Vol.2015-DBS-162 No.26
2015/11/26

品詞データベースを用いた高精度なクエリ語の品詞判別に関する一考察

欅 惇志[1,a)]　宮崎 純[1,b)]

概要：本稿では，クエリ語の品詞付与におけるエラー分析を行い，また品詞データベースを利用した，クエリ語に対する正確な品詞判別手法を提案する．既存の品詞判別手法は文単位を分析対象としているのに対して，クエリの大部分は高々数個の語の羅列であり，正しい文法に則った自然文ではないため，既存手法ではクエリ語に対して正確に品詞を判別することができない．エラー分析の結果，1) 固有名詞が誤って一般名詞と判別される，2) 部分的な文法ルールが適用されて，誤った品詞が付与される，ということが判明した．これらの問題を軽減するため，Web 上の大規模コーパスから文単位の語 · 品詞ペアの組合せを事前に抽出し，クエリが発行されれば，クエリ語の組合せに対して付与される確率の高い品詞の組合せを特定し，付与する．評価実験の結果，提案手法を用いてクエリ語に品詞を付与した場合に，既存手法と比較して，上記の問題は軽減し，品詞判別精度は向上した．

キーワード：クエリ分析，クエリ語の品詞判別，品詞データベース

1. はじめに

ユーザの情報要求は多様化 · 複雑化の一途を辿っているため，クエリの検索意図を推定するためには，深いクエリ分析が重要である．自然言語処理分野において，有用な情報や知識を自動的に抽出することを目的として様々なテキスト分析が行われるが，それらの前段階の処理として，語の分割 (分かち書き) と品詞の判別を行う形態素解析 [1] や *1，語の係り受け関係を明らかにする句構造解析 [2] が行われるがことが一般的である．

その一方で，情報検索システムでは，検索対象テキストやクエリに対して，接辞処理 [3], [4], [5], [6] や原形化 (見出し語化) を行うことで，品詞や時制などの語の揺らぎを除外することが一般的である [7]．これは，検索対象文書とクエリを照合する上で，表層の表現の違いを吸収することで，より正確な検索が可能となるクエリが多数存在することに起因する．事前に定義されたルールに従って語尾を変化させる接辞処理は高速に処理を行うことが可能であるものの，必ずしも全ての語に対して有益な効果を発揮するわけではなく，却って悪影響を及ぼすクエリも存在する．ただし，形態素解析などによる原形化 (見出し語化) は，接辞処理と比較して処理時間が大幅に増大するため，特に大規模テキストを対象とする場合には必ずしも原形化が適用されているわけではない．なお，情報検索システムにおいては，形態素解析の後段の，より高度な自然言語処理技術が適用されることは更に少ない．

上記の通り，情報検索の文脈では，必ずしも自然言語処理技術によるテキスト処理技術が利用されるとは限らない．しかしながら，一般的にクエリは高々数個の語のみから構成され，これら数個のクエリ語の組合せのみを利用した場合には，深い分析を行う上では情報不足に陥る可能性が高い．また，情報検索システムにおいて自然言語処理技術の利用が一般的ではない理由の一因として，クエリに対する適切な自然言語処理技術の適用方法が確立されていないことが考えられる．つまり，検索対象テキストのみに対して自然言語処理技術を利用しても，それらの情報を十分に活かし切れないということである．

そこで本稿では，将来的に深いクエリ分析を行うことを目標に据えて，まずは，自然言語処理技術において最も基礎的なテキスト処理である品詞付与を，クエリ中のクエリ語に正確に行うことを目指す．なお，主要な形態素解析による品詞判別では，実用性を考慮し，「名詞」や「動詞」などといった品詞単位よりも更に細かな，「固有名詞」や「動詞の過去形」などといった粒度に分類されている (表 1 参

[1] 東京工業大学
Tokyo Institute of Technology
[a)] keyaki@lsc.cs.titech.ac.jp
[b)] miyazaki@cs.titech.ac.jp

