スニペットとウェブカウントを用いたウェブ検索クエリの分類

大久保 拓也　颯々野 学
ヤフー株式会社
{tookubo,msassano}@yahoo-corp.jp

1　はじめに

多くの文書が電子化された近年では, 必要な情報を得るために検索システムを利用することが一般的である. 情報は常に増え続けており, 検索システムにはより効率的に必要な情報へ行き着くための工夫が求められている. その工夫の一つとしてユーザが入力するクエリに焦点をおいた研究がある. ユーザが検索システムに入力するクエリは, 一般的に非常に短く, また曖昧性をふくんでおり [1][2], これらがシステムによる検索を困難にしている. こうしたクエリの短さや曖昧性に対して, クエリの表層以外の情報を追加することで, 検索精度の改善が期待できる. 例えば, クエリが製品名だとわかれば, ショッピングサイトやオークションサイトを優先的に表示したり, 地名だとわかれば地図や最寄り駅の情報などを優先的に表示することが出来る. 図 1 は検索クエリ”六本木駅”に対して, 駅周辺の地図を表示する例である.

図 1: 検索クエリに合わせて地図を表示する例

本研究では, クエリを 3 種類の固有表現 (人名, 地名, 組織名) とそれ以外 (その他) の 4 種類に分類することを目的とする. その目的を達成するために二つの手法を考え, 評価実験による効果の検証を行った.

2　クエリのカテゴリ分類手法

本研究では, 以下で説明する二つの手法を利用してクエリのカテゴリ分類を行う.

一つ目は, スニペットを用いる. ここでいうスニペットとは, ある検索クエリでウェブ検索エンジンによる検索を行った際に, 検索結果のページタイトルの下に表示される”検索クエリを含む短い説明文”のことを指す.

二つ目は, ウェブカウントを用いる. ウェブカウントとは, ある検索クエリでウェブ検索した際の hit 数を指すものであり, その検索クエリがウェブ上の文書でどの程度一般的なものかを示す尺度と捉えることもできる.

2.1　スニペットを用いる手法

スニペットを用いる手法では, クエリでウェブ検索を行い, その結果得られるスニペットに着目する. クエリに対する検索結果のスニペットからは, クエリが一般的にどのような文脈で使われるかの情報が得られると考えられる. ここでは, スニペットから情報を抽出する手段として, スニペットに対して固有表現抽出を行い, その結果を利用する. 具体的には, 以下の手順でクエリのカテゴリ分類を行う.

1. クエリでウェブ検索[1]を行う
2. 検索結果の上位 50 件からスニペットを抽出する
3. スニペットを文毎に分割し, クエリを含んだ文を収集する
4. 収集した文に対して, 固有表現抽出を行う
5. 抽出された固有表現のうち, クエリと表記が一致したもののカテゴリを集計する
6. 集計の結果, 最も多いカテゴリを分類結果とする (どのカテゴリも 0 の場合はその他とする)

2.2　ウェブカウントを用いる手法

ウェブカウントを用いる手法では, クエリの周辺テキストに現れる特定のパターンに着目する. ここでいうパターンとは, クエリの前後あるいは周辺に現れる特定の単語を指す. ウェブ上の文章中でクエリがどのようなパターンと同時に使われるかを調べることで, クエリを分類するための情報が得られると考えられる. 例えば, ウェブ文書においてクエリの前後に「株式会社」というパターンが頻出していれば, そのクエリが組織名であることを示す有力な材料となる.

ここでは, パターンをクエリとの文中における位置関係から周辺, 接頭表現, 接尾表現の 3 種類に分けて収集する. 周辺とは, クエリを含む文中に表れる名詞を指す. 接頭表現とは, クエリを含む文中に表れる名詞で, かつクエリの直前にあるものを指す. 接尾表現とは, クエリを含む文中に表れる名詞で, かつクエリの直後にあるものを指す. このようなパターンを集め, クエリとパターンとのウェブカウントから素性を作り, 機械学習[2]によりペアワイズ分類器を作成し, カテゴリ分類を行う.

2.2.1　パターンの収集方法

パターンは以下の方法で収集した.

