2007－NL－182 (9)
2007／11／20

専門用語抽出における助詞情報の利用に関する一考察

村上 浩司　乾 孝司　橋本 泰一
内海 和夫　石川 正道

概要：専門用語抽出において様々な手法が提案されているが，そのほとんどは，コーパス中の名詞の出現頻度を元にした統計情報や，用語候補となる名詞連続中の形態素間の構造を利用するものであった．本稿では，これまで取り扱われてこなかった助詞に着目し，助詞の利用が用語抽出の精度向上に寄与するかを検討した．NTCIR1 の TMREC コーパスの正解用語と共起する助詞情報から計算される助詞スコアを計算し，既に提案されている種々の用語抽出手法と組み合わせてリランキングした結果，用語抽出精度が向上した．さらに，正解用語を用いずに上位の用語候補と共起する助詞情報から求められる助詞スコアを利用しても用語抽出精度が向上することを実験により確認した．

キーワード：専門用語抽出，助詞，リランキングアルゴリズム,TMREC タスク

A Study on Japanese Term Recognition Using Post-position Information

MURAKAMI Koji　INUI Takashi　HASHIMOTO Taiichi
UTSUMI Kazuo　ISHIKAWA Masamichi

Abstract：In this paper, we report the efficiency of using post-position which co-occurs with terms for term recognition task. Various methods of term recognition were already proposed. The majority of them use the statistics of term candidates appearing the corpora. The post-position has not been focused on term recognition task. We propose re-ranking algorithm which combines existing term recognition methods and the score based on frequency of specific post-positions which co-occur with term candidates. The efficiency of using post-position information is evaluated by utilizing TMREC corpus in our experiment. Our method improves the accuracy of term extraction compared to original term recognition methods.

Keywords：Term Recoginition, Post-position, Re-ranking Algorithm, TMREC corpus

1 はじめに

種々の専門分野に関する文書を扱う場合に文書中で用いられる専門用語は，他の一般名詞に比べてその文書の分野を推定できる情報を持っている．更に，その専門文書の中でも特定の話題の情報を表すことから，文書の代表的な語と考えることができる．専門用語の適切な抽出が可能になれば,，様々な応用に利用することができる．例えば，複数文書が与えられ，その文書群がどのようなトピックを持つのかを俯瞰的に知りたい場合が考えられる．この場合，まずクラスタリングを用いて文書を分類するが，このときに行われる素性選択において，適切に専門用語が素性として抽出されれば，より適切なクラスタリングの結果が得られると考えられる．さらに，俯瞰的な理解を促すにはクラスタ毎にクラスタが示すトピックを表すような代表用語を抽出する必要があるが，ここでも，専門用語を一般語よりも優先して考慮することで，よりクラスタに特化した特徴語を選択できると考えられる．

しかしながら，こうした専門用語の抽出は容易ではない．その理由の一つとして，厳密な意味での「専門用語とは何か」という語としての定義が存在

東京工業大学 統合研究院
Integrated Research Institute, Tokyo Institute of Technology
連絡先：murakami@iri.titech.ac.jp

せず，その定義は実際の目的やタスクによって変化することがある．そのため，曖昧な定義を行うと，専門用語だけでなく一般語が多く抽出され，逆に厳密な定義の場合には必要な語彙が抽出できないという問題が生じる．また，実際のテキストに出現する専門用語を捉える場合，専門用語は数が非常に多く辞書だけでは対応できないため，形態素解析や構文解析においては用語として同定できずに名詞や未知語の連続として扱われることも多いことから，容易には抽出できない．しかしながら，文中において専門用語が出現する際の周辺コンテキストと，一般的な単名詞・複合名詞が出現する際の周辺コンテキストは全く同じではないと考えられる．

で我々は，助詞は語の周辺コンテキストとして出現し，その語の「使われ方」を制御することから，この助詞の情報に着目することとした．本稿では何らかの用語抽出法により既にランキングされている，上位の用語候補と共起する助詞の情報から助詞スコアを計算し，元スコアと組み合わせて用語候補全体をリランキングする手法を用いて用語抽出を行った際の抽出精度の変化の検証について報告する．2節で既存の専門用語抽出法について概説し，3節にて助詞スコアを利用したリランキング法について説明する．4節で助詞情報利用の有効性を検証したのち，提案手法と既存の用語抽出法の比較評価実験とその結果について述べる．最後に，5節でまとめと課題について述べる．

