情報獲得支援のための専門用語アノテーション

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として「産学連携支援ツールBluesilk®[1]」[1,2]の開発に取り組んでいる．Bluesilk®は，（技術的な内容の）テキストを入力すると，Webページや論文・特許などの文書集合から要求内容に関連する文書を検索し，さらに指定された属性を持つ語だけを抽出してリストアップすることができるシステムである．プロトタイプでは，人名・組織名・技術名（技術用語）などの属性が指定できるようになっている．

一般的な固有表現(NE)抽出手法によって人名や組織名は認識することができる．しかし，技術用語に関しては，推測するための情報は文中にはあまり存在しないので，基本的には辞書に記載されている語を抽出するしかない．一方で，技術用語は日々新しいものが生み出されるため，それらを辞書に追加していくことが必要となる．

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，ある程度の割合で出現する用語を統計的に獲得した後，人名・地名などの固有表現だけでなく，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てアノテーションを付与することを考えた[3,4]．ここで専門用語とアノテーションされた語は専門用語辞書に追加登録することができる．

ただし，継続運用する応用アプリケーションに対して最新の辞書データを追加していくことを想定しているので，できる限り単純・高速な処理で実現することが必要となる．

処理の前半部分に相当する用語獲得には，文献[3]に述べられた手法・パラメータのうち，最もよい結果を得たものを用いることにする．

本稿では，上記手法により獲得された用語に対して，特に専門用語の判別についての試みを述べる．本研究で用いる手法では，処理の単純化のため，用語リストとその統計的情報だけを受け取り，原文に関する情報は利用しない．

2. 予備実験

本研究の実験では，東京大学の工学部・工学系研究科の Web サイトから収集したホームページ（テキストのみで約 200Mbytes）から，文献[3]で述べられている B1 の手法によって獲得された 6205 語を対象とする．文書集合が工学を対象としたものなので，以降で述べる専門用語とはすべて技術用語を意味する．

まず，一般的な固有表現抽出手法によって容易に固有表現であると判断できる用語を本研究の処理対象から外すことを考える．沖電気で研究開発を行っている固有表現抽出器[5]を用いたところ，6205 語中 1190 語は固有表現であると判定された．語境界の間違いや一般用語を固有表現と判定した間違いなどにより精度は約 9 割であったが，これらの中に専門用語が含まれてしまう間違いはなかった．この結果により，本研究では，処理

[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．

対象用語を残る 5105 語に絞ることにした.

さらに我々は，これらが分野を限定した文書集合から獲得された用語であることから，残った語のほとんどは専門用語ではないかと推測した．もしそうであれば，現状の結果を簡単にチェックするだけで専門用語辞書の見出しとして採用することができることになる.

そこで，実際にこれらのうち専門用語がどれくらいを占めるかを人手で確認したところ，5015 語中 2805 語のみが専門用語である，という結果が得られた．このことは，対象用語をそのまま辞書化するのは現実的ではなく，専門用語か否かの識別を行う必要があることを示している.

3. 単要素属性の拡張実験

用語を構成する各要素については，固有表現抽出器や専門用語辞書とのマッチングにより，固有表現か専門用語かそれ以外の一般的な語であるかを判別した情報を付与することができる．ここで付与される，人名・地名などの固有表現の分類に専門用語という分類を加えたものを，以降の説明では一括して属性ラベルと呼ぶことにする.

ここでは，構成要素の属性ラベルを用語全体に波及させる単要素属性拡張により，専門用語がどの程度判別できるかを確認する.

3.1 手法

処理対象の用語リストに対し，以下の属性拡張ルールを順に適用することによって，専門用語と判定されたものの再現率と適合率を調べる.

(1) 末尾属性拡張ルール

用語を構成する末尾要素の属性ラベルを，用語全体に割り当てる．一般に日本語の複合語においては，末尾要素が複合語全体の品詞・意味を規定するという考えに基づく.

以下の場合はこのルールを無条件に適用する.

- 末尾以外の要素には特定の属性ラベルが付与されていない場合
- 末尾要素属性と同じか同類の属性ラベルを持つ要素のみが存在する場合

末尾要素の属性ラベルと，他の要素の属性ラベルが異なる場合には，相互の相性を記したルールにより割り当てるかどうか判断する.

(2) 専門用語属性拡張ルール

専門用語は固有表現よりも用語全体に与える影響が大きいであろうという仮定に基づいて，（末尾以外に）専門用語属性を持つ構成要素が存在する場合に，上記のルール(1)と同様の条件で用語全体の属性ラベルを割り当てる.

3.2 結果と検討

表 1 に専門用語の判別に関する結果を示す．正解の判定は人手で行った.

ここでルール(1)または(2)が適用できた場合，専門用語判定の適合率は 9 割程度であり，辞書登録前の人手でのチェックは必要であるとしても，これらのルールは実用に耐えうる有効なルールであったと言える.