1. 単一トークンのウェブ検索クエリを使いウェブ検索を行う
2. 検索結果の上位 50 件からスニペットを抽出する
3. スニペットを文毎に分割し, クエリを含んだ文を収集する
4. 収集された文を形態素解析[3]する

[1] http://search.yahoo.co.jp/
[2] 学習には,liblinear[3] を使用
[3] 形態素解析器 Juman[4] を使用

5. 解析結果から周辺, 接頭表現, 接尾表現をそれぞれ抽出する

表 1 に各パターンの例を示す.

表 1: パターンの例
周辺:ファン　接頭表現:法人　接尾表現:医院

2.2.2 ウェブカウントによる素性

分類対象のクエリに対して, 収集したパターンで次のようにウェブカウントの素性を作る.

1. クエリのウェブカウントを得る
2. パターンとクエリを組み合わせた文字列のウェブカウントを得る
   - 周辺: "クエリ"␣AND␣"パターン"
   - 接頭表現: "パターンクエリ"
   - 接尾表現: "クエリパターン"
3. 2 で得たウェブカウントを 1 のウェブカウントで割ったものを本研究でのウェブカウント素性とする

2.3 スニペットとウェブカウントを組み合わせる手法

スニペットの結果とウェブカウントの素性を同時に使う. スニペットの結果は各カテゴリについて集計した値となっているため, 取得したスニペットの数 50 で割ることにより正規化している. こうして得られる正規化されたスニペットの結果の素性と, ウェブカウントの素性で機械学習によるカテゴリ分類を行う.

3 関連研究

クエリ分類の研究では,KDDCUP2005[5] において, 英語の検索クエリを与えられた 67 のカテゴリに分類するタスクが設けられている. ここで最も精度のよかった Shen[6] はクエリに対して検索エンジンが出力したカテゴリリストとの一致, 及びページリストから素性を取り出して,SVM で分類するという方法をとった. 本研究で対象としているクエリは日本語であり, 対象とする言語が異なる.

日本語についての取り組みとしては, 安原らの分野連想語を利用した未知語に対する分野の自動推定 [7][8] があげられる. 彼らは, 未知語と分野連想語を検索エンジンで AND 検索した際の hit 数と OR 検索した際の hit 数をもとに関連度を計算し, 最も関連度の高かった分野連想語から分野の推定を行っている. これは, 本研究で取り上げるウェブカウントを用いた手法に似ているが, 彼らの手法では, 最も関連度の高い分野連想語のみで分野を決定するのに対し, ウェブカウントを用いる手法では, クエリと各パターンのウェブカウントを総合して決定するという点で異なる.

4 評価実験

評価実験には, あらかじめ人手でラベル付けしたクエリを用いる.

4.1 実験用クエリ

実験用のクエリは 2008 年 8 月のウェブ検索のクエリログ上位 10 万件のうち, スペースが入っていないものからランダムにサンプリングした 1,498 個を用いる. このクエリに対し, 人手によるラベル付けを行った. ラベルは固有表現 (人名, 地名, 組織名) とそれ以外 (その他) の 4 種類とし, 固有表現は IREX の定義に基づいてラベル付けを行っている. クエリによっては, 文脈に依存してラベルが複数当てはまるものも存在する. そのようなクエリは, ウェブ検索を行い, その検索結果に含まれるスニペットの文脈から判断し, ラベルを一意にしている. 表 2 に, クエリの内訳を示す. NOTNE であるクエリが半数以上を占め,LOC が少ないが, 実際のウェブ検索に投げられるクエリの分布に対する分類精度をみるためにクエリ数は調整しない.

表 2: 実験データの内訳

ラベル	クエリ数	クエリ例
ORG(組織名)	379	広島市立図書館
PER(人名)	140	末續慎吾
LOC(地名)	65	池袋
NOTNE(その他)	914	貿易実務検定

4.2 固有表現抽出

スニペットを用いる手法における固有表現抽出は,YamCha[4]に類似の固有表現抽出器に日本語文のデータを学習させたものを利用している. 学習データは,IREX[9] の固有表現定義に基づいてアノテーションされている.

