Vol.2019-DBS-169 No.8
Vol.2019-IFAT-136 No.8
2019/9/10

構文木情報と専門用語辞書を用いた医学論文からの未知用語の発見

吉田　恭輔[1,a)]　湯本　高行[1,b)]　金子　周司[2]　礒川　悌次郎[1]　松井　伸之[1]　上浦　尚武[1]

概要：近年，医療や生命科学分野の急激な進歩により，毎年大量の論文が出版され，それに伴って新たな病名や専門用語が生み出されている．専門用語は，その分野における概念を端的に表しており，その自動抽出は新たに生まれた概念を理解するための重要な役割を持つと考えられる．そこで本研究では，機械学習を用いた系列ラベリングによる専門用語抽出の手法を提案する．未知専門用語にも対応した抽出を行うことを目的として，文章中における単語間の関係を表す構文木情報と専門用語辞書を素性として用いる．これらの素性によって学習することで，専門用語の言語的特徴と頻出の表現を考慮したラベリングを行い，専門用語を抽出する．

1. はじめに

近年，医療や生命科学の分野の急激な進歩により毎年大量の論文が出版され，膨大な数の新たな病名や専門用語が生み出されている．論文の多くは英語で書かれており，新たな専門用語は英語の用語として生み出されている．専門用語は，その分野における成果や概念を端的に表しており，その分野のさらなる発展のための重要な役割を担っている．そのため，新たに生まれた概念を説明する要素として，専門用語は重要であり，その専門用語の抽出は，論文の読者にとって理解の手助けになると考えられる．専門用語の抽出の作業は，専門家の手によって行われることが正確であるが，大量の論文を読み，専門用語の抽出を行わなければならないため膨大な時間と労力がかかる．そのため，機械的な処理によって高速かつ効率的な処理が求められる．

そこで，本研究では，未知の専門用語にも対応できるように，機械学習を用いて英語の医学論文から専門用語を抽出する手法を提案する．提案手法では，文章中の単語の関係を表す構文木情報と，専門用語辞書を用いた素性から，専門用語の言語的な特徴と頻出の表現を学習することで系列ラベリングを行う．

また，提案手法によるラベリング結果から専門用語を抽出し，その結果を，既存の手法による専門用語抽出と比較し評価する．

[1] 兵庫県立大学
[2] 京都大学
[a)] ei19w029@steng.u-hyogo.ac.jp
[b)] yumoto@eng.u-hyogo.ac.jp

2. 関連研究

文章中から専門用語を抽出する研究として，統計情報と言語情報をパラメータとして用いた Frantzi らの研究がある [1]．この研究では専門用語の品詞とその出現パターン言語情報とし，出現頻度や他の用語の一部として出現する頻度，種類数を統計情報として組み合わせたスコアである C-value を定義して抽出を行っている．また C-value に文脈情報を組み合わせたスコア NC-value を定義し精度の向上を図っている．この研究では候補用語を言語的なパターンで抽出した後に，スコア付けを行っている．

専門用語を構成する単語における出現頻度と連接頻度を利用した中川らの研究がある [2]．本論文では，この手法をベースラインとして比較実験を行うため，以下にその手法を示す．この手法では，まず候補用語として複合名詞を抽出する．抽出されたある複合名詞を CN としたとき，CN を構成する単名詞 N について左方の単語と連接して複合名詞を形成する回数を FL(N)，右方の単名詞と連接して複合名詞を形成する回数を FR(N) とし，以下の式で単名詞数 L の CN の連接頻度のスコア LR(CN) を算出する．

$$LR(CN) = (\prod_{i=0}^{L}(FL(N_i)+1)(FR(N_i)+1))^{\frac{1}{2L}} \quad (1)$$

このスコア LR(CN) に CN 自体の出現頻度を掛けることで複合名詞にスコア付けを行う．

複合語に対するスコア付けの研究として森山らの研究がある [3]．この研究においても，候補用語として複合語を抽出する．抽出された複合語を構成する単語に対して，連接

する単語の種類数や頻度からパープレキシティを用いることで，複合語のスコア付けを行っている．

他にも，コーパス中の安定して使用される度合いを表すユニット性を用いたスコア付け [4] や，注目する用語と共起する単語の分布を用いたスコア付け [5]，2 言語コーパスを用いたスコア付け [6][7]，単名詞と複合語の関係に着目したスコア付け [8] が試みられた．