2 関連研究

2.1 これまでの研究

専門用語抽出に関しては従来，専門家の人手が中心的な役割を果たしてきた．しかしながら時間が掛かるため，より新しい用語辞書が作れないという問題があった．そこでコーパスから自動的に専門用語や固有名などを抽出する研究が盛んに行われている．Sekine ら [4] は専門用語のタグ付きコーパスから機械学習手法を用いて，専門用語の開始・内部・終了等を学習し，専門用語の抽出を行った．評価型ワークショップである NTCIR においても，用語抽出タスクである TMREC[3] が行われた．中でも Fukushige ら [2] や Uchimoto らは，専門用語は対象の分野に多く出現し，他の分野においては一般的な語でないことからあまり出現しない，などの性質を定義し，統計処理を用いて専門用語抽出を行った．更に Uchimoto ら [5] は，利用するコーパスの種類を増やすことで精度を向上させた．内山ら [6] は英語学習のために必要な単語を，2種類の英語コーパスを比較し，様々な統計的指標を比較，または組み合わせることで分野特徴単語の抽出を行っている．この研究も，分野に特化した用語の抽出を目指していることから，専門用語抽出と言える．長町ら [11] はコーパスから文字 n-gram の統計量を用いて専門用語の抽出を行っている．また，中川らによる FLR 法 [10] は，単語の統計量を用いて抽出を行っている．しかしながら，どの手法にも課題があり，統計量を利用する手法においてはコーパス中での低頻度の用語の抽出は難しい．また，専門用語の候補を構成する単語もしくは文字の n-gram に基づいて専門用語らしさのスコアを設定する場合，スコアの基準がコーパスに依存して変化するために，最終的な専門用語の抽出にはスコアの閾値を経験的もしくは実験によって求める必要がある．低頻度語の抽出に焦点を当てた，文字列パープレキシティを利用した研究も行われている [8]．

2.2 先行研究の分類

専門用語の自動抽出法には，前節で述べたように数多くの手法がある．これらを専門用語性の判断基準で分けると，

1. 語の出現状況をコーパス毎に比較して専門用語性を判断：(文献 [2]，[5]，[6] における手法)
2. 語としての専門用語性を指標とするもの：(文献 [10]，[11] における手法)

の2種類に分けることができる．抽出する用語の単位で分けると，

1. コーパス中の文を解析し，単語を元に抽出する用語候補を決定
2. コーパスを解析せずに，文字 n-gram など元に，統計的な指標のみを用いて抽出する用語候補を決定

の2種類になる．統計的な指標のみを用いて用語を決定する場合は，形態素解析などの解析誤りの影響を受けない利点があるが対象とするコーパスが十分に大きくない場合，適切に用語が抽出できないことが考えられる．これに対し解析を行う手法の場合，抽出される用語候補には句点などの不必要な文字が混入しないため，何らかの語が抽出できるが，解析誤りの影響を大きく受けてしまう．

2.3 代表的な専門用語抽出法

専門用語性の判断基準が異なる用語抽出法のうち代表的な，FLR 法 [10] と Uchimoto らの手法 [5] を対象の用語抽出法とする．これらの手法は，予め抽出した名詞連続などの用語候補に対してそれぞれの専門用語性によりランキングを行うものである．

2.3.1 FLR

FLR は，用語候補を構成する個々の名詞について，その左右来た語基の延べ頻度（連接頻度 LR）または種類数（連接種類 LR）から名詞の重要度を求め，それらと用語候補の頻度を用いて用語候補のスコアを算出する方法である．用語候補 T を構成する単名詞 w_i について，以下の値を求める．

$$l(w_i) = (w_i\text{の左に出現した名詞の頻度}) + 1$$
$$r(w_i) = (w_i\text{の右に出現した名詞の頻度}) + 1$$
$$lr(w_i) = \sqrt{l(w_i)r(w_i)}$$

