クエリログとスニペットの単語連接頻度に基づくWeb検索クエリのセグメンテーション

三宅 純平　　塚本 浩司　　颯々野 学
ヤフー株式会社
{jmiyake, kotsukam, msassano}@yahoo-corp.jp

1　はじめに

Web検索サービスでは，入力されるクエリが適切にセグメンテーションされていないことが原因で，検索精度の劣化が起こることが観察されている．特にゲーム名などのカタカナ文字列は，連続した1語として扱われることが多く，検索精度の劣化の原因となる．そのため，クエリのセグメンテーション誤りの対策として，人手の精査により未知語や複合語の辞書更新で誤り訂正が行われることが多い．しかしながら，これらは非常にコストのかかる作業であり，クエリセグメンテーションの自動化への要求は高い．

そこで，本稿では検索精度の改善を目的として，クエリログとスニペットの単語連接頻度に基づくクエリセグメンテーションの手法を提案する．また，セグメンテーションしたクエリコーパスを学習データとしてSVMの点推定手法を用いたセグメンテーターの実用性に関しても評価した．

2　関連研究

Bergsmaら [1] は，様々な二値の素性や単語および単語連接の対数頻度，境界前後の単語を素性として，SVMによる意味の境界を推定する手法の提案をした．この手法は従来手法である相互情報量による境界推定より大幅な精度改善が報告されており，現段階において最も精度高い手法とされている．Tanら [2] は大規模なWebコーパスから構築した単語5-gram言語モデルを用いて意味の境界を推定する手法を提案した．また，Wikipediaのタイトルやアンカーテキストをコーパスに含めることで大幅な精度改善が報告されている．Wangら [3] は，Microsoft Web N-gramコーパスを用いた言語モデルによる単語分割において，タイトルやアンカーテキストだけを用いたモデルがWeb全体のコーパスを用いることより精度が高いことを報告している．

クエリセグメンテーションにおいて最も簡潔な手法としては，クエリを形態素解析で分割し，クエリカウントやウェブカウントから求めた何らかの尤度に基づいて分割位置を推定する手法が考えられる．しかしながら，日本語は英語とは違い，分かち書きがされていないため，新語や流行語（複合語）などの未知語が多く含まれるWeb検索クエリへの対応を考えると上記の方法では適切なセグメント位置の推定は難しい．

3　クエリログとスニペットの単語連接に基づくクエリセグメンテーション

ここではまず，クエリログにおけるアンド検索のスペース位置の分析結果について述べる，次に検索精度が改善するクエリのセグメンテーション手法を提案する．

3.1　ユーザが入力するクエリの傾向

クエリログの分析では，クエリログからデリミターを削除した時に同一となる異なりセグメント位置を持つクエリの抽出を行ない，ユーザが入力するセグメント位置の傾向について分析した．デリミターは空白と中黒「・」とした．クエリセットの例を表1に示す．クエリログの分析より得た知見を以下にまとめる．

最頻クエリの傾向

最頻クエリには，「無料動画」などのようなクエリ特有の複合語が多い．また，口語表現で使われるフレーズがそのまま入力されることがある．例えば，クエリ「めざまし占い」は，実際の正式名称は「めざましテレビ・今日の占いCountDown」であり，入力クエリと

表 1: クエリログから抽出した異なりセグメント位置を持つクエリセットの例

クエリ	頻度占有率
シェラトングランデ東京ベイ	0.915
シェラトン■グランデ■東京ベイ	0.03
シェラトングランデ■東京ベイ	0.02
シェラトン■グランデ■東京■ベイ	0.013
シェラトン・グランデ・東京ベイ	0.011
...	...