本研究では構文木情報や専門用語辞書を用いた素性から学習を行い，系列ラベリングを行うため，候補用語と専門用語の選定を同時に行う．また，医学分野の専門用語辞書を用いることで，医学分野特有の用語を抽出する．これら点において，本研究は以上の研究とは異なる．

3. 提案手法

3.1 概要

本節では，与えられた文書集合に対して，構文木情報と専門用語辞書を用いた素性から，文書集合における専門用語を，条件付き確率場 (Conditional Random Field，CRF)[9] によって学習し，系列ラベリングを行う手法について述べる．構文木は，文章中の文法的な単語間の関係を関係を表しており，構文木情報を素性に用いることで，専門用語の文法的なパターンを学習することが可能である．また，ライフサイエンス辞書 (Life Science Dictionary，LSD)[*1][10] を専門用語辞書として使用し，LSD に収録されている専門用語の情報を素性として用いることで，専門用語特有の表現の学習を行う．

本手法における系列ラベリングの手順を述べる．まず，与えられた文書集合の全単語に対して素性を付与する．次に文書集合中に存在する，LSD に収録されている専門用語 (以下 LSD 専門用語という) に対して，構文木における出現パターンに基づいた正解ラベルを付与し，CRF によって学習することで，系列ラベリングを行う学習モデルを作成する．今回は，CRFsuite[*2]を用いた．

3.2 構文木情報

正解ラベルと素性に用いる構文木情報について述べる．構文木とは，文章中における品詞情報，単語間の修飾関係などの文法的な関係を木構造で表したものである．本手法では，Stanford Parser[*3]を用いて構文木解析を行う．

例として ‘Cytokine gene transcription have been described.’ という文章に対して，Stanford Parser 用いた構文木解析結果を**図 1** に示す．Stanford Parser による構文木解析では入力文章の単語に対して，1 から順番に番号が割り当てられる．また，修飾関係については，上位ノードから下位ノードに対して矢印で結ばれ，その横に文法的な関係が表示される．例として，図 1 における，‘amod’ は形容詞である下位ノードが上位ノードの名詞を修飾するという意味である．Stanford parser による構文木解析では，入力にかかわらず，最上位ノードとして ‘None’ が 0 番目に割り当てられ，次に上位のノードと ‘root’ で繋がれる．図 1 のように，構文木上では単語間の修飾関係が表される．ここで，本論文では図 1 の，‘Cytokine gene transcription’ における ‘transctiption’ のように，名詞句を形成する場合の意味的に中心の単語を名詞句のコアと定義する．