*1 本研究で対象とするのは英文テキストであるため，分かち書きは不要である．

ⓒ 2015 Information Processing Society of Japan

照)．本稿でもこれに倣い，付与する品詞の種別は表 1 の分類に準拠する．

クエリ語に対して品詞を付与する上で，既存の形態素解析手法を利用した場合には，必ずしも正確に品詞を付与することができない．というのも，形態素解析は周辺の語の生起状況を考慮しつつ文単位に対して品詞の付与が行われるのに対して，クエリは高々数個程度の語から構成され，かつ，正しい文法に則らずに記述される場合が多数であり，十分に周辺の語の情報を利用できないためである．本稿では，まず，予備調査として，既存の形態素解析手法をクエリに適用し，どのような誤りが生じているのか，エラー分析を行った．その結果，下記の二種類の誤りが主要な誤りであると判明した．

(1) 固有名詞が誤って一般名詞と判別される．
(2) 部分的な文法規則が適用され，誤った品詞が付与される．

前述の議論を踏まえて，上記の誤りを軽減する手法を提案する．その際，実用レベルに達した判別精度を示す，文単位に対する形態素解析処理結果を利用して，クエリ中のクエリ語の組合せに対して付与される確率の高い品詞組を特定する．

評価実験では，提案手法を用いることで，既存の形態素解析技術をクエリに適用した場合に観測された，上記の二種類の誤りを軽減することが可能であるのかどうかを検証する．

以降，2 節にて，関連研究に言及しつつ，クエリ語に対して品詞を付与することの利点について議論する．その後，3 節にて，既存の形態素解析技術をクエリに適用した場合のエラー分析を行う．その結果を踏まえて，4 節にて品詞データベースを利用した，クエリに対する品詞判別手法を提案し，限定的な条件における実験結果を紹介する．最後に 5 節で本稿をまとめる．

2. クエリ語に対する品詞の付与

本節では，クエリ中のクエリ語に品詞が付与されることでもたらされる利点について述べる．

まず，クエリに品詞が付与されていれば，直接的に情報検索に役立てることが可能である．例えば，クエリ語が名詞の場合に，一般名詞と固有名詞では，検索における適切な戦略が異なると言われている [7]．具体例を挙げると，固有名詞に対しては，接辞処理や原形化を行えば検索性能が低下する．Web 検索における主要な検索カテゴリの中に，人名検索を含む固有名詞を対象とした検索が含まれている [8] ことからも，正しく品詞を付与して，固有名詞と一般名詞を区別することは重要である．また，近年，Named-entity recognition (NER) [9] や Entity Linking [10] などの研究が盛んに取り組まれているが，それら研究において前提となる Entity の発見には，まずは語が固有名詞であることを判別されていなければ実現できない．

また，正確に語義の曖昧性を解消することで情報検索システムの検索精度が向上した例が多数報告されているが [11], [12], [13], [14]，語義を割り当てる上で，正しい品詞が付与されていることは前提条件である [15]．

その他，Kato ら [16] により，クエリのタスクを推定する際に，クエリの品詞情報は重要であると報告されている．

以上より，クエリ中の各クエリ語に正確な品詞が付与されれば，直接的に情報検索の精度の向上が見込めるだけではなく，様々な応用への貢献が期待される．

3. エラー分析

以降，クエリに対して既存の形態素解析手法を適用した場合に，どのような品詞付与誤りが生じるのか，エラー分析を行う．

3.1 分析準備

以降の議論では，下記の形態素解析ツールと，クエリ，コーパスを利用した．

形態素解析ツール　Stanford Log-linear Part-Of-Speech Tagger[*2][17]

クエリ　TREC Web Track の Web track topics 200 個 (2009–2012 年の各年 50 個 [18], [19], [20], [21])

コーパス　ClueWeb09 Category B[*3] (英文 Web 文書 5,000 万件)

なお，クエリに関して，Web track topics の中には，検索システムに入力する際のキーワード集合，すなわちクエリ (図 1 の `query`) の他に，それぞれのクエリがどのような検索意図で情報検索を行ったのかの記述 (図 1 の `description`) が存在する．