ここで, アノテーションは人名地名組織名に関する定義のみを適用し, 数値表現と固有物名の定義は適用しない. これは, 数値表現については, 単一クエリにおける出現数が非常に少ないためである. また, 固有物名は実在する商品名から抽象的な法律名, 賞名など対象が多岐に渡っているため, 今回の実験では分類対象から除外した. 抽出器の精度は, ニュース記事を主とした評価データに対して,F 値で 87 程度である.

4.3 ウェブカウント用のパターン

2.2.1 で述べたウェブカウント用のパターンは, 以下の条件で収集した.

- 2010/01/01 から 2010/08/15 までの検索クエリからスペースが入っていないもの 265,244 クエリを使用
- 集計したパターンで頻度が上位のものを周辺 : 500 件, 接頭表現 : 500 件, 接尾表現 : 500 件をパターンとする

4.4 実験方法と評価尺度

スニペットを用いた手法については, 実験用のクエリをそれぞれ分類して, その分類精度をみる. また, ウェブカウントを用いた手法と二つを組み合わせた手法については,10 分割交差検定による分類性能をみる.

分類の性能は以下の尺度で評価する.

Accuracy

$$\frac{\text{全てのラベルの正解数}}{\text{全クエリ数}} \tag{1}$$

Recall

$$\frac{\text{対象ラベルに対する正解数}}{\text{対象ラベルのクエリ数}} \tag{2}$$

[4] http://chasen.org/~taku/software/yamcha/

Precision

$$\frac{\text{対象ラベルに対する正解数}}{\text{対象ラベルに分類したクエリ数}} \quad (3)$$

F 値

$$\frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

5 結果

5.1 固有表現全体の分類性能

固有表現全体についての分類性能は, 表 3 のようになった. ここでは,ORG,PER,LOC を対象ラベルとみて,Recall と Precision を計算している. スニペットを用いた分類では Recall が高く, ウェブカウントを用いた分類では Precision が高い. さらに, 二つを組み合わせたものが最も分類性能が良く, 固有表現であるクエリに対して F 値 70 弱で分類できている.

表 3: 固有表現全体の分類性能

分類手法	Acc	Rec	Pre	F 値
スニペット	69.29	**69.35**	53.29	60.27
ウェブカウント	76.03	61.64	66.30	63.89
組み合わせ	**79.64**	65.75	**73.00**	**69.19**

5.2 カテゴリごとの分類性能

カテゴリごとの精度については, 表 4 のようになった. どのカテゴリに対しても, 固有表現全体の精度と同様にスニペットでは Recall, ウェブカウントでは Precision が高い. さらに,ORG の Recall を除いた全ての評価において, 二つを組み合わせたものが良い. また, 各分類手法の結果から,PER は比較的分類しやすく, 逆に ORG は分類が難しいと推測される.

表 4: カテゴリごとの分類性能

分類手法	スニペット			ウェブカウント			組み合わせ		
カテゴリ	Rec	Pre	F 値	Rec	Pre	F 値	Rec	Pre	F 値
ORG	**68.87**	52.20	59.39	58.05	63.22	60.52	62.01	**68.71**	**65.19**
PER	73.57	55.98	63.58	69.29	75.78	72.39	**75.71**	**82.81**	**79.10**
LOC	63.08	53.95	58.16	**66.15**	64.18	65.15	**66.15**	**76.79**	**71.07**

5.3 カテゴリ間の分類結果

表 5,6,7 は, 各手法でのカテゴリ毎の分類数であり, 横軸がシステムによる分類結果で縦軸が正解のラベルである. スニペットのシステム出力結果では, 他の結果に比べて ORG の出力数が多い. NOTNE を誤って ORG や PER と出力することも多いが, 固有表現のクエリに限れば最も多くのクエリをカバーしている. 逆にウェブカウントや組み合わせのシステム出力結果では, カバーできていないクエリもあるが, 固有表現と出力したクエリについては, 精度良く分類できている.