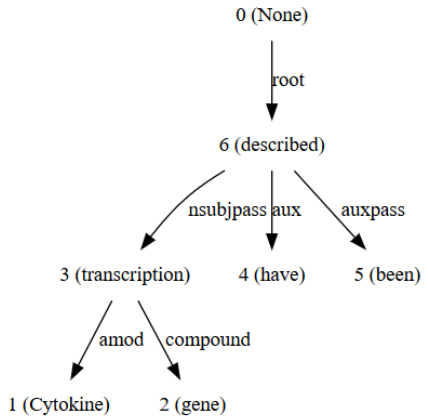

図 1 構文木例

3.2.1 事前調査

LSD 専門用語について，構文木における出現のパターンの事前調査を，実験に用いる文書集合を用いて行った．

文書集合において，LSD 専門用語は 26140 語出現した．このうち，20718 語の LSD 専門用語が，構成する末尾の単語をコアとして，それ以外の単語が修飾する形で出現した．このような専門用語について，コアを親部分，親を修飾する単語を子部分として親子型専門用語と定義し，その例を**図 2**(a) に示す．図 2(a) は LSD 専門用語の ‘citric acid cycle’ であり，末尾の ‘cycle’ がコアとして親部分となり，‘citric’ と ‘acid’ が修飾する形になっている．また，3782 語の LSD 専門用語が，構成する単語とは別の単語をコアとし，構成する全ての単語がコアを修飾する形で出現した．このような専門用語を兄弟型専門用語と定義し，その例を図 2(b) に示す．図 2(b) は LSD 専門用語の ‘acinar cell’ であり，この場合では専門用語を構成する単語ではなく，後ろに続く単語の ‘volume’ がコアとして親部分となり，‘acinar’ と ‘cell’ が修飾する形となっている．それ以外の LSD 専門用語については，出現パターンが判断できなかったため，その他型専門用語と定義し，図 2(c) に例を示す．図 2(c) は LSD 専門用語の ‘foot and mouth disease’ であり，3 層の木構造になっていることがわかる．

3.2.2 ラベルの種類

事前調査によって，LSD 専門用語の出現パターンとして，親子型，兄弟型，その他型が存在していることがわかっ

[*1] https://lsd-project.jp/cgi-bin/lsdproj/ejlookup04.pl2
[*2] http://www.chokkan.org/software/crfsuite/
[*3] https://nlp.stanford.edu/software/lex-parser.shtml

Vol.2019-DBS-169 No.8
Vol.2019-IFAT-136 No.8
2019/9/10

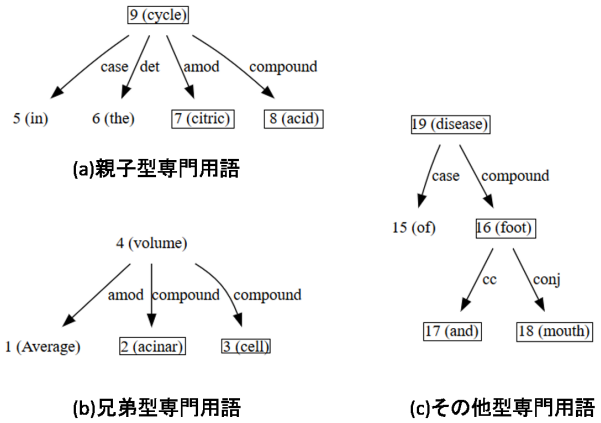

図 2　専門用語の出現パターン

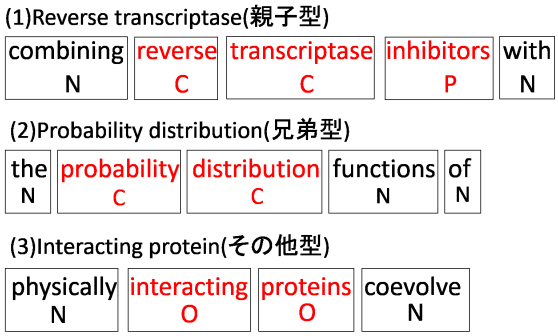

図 3　ラベルの例

た．これらの出現パターンに従うラベルの付与を文書集合に行う．親子型専門用語のコア部分に対しては，‘P’ のラベル，親子型専門用語のそれ以外の単語に対しては ‘C’ のラベルを付与する．兄弟型専門用語に対しては，親子型専門用語の子部分と同様に ‘C’ のラベルを付与し，その他型専門用語に対しては ‘O’ のラベルを付与する．また，上記以外の単語に対しては，専門用語には関係のない単語として，‘N’ のラベルを付与する．

ラベリングの例を図 3 に示す．本手法ではこのラベルに基づいて専門用語を抽出する．図 3(1) では親子型専門用語として，‘C C P’ のラベルに基づき ‘reverse transcription inhibitors’ が抽出される．同様にして，図 3(2) では兄弟型専門用語として，‘C C’ ラベルに基づき ‘probability distribution’ が，図 3(3) ではその他型専門用語として，‘O O’ のラベルに基づき ‘interacting proteins’ が抽出される．

3.3　素性

CRF によって学習を行う際に用いる素性について以下に説明する．

3.3.1　LSD での出現

注目単語が専門用語を構成するような単語かどうかを判定するために，注目単語が LSD に見出し語の一部として出現するかどうかを素性として用いる．注目単語が LSD において出現する場合は ‘True’，出現しない場合は ‘False’ の値をとる．