FLR 法による T のスコア FLR(T) は以下の式で表される．ここで，$f(T)$ は用語候補 T の頻度を表す．

$$FLR(T) = f(T)(lr(w_1)lr(w_2)\cdots lr(w_n))^{1/n} \quad (1)$$

FLR(T) では，用語候補の長さ（単名詞の数）は考慮していない．この手法の場合，左右に多くの語基が出現する単名詞に高いスコアが付与される．

2.3.2 Uchimoto らの手法

Uchimoto らの手法は，対象のコーパス中から抽出した用語候補が専門用語であるためには，他のコーパスには多く出現しないという前提の元，その出現頻度の違いからスコアを算出している．比較するために用いるコーパスは，新聞コーパスおよび，NII が NTCIR で配布した 59 分野の論文抄録コーパスである．Patrik[1] は，異なった分野のコーパスを用語抽出に用いることで，用語候補中の不要な候補を減らすことができることが報告している．スコアリングには，AI 分野の用語候補 w_i について以下の式を使う．

$$F_{AI} = (TF_{w_iAI})^2 \times \frac{1}{FF_{w_i}} \times \frac{TF_{w_iAI}}{TF_{w_iM} + 0.5} \quad (2)$$

TF_{w_iAI} は AI 分野における w_i の出現頻度，FF_{w_i} は分野頻度で，この場合 59 分野のデータを対象とするため最大 59 となる．TF_{w_iM} は，新聞コーパス中の用語候補 w_i の出現頻度である．この式から，他分野にあまり出現せずに対象である AI 分野に数多く出現する用語候補に高いスコアが付与される．

3 助詞情報を利用した専門用語抽出

用語抽出は大きく分けて，コーパスから名詞連続などを用語の候補として抽出する用語候補抽出部と，それらの候補を何らかの専門用語らしさによってランク付けを行う，用語候補ランキング部の 2 つの処理から構成される．本稿で比較する種々の用語抽出手法においては，用語候補抽出の基準が違うためランキングの精度を適切に比較評価できない．そこで統一的な用語候補抽出を行った上でそれぞれのランキング手法を評価する．

3.1 用語候補の抽出

FLR や Uchimoto らの手法ではコーパスに対して形態素解析を行い，品詞情報である “名詞” を手がかりとして，用語候補となる名詞連続を取得する．この時点で取得される複数の名詞からなる文字列は，意味をもつ複合名詞であるか単なる名詞の連続であるかは決定できないため，ここではどちらも “名詞連続” と表記する．しかし単純な名詞連続のみから用語候補を抽出した場合，用語の一部となり得る接頭語や接尾辞が付属する語は候補とできない，副詞として使用可能な名詞が名詞連続に存在する場合，適切な用語候補が抽出できないなどの問題が考えられる．適切な用語候補を抽出するために，接頭語や接尾辞を複合名詞の一部として扱う研究 [7] や，形態素解析結果の名詞の詳細な情報を用いて適切な複合名詞を獲得する研究 [12] がある．

更に，文節を跨って出現する名詞連続が用語候補抽出に与える影響を考慮し，係り受け解析による文節区切りの情報を利用して，同一文節内に出現する名詞，未知語，記号，接頭語，接尾語の連続を用語候補の対象とした．ここで対象とするのは 1 文節内の用語であることから，複数文節から構成される用語「A の B」や「A な B」などは抽出の対象外とした．このとき明らかに用語となり得ない，数字や記号のみから構成される語や，読点を含む用語候補抽出されるが，こうした候補はいくつかの規則によって除外した．

3.2 助詞情報

専門用語は，その分野において専門的な概念や知識を表す語として利用されるため，一般語と異なりどんな文脈においても出現するわけではなく，語の使われ方は限定される．助詞はその種類により，語の使われ方を制限することができることから，専門用語が特定の助詞と頻繁に共起するのであれば，助詞を捉えることにより用語抽出の精度を向上させることができると考えられる．

そこで本稿では，3.1 節で述べた用語候補抽出を行い，その後，既存の用語抽出法の用語候補ランキングアルゴリズムにより順序付けされた用語候補に対し，候補の上位の助詞情報から助詞スコアを計算し，元々のランキングスコアと組み合わせて利用することで用語候補全体をリランキングする手法を提案する．