表 5: スニペットのシステム出力結果

	ORG	PER	LOC	NOTNE
ORG	261	22	20	76
PER	9	103	4	24
LOC	14	5	41	5
NOTNE	216	54	11	633

表 6: ウェブカウントのシステム出力結果

	ORG	PER	LOC	NOTNE
ORG	220	12	11	136
PER	12	97	1	30
LOC	11	1	43	10
NOTNE	105	18	12	779

表 7: 組み合わせのシステム出力結果

	ORG	PER	LOC	NOTNE
ORG	235	7	8	129
PER	5	106	1	28
LOC	13	3	43	6
NOTNE	89	12	4	809

6 考察

以下では, 評価実験の結果から, それぞれの手法の性質について考察する.

6.1 スニペット

6.1.1 文の数と分類精度について

スニペットを用いた手法では, 取得した文中にクエリを固有表現と判断できる文脈が含まれていれば, その固有表現のカテゴリに分類できる. 表 8 は, 分類に使用する文を 10 文にした時と 100 文にした時の分類結果の例である. ”美輪明宏”や”志賀高原”は, 最初の 10 文では, 人名や地名として判断できる文脈のものがなかったため NOTNE と分類されているが,100 文にすると正しい分類ができている. 逆に”諏訪湖花火”の例のような固有表現でないクエリでは,100 文のうちに固有表現と判断される文が含まれていたために誤って分類されてしまうこともある.

表 8: 使用文数による分類結果の例

クエリ	正解ラベル	10 文	100 文
美輪明宏	PER	NOTNE	PER
志賀高原	LOC	NOTNE	LOC
諏訪湖花火	NOTNE	NOTNE	PER

図 2 は, 検索スニペットから取得した文の数を変化させた際の, 分類精度の変化である. 文数を増やすほど Recall が増加していることが確認できる. これらのことから, 処理するスニペットを増やすことで, 多くの固有表現であるクエリをカバーする事ができると考えられる. またその一方で, 固有表現でないクエリを誤って固有表現と分類される可能性も高まる.

6.1.2 固有表現抽出の精度依存について

また, この手法での分類精度は固有表現抽出の精度に大きく依存している. 文法が正確でなく, 表現が不規則な文章であったり, クエリ自身が正しく形態素解析されないような場合, 固有表現抽出に失敗するため, 正しく分類できない. 前者の問題は, スニペットの数を増やして, 固有表現抽出可能なスニペットを探すことで対処できる可能性がある. しかし, 後者のようなクエリ自体が正しく解析されない問題には対処できない.

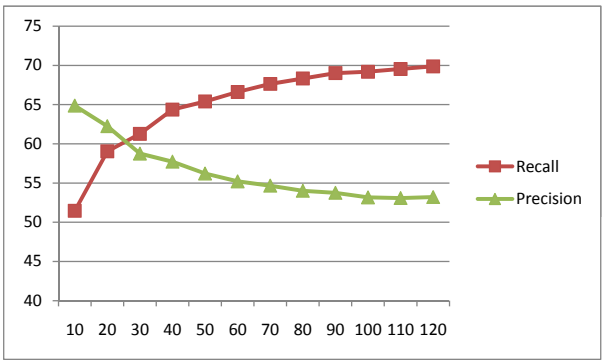

図 2: 分類精度と文数の関係

表 9 に示す”arsenal”や”ほしのまき”のように, アルファベットや平仮名のみのクエリは, 形態素解析の結果が未定義語であったり, 正しい形態素に分けられなかったりで固有表現抽出に失敗している例である. こうした問題には, 形態素解析の結果を修正するなどの別の解決策が必要だと思われる.

表 9: スニペットの分類結果の例

クエリ	正解ラベル	分類結果
arsenal	ORG	NOTNE
ほしのまき	PER	NOTNE
浜松町駅	LOC	LOC
内外タイムス	ORG	ORG

6.2 ウェブカウント

ウェブカウントを用いた手法では, 接頭表現や接尾表現が特定の単語を分類することを目的としているため, 全てのクエリをカバーすることは難しいが, パターンに合致するような特定のクエリについては精度良く分類することができる. 例えば, スニペットで正しく分類できなかった”arsenal”や”ほしのまき”は, ウェブカウントでは正しく分類できている. しかし, 文脈をみていないため,”浜松町駅”のように LOC か ORG を文脈によって判断しなければならないクエリは得意としない.