3.2 分析結果

エラー分析は，下記の手順で行った．

(1) 各クエリに形態素解析を行い，品詞を付与する．その際，クエリに対しては接辞処理などの事前処理は行わない．
(2) 付与された品詞の正否を人手で判断する．その際，`description` を利用することで，クエリ語の持つ曖昧性を解消する．

正否判断において，複合語が固有名詞である場合に，構成される全ての語は固有名詞であるとする．例えば，クエリ *the united states* の *the* は DT (限定詞) であるのに対して，クエリ *the beatles* の *the* は NNP (固有名詞) である．

クエリ語に付与された品詞の個数と判別精度を表 2 に，品詞ごとの正解クエリ語のうち実際に付与できた再現率を表 3 に，品詞が誤って付与された際の，割り当てられた品

[*2] `http://nlp.stanford.edu/software/tagger.shtml`
[*3] `http://lemurproject.org/clueweb09/`

表 1 品詞一覧

関係	例
等位接続詞 (CC)	and, but, &
数値 (CD)	2015, seven, ’60s
限定詞 (DT)	the, all, this
存在の there (EX)	there
外国語 (FW)	ich, de, Monte
前置詞 (IN)	in, on, by
形容詞 (JJ)	big, critical, yellow
形容詞比較級 (JJR)	more, better, cheaper
形容詞最上級 (JJS)	most, best, cheapest
リスト (LS)	First, Second, Third
法助動詞 (MD)	can, must, will
一般名詞単数形 (NN)	car, family, history
一般名詞複数形 (NNS)	doctors, events, members
固有名詞単数形 (NNP)	Barack Obama, Sunday, Swedish Covenant Hospital
固有名詞複数形 (NNPS)	Americans, United Nations, Catholics
前限定詞 (PDT)	both, many, half
所有格 (POS)	’s, ’
代名詞 (PRP)	he, her, itself
所有格の代名詞 (PRP$)	my, mine, their
副詞 (RB)	probably, occasionally, completely
副詞比較級 (RBR)	further, later, earlier
副詞最上級 (RBS)	farthest, latest, earliest
不変化詞 (RP)	up, out, around
記号 (SYM)	%, <, *
不定詞の to (TO)	to
間投詞 (UH)	Wow, Oops, Yeah
動詞原形 (VB)	be, do, know
動詞過去形 (VBD)	was, did, knew
動詞現在分詞形 (VBG)	being, doing knowing
動詞過去分詞形 (VBN)	been, done, known
動詞現在時制 (主語が 1 人称/2 人称，単数) (VBP)	are, obtain, live
動詞現在時制 (主語が 3 人称，単数) (VBZ)	is, does, has
WH 型限定詞 (WDT)	what, which, whatever
WH 型代名詞 (WP)	who, which, whatever
WH 型代名詞所有格 (WP$)	whose
Wh 型副詞 (WRB)	however, whenever, whereever

詞と正解品詞の組合せとその個数を表 4 にそれぞれ掲載する．なお，出現頻度が 0 回の品詞については掲載を省いた．

表 2 の通り，付与された品詞には大きな偏りが存在し，NN (一般名詞単数形) が過半数を占める．その後，NNS (一般名詞複数形)，JJ(形容詞)，IN (前置詞)，DT (限定詞) と続き，その他の品詞はほとんど付与されていない．また，NN の判別精度は 54% 程度，NNS は 75%．JJ は 47%， IN は 75%，DT は 63% と続き，全てのクエリ語の平均は 60% 程度である．これらの値は，文を対象とした形態素解析の精度と比較して明らかに低い [*4]．

[*4] 本研究で扱う文書は Web 文書であるために直接的な比較はできないが，ニュース記事に対する形態素解析の品詞判別精度は 95% 程度である [17].

