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

2007－NL－180 (24)
2007／7／25

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

三浦 康秀† 増市 博†

† 富士ゼロックス株式会社研究本部 〒259-0157 神奈川県足柄上郡中井町境430
E-mail: †{yasuhide.miura,hiroshi.masuichi}@fujixerox.co.jp

あらまし 本稿では，専門分野コーパス内に出現頻度の低い専門用語の候補文字列があるときに，その文字列を構成する部分文字列および専門分野コーパス内での周辺文字列のパープレキシティ用いて，専門用語としてのスコア付けを行う手法を提案する．文字列が与えられたときに，文字列を構成するn-gramの部分文字列を抽出し，それらの専門分野コーパスでのパープレキシティを計算する．また同時に，専門分野コーパス内で文字列の周囲に現れるn-gramの周辺文字列のパープレキシティを計算し，これらの比を文字列のスコアとして設定する．本手法の評価実験として，インターネット上で公開されている病名辞書および解剖学用語辞書の見出し語を構成する文字列で，約6,7000件の医療テキスト内での出現回数が5回以下の文字列についてスコア付けを行い，上位200文字列の用語としての成立の可否を医師が確認した．また，比較のため名詞の出現頻度および連接頻度を用いるTermExtractでも同様の実験を行った．結果として平均で，1-gramでは正解率70.4%，2-gramでは正解率83.5%が得られ，TermExtractによる正解率70.6%と比較して良好な結果が得られた．

キーワード パープレキシティ，専門用語抽出

Extracting Low Frequency Terms Using Substring Perplexities

Yasuhide MIURA† and Hiroshi MASUICHI†

† Corporate Research Group, Fuji Xerox Co., Ltd. 430 Sakai, Nakai-machi, Ashigarakami-gun, Kanagawa, 259-0157 Japan
E-mail: †{yasuhide.miura,hiroshi.masuichi}@fujixerox.co.jp

Abstract This paper describes a extraction method of low frequency domain specific terms, using substring perplexities. When a string is given, n-grams of characters that compose the string are extracted, and their perplexities in a given corpus are calculated. Similarly, n-grams of characters that appear beside the string and their perplexities are extracted. The ratio of these two kinds of perplexities is set as a score that represents the word fitness of the string. As an experiment, n-grams that compose entries in a disease dictionary and a anatomy dictionary, and appear 5 times or less in the corpus of about 67,000 medical texts are scored with the proposed method. In comparison, the same n-grams are scored with TermExtract. The result is, the average accuracy of 70.4% is gained with 1-gram scoring, and 83.5% is gained with 2-gram scoring, and is better compared with 70.6% of that of TermExtract.

Key words Perplexity, Term Extraction, Named Entity Extraction

1. はじめに

自然言語処理の重要な課題の1つに，既存の辞書にはない未知語をコーパスから抽出することがある．未知語の存在は，形態素解析，構文解析等の誤りに繋がり，情報検索，情報抽出等，多くの自然言語処理技術の性能に影響を及ぼす．特に，専門性の高いテキストでは，一般的な辞書には登録されていない多くの専門用語が未知語として現れる．このため，結果として専門分野のテキストでは未知語の割合が大きくなり，一般的なテキストと比べて自然言語処理の適用が難しくなっている．

近年，多くの専門分野の辞書がコンピュータで扱えるように電子化されている．これらの辞書の見出し語を利用すれば，ある程度は専門用語に関する未知語の問題に対処することが可能である．しかし，一般的に専門用語の辞書は自然言語処理のためには作成されてはおらず，単に見出し語を既存の形態素解析の辞書等に追加して利用するには課題もある．例えば，医療分

野の病名辞書として，MEDIS 標準病名マスター(注1)があるが，この辞書では“サイトメガロウイルス病”，“サイトメガロウイルス肺炎”は見出し語として存在しているが，“サイトメガロウイルス”は見出し語として存在していない．これは恐らく，サイトメガロウイルスがウイルス名であり病名ではないためであるが，自然言語処理に適用する観点から見れば，網羅性に問題があるといえる．このため，電子化された辞書が充実してきたとはいえ，コーパスから自動的に専門用語を抽出する技術は必要とされている．