3.3.2　一般語コーパスでの頻度

専門分野の論文においても，一般的にも使われるような単語は多く出現する．そのような一般的な単語を判別するための素性として．実験で用いる文書集合とは別に，特定の分野によらないコーパス (一般語コーパス) における出現頻度を用いる．素性の値としては，注目単語の一般語コーパスにおける出現回数に対して，一般語コーパスで最も頻出する単語の出現回数で割ることで正規化し，その値の常用対数の小数点以下を切り上げた値をとる．この値は-7 から 0 の値をとり，一般語コーパスにおいて出現しない単語には ‘None’ を与えることで，9 段階の値を素性として用いる．

3.3.3　上位ノードとの繋がり方

親子型専門用語において，構文木上における親部分の単語の修飾関係は，その文節自体，言い換えれば専門用語自体の，他の文節との修飾関係を表すことから，親部分の判定は重要である．そこで，注目単語が構文木上において，上位のノードをどのように修飾しているかを素性として用いる．構文木において全ての単語は一つの単語を修飾しているため，Stanford Parser を用いた構文木解析結果から上位ノードとの繋がり方情報を素性の値とし，その値は 38 種類存在する．上位ノードとの繋がり方の一覧を表 1 に示す．

表 1　上位ノードとの繋がり方の一覧

acl	acl:relcl	advcl	advmod
amod	appos	aux	auxpass
case	cc	cc:preconj	ccomp
compound	compound:prt	conj	cop
csubj	csubjpass	dep	det
det:predet	discourse	dobj	expl
iobj	mark	mwe	neg
nmod	nmod:npmod	nmod:poss	nmod:tmod
nsubj	nsubjpass	nummod	parataxis
root	xcomp		

3.3.4　下位ノードとの繋がり方

構文木上において，親子型，兄弟型専門用語の子部分の単語は下位ノードを持たないため，専門用語を抽出する際に，下位ノードとの繋がり方情報は有用な情報である．また，親部分についてもその子部分と名詞同士の繋がり，形容詞と名詞の繋がりをしているため，親部分の判定にも有用であると考えられる．そこで，注目単語の構文木上における下位ノードとの繋がり方情報を素性として用いる．この情報に関しては，構文木上において，単語が下位ノードを持つ個数は同一ではないため，すべての繋がり情報をリストアップし，その単語が持つ下位ノードとの繋がり情報については ‘1’，それ以外には ‘0’ の値をとり，28 次元それ

ぞれを素性として用いる．

下位ノードとの繋がり方の一覧を**表 2** に示す．

表 2 下位ノードとの繋がり方の一覧

neg	mwe	aux	parataxis
cc	conj	nmod	csubj
dobj	case	xcomp	nummod
auxpass	discourse	acl	expl
compound	iobj	cop	advcl
ccomp	nsubj	advmod	det
appos	dep	anod	nsubjpass

3.3.5 n-gram の頻度

専門用語辞書における，連続する単語の組み合わせが頻出する場合，専門用語でよく使用される表現であると考えられ，専門用語らしさの指標になると考えられる．そこで，注目単語に対して，前の単語との 2-gram，後ろの単語との 2-gram，そして前後の単語との 3-gram の 3 つの複数単語の LSD における出現頻度の情報をそれぞれ素性として 3 次元追加する．例として，‘soft tissue therapy’ という 3 単語の ‘tissue’ が注目単語だった場合，前の単語との 2-gram は ‘soft tissue’，後ろの単語との 2-gram は ‘tissue therapy’，3-gram は ‘soft tissue therapy’ である．

n-gram の頻度を用いた素性としては，LSD における出現回数の常用対数の小数点を切り捨てた値をとる．また，先頭の単語に対しては，前の単語との 2-gram と 3-gram が存在しないため ‘BOS’ を与える．末尾の単語にも同様に，後ろの単語との 2-gram と 3-gram には ‘EOS’ を与える．2-gram，3-gram のいずれにおいても，LSD に出現しない場合は ‘None’ を与える．2-gram は 0 から 3 の値に ‘None’ を合わせた 5 段階，3-gram は 0 から 2 の値に ‘None’ を合わせた 4 段階の値を取る．

