Vol.2010-NL-199 No.4
2010/11/18

検索クエリログとクリックスルーログを用いた略語の展開候補獲得

内海 慶[†1] 小町 守[†2] 町永 圭吾[†1]
前澤 敏之[†1] 佐藤 敏紀[†1] 小林 義徳[†1]

我々は，クエリ訂正を統一的に行う手法として，検索クエリログとクリックスルーログを用いたグラフに基づく手法を提案する．提案手法では，クリックスルーログを用いたラベル伝播により，入力されたクエリで検索を行った場合と同一のページに到達するクエリを獲得し，これをクエリの訂正候補とした．次に，獲得した訂正候補に対して，検索クエリログから生成した言語モデルを用いて尤度を計算し，ラベル伝播時のスコアとあわせて候補のランキングを行った．これによって，人手による学習コーパスを必要とせずに，入力されたクエリと高く関連し，かつクエリとして適切な候補をログから抽出できることを示す．

Abbreviation Expansion with Query and Click Through Logs

Kei Uchiumi,[†1] Mamoru Komachi,[†2]
Keigo Machinaga,[†1] Toshiyuki Maezawa,[†1]
Toshinori Satou[†1] and Yoshinori Kobayashi[†1]

In this paper, we propose a new method to refine web search queries. This method is based on a graph theoretic label propagation and uses web search query and clickthrough logs. Our method first enumerates query candidates with common landing pages with regard to the given query. Then it calculates likelihoods of the candidates, making use of language model generated from web search query logs. Finally the candidates are sorted by their scores calculated from the likelihoods and the label propagations. As a result, we are able to extract appropriate candidates from web search query and clickthrough logs, without using hand-crafted training data.

1. はじめに

近年の検索エンジンでは，ユーザの入力したクエリ（入力クエリ）に対して，誤りが含まれる場合にはスペル訂正をしたり，同義語や異表記がある場合にはそれも含めた検索を行っている．これらの機能はスペラーやクエリリライトなどと呼ばれており，普段ユーザは意識することなく利用している．最も初期のスペラーでは，専門知識を持つ専門家が，入力クエリに対する訂正候補を決めていた．しかし，入力クエリが多様化した現在では，人手で訂正候補を作成することは困難である．このため，検索ログを用いてクエリの訂正候補を自動獲得する手法の研究・開発が進められている．

Cucerzan ら[7] は，検索クエリログと Noisy Channel Model の枠組みを用いて検索クエリの訂正を行った．彼らは，言語モデルにクエリの単語 bigram を，また変換モデルとして重み付き編集距離を用いた．Gao ら[8] は Cucerzan らの手法をベースに，取り出した訂正候補にニューラルネットによるリランキングを導入することで拡張した．

しかしながら，これらの先行研究では単語の接続や分割，文脈を考慮した修正の変更なども行っているものの，基本的にはスペル訂正のみにフォーカスしているため，略語などの同義語は扱っていない．また，同義語を扱う場合には，入力クエリと検索クエリログに含まれるクエリとの類似度に，先行研究で用いられていた編集距離のような尺度は単純には利用できない．

そこで我々は，同義語の獲得に検索クリックスルーログを用いることにした．検索クリックスルーは，入力クエリに対して検索エンジンが返した結果から，ユーザがタイトル・アドレス・要約（スニペット）を見てそのアドレスをクリックしたという事を表す．そのため，検索クリックスルーにはユーザの意図が直接反映されていると考えられる．つまり，同じアドレスに到達する入力クエリは同じ意図で検索された可能性が高く，異なる 2 つのクエリは同義語である可能性が高いと考えられる．

検索クリックスルーログを利用した固有表現の獲得に，小町ら[10] の研究がある．小町らは，ラベル伝播手法による意味カテゴリ獲得（ここでは，あるカテゴリに属する固有表現の獲得）に検索クリックスルーログを利用し，検索クエリログを用いた場合に比べて高精度に

†1 ヤフー株式会社
Yahoo Japan Corporation
†2 奈良先端科学技術大学院大学
NARA INSTITUTE of SCIENCE and TECHNOLOGY