要求する文書に含まれる表現とが異なることがある．

カタカナ文字列における中黒「・」

カタカナ文字列は，複数の単語が繋がるものでも連続した 1 語として入力される傾向が高い．また，デリミターとして空白の代わりに中黒「・」が挿入されていることが多く，これは正しいセグメント位置であることが多い．

英数字文字列

英語文字列の最頻クエリは，正しくセグメントされていることが多い．また「iphone4，iphone 4」のような型番を含むクエリは，カタカナ文字列の場合と比べて，スペース位置が検索ランキングに大きく影響することが確認された．

これらの分析より，最頻クエリは必ずしも検索の精度改善に適切なセグメント位置ではないことが確認された．しかしながら，中頻度クエリは適切にセグメントされているクエリも多く，適切なセグメント位置の手掛かりになると考えられる．そこで，われわれは異なりセグメント位置を持つクエリセットを用いて，検索精度を改善するセグメント位置を推定する手法を提案する．

3.2 クエリログとスニペットを用いたクエリセグメンテーションの提案

3.2.1 クエリのセグメント位置の推定手法

図 1 に提案手法を示す．提案手法は 2 段階に分かれている．1 段目では，クエリセットから誤りセグメント位置を含まず，よりセグメント数の多いクエリの選択を行なう．2 段目では，選択したクエリの Web 検索結果のスニペットを取得し，各単語の頻度と単語連接頻度を算出し，シンプソン係数に基づいてクエリのセグメント位置の再判定を行なう．これにより，実際の Web ページに含まれるクエリの単語を考慮したクエリセグメンテーションが行われる．

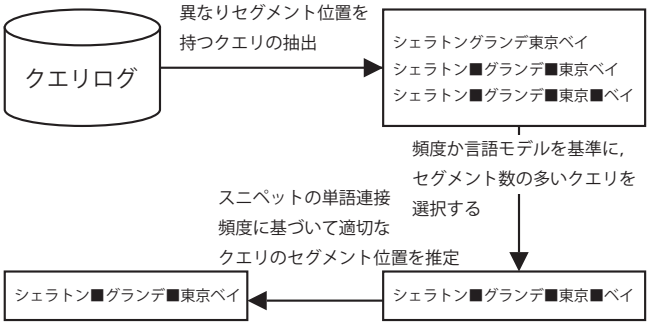

図 1: 検索精度が改善するセグメント位置の推定

われわれは提案手法の 1 段目であるセグメント数の多いクエリを選択する手法として，以下の 2 つの手法を用いた．

最多セグメント数による選択

最多セグメント数による選択は，ヒューリスティックな方法である．異なりセグメント位置を持つクエリセットから頻度占有率 0.1% 以上のものを対象に最もセグメント数の多いクエリを選択する．セグメント数の同じクエリが複数ある場合は頻度占有率の高いクエリを選択する．

言語モデル尤度による選択

言語モデル尤度による選択では，クエリセットから文字 3-gram 言語モデルの尤度を用いて最尤のクエリを選択する．これは誤りセグメント位置の棄却に対応する他，アンド検索がされ易い単語が含まれるクエリ対してはよりセグメント数の多いクエリを選択する効果がある．ただし，英数字文字列のみで構成されているクエリに関しては最頻のクエリを選択している．これは，3-gram モデルでは適切なクエリ選択がされなかったためであり，英数字文字列への対応にはより高次の n-gram が必要であると考えられる．各クエリ q のクエリセットを $\boldsymbol{Q}$，各クエリ q の文字 x_i の長さを N とおく．

$$\boldsymbol{q} = \{x_0, x_1, x_2, ..., x_N\}, \qquad q \in \boldsymbol{Q}$$

この時，クエリ文字列に対する 3-gram 言語モデルの対数尤度の相加平均よりクエリセット $\boldsymbol{Q}$ における最尤の q を選択する．

$$\max_{q \in \boldsymbol{Q}} \frac{\sum_{i=1}^{N} \log P(x_i | x_{i-2}, x_{i-1})}{N-1}$$

