を重ねる予定である.

参考文献

[1] 中村達生, 産学連携支援ツール（Bluesilk®）の仕組み, 情報管理, Vol.46, No.7, pp.455-462, Oct.2003.

[2] 産学連携支援ツール Bluesilk®, http://www.bluesilk.biz/

[3] 山本英子，池野篤司，濱口佳孝，井佐原均, “検索支援に向けた Web 文書集合からの用語獲得,” 情報処理学会研究報告 04-NL-164-29.

[4] K. W. Church, “Empirical Estimates of Adaptation: The chance of Two Noriega’s in close to p/2 than p^2,” Coling2000, pp.180-186, 2000.

[5] K. Kageura and B. Umino, “Methods of Automatic Term Recognition: A Review,” Terminology, Vol.3, No.2, pp.259-289, 1996.

[6] K. Frantzi and S. Ananiadou, “Extracting Nested Collocations,” COLING96, pp41-46, 1996.

[7] 田中久美子，山本真人，中川裕志：web 検索に基づく多言語動的 KWIC，情報処理学会研究報告，自然言語処理研究会，02-NL-152-17.

[8] 池野篤司, 濱口佳孝, 山本英子, 井佐原均, “統計的に獲得された用語への属性ラベル付与,” 情報処理学会研究報告, 04-NL-164-30.

[9] 池野篤司, 濱口佳孝, 山本英子, 井佐原均, “情報獲得支援のための専門用語アノテーション,” 言語処理学会，第１１回年次大会, B4-3, 2005.

[10] 池野篤司, 濱口佳孝, 山本英子, 井佐原均, “属性影響語を用いた専門用語判別,” 情報処理学会研究報告, 05-NL-168-14.