意味カテゴリ獲得が行える事を示した．ここで言う意味カテゴリには，同義語も含まれる．

本稿では小町らの手法と Noisy Channel Model を用いた，検索ログを用いたグラフに基づくクエリ訂正手法の提案を行う．

2. 関 連 研 究

検索クエリの訂正では，常に新しく現れる新語を考慮する必要がある．そのため従来の辞書を必要としたスペル訂正手法では対応が難しい．また，検索クエリの訂正は，スペルの訂正以外にも，単語の追加・削除，単語の分割・結合，文脈を考慮した単語の修正，接尾辞処理，略語・頭字語の展開といったタスクを含んでいる．従来の研究では，これらの各問題に対して，個別に対処がなされてきた[1)6)2)12)14)15)]．

Cucerzan ら[7)] は，検索クエリの訂正について，従来のスペル訂正手法では難しい問題を明らかにし，検索クエリログと Noisy Channel Model を用いることで各種問題に対処した．Gao ら[8)] は Cucerzan らの手法をベースに，取り出した訂正候補にニューラルネットによるリランキングを導入することで拡張した．また，彼らはクエリの訂正候補の獲得と限定的にではあるものの，検索クリックスルーログを利用している．しかし，これらの手法では略語や頭字語の展開といった問題には対処していない．

Guo ら[9)] は，それまで生成モデルが主流だった検索クエリの訂正に識別モデルを用いた．彼らは CRF[11)] の観測素性にオペレーションを加え，素性とラベルとオペレーションからなる 3 つ組みとすることで，検索クエリ訂正の各種問題を統一的な枠組みで行えるように拡張した．オペレーションは，各検索クエリの訂正タスクにおける処理で，スペル訂正であれば**削除**，**挿入**，**置換**，クエリ分割であれば**分割**，**結合**，頭字語の展開であれば**展開**などが挙げられる．しかし，この手法では学習コーパスに含まれる単語についてしか訂正は行えないため，新語への対応のためには常に学習コーパスを更新する必要がある．

検索ログを用いた検索クエリに関する研究には，他にもクエリ推薦がある[13)5)]．クエリ推薦では，推薦されるクエリがユーザの入力したクエリと意味的に異なる事もあるため，我々とは目的が異なる．

3. 提 案 手 法

本節では，我々の提案する検索クエリ訂正手法について述べる．本研究では検索クエリの訂正手法として，Noisy Channel Model を使用した．我々のタスクにおける Noisy Channel Model の考え方を以下で説明する．

検索クエリ q が与えられた時，条件付き確率 $P(c|q)$ が最大となる訂正候補 c^* を求めたい．ベイズの定理を用いれば，式（1）のように表せる．

$$\begin{aligned} c^* &= \operatorname*{argmax}_c P(c|q) \\ &= \operatorname*{argmax}_c \frac{P(c)P(q|c)}{P(q)} \\ &= \operatorname*{argmax}_c P(c)P(q|c) \end{aligned} \tag{1}$$

式（1）における $P(c)$ は言語モデルであり，本研究では検索クエリログから作成する．そのため，$P(c)$ は訂正候補 c のクエリらしさを表す．$P(q|c)$ は c から q への変換モデルと考えることができ，我々はここに小町ら[10)] の提案したラベル伝播手法を用いた．

小町らは，意味カテゴリ学習のタスクで初めて検索クリックスルーログを用いた．彼らは，ラベル伝播に正則化ラプラシアンを用いることで意味カテゴリ学習で大きな問題となる意味ドリフトに，特別なヒューリスティックを用いることなく対処している．

図 1 に，我々の提案手法のフレームワークを示す．

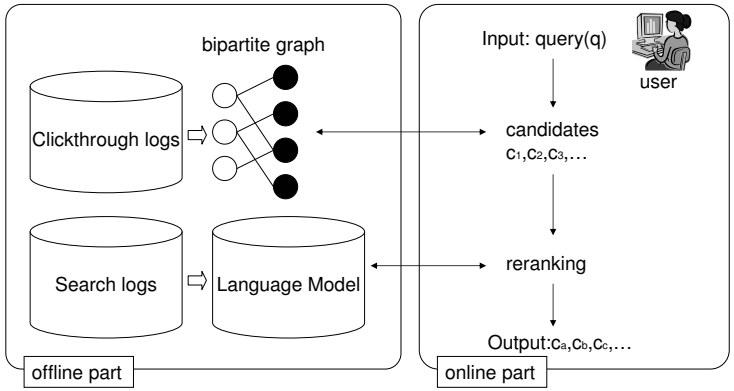

図 1 framework

入力されたクエリに対する訂正候補を得るために，検索クリックスルーログからクエリと，そのクエリで検索された際にクリックされたアドレスからなる 2 部グラフを作る．この