3.3.6 入力範囲

文章中における単語の前後関係も考慮した学習を行うため，上記の情報と対象単語自体も素性とした，37 次元の素性に，対象単語の前後 2 語の素性も加えた $35 \times 5 = 175$ 次元の素性を CRF による学習に用いる．

4. 実験

4.1 データセット

本論文では，手法によって専門用語を抽出する文書集合として，米国科学アカデミー紀要 (Proceedings of the National Academy of Sciences of the United States of America, PNAS) の抄録 2482 件集めたデータを使用する．また，LSD 専門用語として，LSD に収録されている用語のうち名詞句として収録されている用語を使用し，PNAS の抄録データに出現する LSD 専門用語を正解用語として正解データを作成する．一般語コーパスとしては，Wikipedia に対して前処理を行った後 100MB で切り出したデータである text8[*4] を使用する．

4.2 評価指標

本論文では，手法に対しての評価指標として適合率 (Precision)，再現率 (Recall)，F 値 (F-measure) を使用する．出現頻度と連接頻度による手法では，1 つの複合名詞に対して，1 つのスコアが計算されるため，単純に正解用語についてどれだけの種類を抽出されているかで適合率，再現率，F 値を算出する．提案手法に関しては，単語の前後関係も考慮しているため，同じ複数単語でも，素性によっては抽出される場合とされない場合が存在するため，抽出した用語に対して，その抽出位置も考慮して適合率，再現率，F 値を算出する．例として，‘Escherichia coli’ という専門用語は，‘Similarly, the evolution of Escherichia coli from Salmonella’，‘When expressed in Escherichia coli, the NPC1 protein exhibits lipid’ といったように複数の文章で出現する．このような場合はそれぞれ別の用語として正解数をカウントする．

表 3 の混同行列における各指標の定義式を以下に示す．

表 3 混同行列

		正解データ	
		正	負
抽出データ	正	TP	FP
	負	FN	TN

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

4.3 比較実験

本論文では，2 節で述べた，出現頻度と連接頻度に基づく手法をベースラインとして用いて，提案手法との比較実験を行う．各手法を用いて，PNAS の抄録データからの専門用語抽出を行い，作成した正解データによって抽出精度を算出する．

4.3.1 実験方法

3 節で述べた手法により，PNAS の抄録データに対してラベル，素性付与を行い，CRFsuite に入力することで学習モデルを作成する．作成した学習モデルによってラベリングを行い，評価を行う．また，ラベリング結果から専門用語を抽出した場合の評価も行う．専門用語抽出，ラベリングそれぞれの評価について，5 分割交差検証によって行う．

ベースラインとして，出現頻度と連接頻度に基づく手法を用いた termextract[*5] を使用して専門用語抽出を行う．

*4 http://mattmahoney.net/dc/textdata.html
*5 http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

出現頻度と連接頻度に基づく手法では専門用語を抽出する際，PNAS の抄録データに出現する全複合名詞に対してスコア付けがされる．この複合名詞群について，スコアの降順に出力し，その上位 N 件の複合名詞を専門用語として抽出し，それぞれ精度を算出する．上位 N 件については，N=3000，6000，9000，12000，15000 の 5 パターンで抽出を行う．

4.3.2 実験結果と考察

提案手法によるラベリング結果の精度を 5 分割交差検証で算出した結果を**表 4** に示す．また，ラベリング結果から専門用語を抽出した結果を**表 5** に示す．

表 4 ラベリング結果

ラベル	適合率	再現率	F 値
C	0.908	0.876	0.888
N	0.988	0.990	0.990
O	0.788	0.602	0.680
P	0.884	0.850	0.868