本稿では，このようなコーパスから専門用語を自動的に抽出する手法，特に，医療分野のような，一般的には用いられない専門用語多く出現し，また低頻度の単語の抽出が重要となるテキストからの抽出に適した手法について述べる．本稿の構成としては，2. で既存の専門用語，固有名抽出手法を概説し，3. で本稿が対象としている医療分野のテキストからの専門用抽出の方針について述べる．4. で本稿で提案する抽出手法の詳細を説明し，5. に提案手法に関して実施した評価実験とその結果を記す．最後に，6. で考察および今後の展望を述べる．

2. 関連研究

コーパスから自動的に専門用語，固有名等を抽出する研究は，従来より盛んに行われている [1]〜[8]．福田ら [1] は，医学生物学分野のコーパスから人手で設定されたルールを用いた，高精度のタンパク質名抽出手法を報告している．しかし，この手法を他の分野に適用するには，ルールを設定できる専門家の協力が必要となる．Sekine ら [2]，山田ら [3] 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ている．これらの手法はタグ付きコーパスが存在する分野では容易に適用することができるが，タグ付きコーパスの存在しない分野ではコーパス作成の問題がある．Nagao ら [4]，長町ら [5] は，生コーパスから文字の n-gram の統計量を用いて専門用語の抽出を行っている．Shimohata ら [6]，中川ら [7] は，単語の統計量を用いて専門用語の抽出を行っている．小山ら [8] は，名詞形単語の連接から，非語/非用語を構成しやすいパターンを排除することにより，専門用語の抽出を行っている．これらは，ある程度の規模の生コーパスがあれば専門用語の抽出が行え，幅広い分野への適用が容易に可能である．しかし，これらの手法にも課題はあり，Nagao ら [4]，Shimohata ら [6] の手法は，専門用語の候補の文字列もしくは連語に対して，その直接の統計量を用いて抽出を行うため，コーパス内での出現頻度の低い専門用語の抽出が難しい．長町ら [5] ら，中川ら [7] の手法は，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が，スコアの基準がコーパスに依って変化するため，専門用語の自動抽出には別途，スコアの閾値を経験もしくは実験により定めなければならない．小山ら [8] の手法は，特定のパターンに合致する単語を非語/非用語として排除しており，例えば，単語を構成する形態素列の接頭辞が“各，御，今，他，第，同，本，約”のいずれかであるものは，用語として用いられる可能性が極めて低いと述べている．しかし，医療分野では“第十二助骨”，“同側性片麻痺”等，上記の接頭辞を持つ語が散見され，分野によってはパターンの修正を行う必要性がある．

多発	タハツ	タハツ	名詞-サ変接続
性	セイ	性	名詞-接尾-一般
助	ジョ	助	接続詞-名詞接続
骨	ホネ	骨	名詞-一般
骨折	コッセツ	骨折	名詞-サ変接続

図 1 形態素解析の誤り例
Fig. 1 The example of a morphlogy analysis error

3. 専門用語抽出の方針

本稿では，医療分野のテキストから病名・部位名等の専門用語を抽出することを目的としている．本稿が対象とする医療分野には，一般的に利用可能な専門用語のタグ付きのコーパスは存在していない．また，抽出の方針として，不足なく専門用語の抽出を行うことを目的としている．これは，例えば病名については，重要性が高いにも関わらず病気の発生が稀であるために，コーパス中での出現頻度が低くなるものが多数存在するためである．

生コーパスから専門用語を抽出する手法については，2. 節で述べたように，文字単位で行う手法と単語単位で行う手法が考えられる．中川ら [7] は，専門用語の多くが複合名詞であることに注目し，名詞の出現頻度および連接頻度を用いて，高い精度での抽出を実現している．しかし，日本語ではテキストには明確な境界がないため，単語単位での統計量を用いるには，形態素解析等でテキストを分かち書きにする必要がある．また，一般的に形態素解析の精度は辞書に依存し，分野独自の専門用語が多い医療分野のテキストでは精度が低くなってしまう．例えば，医療分野の用語として“多発性助骨骨折”があるが，これを形態素解析器の茶筌(注2)で解析すると図 1 のようになり，細かく分けられてしまう上に，“助骨”の“助”が接続詞となってしまう．

