大規模 Web コーパスを利用した Web 検索クエリの品詞付与

欅 惇志　　宮崎 純
東京工業大学 情報理工学研究科
keyaki@lsc.cs.titech.ac.jp, miyazaki@cs.titech.ac.jp

1 はじめに

言語処理において語の品詞付与は最も基盤的な技術である．より高度な自然言語処理技術の適用において必須であるだけではなく，Web 検索においても品詞情報は有用な情報である．情報検索では，検索対象文書に対して，不要語処理 [15] や接辞処理 [9, 13]，見出し語化 [5, 16] などの事前処理が行われることが一般的である．クエリ語の品詞ごとに適切な事前処理は異なるため，品詞ごとに検索索引を使い分けることでより高精度な検索が可能となる．また，クエリに付与された品詞は，クエリ拡張やクエリ整形における制約として利用することで高精度検索に貢献可能である [2]．その他，語義の曖昧性解消を用いた情報検索 [17] やセマンティック・タギング [12, 11]，クエリのタスク分類 [8] などの研究においても正確な品詞の付与は前提条件であるため，クエリに対する品詞付与は極めて有用である．

その一方で，検索対象文書に対する品詞の付与は実用レベルに達しているものの，クエリに対する品詞の付与は容易ではない．これは，Web 検索クエリの持つ，1) クエリに含まれる語は高々数単語である，2) 大文字情報が欠落している，3) クエリ語の語順が変則的である，といった特徴 [6] に起因し，既存の自然言語処理ツール，すなわち，形態素解析ツールによる品詞付与が困難であるためである．

本稿では，まず，予備調査として，既存の形態素解析ツールを用いてクエリに対して品詞を付与し，誤り分析を行う．次に，分析結果を踏まえて，大規模 Web コーパス中の文単位における語の共起に着目した品詞付与手法を提案する．その際，大規模 Web コーパスから事前に作成した単語–品詞データベースを利用する．

2 予備調査

2.1 実験準備

予備調査として，Web 検索クエリに対して既存の形態素解析ツールを用いて品詞を付与する．Web 検索クエリは，TREC Web Track[1] の Web track topics 200 個 (2009–2012 年の各年 50 個) を用いた．既存の形態素解析ツールとして，Stanford Log-linear Part-Of-Speech Tagger[2] [16] を用いた．なお，形態素解析ツールの出力は,「動詞」や「名詞」といった品詞単位よりも細かな粒度であるが (表 1 参照)，実用性を考慮し，本稿においてもこれら細かな粒度における品詞の分類を採用する．利用するモデルに関して，デフォルトの英語用モデルに加え，テキスト中の大文字情報を無視して処理するモデル (Caseless モデル) も用いる．

クエリ語の正解品詞は，各クエリの情報要求や検索背景を記述した `description` とその形態素解析結果を参照しつつ，適切な品詞を人手で決定した．

表 1: デフォルトモデルによる形態素解析結果

品詞	精度	再現率
NN (一般名詞 · 単)	.521 (147/282)	.974 (147/151)
NNS (一般名詞 · 複)	.731 (49/67)	1.0 (49/49)
JJ (形容詞)	.453 (24/53)	.960 (24/25)
VB (動詞 · 原形)	1.0 (8/8)	.8 (8/10)
VBG (動詞 · 現在分詞)	.857 (6/7)	1.0 (6/6)
VBN (動詞 · 過去分詞)	1.0 (5/5)	1.0 (5/5)
NNP (固有名詞 · 単)	1.0 (2/2)	.010 (2/197)
NNPS (固有名詞 · 複)	N/A	.0 (0/6)
全クエリ語	.580 (287/495)	.580 (287/495)

2.2 デフォルトモデルによる形態素解析

まず，デフォルトモデルを用いて形態素解析を行った結果について述べる．品詞ごとの精度と再現率を表 1 に掲載する．ただし，機能語及び付与頻度が少ない品詞の一部は省略した．全てのクエリ語のうち正しい品詞を付与できたのは 6 割弱であり，文単位を対象とした形態素解析の精度よりも大幅に低く，実用的なレベルに達しているとはいえない．

過半数のクエリ語 (58%) が NN (一般名詞単数形) として判別された．NNP (固有名詞単数形) の再現率が僅か 1% であったことからも，ほぼ全ての固有名詞が一般名詞と判別された．これは，Web 検索クエリの語の大部分は大文字情報が欠落していることが原因であると考えられる．

[1] `http://trec.nist.gov/data/webmain.html`
[2] `http://nlp.stanford.edu/software/tagger.shtml`