表 5 提案手法の抽出結果

適合率	再現率	F 値
0.825	0.772	0.798

表 4 より，‘O’ 以外のラベルでは適合率，再現率ともに 0.8 を超えており，高い精度でラベリングができていると考えられる．しかし，‘O’ のラベルのその他型専門用語に対しては，再現率が低くなっていることがわかる．これは，提案手法では，その他型の専門用語の出現パターンについて，規則性が存在しない，または複雑であるために発見できていないことが原因である．正解データにおける専門用語のラベルについて，それぞれの出現回数を**図 6** に示す．図 6 より，‘O’ のラベルは専門用語のラベル全体の 7%ほどであり，その影響はそれほど大きくないとわかる．その他型専門用語の構成として，**図 4** のような構文木解析のミスがある．図 4 は，‘confocal laser scanning microscopy’ という LSD 専門用語の構文木解析結果であり，‘laser’ がコアとなり ‘microscopy’ のノードが離れていることがわかる．また，‘and’ や ‘of’ などの前置詞を含む専門用語もその他型に分類され，再現率を向上させるためにはその他型専門用語の出現パターンの調査が必要である．

表 6 ラベルごとの出現回数

ラベル	出現回数
C	30154
P	20718
O	3940

一方，ベースラインの手法では，77817 種類の複合語が抽出された．このうち複合語のスコア上位 N 件の抽出を行った結果を**表 7** に示す．

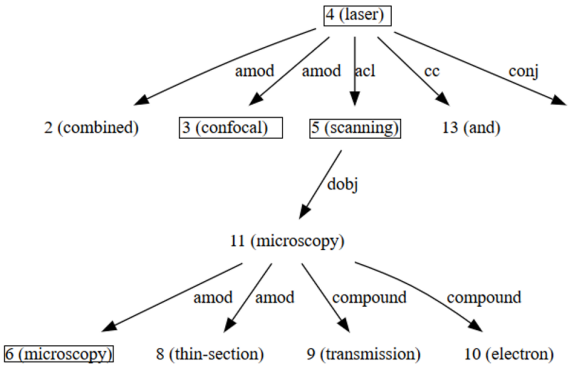

図 4 構文木解析のミス

表 7 ベースラインの抽出結果

N	適合率	再現率	F 値
3000	0.427	0.174	0.247
6000	0.292	0.239	0.263
9000	0.241	0.294	0.265
12000	0.206	0.336	0.255
15000	0.177	0.361	0.238

表 7 より，複合名詞のスコア上位 15000 件を抽出した場合でも，再現率は 0.361 であり，既存の専門用語に対しての網羅性に欠けていることがわかる．そこで，正解用語 7355 語のうち，複合語抽出の時点で，候補用語としてどれだけの正解用語を抽出できているかの調査を行った結果，4857 語抽出されていることがわかった．このことから候補用語の時点で正解用語を 66%ほどしか網羅していないことがわかる．次に，抽出した候補用語の中で正解用語と部分一致している用語の調査を行った結果，6832 語抽出されていた．このことから候補用語のうち，1975 語は正解用語を含むが，完全一致はしない用語であるとわかる．以上のことから，兄弟型専門用語を親子型専門用語と誤認識する場合と同様の誤抽出が行われており，これにより再現率が低くなっていると考えられる．

表 7 と表 5 より抽出結果を比較すると，提案手法において，各評価指標が高くなっていることがわかる．提案手法による抽出が優れた結果になった要因として，文章中における単語間の関係を考慮したことが挙げられる．単語間の関係を考慮することによって，専門用語を文節区切りで判断するのではなく，文節内において不必要な単語は除くような，より専門用語的なパターンで抽出できていると考えられる．兄弟型は実際には，後ろに名詞を加えた親子型であると述べたが，そのような場合に後ろに続く名詞は，一般的によく使われるような単語であることが多かった．例として，‘single amino acid’ という専門用語は，文章中では ‘single amino acid change’ という形で出現し，‘change’ をコアとした兄弟型専門用語である．そのような兄弟型専門用語に対しては，提案手法の素性による一般語除去に

よって正しく抽出されていると考えられる．また，提案手法の素性である n-gram の頻度を考慮したことで，専門用語らしさを学習できたと考えられ，その結果として親子型，兄弟型の判別や，再現率の向上に繋がったと考えられる．しかし，一般語としても専門用語としても頻出するような単語が後ろに続いている場合には親子型と判別される．例として，‘protein interaction’ は PNAS の抄録データには ‘protein interaction data’ として出現する．この場合は ‘data’ がコアとなるが，LSD で頻出するため，除去されず親子型の ‘protein interaction data’ として抽出される．そのため，抽出できなかった兄弟型専門用語について，コアの統計的な調査などを行って，対策を考える必要がある．

