社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＮＬ－168（14）
2005／7／23

属性影響語を用いた専門用語判別

池野 篤司[1,3]　濱口 佳孝[1]　山本 英子[2]　井佐原 均[2,3]

[1]沖電気工業株式会社 〒541-0053 大阪市中央区本町 2-5-7 丸紅大阪本社ビル 4F
[2]独立行政法人　情報通信研究機構 〒619-0289 京都府相楽郡精華町光台 3-5
[3]神戸大学大学院　自然科学研究科　〒657-8501 神戸市灘区六甲台町 1-1
E-mail: [1]{ikeno546, hamaguti662}@oki.com, [2]{eiko, isahara}@nict.go.jp

あらまし　Web ページから統計的に獲得された用語候補に対して，専門用語判別を行う手法の開発に取り組んでいる．属性影響語の概念を導入した手法により再現率において一定の成果を挙げていたが，適合率が低下するという問題があるため，属性影響語を適切に選別する方法を検討した．属性影響語候補の，用語全体集合と専門用語集合での出現確率を比較した結果，専門用語性を減じる効果のある語が存在する可能性があることがわかったが，具体的な効果の確認は継続検討課題とする．また，idf 値を用いた選別を試みたが，単純な idf 値のみでの選別は効果が少ないことがわかった．また，現在把握している他の問題点に関する判別性能の検討内容についても記述した．

キーワード　情報抽出，固有表現，属性ラベル，属性影響語，辞書，出現確率，Idf

Technical Term Discrimination using "Label Affecters"

Atsushi IKENO[1,3]　Yoshitaka HAMAGUCHI[1]　Eiko YAMAMOTO[2]　Hitoshi ISAHARA[2,3]

[1]Oki Electric Industry Co., Ltd.
4F Marubeni Building, 2-5-7 Honmachi, Chuo-ku, Osaka-shi, Osaka, 541-0053 Japan
[2]Independent Administrative Institution, National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
[3]Graduate School of Science and Technology, Kobe University
1-1 Rokkodai-cho, Nadaku, Kobe 657-8501 Japan
E-mail: [1]{ikeno546, hamaguti662}@oki.com, [2]{eiko, isahara}@nict.go.jp

Abstract　We have been developing methods to label "technical terms" to the statistically acquired terms from the Web pages, and have found simple expansion method with the introduction of "label affecters" is able to raise the recall, but the decline of precision has been a problem. This paper describes the experimental results to solve the problem by selecting the label affecters with statistical indices. Probability of appearance among term candidates shows that some affecter candidates are rather useful as "de-affecters." Selecting affecters by idf value among term candidates do not appear as a significant approach.

Keyword　Information Extraction, Named Entity, Label Attachment, Label Affecter, Dictionary, Probability, Idf

1. はじめに

文章の解析においては，日々生み出される新しい用語をいちはやく辞書に取り入れることが精度に影響を及ぼす．従来は，新聞・論文・マニュアルなど，ある程度定められた用語で記述される文章を対象にすることが多かったため，用語が十分に確立されたことを人手で確認して辞書に登録する方法がとられてきており，その方法で大きな問題になることはなかった．

しかし，インターネットでの情報検索が広く使われるようになった昨今では，Web ページが重要な解析対象の一つである．Web においては新しい用語が使われ始めてから短期間でその語を用いたページは急増する．つまり，Web においては，用語は急速に確立されるので，自動的な用語獲得の手段を用意することが重要となる．

一方，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．出力の人名リストや組織名リストは連携のためのコンタクト先を検討するために，技術用語は技術理解や発想支援のために，それぞれ用いられることを想定している．

人名や組織名は一般的な固有表現(NE)抽出手法によって未知のものでもある程度認識することができる．しかし，技術用語に関しては，未知のものを推測するための情報は文中にはほとんど存在しないので，基本的には辞書に記載されている語をそのまま抽出する方法をとることになる．

そこで我々は，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[3]，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する[4,5]研究に取り組んでいる．

現在までの成果として，獲得された用語の判別に関して，基本的な手法で適合率9割を達成し，さらに「属性影響語」の概念の導入により再現率を向上させられることを確認した．再現率を上げる際に適合率を維持することが課題として残されている．

本研究の最終成果としては，Bluesilk®などの情報検索・抽出のシステムの辞書に対する継続的な用語の供給を行うシステムを想定している．供給すべき用語の決定にあたっての人手の関与は妨げないが，用語の獲得と辞書への供給を短期間のサイクルで回せるようにするために，できる限り人間の労力を減らしたい．そのためには現在の適合率と再現率のさらなる向上が望ましい．