3.3 助詞情報を用いた用語候補のリランキング手法

提案手法は何らかの用語候補抽出と用語候補ランキングにより既に順序付けされた用語候補に対し，候補の上位を正解用語と仮定して，正解用語とした用語候補と共起する助詞の情報から助詞スコアを計算し，元のランキングスコアと組み合わせて用いることで用語候補全体をリランキングを行う．基本的な考えとしては，何らかの用語候補ランキングにより上位にランクされる用語候補は最も専門用語である確率が高いことから，そうした上位の用語候補と共起する助詞から得られる情報は，局所的な「専門用語としての語の使われ方」を表すと仮定している．

まず，何らかの手法によりスコア付けされた用語候補 $t_i \in T$ を対象とする．ここで T は全体の用語候補集合である．各 t_i は元スコア $S_{method}(t_i)$ が計算されている．この用語候補集合の上位 N 位を正解と仮定して抽出する．正解と仮定した用語候補集合を $T_{sub}(T_{sub} \subset T)$ とする．次にコーパス中において，この疑似正解用語集合中の用語候補と共起して出現した助詞に着目し，各々の助詞が候補に後続して共起した用語候補数を計測する．このとき得られる助詞を p_j，全体集合を P と表す (このとき，$p_j \subset P$)．助詞 p_j と共起した用語候補数を $nt(p_j)$ と表す．そして各助詞に関して，用語候補全体における正解確率 $Prob_c(p_j)$ を計算する．

$$Prob_c(p_j) = \frac{nt(p_j)}{|T|} \tag{3}$$

助詞集合 (P) 中で，低い確率値を持つ助詞 p_j は，正解の用語候補とは多く共起していないことから，「専門用語としての語の使われ方」をより適切に表現する，高い正解確率を持つ助詞 p_j のみに着目する．しかしながら，全体の用語候補 $|T|$ や抽出する上位 N の大きさにより確率値は依存するため，絶対的な値そのものを閾値として決定することは難しい．そこで正解確率値で各助詞を順位付けし，上位 $M\%$ の助詞を「専門用語としての語の使われ方」を表す助詞として取り扱う．得られる助詞集合を P_{sub} とする．この上位の助詞 $p_j \subset P_{sub}$ を含む用語候補に対して，助詞スコア $S_{P_{sub}}(t_i)$ を以下の式で計算する．

$$S_{P_{sub}}(t_i) = log(num(t_i, P_{sub})) \times \frac{num(t_i, P_{sub})}{t_i\text{と共起する助詞の種類数}} \tag{4}$$

$$num(t_i, P_{sub}) = t_i\text{と共起する}\ P_{sub}\text{中の助詞の種類数}$$

次に，全体の用語候補 t_i に対して，t_i の持つ元スコアと，計算された助詞スコアを組み合わせることでリランキングを行う．このとき，以下の式を用いてリランキングを行って得られる用語候補の上位について，実験で評価する．

$$S(t_i) = log(S_{method}(t_i)) + S_{P_{sub}}(t_i) \tag{5}$$

4 評価実験

助詞情報の利用が用語抽出の精度向上に寄与するかを実験によって検証する．実験は 2 種類行った．

1. 正解データを用いた，助詞情報の利用による用語抽出精度への影響の検証
2. 正解データを用いない，上位候補の助詞情報を利用した用語抽出精度の向上の検証

4.1 データ

NTCIR-1 の TMREC タスク [3] で配布されたテストコレクションを利用する．TMREC タスクに従い，情報処理分野 1,870 の論文抄録を対象コーパス，その他 58 分野の論文抄録を比較コーパスとして利用し，各用語抽出法の評価を行う．また，式 2 の計算のために，追加の比較コーパスとして，毎日新聞 94，95 年の 2 年分を用いた．テストコレクション中の正解の用語は予め与えられる 8,834 語である．

各々のコーパスに対して，日本語係り受け解析システム CaboCha[9] を用いて 1 文節内の名詞・未知語・記号・接頭語・接尾語の連続を抽出し，3.1 節で述べたようにフィルタリングを施し，用語候補を抽出した．得られた用語候補は 16,943 語であり，そのうち 7,613 語が正解であった．用語候補抽出部の評価としては，再現率 0.861，適合率 0.449，F 値 0.590 となる．正解の専門用語 8,834 語には，複数文節からなる用語が 349 個ある．こうした用語は本稿での実験の対象外であることから，本稿で用いる用語候補抽出部が抽出できる最大用語数はは 8,440 個であり，カバレッジを計算すると 0.902 となる．