表 2: デフォルトモデルの品詞判別の誤りパターン

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
JJ	NN	2	0.010
NNS	NN	2	0.010
NN	JJ	1	0.005
NN	VB	1	0.005
VBD	NNP	1	0.005
VBG	NNP	1	0.005
VBP	NN	1	0.005
VBP	NNP	1	0.005

固有名詞の再現率が極めて低い一方で，一般名詞をはじめとするその他の品詞の再現率は概ね高い．このことから，クエリ語の品詞付与において最も重大な課題は，固有名詞を適切に判別することだといえる．

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 2 に纏めた．誤り全体のうち 64% は，NNP (固有名詞) が NN (名詞) として付与された誤りである．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も，NNP と判別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* などである．上記の例と同様，クエリ語が固有名詞の一部であることを判別できなかった．

また，固有名詞に関連する誤りの他に，部分的な文法規則が適用された結果発生した誤りも存在する．具体例を挙げると，*lower heart rate* の *lower* は，`description` によれば実際は動詞であるものの，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．

2.3 Caseless モデルによる形態素解析

続いて，Caseless モデルを用いた形態素解析結果について議論する．表 3 に示す通り，デフォルトモデルと比較し，全体的に精度，及び，再現率が向上した．デフォルトモデルにおいてほとんど発見できなかった固有名詞も数多く判別できた．しかしながら，表 4 の通り，全ての固有名詞を発見できたわけではなく，また，一般名詞を固有名詞と誤判別するケースが生じた．

表 3: Caseless モデルによる形態素解析結果

品詞	精度	再現率
NN (一般名詞 · 単)	.773 (116/150)	.768 (116/151)
NNS (一般名詞 · 複)	.975 (39/40)	.796 (39/49)
JJ (形容詞)	.677 (21/31)	.84 (21/25)
VB (動詞 · 原形)	.778 (7/9)	.7 (7/10)
VBG (動詞 · 現在分詞)	.833 (5/6)	.833 (5/6)
NNP (固有名詞 · 単)	.776 (142/183)	.721 (142/197)
NNPS (固有名詞 · 複)	.3 (3/10)	.5 (3/6)
全クエリ語	.77 (381/495)	.77 (381/495)

表 4: Caseless モデルの誤りパターン

付与された品詞	正解品詞	語数	全体に対する割合
NNP	NN	32	0.281
NN	NNP	31	0.271
JJ	NNP	7	0.061
NNPS	NNS	7	0.061

具体的な誤り例を挙げると，*discovery channel store* の *store* は一般名詞であるものの，固有名詞として判別された．クエリには (複合) 語と (複合) 語の間の明示的な分割点が存在しないため，本来は *discovery channe* と *store* から構成されるクエリが，全て一つの固有名詞であると判定されたことが原因であると考えられる．従って，大文字化欠落の問題を回避して固有名詞の発見ができたとしても，部分的な文法規則が適用された結果の誤りは依然として課題として残る．

3 提案手法

3.1 提案手法の手順

2 節の予備調査から，クエリに対して直接形態素解析を行った場合に正確な品詞の付与が困難であると判明した．従って，我々は，既に実用的なレベルの精度を達成している，文単位に対する形態素解析結果を利用して，クエリに品詞を付与する．その際，部分的な文法規則の影響を軽減するため，語の出現順序は反映させないようにする．これらを踏まえ，大規模 Web コーパス中からクエリ語の組合せを含む文を抽出し，それらの語に対して付与される確率の高い品詞をクエリ語に対する品詞として付与する．

提案手法は下記の二つの手順によって構成される．

単語–品詞データベース構築 大規模 Web コーパスの各文に対して形態素解析を行い，単語–品詞ペアの組合せを単語–品詞データベースに格納する．情報検索システムでは，テキストに対する事前処理として，形態素解析による見出し語化を行うことが一般的であり，単語–品詞データベースの構築は見出し語化と同時に行うことが可能である．

クエリ語の品詞付与 クエリが発行されれば，2 語以上のクエリ語を含む文を単語–品詞データベース

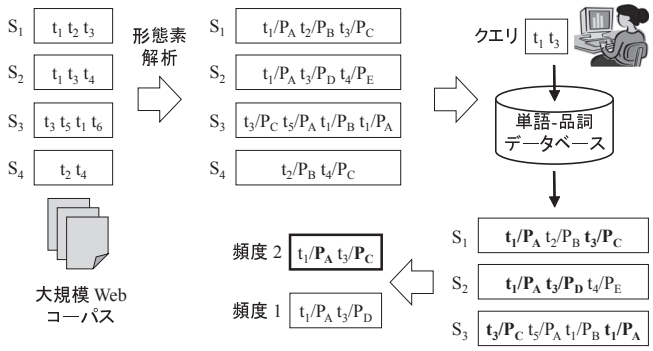

図 1: 提案手法によるクエリ語への品詞付与

から取り出し，それらを用いてクエリ語に適切な品詞を付与する．ただし，クエリが 1 語から構成されるクエリについては，コーパス中においてクエリ語に対して最も付与された品詞を付与する．