3.2.2 提案手法による検索精度改善の検証

提案手法 (言語モデル+スニペット，セグメント数+スニペット) が検索精度を改善するものであるかを検証

表 2: クエリログとスニペットを用いたクエリセグメンテーションの実験条件

正解データの期間	2010 年 10 月 1 日〜31 日
サンプル数	615 件
人手正解データの一致率	82.4 %
言語モデルの学習データ	2010 年 10 月 1 日〜31 日
検索結果取得数	20

表 3: 提案手法と比較手法のクエリセグメンテーションの実験結果

	Qry-Acc	Seg-Acc
最頻クエリ	0.645	0.937
形態素解析	0.617	0.923
言語モデル	0.731	0.951
セグメント数	0.732	0.953
形態素解析+スニペット	0.739	0.952
言語モデル+スニペット	0.773	0.962
セグメント数+スニペット	0.781	0.962

するために，人手で Web 検索に適切なクエリのセグメント位置を付与した正解データを作成し，提案手法によるセグメント位置との一致率を評価した．比較手法として，最頻クエリのセグメント位置と，最頻クエリを形態素解析[1] し，スニペットの単語連接頻度に基づいてセグメント位置を推定する手法 (形態素解析+スニペット) を扱う．評価基準は，Bergsma らが用いた Query Accuracy(Qry-Acc) と Segment Accuracy(Seg-Acc) を用いる．Qry-Acc はクエリの完全一致率であり，Seg-Acc は文字列境界の正解率である．シンプソン係数は精度が最高となる 0.9 を用いる．

表 2 に実験条件を示す．正解データは，1ヵ月分のクエリログの上位 10 万件のクエリから頻度 2 以上の異なりセグメント位置を持つクエリセットを抽出し，最頻クエリの頻度占有率の 100%から 5%で間隔でランダムサンプリングを行ない，合計が 600 件に近くなるように均等に選んだ．正解データのタグ付与は 2 名で行ない，実験結果の正解率では平均をとった．

表 3 の実験結果より，提案手法である「セグメント数+スニペット」が Qry-Acc と Seg-Acc において一番精度が高い.「セグメント数+スニペット」ではカタカナ文字列の分割精度の改善が確認されている．言語モデル尤度ではある程度多くセグメントされているクエリを選ぶ傾向があるが，クエリに特化したモデルであるため，クエリ特有の複合語はセグメントされないままのクエリは選択されてしまう．言語モデルでより適切なクエリ選択を行うためには，クエリコーパス以外に Wikipedia など他コーパス資源とのマージが必要であると考えられる．

[1] Yahoo! Japan デベロッパーネットワーク日本語形態素解析 Web API と同等のもの
http://developer.yahoo.co.jp/webapi/jlp/ma/v1/parse.html

4 SVM の点推定手法を用いたクエリセグメンテーターの実用性評価

前節の提案手法では，クエリログやウェブの頻度情報に基づいてクエリのセグメンテーションを行なったが，文字や文字種などの素性を用いることで精度の良いセグメンテーションができることも期待される．そこで，われわれはクエリコーパスを SVM の点推定による単語分割手法に適用したクエリセグメンテーターを実装し，実用性の評価を行なった．

4.1 SVM の点推定による単語分割手法

SVM の点推定による単語分割手法は Sassano[4] や Neubig ら [5] より提案されており，高精度に単語分割が行えることが報告されている．これらの手法は，各文字列間が単語境界であるかどうかの二値分類問題としてとらえたものであり，注目している文字列間の前後の文字 n-gram や文字種 n-gram，辞書単語の始端終端であるかなどを素性として組み込み，SVM による学習を行なっている．

4.1.1 素性

SVM で用いる素性は以下のものである．

文字 n-gram
セグメンテーションを識別する x_i, x_{i+1} における窓幅前後それぞれ w 文字の文字列 $x_{i-w+1}, .., x_i, x_{i+1}, ..., x_{i+w}$ の文字 n-gram