異なる観点から結果を分析するため，各品詞ごとに，正解クエリ語のうちどの程度のクエリ語を判別ができたのか，再現率を表 3 に示す．この結果から，今回利用したクエリ語集合中には NNP (固有名詞単数形) の比率が高く (全体の 40% 程度)，しかしながら，1% 程度しか発見できていないことが分かる．実際，表 2 の通り，NNP を割り振られたクエリ語は 2 語のみであり，ほとんどの場合において NNP が割り振られていないということが分かる．また，NNPS (固有名詞複数形) に至っては一度もクエリ語に付与されておらず，当然ながら再現率も 0 である．

NNP の再現率が極めて低い一方で，NN をはじめとする他の品詞についての再現率は極めて高い．このことから，クエリ語の品詞付与にて最も重大な課題は，固有名詞を適

ⓒ 2015 Information Processing Society of Japan

```
<topic number="7" type="faceted">
    <query>air travel information</query>
    <description>
        Find information on air travel, airports,
        and airline companies.
    </description>
    <subtopic number="1" type="inf">
        What restrictions are there for checked baggage
        during air travel?
    </subtopic>
    <subtopic number="2" type="inf">
        What are the rules for liquids in carry-on luggage?
    </subtopic>
    <subtopic number="3" type="inf">
        Find sites that collect statistics and reports
        about airports, such as flight delays,
        weather conditions, etc.
    </subtopic>
    <subtopic number="4" type="nav">
        Find the AAA's website with air travel tips.
    </subtopic>
    <subtopic number="5" type="nav">
        Find the website at the Transportation Security
        Administration (TSA) that offers air travel tips.
    </subtopic>
</topic>
```

図 1　クエリ例

表 2　形態素解析による品詞判別精度

品詞	クエリ語数 (割合)	正解語数	精度
NN	282 (0.570)	152	0.539
NNS	67 (0.135)	50	0.746
JJ	53 (0.107)	25	0.472
IN	24 (0.048)	18	0.750
DT	16 (0.032)	10	0.625
VB	8 (0.016)	8	1.0
CC	7 (0.014)	7	1.0
VBG	7 (0.014)	6	0.857
VBN	5 (0.010)	5	1.0
RB	4 (0.008)	2	0.5
TO	3 (0.006)	3	1.0
VBP	3 (0.006)	1	0.333
FW	3 (0.006)	0	0.0
NNP	2 (0.004)	2	1.0
PRP	2 (0.004)	1	0.500
VBD	2 (0.004)	1	0.500
CD	2 (0.004)	0	0.0
VBZ	1 (0.002)	1	1.0
WP	1 (0.002)	1	1.0
WRB	1 (0.002)	1	1.0
JJR	1 (0.002)	0	0.0
PRP$	1 (0.002)	0	0.0
全クエリ	495 (1.0)	294	0.594
二語以上のクエリ	442 (0.893)	268	0.606

表 3　品詞ごとの正解品詞数とその再現率

品詞	正解クエリ語数 (割合)	付与成功クエリ語数	再現率
NNP	189 (0.382)	2	0.011
NN	157 (0.317)	152	0.968
NNS	50 (0.101)	50	1.0
JJ	26 (0.053)	25	0.962
IN	18 (0.036)	18	1.0
DT	10 (0.020)	10	1.0
VB	10 (0.020)	8	0.8
CC	7 (0.014)	7	1.0
NNPS	6 (0.012)	0	0.0
VBG	6 (0.012)	6	1.0
VBN	5 (0.010)	5	1.0
TO	3 (0.006)	3	1.0
RB	2 (0.004)	2	1.0
PRP	1 (0.002)	1	1.0
VBD	1 (0.002)	1	1.0
VBP	1 (0.002)	1	1.0
VBZ	1 (0.002)	1	1.0
WP	1 (0.002)	1	1.0
WRB	1 (0.002)	1	1.0

表 4　品詞判別の誤りパターンとのその語数

付与された品詞	正解品詞	語数	全体に対する割合
NN	NNP	128	0.637
JJ	NNP	26	0.129
NNS	NNP	9	0.045
NNS	NNPS	6	0.030
IN	NNP	6	0.030
DT	NNP	6	0.030
FW	NNP	3	0.015
CD	NNP	2	0.010
JJ	NN	2	0.010
NNS	NN	2	0.010
RB	NNP	2	0.010
JJR	VB	1	0.005
NN	JJ	1	0.005
NN	VB	1	0.005
PRP	NNP	1	0.005
PRP$	NNP	1	0.005
VBD	NNP	1	0.005
VBG	NNP	1	0.005
VBP	NN	1	0.005
VBP	NNP	1	0.005