図 1 を用いて，具体例を説明する．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．これらに対して形態素解析を行い，品詞を付与する．続いて，文ごとの単語–品詞ペアの組合せをデータベースに格納する．ここで，クエリ t_1 t_2 が発行されれば，データベース中から t_1 t_2 を含むデータ，すなわち，S_1, S_2, S_3 を取り出す．t_1 と t_3 に着目すると，付与されている品詞の組合せは，t_1/P_A t_3/P_C が頻度 2，t_1/P_A t_3/P_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，t_1/P_A t_3/P_C となる．

なお，クエリ語が 3 語以上の語から構成されるクエリを想定したときに，同一文中で全クエリ語が頻繁に共起する場合には，品詞データベースから全クエリ語を含むデータを取り出し，付与頻度の高い品詞を選択すればよいと考えられる．しかしながら，全クエリ語が一文中で頻出することは稀であり，特定の 2 語のみが頻繁に共起する状況においては，それら 2 語の影響力を大きくすることが妥当であると考えられる．このような状況においても適切に品詞を判別することを目的として提案したクエリ語への品詞付与手法について，次節にて述べる．

3.2 クエリ語への品詞付与

単語–品詞データベースから抽出された単語–品詞ペアを，クエリ語ペアごとに分類する．3 語以上のクエリ語が頻出する場合に，その任意の組合せのクエリ語 2 語も頻出するという事実を踏まえ，任意のクエリ語 2 語における品詞の結果を総合的に考慮し，クエリ語に適切な品詞を付与する．

続いて，各クエリ語ペアにおいて，品詞の組合せごとに，各単語–品詞ペアの頻度をカウントする．図 2 はクエリ t_A t_B t_C における例である．クエリ語 t_A と

query {t_A t_B t_C}

t_A:t_B		頻度	標準化頻度
t_A/P_1	t_B/P_2	5	0.33
t_A/P_1	t_B/P_3	3	0.20
t_A/P_2	t_B/P_4	7	0.47

t_A:t_C			
t_A/P_1	t_C/P_2	3	0.43
t_A/P_3	t_C/P_3	4	0.57

t_B:t_C			
t_B/P_1	t_C/P_2	5	0.5
t_B/P_4	t_C/P_2	5	0.5

図 2: クエリ語ペアごとの単語–品詞ペアの頻度

表 5: 品詞の判別精度と精度改善率

	精度	精度改善率
MaxFreq	.819	1.50
MostLikelihood	.798	1.46
AllCombi	.826	1.51
SingleFreq	.702	1.29
Caseless	.770	1.41
Stanford	.547	1.00

t_B の組合せ $t_A : t_B$ に関して，$t_A/P_1, t_B/P_2$ である場合の頻度は 5，クエリ語 2 語の全出現回数で割った標準化頻度は 0.33 ($\frac{5}{5+3+7}$) である．これらを踏まえ，品詞を付与する手法を三種類提案する．

MaxFreq 全てのクエリ語ペア中において，最も高頻度に付与される品詞をクエリ語の品詞とする．t_A が最も多く出現するのは全組合せのうち，$t_A/P_2, t_B/P_4$ の 7 回であるため，t_A の品詞は P_2 となる．

MostLikelihood 全てのクエリ語ペアにおいて，最も標準化頻度の高い品詞をクエリ語の品詞とする．t_A の標準化頻度が最も高いのは $t_A/P_3, t_C/P_3$ の 0.57 であるため，品詞は P_3 となる．

AllCombi 全てのクエリ語ペアにおける単語–品詞ペアの出現頻度の総和が最も大きな品詞をクエリ語の品詞とする．t_A/P_1 の合計頻度は 11 (5+3+3) であるため，品詞は P_1 となる．

4 評価実験

提案手法の有用性を検証するため，評価実験を行う．大規模 Web コーパスは，Web track topics の検索対象文書である，ClueWeb09 Category B (英文 Web 文書 5,000 万件) を用いた．

比較手法として，Stanford, Caseless，SingleFreq を設定した．Stanford はデフォルトモデルを用い，Caseless は Caseless モデルを用いる手法である．SingleFreq は，文書中において最も出現頻度の高い品詞を付与する手法で，文献 [2] にて高精度品詞付与に寄与したと報告されている．

評価実験の結果，表 5 の通り，提案手法の中では，AllCombi が最も高精度であり，従来手法と比較してより正確に品詞を付与することに成功した．なお，符号検定の結果，AllCombi は有意水準 1% で Stanford, Caseless，SingleFreq より有意に精度が高かった．