このため，本稿では文字の n-gram に基づく専門用語の抽出を試みる．また，連接する単語 [7] や，用語の区切り [2], [3] を考慮して抽出を行うことにより良い結果が得られていることを踏まえ，文字列のコーパス内でのパープレキシティを利用することを考えた．これは，単語のエントロピーを用いて連語の抽出を試みた Shimohata ら [6] の手法を，文字の n-gram で置き換えた形に近く，文字列を s として，連接する n-gram の集合を C_n としたとき，文字列のパープレキシティ $PP(s)$ は式 (2) のように計算される [9]．

(注1)：(財) 医療情報システム開発 (http://medis.or.jp/) の標準マスターにて公開されている．

(注2)：http://chasen-legacy.sourceforge.jp/にて公開されている．なお，本稿の例は ver. 2.3.3 に基づいている。

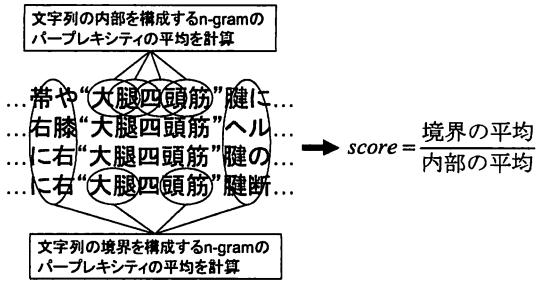

図 2　基本的な考え方

Fig. 2　The Basic Concept

$$H(s) = -\sum_{C_n} P(c_n)\log P(c_n) \tag{1}$$

$$PP(s) = 2^{H(s)} \tag{2}$$

しかし，この手法には 1 つの問題がある．それは，文字列のパープレキシティの上限が，その性質上 s の出現頻度になり，出現頻度の低い s に関しては値の信頼性が低くなることである．そこで本稿では，コーパス内での出現頻度の低い文字列に対してもパープレキシティに基づく専門用語抽出を行えるように，文字列を構成する n-gram の部分文字列およびコーパス内で文字列の周囲に現れる n-gram の周辺文字列のパープレキシティを用いた，低頻度文字列に対しても有効なスコア付け手法を提案する．

4. 提案手法

提案手法は，文字列と専門分野の生コーパスが与えられたときに，文字列を構成する n-gram の部分文字列のパープレキシティ，およびコーパス内で文字列の周囲に現れる n-gram の周辺文字列のパープレキシティを用いて，文字列の内部と境界のパープレキシティを計算し，その比を文字列の専門用語らしさのスコアとして与える．基本的な考え方としては，文字列を構成する n-gram の文字列の内部のパープレキシティより，文字列の境界のパープレキシティが大きいほど，文字列は専門用語らしいと仮定している（図 2）．スコア付け手法の詳細手順は，以下のようになる．

（1）　長さ l の文字列 s が与えられたとき，s に含まれる n_s 文字（n_s は 1〜2 程度の小さな値を想定）の部分文字列 I_i（$I_i = s_i^{i+n_s-1}, 1 \leqq i \leqq l-n_s+1$）を全て抽出する．

（2）　部分文字列 I_i の，1 つ左の n_s-gram が s に含まれるもの（$i-1 \geqq n_s$）は，左側の n_p 文字（$n_s \geqq n_p$）のパープレキシティ $PP_L(I_i)$ を，1 つ右の n_s-gram が s に含まれるもの（$(l-n_s+1)-i \geqq n_s$）は，右側の n_p 文字のパープレキシティ $PP_R(I_i)$ を計算する．

（3）　全ての $PP_L(I_i)$ と $PP_R(I_i)$ の平均を計算し，文字列の内部のパープレキシティ $\bar{PP}_{in}(s)$ とする．

（4）　部分文字列 I_i の，s の先頭となるもの I_1 の左側 n_p 文字のパープレキシティ $PP_L(I_1)$，s の末尾となるもの I_{l-n_s+1} の右側 n_p 文字のパープレキシティ $PP_R(I_{l-n_s+1})$ を計算する．