Vol.2010-NL-199 No.4
2010/11/18

2 部グラフ上でのクエリ間の関連度の計算を行い，与えられたクエリに対する訂正候補のリスト C を抽出する．抽出には，小町らの用いたラプラシアンラベル伝播を用いる．ラプラシアンラベル伝播は，正規化した頻度を 2 部グラフのエッジの値に用いて 2 部グラフ上のランダムウォークのように解釈すると，値は直接 $P(q|c)$ をモデル化していると考えることができる．そのため，ここで得られたスコアを，与えられたクエリ q と訂正候補 $c, c \in C$ の類似度 $sim(q, c)$ とすれば，式 (1) を最大化する c は $P(c) \times sim(q, c)$ となる．従って，与えられたクエリと訂正候補の類似度，及び訂正候補のクエリらしさの 2 つを用いて，以下の式を用いてリランキングを行い，最終的な訂正候補の出力を行う．

$$score(q, c) = P(c) \times sim(q, c) \tag{2}$$

これによって人手による学習コーパスを作成したり，特別なルールなどを用いることなく，検索ログのみを用いてクエリ訂正を行うことができる．また，本手法は汎用的な訂正のための枠組みであり，特定の誤りに対する訂正手法ではない．以降では，小町らの提案した *Quetchup* [⋆1] アルゴリズムと，本手法で使用した言語モデルについて説明する．

4. *Quetchup* アルゴリズム

小町ら[10] の提案手法である，*Quetchup* アルゴリズムについて説明する．*Quetchup* アルゴリズムは，ラベル伝播に基づく手法である．ラベル伝播をはじめとする，グラフに基づく手法は，少数のシードを用いても比較的高い精度が得られ，大規模化が容易であるという特徴がある．

図 2 に，シードクエリ「tx」を与えた時に構築されるインスタンス・パターン共起グラフとラベル伝播の様子を示す．

インスタンス・パターン共起グラフは左側のノードがインスタンス，右側のノードがそのクエリと共起するパターンとなっている 2 部グラフで，エッジの強さは 2 つの関係の強さを表す．ここでは，インスタンスとしてクエリ，パターンにはクリックスルーを用いている．インスタンスにおけるノードの濃さはシードインスタンスとの類似度を，パターンにおけるノードの濃さはそのパターンの特徴度を表している．「tx」と関連の強いアドレス「http://www.mir.co.jp/」は特徴的なパターンであるとして，「つくばエクスプレス」にラベルが伝播される．一方，「http://ja.wikipedia.org/」は「tx」と関わりの低い「常磐線」とも繋がっており，比較的中

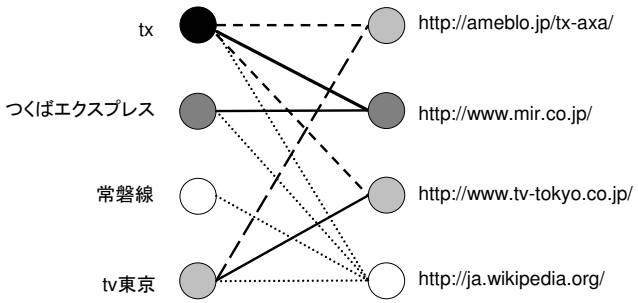

図 2 インスタンス・パターン共起グラフとラベル伝播の様子

立なパターンである．インスタンス「tv 東京」は，パターン「http://ameblo.jp/tx-axa/」，「http://www.tv-tokyo.co.jp/」及び「http://ja.wikipedia.org/」の 3 つのパターンを「tx」と共有しているため，シードクエリ「tx」のラベルが伝播して類似したインスタンスであると見なされる．

ラベル伝播はこのようにシードとして与えるノードのラベルを順次隣接ノードに伝播していく手法である．最適なラベルは，ラベル伝播のプロセスが終了した状態のラベルとして与えられる．

意味カテゴリ獲得を目的とした場合，図 2 のようにシードを与えては恐らく望ましい結果は得られない．「tx」は「鉄道」と「マスメディア」の 2 つのカテゴリにまたがる単語であり，ラベル伝播を繰り返すと，獲得されるインスタンスは 2 つのカテゴリに属するものが混在すると予想できる．このように，本来シードインスタンスを決める際には，複数のカテゴリに属さないようなインスタンスを選ぶ必要があり，カテゴリに対する知識を要する．しかし，検索クエリの訂正においては，シードインスタンスは常に入力された単一のクエリとなるため，特別な知識は必要としない．