本稿の実験で用いる用語抽出法は全て，3.1 節で述べた用語候補抽出を行った後，それぞれの手法によりランキングを行った．

4.2 実験 1 : 助詞情報の利用による用語抽出精度への影響の検証

助詞の情報を利用することそのものが，用語抽出タスクにおいて有効であるかを検証する必要がある．そこでまず，TMREC データセット中の 8,834 個の正解用語と共起する助詞の情報を利用すること

表 1　実験 1：各手法により抽出された正解用語数

候補語数	UCH	UCH+	FLR	FLR+	FLR[10]
3000	2238(.378)	2280(.385)	1976(.333)	2026(.342)	1970(.333)
6000	3843(.518)	3891(.524)	3381(.455)	3447(.464)	3456(.466)
9000	4952(.555)	5159(.578)	4685(.525)	4720(.529)	4866(.545)
12000	6026(.578)	6216(.596)	5886(.565)	5921(.568)	6090(.584)
15000	7142(.599)	7147(.599)	6930(.581)	6958(.583)	7081(.593)

表 2　実験 1：正解用語と共起する助詞とその確率の例

助詞	共起確率	助詞	共起確率
にとって	0.838	につき	0.333
のみ	0.703	でも	0.250
によって	0.679	にて	0.250
により	0.672	ほど	0
に対する	0.669	にわたって	0

で，助詞が「専門用語としての語の使われ方」に寄与しているか，そして用語抽出精度に対してどのような影響があるかを実験により検証する．

4.2.1　実験条件

ここでは，語の出現頻度を元にした統計情報により専門用語性を導出する Uchimoto らの手法と，語としての専門用語性を導出する FLR，およびこれらの手法に助詞スコアを組み合わせたリランキング手法の精度比較を行う．比較する用語抽出法は，以下のものとなる．

- FLR におけるランキングアルゴリズム
- Uchimoto らの手法（以後，本稿では UCH と表記する）におけるランキングアルゴリズム
- 提案手法（助詞スコアを用いたリランキング手法との組み合わせ: FLR+，UCH+）

提案手法である助詞スコアを利用したリランキング手法では，上位 N 位を正解用語と仮定して助詞を抽出，スコアリングを行う．そのため，16,943 個の全用語候補に含まれる正解用語である 7,613 個を上位の用語として，共起する助詞を調査する．正解用語と共起する助詞は全部で 65 種類 ($|PP| = 65$) あった．ここでは，助詞集合 PP から上位 50%($M = 50$) の助詞を選別，全用語候補に対して助詞スコアを計算して FLR，UCH それぞれで得られている元スコアと組み合わせてリランキングを行う．

4.2.2　実験結果

実験結果の詳細を表 1（括弧内は F 値）と図 1 に示す．FLR，UCH ともに助詞スコアを付加してリランキングすることで，どの用語候補数においても正解用語数が増加することが分かる．候補数 15000 の場合，正解用語数に大きな変化はないが，候補数 3000，6000，9000，12000 の場合においてはリランキングにより，抽出できる正解用語数に大きな差がある．手法間の精度差を見やすくするため，UCH との正解用語数の差を図 2 に示す．FLR に比べて，