4.4 品詞情報の影響の分析

機械学習を用いて系列ラベリングを行う際，素性としてよく用いられるのが単語の品詞情報情報である．品詞情報を素性に追加して，CRF による学習を行い，ラベリングを行うことで，品詞情報の有用性を分析する．

4.4.1 実験方法

提案手法の素性に，注目単語とその前後 2 語の品詞情報 5 次元を追加して，CRF による学習を行うことで系列ラベリングを行い，専門用語を抽出する．4.3.1 と同様にして 5 分割交差検証によって評価を行う．

4.4.2 結果と考察

品詞情報も素性に追加して系列ラベリングを行い，専門用語を抽出した結果を**表 8** に示す．

表 8 品詞情報を考慮した抽出結果

適合率	再現率	F 値
0.825	0.754	0.787

表 5 と比べて，再現率が少し下がっていることがわかる．このことについて，親子型専門用語の親部分と兄弟型専門用語のコアの品詞は動詞であり，同じ品詞情報を持つ．そのため，学習を行うことで，兄弟型の区別ができない専門用語が増加していると考えられる．

5. おわりに

文書集合からの専門用語抽出の手法として，文書集合に対して，専門用語辞書に存在する用語を正解用語とし，構文木情報と専門用語辞書を用いた素性から，CRF を用いて学習することで系列ラベリングを行う手法を提案した．

実験において，提案手法によるラベリング結果から専門用語を抽出し，既存の手法である，出現頻度と連接頻度に基づいた手法による専門用語抽出の結果と比較すると，提案手法の方が優れた結果となった．この要因について，構文木情報を素性に用いて系列ラベリングを行うことで，文章中における専門用語の言語的な特徴を学習することができたためと考えられる．また，専門用語辞書を素性に用いることで，専門用語で頻出の表現の学習を行ったことも要因のひとつと考えられる．しかし，その他型専門用語については，その出現パターンを明確に定義できておらず，その再現率が低くなった．再現率を向上させるためには，その他型専門用語の出現パターンの調査が必要である．また，親子型専門用語と兄弟型専門用語の区別も完全に行えてはおらず，兄弟型のコアの単語が専門用語でも一般的にもよく使われるような単語の場合，コアを親部分とした親子型専門用語としてラベリングしてしまう．そのため，そのような単語の統計的な調査を行い，対処を検討する必要がある．

謝辞 本研究は JSPS 科研費 JP17K00429 の助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Katerina Frantzi, Sophia Ananiadou, Hideki Mima. “Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method”. International Journal on Digital Libraries Vol.3, pp.115-130, 2000
[2] 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
[3] 森山 聡, 吉田 稔, 中川 裕志. 複合語のパープレキシティに基づく重要語抽出の研究. 情報処理学会研究報告自然言語処理, pp.55-60, 2006
[4] Kyo Kageura, Bin Umino. “Methods of Automatic Term Recognition”. Terminology, pp.259-289, 1996
[5] Toru Hisamitsu, Yoshiki Niwa, Junichi Tsuji. “A Method of Measuring Term Representativeness -Baseline Method Using Co-occurrence Distribution-”. COLING 2000, pp.320-326, 2000
[6] Kyo Kageura, Keita Tsuji, Akiko Aizawa. “Automatic Thesaurus Generation through Multiple Filtering”. COLING 2000, pp.397-403, 2000
[7] Beatrice Daille, Eric Gaussier, Jean Marc Lange. “Toward Automatic Extraction of Monolingual and Bilingual Terminology”. COLING 94, pp.515-521, 1994
[8] Hiroshi Nakagawa, Tatsunori Mori, “Nested Collocation and Compound Noun For Term Extraction”, Terminology. pp.64-70, 1998
[9] John Lafferty, Andrew Mccallum, Fernando Pereira. “Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data”. ICML 2001, pp.282-289, 2001
[10] 金子 周司. ライフサイエンス辞書とは. 情報管理, Vol49 No.1, pp.24-35, 2006