現時点で改善の可能性がある点は2つ考えられる．1つは，文献[4,5]において専門用語判別のために導入した属性影響語の認定基準を検討することによる改善である．もう1つは，基本手法によって誤判定してしまったものの原因と対策の発見による改善である．

本稿では，第1の改善点に関連して，属性影響語を統計指標により選別する試みについて述べる．さらに，第2の改善点に関連して，今回はまず現状の問題点の分析について述べる．

2章でwebページからの用語獲得と判別の必要性について述べた後，3章で基本判別手法について概説し，4章で属性影響語の選別の実験結果を記し，考察する．5章では現状の問題点の分析について簡単に説明する．6章で総括する．

2. Webページからの用語獲得

2.1. 形態素ベースの頻度による獲得

用語獲得に関しては，様々なアプローチから研究が行われている[6,7,8]．中川ら[8]は，専門用語には複合名詞が多いことを考慮して，複合名詞を獲得対象する研究を行い，良好な結果を得ている．また形態素ではなく文字列単位で連接を見ることによる方法もある．

我々は，アプリケーションへの適用を考えて，「～の法則」や「…を用いた○○○」のような用語も獲得対象とすることを想定しているため，用語を獲得する段階では複合名詞に限定しない手法を用いる．

また，1ヶ月程度で最新のページ群から用語を獲得しつづけることを想定しているので，計算コスト面から，連接を見る単位は文字列ではなく，形態素をベースとする手法を採用する．

以降の専門用語判別の入力となる用語候補群は，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N-1の手法による出力結果を用いた．

2.2. 獲得用語に対する判別の必要性

文献[3]における実験の対象テキストは東京大学の工学部・工学研究科のホームページを収集したものである．このようにドメインがある程度限定されたテキストから獲得された用語候補は，そのほとんどが専門用語であろうと当初想定していた．しかし，人手による判定の結果，専門用語（ここでは工学全般に関係す

[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．

る用語）はそのうち 5 割弱であることがわかった[5]. ちなみにそれ以外の語は基本的な固有表現に相当するものや，一般的な語の域を超えないものであった.

また，そもそもどういう語を専門用語と認めてよいかという定義の問題もある．これに関してはまだ結論が出せていないが，例えば獲得された用語をどのように利用するかによって定義が変わる可能性があることが，いくつかの議論[7,9]から読み取ることができる.

本研究においては，前述のように，Bluesilk®の出力として検索結果文書中の専門用語（技術用語）リストとしてユーザに提示することを想定している．そのため，ユーザに提示して有用な語という観点で，分野全般にある程度の知識がある人間の評価者が認定できたものを専門用語であるとする.

3. 専門用語判別

専門用語の判別対象となる用語は複合語である．用語を構成する各要素には，NE 抽出器や辞書とのマッチングにより，人名・地名や専門用語であるというラベルがつけられている．このラベルを本稿では属性ラベルと呼ぶ.

我々は，用語全体の属性ラベルを推定するために，用語を構成する要素の属性ラベルを全体に波及させる手法を提案した．ただし，構成要素に属性ラベルが付与されていない場合には手がかりがなく，推定に失敗する．そこで，用語全体の属性ラベルに対して強い影響を与えていると考えられる構成要素を「属性影響語」として，それらに一時的に属性ラベルを与えた後，再度ラベル波及ルールを適用するという手法を用いている．以下にその手法について概説する.

なお，実際には専門用語以外の属性ラベルについても同様の手法を適用する実験も行っているが，以降の説明では専門用語という値のラベルを持つものの処理についてのみ言及する.

3.1. 単要素属性の拡張

処理対象の用語候補集合に対し，まず以下の属性拡張ルールを順に適用する.

(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．本稿における実験の段階では，日時に関連するラベル，および，数値＋単位のラベルを持つ要素が存在する場合は，用語境界を誤っていることが多いので，末尾要素のラベルの全体への割り当てを保留するというルールのみ準備している.

(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. 属性影響語による拡張

上記の処理を適用しても用語全体の属性ラベルを付与することができなかったものに対して以下の処理を行ってラベル付与を試みる.

(1)「属性影響語」の抽出

特定の属性を選択し（ここでは専門用語属性），先の処理の結果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する．それらの構成要素を「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に対して，(1)で選択した特定の属性（ここでは専門用語属性）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仮設定した状態で，3.1 節の各ルールを再度適用する.