切に判別することである．

より詳細にエラー分析を行うため，品詞が誤って付与された際に，いずれの品詞がいずれの品詞として付与されたのを表 4 にまとめた．誤り全体の 64% が NNP を NN として付与されている誤りである．具体例を挙げると，*obama* などの人名，*india* などの地名，*ritz carlton* などの施設名などである．また，テレビ番組である *discovery channel* のよ

ⓒ 2015 Information Processing Society of Japan

うに，各語そのものは一般名詞である場合や，サンフランシスコの略語である *sf* のように，コンテキストや外部知識を有効に利用しなければ判別が困難であると思われるクエリ語も見受けられた．

その他の誤りのうちの多くが固有名詞と判断できなかった場合の誤りである．NNP を JJ (形容詞) と判別したケースとしては，*the united states* の *united* や，研究所である *pacific northwest laboratory* の *pacific* である．固有名詞の一部であることを判別できなかったという点では，上記の例と同様である．

また，固有名詞に由来しない誤りの例を挙げると，*lower heart rate* がある．*lower* は下げると意味する動詞であるものの，形容詞の比較級であると判別された．また，*gs pay rate* (*gs* は General Schedule の略語) の *pay* は，名詞を意図して問い合わせられたにも関わらず，動詞と判別された．これらは，名詞の前には形容詞が配置される確率が高い，主語の後ろには動詞が配置される確率が高い，などといった，部分的な文法規則によって品詞を判別したことによって起こる誤りであると考えられる．

4. 品詞データベースを利用したクエリの品詞判別

前節の調査の結果より，クエリに対して形態素解析によって品詞を判別することは必ずしも適切ではないという結果が得られた．そこで本節では，クエリに直接形態素解析を行うというアプローチを避けながらも，クエリ語に対して正確な品詞付与を目指す．その際，文単位に対する形態素処理は既に実用レベルに達しているという事実を考慮すると，文単位に対して付与された品詞情報を最大限に活用してクエリ語に品詞を付与することが妥当である．

4.1 提案手法の概要

Yarowsky [22] は，語義の曖昧性を解消する上で，テキスト中で近接する語は語義を特定する上で大きな手掛かりになるという仮説を提唱した．具体的には，同一文中において，特定の語義同士は共起しやすいという知見が得られている [15].

本研究における取り組みは語義の曖昧性解消ではなく品詞の付与であるものの，上記の仮説は品詞の判別においても当てはまると仮定する．すなわち，同一文中において，特定の語の組合せはそれぞれ同一の品詞が付与される確率が高いということである．仮に仮説が正しいとすれば，クエリに対しても同様に，語の組合せに対して付与される確率の高い品詞をクエリ語に付与することが可能となる．