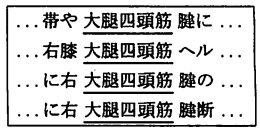

図 3　コーパス上での出現例

Fig. 3　The example of contexts in a corpus

（5）　コーパス内で文字列 s の左側に現れる n_s 文字および右側に現れる n_s 文字を抽出し，これらを L および R とする．

（6）　L に含まれる周辺文字列 L_i の右側 n_p 文字のパープレキシティ $PP_R(L_i)$，R に含まれる周辺文字列 R_i の左側 n_p 文字のパープレキシティ $PP_L(R_i)$，を計算する．

（7）　全ての $PP_R(L_i)$ の平均と，全ての $PP_L(R_i)$ の平均を計算し，$\bar{PP}_{R,out}(s)$ と $\bar{PP}_{L,out}(s)$ とする．

（8）　$\bar{PP}_{R,out}(s)$ と $PP_L(I_1)$ の最小値と，$PP_R(I_{l-n_s+1})$ と $\bar{PP}_{L,out}(s)$ との最小値を取り，2 つの平均を文字列の境界のパープレキシティ $PP_{thresh}(s)$ とする．

（9）　s のスコアとして，n_s-gram の境界のパープレキシティと内部のパープレキシティの比，$score(s) = PP_{thresh}(s)/\bar{PP}_{in}(s)$ を設定する．

n_s が小さければ，文字列の全体としての出現頻度が低くても，部分文字列はパープレキシティの計算に十分な頻度が得られる可能性が高くなる．このため，出現頻度の低い文字列に対しても，全体を用いるのと比較して有意なスコアを与えることができる．また，パープレキシティの比を用いるため，スコアの基準がコーパスに依存せず，スコアの閾値をコーパス毎に設定する必要がない．

例として，s = “大腿四頭筋”，$n_s = 2$，$n_p = 1$ の場合のスコア設定手順を述べる．なお，例中の 2-gram のパープレキシティ値は説明のために付した参考値であり，実際にはコーパスを基に式 (2) によって算出される．

（1）　s = “大腿四頭筋”，$n_s = 2$，$n_p = 1$ として，I = {“大腿”, “腿四”, “四頭”, “頭筋”} が抽出される．

（2）　1 つ左の 2-gram が “大腿四頭筋” に含まれる，PP_L(“四頭”) = 1.00，PP_L(“頭筋”) = 4.25 が計算される．また，同様に 1 つ右の 2-gram が含まれる，PP_R(“大腿”) = 5.73，PP_R(“腿四”) = 1.00 が計算される．

（3）　“大腿四頭筋” の内部のパープレキシティとして，$\bar{PP}_{in}$(“大腿四頭筋”) = 3.00 が計算される．

（4）　“大腿四頭筋” の，先頭の 2-gram の PP_L(“大腿”) = 9.35，末尾の 2-gram の PP_R(“頭筋”) = 11.70 が計算される．

（5）　コーパス内で “大腿四頭筋” が図 3 のように現れるとして，L = {“帯や”, “右膝”, “に右”, “に右”}，R = {“腱に”, “へル”, “腱の”, “腱断”} が抽出される．

（6）　“大腿四頭筋” の左側に現れる文字列の，PP_R(“帯や”) = 8.53，PP_R(“右膝”) = 13.20，PP_R(“に右”) = 19.21 が，右側に現れる文字列の，PP_L(“腱に”) = 2.61，PP_L(“へル”) = 12.92，PP_L(“腱の”) = 4.97，PP_L(“腱断”) = 2.31 が計算される．

（7）　“大腿四頭筋” の $\bar{PP}_{R,out}$(“大腿四頭筋”) = 15.04 と，

他に分類される感染症における腹膜の障害
尿の微生物学的検査における異常所見
ヘパリン・コファクターII欠乏症
アクチバトール（FKO）不適合
脳梗塞の続発・後遺症
躁状態双極性感情障害
COLLES骨折
リンパ肉腫（症）
血管リンパ管腫
肩甲筋痛

図4 MEDIS標準病名マスターの見出し語例

Fig. 4 The example of entries in "MEDIS Standard Disease Master"

仙骨神経および尾骨神経［S1－S5，Co］
小脳半球（HII）～小脳半球（HX）
抹消自律神経叢と抹消自律神経節の胸部
涙腺動脈の中硬膜動脈との吻合枝
小脳半球小葉（H－VIIa）
S状結腸リンパ節
三叉神経運動核
頭蓋，ズガイ
網嚢の上陥凹
肺静脈口