実験の結果，属性影響語の導入による再現率の向上

が確認されると同時に，適合率を大きく下げてしまうという傾向も明らかになった．

4. 属性影響語の選別

従来の実験では，出現頻度に閾値を設けて属性影響語を選別して評価を行ったが，差ははっきりと見られなかった[4]．実際に属性影響語とされた構成要素を頻度順に並べてみても，妥当と思われるものとそうでないものが混在していた．

例えば，高頻度な要素には「情報」「研究」など，専門用語属性を仮設定してしまうと影響が大きすぎると思われる語が存在する．一方で同じ高頻度な要素の中に「磁場」のように，あまり一般的な用語には含まれないような語も含まれる．低頻度な要素には「励起」「稜線」など専門的とみなすことができる語に混じって「表現」「力」などの一般用語して使われることも多いものが存在する．

属性影響語の概念の導入には一定の効果があったが，その選別基準を単純に用語候補集合中の出現頻度だけで行おうとした点に問題があったと我々は考えた．

そこで，本稿における実験では，属性影響語の候補を term，用語候補集合を全体集合とみなして，出現確率および idf を求め，属性影響語の選別への効果を検討する．

4.1. 出現確率の比較

まず，属性影響語候補の全体集合での出現確率と，専門用語（と判定された語）集合での出現確率を比較する．出現確率の出現確率の異なり度合いは属性影響語の選別に利用できると考えられる．

図 1 に比較図を示す．ここで，x 軸は全体集合での確率，y 軸は専門用語集合での確率である．

この結果を見ると，専門用語集合にのみ顕著に現れる語よりも，専門用語集合にはあまり現れない傾向を持つ語の存在の方が目立つ．このことは，用語候補がある種の構成要素を含んでいる場合，専門用語性が減じる可能性があることを示していると思われる．

このような，言わば属性に負の影響を与える語については，扱うための枠組みをまだ用意していないため，今後手法に取り入れることを検討する．

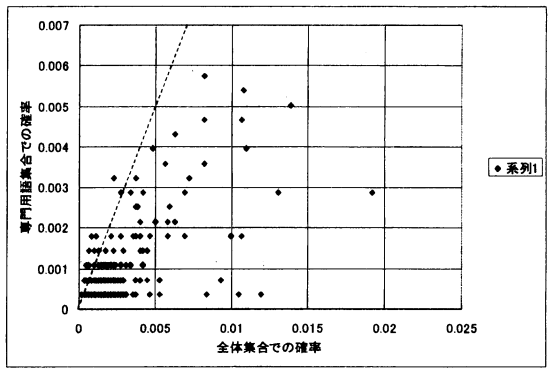

図 1　属性影響語候補の全体集合と専門用語集合での出現確率の比較

4.2. idf による選別実験

属性影響語候補の idf 値に対して閾値を設けて選別し，3.2 節の手順を繰り返した場合の実験結果を以下の表に記す．

この結果を見ると，idf 値の低い方の 10％程度の候補を落とした（表 2）だけで，属性影響語導入の本来の再現率の改善効果が大幅に減じたにもかかわらず，適合率に関しての改善はあまり見られない状態となっている．このことは，属性影響語を idf 値の大小だけで選別する方法はあまり効果的でないことを示していると思われる．

ここで落とされた候補には「研究」「システム」「環境」「技術」「情報」「構造」などがある．「システム」や「技術」を構成要素に持つラベル未付与の専門用語はあまり多くないため効果的に働いていると言える．その一方で，「環境」や「構造」を含む用語は，例えば「環境保全」や「平面構造」などラベル未付与の専門用語中に数多く存在する．これらに対するラベル付与ができなくなってしまうため再現率が下がるという結果となっている．

idf の値が大きい候補については，専門用語属性を仮設定しても問題ないと思われるものが多く並んでいる．適合率の向上に寄与するという予想通りの結果であるが，このような語はラベル未付与の専門用語中にはあまり存在しないため，再現率の向上にはほとんど効果がない．

表 1　属性影響語候補すべて（556 語）を利用した場合の専門用語判別

	適用語数	判別語数	うち正解	正解総数	再現率	適合率
(1)末尾	1476	1476	975	(975)	(1.0000)	0.6605
(2)専門	641	641	378	(378)	(1.0000)	0.5897
未適用	1422	-	-	209	-	-
合計	3539	2117	1353	1562	0.8662	0.6391

表 2　属性影響語（idf 5.8 以上，500 語）を利用した場合の専門用語判別

	適用語数	判別語数	うち正解	正解総数	再現率	適合率
(1)末尾	764	764	494	(494)	(1.0000)	0.6466
(2)専門	615	615	411	(411)	(1.0000)	0.6683
未適用	2160	-	-	657	-	-
合計	3539	1379	905	1562	0.5794	0.6563