これらを踏まえて，提案するクエリ語に対する品詞判別手法の手順を下記に示す．また，提案手法の概要は図 2 の通りである．

(1) 大規模 Web コーパスに対して形態素解析を行い，品詞を付与する．

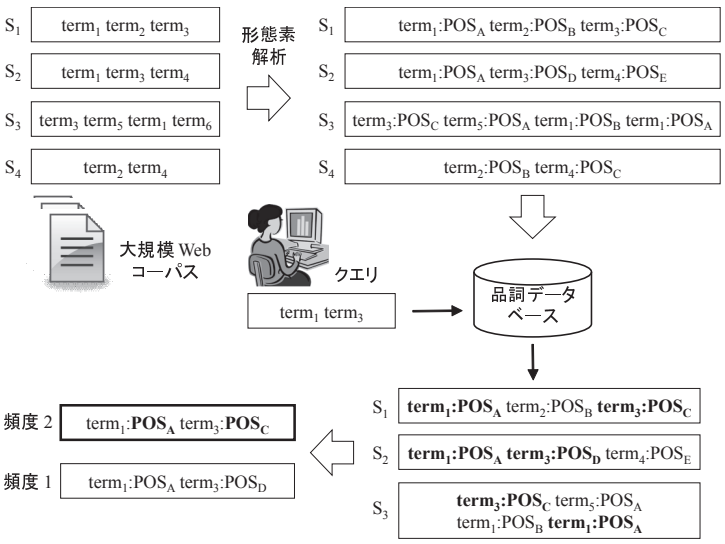

図 2 提案手法のフレームワーク

(2) 文ごとに語 · 品詞ペアの組合せを品詞データベースに保存する．

(3) クエリが問い合わせられると，品詞データベースを走査し，クエリ語の組合せに対応するデータを取り出す．

(4) 取り出されたデータの語 · 品詞を確認し，最も頻繁にクエリ語の組合せに対して付与されている品詞を特定し，それらの品詞をクエリ語に付与する．

図 2 を用いて，具体例を用いて説明する．大規模 Web コーパス中に S_1, S_2, S_3, S_4 の四つの文が存在するとする．これらに対して (1) 形態素解析を行い，品詞を付与する．続いて，(2) 語 · 品詞ペアの組合せを，品詞データベースに格納する．そして，(3) クエリ $term_1$ $term_2$ を受け取れば，品詞データベース中から $term_1$ $term_2$ を含むデータ，すなわち，S_1, S_2, S_3 を取り出す．(4) $term_1$ と $term_2$ に着目すると，付与されている品詞の組合せは，$term_1{:}POS_A$ $term_3{:}POS_C$ が頻度 2，$term_1{:}POS_A$ $term_3{:}POS_D$ が頻度 1 であるため，最終的にクエリ語に付与される品詞は，$term_1{:}POS_A$ $term_3{:}POS_C$ となる．

また，語義の曖昧性解消においては，共起語を抽出する上で，内容語，すなわち，名詞，動詞，形容詞，副詞のみが対象となる [15]．これは，機能語は他の語の語義を特定する上で有用な情報源にならないためであるが，品詞判別においても，互いの品詞を判別する上で有用とは考えれないクエリ語を除外する．具体的には，情報検索において，情報量のない語を除外する不要語処理 [23] を行う．

4.2 仮説の検証

前節における仮説が正しいのかどうか検証するため，限定的な条件における実験を行い，品詞判別精度を計測する．

表 5　提案手法による品詞判別精度 (全品詞)

手法	正解クエリ語数 (全クエリ語数)	判別精度
提案手法	104 (126)	0.825
形態素解析	89 (126)	0.706

表 6　提案手法による品詞判別精度 (NNP/NNPS)

手法	正解クエリ語数 (NNP/NNPS の語数)	判別精度
提案手法	26 (37)	0.702
形態素解析	2 (37)	0.0541

4.2.1　実験準備

形態素解析対象の大規模 Web コーパスとして, ClueWeb09 Category B を用いる．当コーパスは本研究で使用するクエリである Web track topics の 2009 年から 2012 年までの検索対象文書であるため，コーパスとクエリの語彙のミスマッチが少ないと考えられる．

また，今回の検証においては，2 語のクエリ語のみから構成されるクエリのみを対象とする．これは，仮に 3 語から構成されるクエリを想定した場合に，同一文中でクエリ語の 3 語が頻繁に共起する場合には品詞データベースから 3 語全てを含むデータを取り出せばよい．しかしながら，3 語全てが一文中で頻出することは稀であるものの，特定の 2 語のみが頻繁に共起する状況も起こりうる．そのような状況においていずれの方針で最適な品詞の組合せを判別するのかについては，現時点では未知数であるため，3 語以上のクエリ語を含む場合における品詞付与手法の考案は今後の課題とする．

不要語処理を行った上で 2 語から構成されるクエリは合計で 64 個のクエリであった．クエリ語の合計は 126 個であり，その中の NNP/NNPS の個数は 37 個である．次節の実験ではこれらのクエリに対して提案手法を適用し，品詞の判別精度を計測する．