図5 最新解剖学用語集の見出し語例

Fig. 5 The example of entries in "Terminology of Anatomy"

$\bar{PP}_{L,out}$("大腿四頭筋") = 5.70 が計算される．

（8） "大腿四頭筋" の境界のパープレキシティとして，$\bar{PP}_{R,out}$("大腿四頭筋") と PP_L("大腿") の最小値 9.35 と，PP_R("頭筋") と $\bar{PP}_{L,out}$("大腿四頭筋") の最小値 5.70 の平均，PP_{thresh}("大腿四頭筋") = 7.53 が設定される．

（9） "大腿四頭筋" のスコアとして，$score$("大腿四頭筋") = 2.51 が設定される．

5. 評価実験

専門用語を抽出するテキスト T として，MEDIS標準病名マスター ver. 2.53 および最新解剖学用語集(注3) の見出し語を用いた．それぞれのフォーマットは，図4，図5のようになっており，病気の説明文的なものも含まれている．パープレキシティを計算するためのコーパス C としては，実際の医療レポート約6,7000件から抽出したテキストを用いた．なお，テキストは全て医師が自由記述により入力したものである．

評価実験は，これらの T, C を用いて，1-gram と 2-gram の2通りについて，以下の手順で行った．

（1） T を構成する，全ての異なり文字列集合 S を抽出する．

（2） S の中で，C での出現頻度 f が5回以下である文字列を抽出し，これらを $S_{f\leqq 5}$ とする．

（3） $S_{f\leqq 5}$ 内の文字列に対して，以下の4種類の簡単なフィルタリングを行い，専門用語の候補となりえないものを削除し，これらを $S'_{f\leqq 5}$ とする．

（a） 見出し語中に定型的に使用される，"が，を，する，と，に，または，による，の，される，のための，および，からの，における，との，への" らの，平仮名文字列を含むもの．

（b） 文字列の先頭が捨て仮名，句読点，中点，等であるもの．

（c） 文字列の末尾が読点，中点，等であるもの．

（d） 茶筌の標準辞書である，IPADIC 2.7.0 の見出し語を構成するもの．

（4） $S'_{f\leqq 5}$ 内の文字列に対して，提案手法でスコアを付け，これを S_{ranked} とする．

（5） S_{ranked} の上位200を，医師が用語として適切であるか確認する．

また，比較のため，中川ら[7]の手法に基づくTermExtract(注4)を用いて，同様に $S'_{f\leqq 5}$ 内の文字列に対してスコアを付け，上位200を医師が確認した．なお，TermExtractについてもコーパスの統計量を用いて解析を行うように，予めTermExtractの持つ学習機能を用いて C に対して学習を行わせた．

結果は，MEDIS標準病名マスターに関しては表1，最新解剖学用語集に関しては表2のようになった．なお，表中の RR_{sum} はランキング性能を現すための，Reciprocal Rank（正解の順位の逆数）の和である．なお，3つの手法での上位20およびそのスコアは，1-gramでは表3，4のようになり，2-gramでは表5，6のようになり，TermExtractでは表7，8のようになった．結果として，2種類のテキストの平均で，1-gramでは正解率70.4%であったが，2-gramでは正解率83.5%が得られ，TermExtractによる正解率70.6%と比較して良好な結果が得られた．

6. おわりに

本稿では，専門用語の候補文字列があるときに，その文字列を構成するn-gram部分文字列および専門分野コーパス内でのn-gram周辺文字列のパープレキシティ用いて，専門用語としてのスコア付けを行う手法を提案した．2種類の医療テキスト

表1 MEDIS標準病名マスターからの上位200の評価結果

Table 1 Evaluation results of top 200 strings from "MEDIS Standard Disease Master"

	TermExtract	1-gram	2-gram
正解率	69.85%	64.82%	83.50%
RR_{sum}	4.05	3.47	4.97

表2 最新解剖学用語集からの上位200の評価結果

Table 2 Evaluation results of top 200 strings from "Terminology of Anatomy"