表 3　属性影響語（idf 6.4 以上，437 語）を利用した場合の専門用語判別

	適用語数	判別語数	うち正解	正解総数	再現率	適合率
(1)末尾	500	500	341	(341)	(1.0000)	0.682
(2)専門	477	477	316	(316)	(1.0000)	0.6625
未適用	2562	-	-	905	-	-
合計	3539	977	657	1562	0.4206	0.6727

表 4　属性影響語（idf 6.9 以上，376 語）を利用した場合の専門用語判別

	適用語数	判別語数	うち正解	正解総数	再現率	適合率
(1)末尾	370	370	266	(266)	(1.0000)	0.7189
(2)専門	333	333	226	(226)	(1.0000)	0.6787
未適用	2836	-	-	1070	-	-
合計	3539	703	492	1562	0.3150	0.6999

先の出現確率の指標も含めて複数の統計的指標を組み合わせて属性影響語を選別する方法に今後取り組むことにしたい．

5. 現状の問題点の分析

属性影響語の概念を導入し，それがある程度うまく働いたとしても，誤ったラベルを付与してしまったりする用語や，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．全体の判別性能を向上させるため，まず今回は現状まだ手がつけられていない問題点について検討している内容についてふれる．

5.1. 専門用語と一般名詞からなる用語

例えば「都市工学/実習」「機械的/性質」「直角/方向」などは専門用語と判定したくない．一方で「空隙/構造」は専門用語であると判定したい．ところが，現在与えられている情報は，いずれも「専門用語＋一般名詞」であるということだけである．

末尾要素を重視するという観点からは，「実習」と「構造」との間に何らかの違いを見出せなければ判別は困難である．あるいは，同じ専門用語属性であって

も，「都市工学」と「空隙」との間に専門用語としての強さの違いが存在すると考えるべきだろうか．

属性影響語の導入に伴う議論はここでも当てはまるかもしれないが，ここまでの結果からは，統計情報は決め手とするには弱いように思われる．この問題に関しては，今後，用語の語構成の果たす役割について詳細に分析する予定である．

5.2. 再現率の向上に関して

「公開/鍵」（および「秘密/鍵」）という暗号技術に関する専門用語は，どちらの構成要素も一般名詞であるので，それだけの情報から，これを技術用語と判定するのは不可能である．この語を知らない人が，これを技術用語とわかるのは周囲の語（共起語）との関係を見て初めて判断できるのではないだろうか．

実際に「最終処分場」という，これも一般名詞の組み合わせでなる用語があるが，これは「廃棄物」にかなり強く共起することにより（環境関連の）専門用語であると判断することができる．

現時点では，計算コストが高いため，共起語を判別に用いることは考えていない．しかし，どれくらいの割合でこのような問題を生じる用語が存在するかを把握した結果，対応が必要になる可能性があるので，今後も検討を継続する．

6. まとめ

統計的に獲得された用語候補に対して，特に専門用語であるかどうかの判別を行う手法を開発中である．属性影響語の概念を導入した手法により再現率の向上の点で一定の成果を挙げていたが，適合率が低下するという問題があるため，属性影響語を適切に選別する方法を検討した．

属性影響語候補の，用語全体集合と専門用語集合での出現確率を比較した結果，専門用語性を減じる効果のある語が存在する可能性があることがわかったが，具体的な効果の確認は継続検討課題とする．

また，idf 値を用いた選別を試みたが，単純な idf 値のみでの選別は効果が少ないことがわかった．

今後は各種の統計指標の組み合わせによる選別について引き続き検討する．また，現状の問題点について他のアプローチで，判別性能を向上させる手法についても検討を行っていく予定である．

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.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.

[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] 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，Mar.2005.

[6] 下畑さより，杉尾俊之，“隣接文字情報を用いた n-gram 抽出文字列からの名詞句の自動抽出，”情報処理学会研究報告，自然言語処理研究会，Vol.114, No.3, Jul.1996.

[7] 久光徹，丹羽芳樹，辻井潤一，“タームの representativeness を測る，”情報処理学会研究報告，自然言語処理研究会，Vol.133, No.16, Sep.1999.

[8] 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.

[9] 九津見毅，吉見毅彦，小谷克則，佐田いち子，井佐原均，“サポートベクターマシンを用いた対訳表現の機械翻訳辞書登録適切性の自動判定，”言語処理学会，第１１回年次大会，Mar.2005.