4.2.2　実験結果

全品詞に対して行った評価実験の結果を表 5 に示す．提案手法を適用することで，形態素解析手法と比較し，17% 高精度に品詞を判別できた．

続いて，3 節の分析によって，その判別精度の低下が問題として明らかになった，固有名のみに対する評価実験を行った．表 6 に示す通り，提案手法を用いることで大幅に品詞判別精度が改善し，形態素解析手法として比較して，12 倍程度の精度を示した．

提案手法によって正しい品詞を発見できるようになった例を挙げる．*university of phoenix* の *university* は，形態素解析で NN と判別されていたが，提案手法を適用することで NNP と判別することに成功した．その他，*south africa* における *south* や *neil young* における *young* は，形態素解析では JJ (形容詞) と判定されていたところを NNP と判別できた．また，人名や略語などに対しても NNP と判別することに成功した．

以上の結果から，形態素解析によってクエリ語に品詞を付与した際に問題となった，固有名詞を正しく判別することや，部分的な文法規則によって誤った品詞に判別されるという問題を軽減することに成功した．

その一方で，形態素解析では正しい品詞を付与できていたが，提案手法によって誤った品詞を付与されたクエリ語も存在する．例えば，*jobs* である．正しい品詞は NNS (一般名詞複数形) であるものの，コーパス中に人名の *Steven Paul Jobs* が頻出した結果，*jobs* は NNP として判別された．このような問題を抑制するためには，語の重要度に対して何らかの正規化を行う必要があり，今後の課題である．

4.3　課題と今後の展望

前述の課題以外にも，本手法には処理速度における課題を抱える．クエリ語の個数が増えるほど，クエリ語の組合せの候補が増え，更に，その冪集合に一致するデータの抽出を行う可能性もある．このような処理を行うには膨大なクエリ処理コストを要するため，効率的な検索手法を提案する必要がある．そのためには，前述の処理に相応しい構造のデータベース (索引) の構築や，パターンマイニング技術 [24] の利用などを検討する．

また，本稿における実験では固有名詞の比率が高いクエリ集合を利用した．提案手法を利用することで，固有名詞に対しても，従来手法よりも正確に品詞判別を行えることを示したものの，このような問題は，固有名詞辞書や知識ベースを利用することでも解決可能であることを予想される．実際，文献 [9] では，実社会の Entity を格納した知識ベースを利用することで，クエリやマイクロブログ記事といった短文に対する Named-entity recognition (NER) を実現している．

それに対して，本研究でも言及した，部分的な文法ルールによって誤って付与される品詞の問題は，辞書や知識ベースで解決することが困難である．つまり，様々な品詞が混合したクエリでこそ提案手法の長所を活かすことができると考えられるため，今後は Web track topics 以外のクエリを用いた実験を行う予定である．

5.　まとめ

本稿では，既存の形態素解析技術をクエリに適用した場合に，どのような誤りが発生するのかエラー分析を行った．その結果，1) 固有名詞を判別することができない，2) 部分的な文法知識が反映されて，クエリ語に対して正しく品詞を付与することができない，といった問題点が明らかになった．

クエリ語に対して品詞を割り当てる上で，大規模 Web コーパスに対して付与された語と品詞の組合せ情報を利用する．文ごとの語 · 品詞ペアの組合せを予め品詞データベースに格納し，クエリ語の組合せに対して，付与される

ⓒ 2015 Information Processing Society of Japan

確率の高い品詞組を特定する．

2 語から構成されるクエリに対して行った評価実験の結果，提案手法を利用することで，従来手法よりも品詞の判別精度が向上し，固有名詞の低判別率の問題や，部分的な文法知識の影響による品詞の誤判別の問題を軽減することができた．

今後の課題として，3 語以上から構成されるクエリへの対応や，高速な処理技術の提案などが挙げられる．また，TREC 以外のテストレコクションにおける提案手法の有効性の検証も行う予定である．

謝辞 本研究の一部は，JSPS 科研費若手研究 (B) (課題番号:15K20990)，東京工業大学基金による研究助成の支援による．ここに記して謝意を表す．

参考文献

[1] 松本裕治， 北内啓，山下達雄，平野善隆， 松田寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』version 2.2.7 使用説明書，pp. 1–21 (2000).