	TermExtract	1-gram	2-gram
正解率	71.36%	76.02%	83.50%
RR_{sum}	4.71	4.30	5.49

(注3)：http://web.sc.itc.keio.ac.jp/~funatoka/anatomy.html

(注4)：http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html にて公開されている．

表 3 MEDIS 標準病名マスターからの 1-gram での上位 20

Table 3 Top 20 results from "MEDIS Standard Disease Master" using 1-gram perplexity

	String	Score		String	Score
1	背後	19.08	11	シニア	13.87
2	脂質	16.34	12	末節	13.67
3	脳死	15.37	13	背挫傷	13.37
4	転子骨折	15.01	14	甲殻	13.24
5	転子部	14.95	15	リト	12.77
6	リンパ球	14.72	16	食後	11.97
7	静脈塞栓	14.21	17	単胞	11.16
8	高張	14.20	18	軽症	11.16
9	仮死	14.16	19	自己免疫疾患	11.02
10	小胞	13.95	20	転症候群	10.98

表 4 最新解剖学用語集からの 1-gram での上位 20

Table 4 Top 20 results from "Terminology of Anatomy" using 1-gram perplexity

	String	Score		String	Score
1	末節	13.67	11	頭室	9.13
2	脊柱部	12.17	12	骨盤縁	9.11
3	篩骨胞	11.81	13	膜張	8.97
4	横走部	10.90	14	脳梁縁	8.58
5	閉鎖管	10.19	15	腹静脈	8.40
6	神経後	9.83	16	下腿部	8.38
7	索リンパ節	9.81	17	橈尺関節	8.22
8	白膜	9.56	18	海綿体	8.19
9	下顎後	9.56	19	海綿骨	8.19
10	下顎リンパ節	9.29	20	下腿骨	7.82

表 5 MEDIS 標準病名マスターからの 2-gram での上位 20

Table 5 Top 20 results from "MEDIS Standard Disease Master" using 2-gram perplexity

	String	Score		String	Score
1	肺動脈瘤	16.24	11	リンパ球性下垂体炎	11.18
2	上大動脈	14.42	12	石灰化症	11.07
3	リンパ嚢腫	14.11	13	総胆管癌	10.97
4	尿膜管癌	13.70	14	マイヤー	10.66
5	冠動静脈瘻	13.35	15	仙腸関節炎	10.37
6	冠状静脈洞	12.57	16	リンパ球性下垂体	10.29
7	冠動脈瘤	12.41	17	肺炎桿菌性	10.27
8	肺気管支	12.03	18	足関節部	10.26
9	十二指腸傍ヘルニア	12.00	19	前脛骨筋	9.92
10	リンパ管炎	11.50	20	耳下腺癌	9.91

表 6 最新解剖学用語集からの 2-gram での上位 20

Table 6 Top 20 results from "Terminology of Anatomy" using 2-gram perplexity

	String	Score		String	Score
1	冠状静脈洞	12.57	11	股関節部	9.06
2	奇静脈弓	12.17	12	眼窩下動脈	8.94
3	十二肋骨	11.96	13	下顎後静脈	8.71
4	左外側区	11.41	14	三叉神経節	8.61
5	脳梁縁動脈	10.75	15	肺動脈弁	8.55
6	後脛骨筋	10.43	16	後半規管	8.15
7	外側胸動脈	10.28	17	腹腔神経叢	8.05
8	前脛骨筋	9.92	18	左尾状葉	8.03
9	反回動脈	9.41	19	動脈管索リンパ節	7.99
10	外側上顎	9.09	20	左胃大網	7.61

表 7 MEDIS 標準病名マスターからの TermExtract での上位 20

Table 7 Top 20 results from "MEDIS Standard Disease Master" using TermExtract

	String	Score		String	Score
1	閉鎖性	102654.28	11	脳障害	32811.31
2	結節性甲状腺腫	77993.12	12	脳内出血	32169.84
3	神経性	49648.43	13	側性腎	30810.98
4	腫腎	47349.98	14	脂血症	28696.61
5	骨症	43200.52	15	上皮内癌	28517.50
6	多発性損傷	35690.22	16	腎性骨	27647.64
7	縦靭帯骨化症	35645.12	17	間性	27169.26
8	状リンパ管腫	35483.25	18	骨化性節炎	26000.21
9	脳室腫瘍	33545.74	19	膜外出血	25219.48
10	胃性	32908.10	20	急性リンパ節炎	25171.78