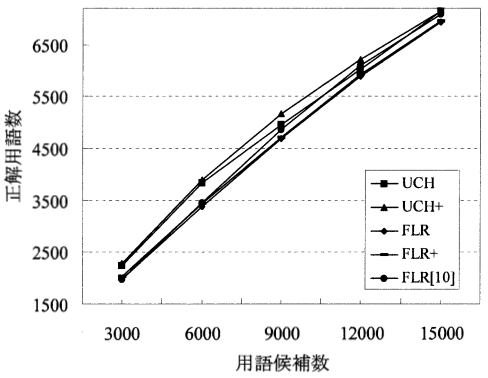

図 1　各手法の用語候補数と正解用語数

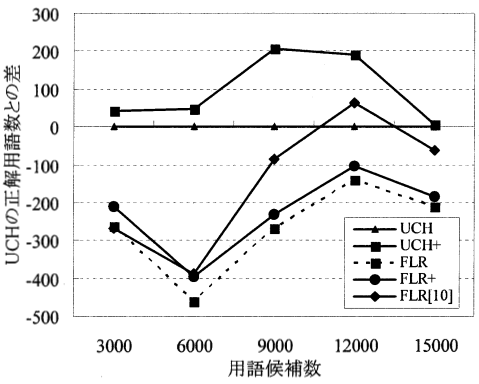

図 2　各手法の用語候補数と UCH との正解用語数との差

全体的に UCH の正解用語数が多く，UCH+ が比較した 5 種類の中で最も正解語数が多い．

この実験結果から，正解用語と高い確率で共起する助詞が存在することが確認された．用語候補とそれらの助詞との共起を考慮して助詞スコアを計算し，用語候補全体のリランキングに用いることで，より高い精度の用語抽出が可能になることが示された．そして FLR，UCH のような抽出すべき用語の専門用語性の判断基準が異なっていても，それとは独立して助詞情報が用語抽出に貢献できることを示した．表 2 に，(3) 式で求められる正解用語と共起する助詞との確率のうち，高いものと低いものをを示す．

4.3 実験 2 : 上位候補の助詞情報を利用した用語抽出精度の向上の検証

実験 1 により，助詞情報が用語抽出の精度向上に影響を与えることが確認された．しかしながらこの実験は正解用語が与えられている条件の元，正解用語と高い確率で共起する助詞に着目し，助詞のスコアを計算した．実際の用語抽出タスクにおいて抽出すべき正解用語は未知であるため，助詞スコアを計算することはできない．そこで，同じ TMREC コーパスで正解用語を用いずに，上位の用語候補と共起する助詞情報を利用した場合の用語抽出精度の向上を実験により検証する．

4.3.1 実験条件

実験 1 の結果から，より高い精度で用語を抽出できた UCH のみに着目する．3.3 で述べたように，助詞スコアを計算する上で考慮すべきパラメータには，助詞との共起を調べるための用語候補集合 T における上位の用語候補個数 N，正解確率により順位付けされた助詞集合 PP において助詞を選別する割合 (%) である M の 2 つがある．実験 1 では N を正解用語の個数を与えたが，T の大きさにより N の大きさを変更する必要がある．そのため，N を用語個数としてではなく，用語候補全体の大きさ $|T|$ に対する割合とした．本実験では，用語候補全体の 30% と 10% の 2 種類を用いることとした．これにより，正解用語と共起する助詞は $N = 30\%$ の場合，50 種類 ($|PP| = 50$)，$N = 10\%$ の場合，48 種類 ($|PP| = 48$) となった．M に関しては，50% と固定にした．

比較する用語抽出法は，以下のものとなる．

- UCH におけるランキングアルゴリズム
- UCH+ (実験 1 で用いた，正解用語を用いた助詞スコアリングによるリランキング手法)
- 提案手法 (UCH+(A)(N=30%，M=50%)，UCH+(B)(N=10%，M=50%))

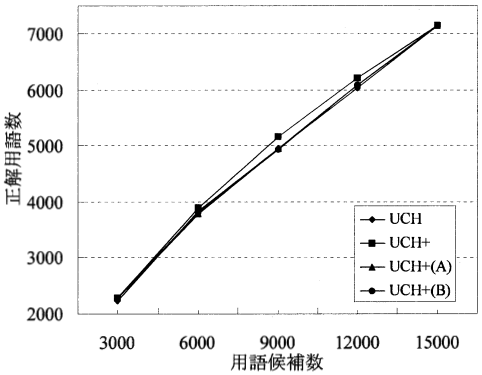

図 3　各手法の用語候補数と正解用語数

4.3.2 実験結果

実験結果の詳細を表 3 (括弧内は F 値)，および図 3 に示す．正解用語を利用して得られた助詞スコアを利用した UCH が最も高い精度を示しているが，疑似正解用語集合を用いて助詞スコアを計算してリランキングした手法 (UCH+(A)，UCH+(B)) でも，助詞情報を利用しない場合に比べて，全体的には多くの正解用語を抽出できたことが分かる．UCH+(A) と UCH+(B) を比較した場合，全体の精度に大きく差はないが，候補語数によって正解用語数の差に傾向が見られる．特に候補語数が少ない 3000 の場合は，UCH の適合率 (0.746:候補数 3000) が高いことから，その上位の用語候補を用いて得られた助詞スコアはリランキング後の抽出精度に大きく貢献し，UCH+ よりも多く正解用語が抽出された．また，UCH+(A)UCH+(B) を比較すると，少ない割合 (10%) の用語候補から助詞情報を利用した (B) では，ノイズとなる低正解率の助詞が助詞スコアに殆んど反映されなかった結果，より多くの正解用語を抽出できたと考えられる．それに対して候補語数が 6000 以上において UCH+(A) と UCH+(B) の抽出精度を比較した場合，適合率が UCH の適合率が 0.7 以下と減少する．これにより上位 10% の用語候補に含まれる正解用語数も減少し，助詞スコアの計算に影響したことが原因でリランキング後に抽出した正解用語数が減少したと考えられる．