文字種 n-gram
上記の文字 n-gram における文字種（ひらがな，カタカナ，漢字，アルファベット，数字，その他）の n-gram

辞書単語素性
文字 n-gram において辞書に含まれる単語．ただし，

表 4: SVM の点推定手法の適用によるクエリセグメンテーターの実験条件

クエリログの期間	2010 年 10 月 1 日〜31 日
サンプル数	10 万件
言語モデルの学習データ	2010 年 10 月 1 日〜31 日
検索結果取得ページ数	20
SVM 学習器	liblinear

セグメンテーションを識別する x_i, x_{i+1} において，始端になっている単語には R，終端になっている単語には L，x_i, x_{i+1} を跨いでいる単語には I のフラグも共に付与している．
辞書単語には，ipadic-2.7.0-20070801 と日本語・英語 Wikipedia のアブストラクトから英数字単語のみをカウントして作成した辞書を用いた (日本語 Wikpedia: 頻度 2 以上，英語 Wikipedia:頻度 10 以上).

4.2 実験条件

評価実験ではSVM の点推定手法よるクエリセグメンテーションと正解データとの一致率を評価した．SVM の学習データは，1ヵ月分の Web 検索のクエリログにおける上位 10 万件（正解データは含まない）を提案手法 (言語モデル+スニペット，セグメント数+スニペット) でセグメントしたクエリコーパスを用いた．表 4 に実験条件を示す．評価基準は，Qry-Acc と Seg-Acc を用いる．SVM の学習器としては，liblinear[6] を用いた．また，点推定手法の窓幅は 5，n-gram のサイズは 3 を用いた．

4.3 実験結果

実験結果を表 5 に示す．前節同様に「セグメント数+スニペット」の手法が最も良い精度を示した．点推定手法では局所的にセグメントの識別を行うため，クエリに対し多くセグメントされる傾向が見られる．先行研究よりクエリカウントやウェブカウントがセグメント位置推定に有効であるという報告が多くされている他，離れた単語の組み合わせによるセグメント位置の変化やクエリ全体から適切なセグメント位置の推定などを考慮することで更なる精度改善も期待できる．

表 5: SVM の点推定手法の適用によるクエリセグメンテーターの精度

	Qry-Acc	Seg-Acc
言語モデル+スニペット	0.659	0.943
セグメント数+スニペット	0.667	0.945

5 おわりに

本報告では，入力クエリにおけるセグメント位置の違いによる検索精度劣化への対策のため，クエリログのアンド検索のスペース位置とスニペットの単語連接頻度を用いて，Web ページに出現するクエリの単語を考慮したセグメント位置を推定する手法を提案した．実験結果より，クエリ選択として異なりセグメント位置を持つクエリセット内の最多セグメント数を用いたものが最も良い精度であった．また，クエリコーパスから SVM の点推定手法を用いたクエリセグメンテーターの実用性の評価を行なった．今後はクエリセグメンテーターの精度改善を目指すとともに未知語分割器としての応用にも取り組む．

参考文献

[1] S. Bergsma and Q.I. Wang. Learning noun phrase query segmentation. In *Proc. of EMNLP-CoNLL*, 2007.

[2] B. Tan and F. Peng. Unsupervised query segmentation using generative language models and wikipedia. In *Proceeding of the 17th international conference on World Wide Web*, pp. 347–356. ACM, 2008.

[3] K. Wang, C. Thrasher, E. Viegas, X. Li, and B.P. Hsu. An overview of Microsoft web N-gram corpus and applications. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pp. 45–48. Association for Computational Linguistics, 2010.

[4] M. Sassano. An empirical study of active learning with support vector machines for Japanese word segmentation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 505–512. Association for Computational Linguistics, 2002.

[5] Graham Neubig, 中田陽介, 森信介. 点推定と能動学習を用いた自動単語分割器の分野適応. 言語処理学会第 16 回年次大会 (NLP2010), 東京, 3 2010.

[6] 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.