提案手法によって正しい品詞を付与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 (一般名詞単数形) と判別されていたが，提案手法を適用することで NNP (固有名詞単数形) 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．

以上より，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できないことや，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，提案手法を適用することで，却って悪影響を及ぼした例が存在する．例を挙げると，*president united states* の *president* である．形態素解析の結果，*president* は正しく一般名詞であると判別されていたが，大規模 Web コーパスにおいて *president* が頻繁に固有名詞であると判別された結果，提案手法によって誤って固有名詞として判別された．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．

5 関連研究

クエリに対する品詞付与に関する研究には，1) 知識ベースを利用するアプローチ [7]，2) クエリの構造を理解するアプローチ [1, 2, 10, 4, 14]，3) クエリ語を含む文に対する形態素解析結果を利用するアプローチ [3, 6] が存在する．1) のアプローチでは，大規模な知識ベースが必要であり，2) は，人手で付与した正解ラベルが必要となる．その点，我々の手法が含まれる 3) では知識ベースや教師情報が不要である．

3) のアプローチに関する既存の研究では，擬似適合性フィードバックによって得られた検索結果上位の結果を利用したり [3]，クエリログからユーザが閲覧した文書のスニペットを利用しており [6]，少量のデータを利用している．また，品詞は名詞，動詞，その他の三種類に分類している．本研究では大規模データの統計量を扱う点や，実用性を考慮してより細かな粒度にて品詞の分類を行う点で異なる．本稿の実験結果から，大規模 Web コーパスによって得られた大域的な情報を利用することで，より高精度にクエリ語の品詞付与が実現できることを示唆している．また，クエリ語に対する品詞付与において，固有名詞を一般名詞と誤る例が極めて多いことからも，本稿で採用したより細かな粒度の品詞分類はより適切であると考えられる．

6 おわりに

本稿では，大規模 Web コーパスに対する形態素解析結果を利用して，クエリ語に対して，品詞の付与を行った．評価実験の結果，提案手法は既存の形態素解析ツールよりも正確に品詞を付与することに成功した．

今後の課題として，提案手法によって悪影響が及ばないための語の重みの正規化や，異なるクエリ集合による評価を行う．

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] James Allan and Hema Raghavan. Using Part-of-speech Patterns to Reduce Query Ambiguity. In *Proc. of SIGIR*, pp. 307–314, 2002.

[2] Cory Barr, Rosie Jones, and Moira Regelson. The Linguistic Structure of English Web-Search Queries. In *Proc. of EMNLP*, pp. 1021–1030, 2008.

[3] Michael Bendersky, W. Bruce Croft, and David A. Smith. Structural Annotation of Search Queries Using Pseudo–Relevance Feedback. In *Proc. of CIKM*, pp. 1537–1540, 2010.

[4] Michael Bendersky, W. Bruce Croft, and David A. Smith. Joint Annotation of Search Queries. In *Proc. of HLT*, pp. 102–111, 2011.

[5] Eric Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Journal Computational Linguistics*, Vol. 21, pp. 543–565, 1995.

[6] Kuzman Ganchev, Keith Hall, Ryan McDonald, and Slav Petrov. Using Search-Logs to Improve Query Tagging. In *Proc. of ACL*, pp. 238–242, 2012.

[7] Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. Short Text Understanding Through Lexical-Semantic Analysis. In *Proc. of ICDE*, pp. 495–506, 2015.

[8] In-Ho Kang and GilChang Kim. Query Type Classification for Web Document Retrieval. In *Proc. of SIGIR*, pp. 64–71, 2003.

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

[10] Xiao Li. Understanding the Semantic Structure of Noun Phrase Queries. In *Proc. of ACL*, pp. 1337–1345, 2010.

[11] Xiao Li, Ye-Yi Wang, and Alex Acero. Extracting Structured Information from User Queries with Semi-Supervised Conditional Random Fields. In *Proc. of SIGIR*, pp. 572–579, 2009.

[12] Mehdi Manshadi and Xiao Li. Semantic Tagging of Web Search Queries. In *Proc. of ACL*, pp. 861–869, 2009.

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

[14] Jeffrey Pound, Alexander K. Hudek, Ihab F. Ilyas, and Grant Weddell. Interpreting Keyword Queries over Web Knowledge Bases. In *Proc. of CIKM*, pp. 305–314, 2012.

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

[16] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *in Proceedings of NAACL*, pp. 173–180, 2003.

[17] Zhi Zhong and Hwee Tou Ng. Word Sense Disambiguation Improves Information Retrieval. In *in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL '12)*, pp. 273–282, 2012.