候補数が 6000 および 9000 の場合，UCH+(A) および UCH+(B) の正解用語数は UCH に比べ減少している．助詞の正解確率と順位に着目したと

表 3　実験 2：各手法により抽出された正解用語数

候補語数	UCH	UCH+	UCH+(A)	UCH+(B)
3000	2382(.378)	2280(.385)	2285(.386)	2288(.386)
6000	3843(.518)	3891(.524)	3783(.510)	3819(.514)
9000	4952(.555)	5159(.578)	4947(.554)	4944(.554)
12000	6026(.578)	6216(.596)	6090(.584)	6087(.584)
15000	7142(.599)	7148(.599)	7144(.599)	7139(.599)

表 4　助詞と共起する用語候補数に閾値を導入した場合の用語抽出結果

UCH+(C)	UCH+(D)
2257(.381)	2264(.382)
3838(.517)	3820(.515)
4993(.559)	5004(.561)
6133(.588)	6131(.588)
7144(.599)	7132(.598)

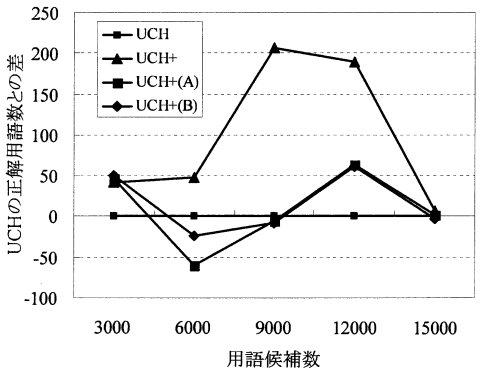

図 4　各手法の用語候補数と UCH との正解用語数との差

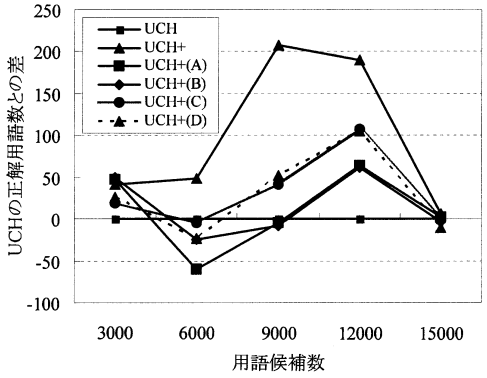

図 5　各手法の用語候補数と UCH との正解用語数との差

ころ，正解確率は高いが数個の用語候補とのみ共起する助詞がいくつか存在し，高い順位となっていた．そのため助詞の正解確率を求める際に，共起する用語候補数が小さすぎるときには，計算される確率に拘らず順位を低くする補正を行うため，共起する用語候補数に閾値を設けた．ここでは，閾値を 5 とし，閾値以上の用語候補と共起する助詞のみを対象とすることとした．この条件を加えて，改めて（UCH+(C)(N=30%, M=50%, 閾値:5), UCH+(D)(N=10%, M=50%, 閾値:5)）の 2 セットを作成し用語抽出実験を行った．

実験結果の詳細を 4 と 5 に示す．この図は図 4 に UHC+(C) と UHC+(D) の結果を加えたものである．どちらの手法も (A) と (B) に比べて，候補数 3000 を除くとおおよそ高い精度で用語抽出が行われることが確認できる．候補数が 3000 の場合は (A) や (B) に比べると，着目する助詞を更に強い制限により助詞スコアを算出するため，正解用語数は減少しているが，UCH に比べると高い精度で抽出できていることが分かる．候補数が 6000，9000，12000 の場合においては，正解用語数が大幅に増加した．このことから候補数がある程度大きい場合には，助詞スコアを計算するための助詞の選別を適切に行うことで，リランキングによりより多くの正解用語を抽出できることが確認された．