表 8 最新解剖学用語集からの TermExtract での上位 20

Table 8 Top 20 results from "Terminology of Anatomy" using TermExtract

	String	Score		String	Score
1	葉動脈	168468.93	11	骨間動脈	44009.56
2	肺底静脈	132312.23	12	背枝	39370.08
3	体神経節	93550.87	13	右冠状動脈	39014.01
4	背側核	67682.71	14	葉間動脈	38564.53
5	左冠状動脈	62305.74	15	腹側核	38242.97
6	肺底動脈	58308.92	16	腹側視床	38213.80
7	外側核	52978.12	17	外側肺底	37928.97
8	篩骨動脈	52398.18	18	膜枝	35625.10
9	背内側核	49120.99	19	筋枝	34462.73
10	間脳	48601.46	20	外側枝	34140.40

を対象に，コーパス内での出現頻度の低い文字列に対して評価実験を行ったところ，TermExtract と比較して，1-gram でほぼ同等，2-gram で約 13%の正解率の向上が得られ，提案手法の有効性を確認することができた．

本手法の今後の課題としては，以下の点が挙げられる．

（1） 高頻度用語の抽出性能の評価

（2） 一般的なテストセットを用いての性能評価

(1) については，本手法はコーパス内で低頻度な文字列に対するスコア付けしか評価を行っておらず，高頻度な文字列のスコアが適切に設定されるかは確認していない．無論，出現頻度が高ければ文字列全体を用いてのパープレキシティ（もしくはエントロピー）での評価が可能であるが，本手法の出現頻度によらない性能を評価するためにも行いところである．

(2) については，本手法は医療テキストという，極めて専門性の高いテキストでの評価しか行っておらず，一般的なテキストでの用語抽出性能の評価は行っていない．森山ら [10] は，特性の異なる複数のテストコレクションに対する抽出手法の比較を行っているが，手法によってはテストコレクション間の精度

性の異なる複数のテストコレクションに対する抽出手法の比較を行っているが，手法によってはテストコレクション間の精度が大きくばらついており，本手法のより正確な性能評価には特性の異なるテストコレクションを用いた評価実験を予定している．このため，長町ら [5]，中川ら [7]，小山ら [8] が評価に用いている，NTCIR1 の TMREC テストコレクション(注5)等を用いた評価実験を検討している．

文　献

[1] 福田賢一郎, 角田達彦, 田村あゆち, 高木利久，医学生物学文献からの専門用語の抽出に向けて：タンパク質名の自動抽出, 情報処理学会論文誌, **39**, 8 (1998).

[2] S. Sekine, R. Grishman and H. Shinnou, A decision tree method for finding and classifying names in japanese texts, Proceedings of the 6th Workshop on Very Large Corpora (1998).

[3] 山田寛康, 工藤拓, 松本裕治，単語の部分文字列を考慮した専門用語抽出と分類, 情報処理学会研究報告．自然言語処理研究会報告, **2000**, 107 (2000).

[4] M. Nagao and S. Mori, A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese, Proceedings of COLING-1994 (1994).

[5] 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, **14**, 1, pp. 67 86 (2007).

[6] S. Shimohata, T. Sugio and J. Nagata, Retrieving collocations by co-occurrences and word order constraints, Proceedings of ACL/EACL-97 (1997).

[7] 中川祐志, 湯元紘彰, 森辰則，出現頻度と連接に基づく専門用語抽出, 自然言語処理, **10**, 1, pp. 27 45 (2003).

[8] 小山照夫, 影浦峡, 竹内孔一，日本語専門分野テキストコーパスからの複合語用語の抽出, 情報処理学会研究報告．自然言語処理研究報告, **2006**, 124 (2006).

[9] 北研二，確率的言語モデル, 東京大学出版会 (1999).

[10] 森山聡, 辻河亨, 吉田稔, 中川祐志，専門用語抽出手法のテストコレクション依存性, 情報処理学会研究報告．自然言語処理研究報告, **2004**, 161, pp. 9 15 (2004).

(注5)：NTCIR のサイト (http://research.nii.ac.jp/ntcir/) より入手可能.