[2] Miyao, Y., Saetre, R., Sagae, K., Matsuzaki, T. and Tsujii, J.: Task-Oriented Evaluation of Syntactic Parsers and Their Representations, *in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL '08)*, pp. 1–21 (2008).

[3] Lovins, J. B.: Development of a Stemming Algorithm, *Mechanical translation and computational linguistics*, Vol. 11, pp. 22–31 (1968).

[4] Paice, C. D.: Another Stemmer, *in Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 56–61 (1990).

[5] Krovetz, R.: Viewing morphology as an inference process, *in Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–202 (1993).

[6] M.F.Porter: An Algorithm for Suffix Stripping, *Readings in Information Retrieval*, pp. 313–316 (1997).

[7] Manning, C. D., Raghavan, P. and Schutze, H.: *Introduction to Information Retrieval*, pp. 157–159, Cambridge University Press (2008).

[8] Huffman, J. L. S. and Tokuda, A.: Viewing morphology as an inference process, *in Proceedings of the 32th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–50 (2009).

[9] Hua, W., Wang, Z., Wang, H., Zheng, K. and Zhou, X.: Short Text Understanding Through Lexical-Semantic Analysis, *Proceedings of the 31st International Conference on Data Engineering (ICDE '15)* (2015).

[10] Chen, Z. and Ji, H.: Collaborative Ranking: A Case Study on Entity Linking, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)* (2011).

[11] Mihalcea, R. and Moldovan, D.: Semantic Indexing using WordNet Senses, *in Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval (RANLPIR '00)*, pp. 35–45 (2000).

[12] Stokoe, C., Oakes, M. P. and Tait, J.: Word Sense Disambiguation in Information Retrieval Revisited, *in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 159–166 (2003).

[13] Liu, S., Yu, C. and Meng, W.: Word Sense Disambiguation in Queries, *in Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM '05)*, pp. 525–532 (2005).

[14] Zhong, Z. and Ng, H. T.: Word Sense Disambiguation Improves Information Retrieval, *in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL '12)*, pp. 273–282 (2012).

[15] 欅 惇志，荒瀬由紀，山本風人，辻井潤一：語義グラフを用いた整数線形計画法による
語義曖昧性解消手法の提案，Web とデータベースに関するフォーラム (WebDB Forum 2014) 論文集 (2014).

[16] Kato, M. P., Ekstrand-Abueg, M., Pavlu, V., Sakai, T., Yamamoto, T. and Iwata, M.: Overview of the NTCIR-10 1CLICK-2 Task, *in Proceedings of 10th NTCIR Conference*, pp. 182–211 (2013).

[17] Toutanova, K., Klein, D., Manning, C. and Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, *in Proceedings of the the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03)*, pp. 173–180 (2011).

[18] Clarke, C. L. A., Craswell, N. and Soboroff, I.: Overview of the TREC 2009 Web Track., *in Proceedings of the 18th Text Retrieval Conference (TREC '09)*, pp. 1–9 (2009).

[19] Clarke, C. L., Craswell, N., Soboroff, I. and Cormack, G. V.: Overview of the TREC 2010 Web Track., *in Proceedings of the 19th Text Retrieval Conference (TREC '10)*, pp. 1–9 (2010).

[20] Clarke, C. L. A., Craswell, N., Soboroff, I. and Voorhees, E. M.: Overview of the TREC 2011 Web Track., *in Proceedings of the 20th Text Retrieval Conference (TREC '11)*, pp. 1–9 (2011).

[21] Clarke, C. L. A., Craswell, N. and Voorhees:, E. M.: Overview of the TREC 2012 Web Track., *in Proceedings of the 21th Text Retrieval Conference (TREC '12)*, pp. 1–8 (2012).

[22] Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Method, *in Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL '95)*, pp. 1–21 (1995).

[23] Salton, G.: *The SMART Retrieval System–Experiments in Automatic Document Processing*, Prentice-Hall (1971).

[24] Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases, *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, pp. 487–499 (1994).

ⓒ 2015 Information Processing Society of Japan