表 3 および表 4 の結果から，実験 1 のように正解用語が既知でない場合においても，何らかのスコアリングによってランキングされている用語候補が存在すれば，それらの上位を正解用語と仮定して，そこから得られる助詞スコアを用いて用語候補全体をリランキングすることで，実験 1 と同様，より高い精度の用語抽出が可能になることを示した．

5　おわりに

専門用語抽出において，これまで様々な手法が提案されてきたが，そのほとんどは，コーパス中の語の出現頻度を元にした統計情報や，用語候補となる名詞連続中の形態素間の構造を利用するものであった．我々は，これまで取り扱われてこなかった助詞に着目し，助詞の利用が用語抽出の精度向上に

寄与するかを検討するために，リランキング手法を提案し評価を行った．まず，助詞の情報が用語抽出に有効であるかを検証するために，NTCIR1 の TMREC コーパス中の正解用語と共起する助詞情報から計算される助詞スコアを，既に提案されている用語抽出法である FLR，UCH と組み合わせてリランキングした結果，用語抽出精度が向上した．また，FLR，UCH のどちらと組み合わせても精度が向上したことから，助詞情報は手法に依存せずに利用できることを確認した．さらに，正解用語を用いずに上位の用語候補と共起する助詞情報から得られる助詞スコアを利用しても，用語抽出精度が向上することを実験により確認した．特に低用語候補数の場合にリランキングにより精度が大きく向上した．

今後の予定として，本稿においては用語抽出部において導入した係り受け解析の有効性についてはまだ評価していないため，この用語抽出部の定量的な評価を行う必要がある．そして，助詞スコア計算では単純な式を用いたが，各助詞の正解確率を利用できるようなモデルの適用を検討している．また，これまで専門用語抽出タスクで扱われてこなかった情報で，助詞のように精度向上に寄与するものがあるかを検討したい．

謝辞

本研究は，文部科学省科学技術振興調整費「戦略的研究拠点育成プログラム」の支援の下に実施した．

参考文献

[1] Patrick Drouin. Term extraction using non-technical corpora as a point of leverage. *Terminology*, Vol. 9, No. 1, pp. 99–117, 2003.

[2] Y. Fukushige and N. Noguchi. Ntcir experiments at matsushita: Tmrec task. In *Proc. of the 1st Conference of NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 467–474, 1999.

[3] K. Kageura, M. Yoshioka, K. Takeuchi, and T. Koyama. Overview of the tmrec tasks. In *Proc. of the 1st Conference of NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 415–415, 1999.

[4] S. Sekine, R. Grichman, and Hiroyuki Shinnou. A decision tree method for finding and classifying names in japanese texts. In *Proc. of the Sixth Workshop on Very Large Corpora*, 1998.

[5] K. Uchimoto, S. Sekine, M. Murata, H. Ozaku, and Isahara H. Term recognition by using corpora from different field. *Terminology*, Vol. 6, No. 2, pp. 233–256, 2001.

[6] 英語教育のための分野特徴単語の選定尺度の比較. 内山将夫 and 中條 清美 and 山本 英子 and 井佐原 均. 自然言語処理, Vol. 11, No. 3, pp. 165–198, 2004.

[7] 辻河亨, 吉田稔, 中川裕志. 語彙空間の構造に基づく専門用語抽出. 情報処理学会 研究報告 159, pp. 155–162, 2004.

[8] 三浦康秀, 増市博. 部分文字列のパープレキシティを利用した低頻度専門用語抽出. 情報処理学会 研究報告 180, pp. 139–144, 2007.

[9] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[10] 中川裕志, 森辰則, 湯本紘彰. 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.

[11] 長町健太, 武田善行, 梅村恭司. 文書拡張によるキーワード抽出. 自然言語処理, Vol. 14, No. 1, pp. 67–86, 2004.

[12] 木浪孝治, 池田哲夫, 高山毅, 武田利明. 品詞の組み合わせの拡張による看護分野での専門用語抽出性能の改善. In *DEWS 2006 1B-i8*, 2006.