表 10: ウェブカウントの分類結果の例

クエリ	正解ラベル	分類結果
arsenal	ORG	ORG
ほしのまき	PER	PER
浜松町駅	LOC	ORG
内外タイムス	ORG	LOC

表 11 は, ウェブカウントを用いた学習で作成された各分類器について, 重みが大きい素性上位 5 件づつを並べたものである. 例えば,PER vs ORG をみると, クエリの周辺に”ファン”というパターンや, 接尾表現に”医院”というパターンが頻出している場合は, そのクエリが ORG よりも PER に近いと分類される.

各カテゴリ別にみると,ORG や NOTNE かどうかを分類するためには周辺のパターン,LOC は接頭表現や接尾表現,PER に対しては比べるカテゴリに応じて様々なパターンが必要であることが確認できる.

表 11: 各分類器における重みが上位の素性
(C=周辺,P=接頭表現,S=接尾表現)

ORG vs PER	PER vs ORG	LOC vs ORG	NOTNE vs ORG
C:１０	C:ファン	S:駅	C:ファン
P:社	S:医院	S:周辺	C:作成
C:各種	C:美	S:史上	C:解説
C:産業	S:製造	P:祭	C:毎日
C:協会	C:作品	S:プロジェクト	C:管理
ORG vs LOC	**PER vs LOC**	**LOC vs PER**	**NOTNE vs PER**
C:１０	C:ファン	S:プロジェクト	C:１０
C:館	S:製造	S:担当	C:各種
C:電気	C:赤	P:祭	C:ところ
C:協会	C:１０	P:東海道	P:番組
P:鉄道	S:製菓	S:メンバー	C:携帯
ORG vs NOTNE	**PER vs NOTNE**	**LOC vs NOTNE**	**NOTNE vs LOC**
C:産業	P:アーティスト	S:史上	C:１０
C:文化	P:愛犬	S:プロジェクト	C:赤
C:協会	P:リン	C:寺	P:東京
P:アーティスト	C:日比	P:祭	C:ところ
C:マーク	S:流	S:医院	P:飯田

7 まとめ

クエリの分類問題に対して, 二つの手法を試し, その効果を検証した. スニペットを用いた手法とウェブカウントを用いた手法について, それぞれの特徴を述べた. スニペットを用いた手法では, 使用する文の数を増やすことで Recall を高めることができた. ウェブカウントを用いた手法では, 高い Precision を出せた.

参考文献

[1] B.J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management*, Vol. 36, No. 2, pp. 207–227, 2000.

[2] S.M. Beitzel. *On understanding and classifying web queries*. PhD thesis, Citeseer, 2006.

[3] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.

[4] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version6.0. 京都大学大学院 情報学研究科 `http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html` より入手, 2009.

[5] Y. Li, Z. Zheng, and H.K. Dai. KDD CUP-2005 report: Facing a great challenge. *ACM SIGKDD Explorations Newsletter*, Vol. 7, No. 2, pp. 91–99, 2005.

[6] D. Shen, R. Pan, J.T. Sun, J.J. Pan, K. Wu, J. Yin, and Q. Yang. Q 2 C@ UST: our winning solution to query classification in KDDCUP 2005. *ACM SIGKDD Explorations Newsletter*, Vol. 7, No. 2, pp. 100–110, 2005.

[7] 橋本力, 黒橋禎夫. 基本語ドメイン辞書の構築と未知語ドメイン推定を用いたブログ自動分類法への応用. 自然言語処理, Vol. 15, No. 5, p. 10, 2008.

[8] 安原寛之, 森田和宏, 泓田正雄, 青江順一. 分野連想語を利用した未知語に対する分野の自動推定. *The 23rd Annual Conference of the Japanese Society for Artificial Intelligence*, 2009.

[9] IREX Homepage. `http://nlp.cs.nyu.edu/irex/index-j.html`.