一方で，どちらのルールも適用できなかったも

表 1　専門用語判別の結果

	ルール適用語数	うち専門用語と判別された数	うち正解であった数	専門用語正解の総数	再現率	適合率
(1)末尾属性	864	745	669	672	0.9955	0.9020
(2)専門用語属性	612	612	553	553	(1.0000)	0.9036
未適用	3539	-	-	1562	-	-
合計	5015	1357	1222	2787	0.4385	0.9005

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，この二つのルールだけでは十分ではないことがわかる．

4. 属性影響語による拡張実験

先のルール未適用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，先の実験で判別できた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．

4.1 手法

(1) 「属性影響語」の抽出

特定の属性を選択し，その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，現在は属性ラベルを持っていない語（固有表現でもなく専門用語でもない語）に対して，(1)で選択した特定の属性を一時的に設定する．

(3) 拡張ルールの適用

属性影響語の属性を設定した状態で，3.1 節の各ルールを再度適用する．

4.2 結果と検討

ここでの属性影響語の抽出は，属性ラベルが専門用語であるものを対象に行った．一時的に属性ラベルを付与する属性影響語として556語が得られた．ルール適用後の結果を表 2 に示す．

この手法により，3 節の単要素属性による拡張実験の際に取りこぼした専門用語のうち約 87%を救済できたことは評価できる．

しかし，その一方で適合率は大きく低下した．属性影響語には，専門用語として妥当と考えられるものも多数あるが，一般名詞としてよく使われるものも多く含まれていることが原因であると思われる．特に高頻度の語に関して一般名詞として使われるものが多い傾向が見られるが，さらなる分析が必要である．

属性影響語の例を以下に示す．

（高頻度語）解析，構造，システム，制御，工学，計算，情報，光，特性，ファイル，冷却，熱，実験，要素，変換，表示，生成，…
（低頻度語）連立，列，力学，稜線，面積，未知数，法線，溶液，補強，濾過，攪乱，炉心，励起，領域，抑制，溶融，名前，…

4.3 属性影響語の選択基準に関する検討

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．4.2 節の検討により頻度を唯一の基準とする方法は有効ではないと考えられる．

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

表 2　属性影響語を利用した専門用語判別の結果

	ルール適用語数	うち専門用語と判別された数	うち正解であった数	専門用語正解の総数	再現率	適合率
(1)末尾属性	1476	1476	975	975	(1.0000)	0.6605
(2)専門用語属性	641	641	378	378	(1.0000)	0.5897
未適用	1422	-	-	209	-	-
合計	3539	2117	1353	1562	0.8662	0.6391

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. 考察と今後の課題

実験全体を通じて，再現率を高めることに成功した点は実用化の観点から意義がある．ただ，新しい用語として辞書化するためには，人手でのチェックを欠かすことはできないため，適合率を下げてしまった点は改善の余地がある．

それぞれの手法は良好な結果を得ており，適合率維持を目的として，属性影響語の選択基準に関しては，継続して検討したいと考えている．既知の専門用語の統計的な振る舞いを参考にする方法に取り組む予定である．

その他に，接辞から始まるものや接辞で終わる用語は，実験においてはすべて誤りであると評価している．これらは品詞や接辞見出しの情報により容易に排除することが可能であるので，適合率の底上げが見込める．

また，現在は属性を全体に波及させる際の「相性」ルールの設定が緩いので，これを厳しく設定することによっても適合率を向上させることは可能である．

本研究の手法を適用してもなお，専門用語と判別することができなかった語には，例えば「公開鍵」といったような語が挙げられる．これは形態素では「公開」と「鍵」として認識されるが，暗号に関連する 1 語の専門用語として提示されてほしい語である．このように構成要素のいずれも一般的な語である場合は，（コストを高くしたくないという当初の要求からは反するが）原文における周囲の語との共起関係を見るよりほか方法がないように思われる．再現率のさらなる向上のためにはこの問題も避けて通れないと考えている．

6. まとめ

情報抽出をベースとしたアプリケーションに利用される辞書作成のために，Web ページなどの大規模文書集合から統計的に獲得された用語に対して専門用語判別を試みた．

与えられた用語の構成要素のうち，末尾要素と，属性ラベルが専門用語である要素とに着目し，まず単純な属性拡張ルールを適用した．その結果，適合率は高いものの，再現率が十分ではないことがわかった．

そこで，専門用語と判定されたものの構成要素を「属性影響語」とする概念を導入し，属性影響語に専門用語属性を仮設定して，再度属性拡張ルールを適用した．その結果，再現率を 9 割近くまで引き上げることに成功した．ただし，この手法による専門用語判定の適合率を高く維持することができなかった．

そのため，属性影響語の選択に際して idf 値を用いることを検討したが，現時点では明確な基準とすることが困難であることがわかったため，検討を継続している．

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，自然言語処理研究会， Vol.164, No.29, (電子情報通信学会言語理解とコミュニケーション研究会), Nov.2004.

[4] 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，自然言語処理研究会，Vol.164, No.30, (電子情報通信学会言語理解とコミュニケーション研究会), Nov.2004.

[5] 大沼宏行，池野篤司，“ホームページやメールを対象とした質問応答システム，”情報アクセスのためのテキスト処理シンポジウム発表論文集， pp.89-95, Feb.2003.