また，我々の手法では，ラベル伝播における伝播の回数を 1 回とする[⋆2] ことで，入力クエリと同一ページに到達するクエリのみを検索クリックスルーログから抽出する．こうして抽出されたクエリは，入力クエリと同義語の関係である可能性が高いため，入力クエリから

⋆1 Query Term Chunk Processor

⋆2 ラベル伝播の k-step 近似において $k = 1$ とした場合に相当する．

意味的な飛躍なしでクエリの訂正が行えると考えられる．

Quetchup アルゴリズムを図 3 に示す．

入力：
　シードインスタンスベクトル $F(0)$
　インスタンス類似度行列 A
出力：
　インスタンススコアベクトル $F(t)$
1: 正規化ラプラシアン行列 $L = I - D^{-1/2}AD^{-1/2}$
　を作成する
2: 収束するまで $F(t+1) = \alpha(-L)F(t) + (1-\alpha)F(0)$
　を繰り返す

図 3 ラプラシアンラベル伝播アルゴリズム

今，インスタンス集合 $\mathcal{X} = \{x_1, ..., x_l, x_{l+1}, ..., x_n\}$ と，ラベル集合 $\mathcal{L} = \{1, ..., c\}$ が与えられ，かつ l 個のインスタンス $x_i (i \leq l)$ はラベル $y_i \in \mathcal{L}$ が付けられているとする．ラベル伝播アルゴリズムでは，これらからラベル付けされていないインスタンス $x_u (l+1 \leq u \leq n)$ のラベル付けを行う．

$\mathcal{F}$ は $n \times c$ の行列で，要素には正の値が入る． $F = [F_1, ..., F_n]^T \in \mathcal{F}$ は，ラベル付けされたインスタンス集合 $\mathcal{X}$ に該当する． 各インスタンス x_i のラベルは， $y_i = argmax_{j \leq c} F_{ij}$ で決定される．F の初期値は，$F_{ij} = 1(if\ y_i = j), F_{ij} = 0(if\ y_i \neq j)$ で与えられる．

出力として得られるインスタンスのスコアベクトルは，シードとして与えたインスタンスに対する類似度順に整列したベクトルになっている．我々はこのスコアを，式（2）の類似度 $sim(q, c)$ とした． q はシードインスタンス，c はクエリの訂正候補で，ラベルの伝播した各インスタンスである．

A はインスタンス類似度行列で，インスタンス・パターン行列 W を用いて $A = W^T W$ とする．W_{ij} はインスタンス x_i とパターン p_j の共起スコアである．小町らはこれに行ごとに正規化した共起頻度を用いたが，出現回数の多いパターンの影響から意味ドリフトが起きる事を確認している．これは我々の予備実験でも同様の事が確認できた．小町らはこれに対処する方法として，自己相互情報量（PMI）や対数尤度比といった相対頻度の使用を挙げている．

$D(N)$ は $D(N)_{ii} = \sum_j N_{ij}$ で定まる次数対角行列である．

ラベル伝播手法はシードのラベルとグラフ構造どちらを重視するかというパラメータ α（$0 \leq \alpha < \lambda^{-1}$ ただし λ は A の主固有値）を持ち，α が λ^{-1} に近づけばラベルなしデータから作成されるグラフ構造を考慮した結果となる．

4.1 Normalized PMI

前節で述べた意味ドリフトの問題に対応するために，我々は W_{ij} に，式（3）に示す PMI を用いることにした．

$$PMI(x, p) = ln \frac{P(x, p)}{P(x)P(p)} \tag{3}$$

PMI では，低頻度なペアに対して大きな値を与えることが知られている．例えば，ある x, p がデータ中にそれぞれ一度だけ，かつ共起して出現している場合，PMI は $-lnP(x, p)$ となる．また，一度も共起しないペアに対しては $-\infty$ となる．そのため，単純に PMI を用いると，共起頻度を用いた際には疎行列であった W が密になり，かつデータ中の低頻度なパターンに対する値が大きくなりすぎてしまう．そこで，我々は式（4）に示す， PMI を正規化した Normalized PMI（NPMI）[3] を用いた．

$$NPMI(x, p) = \left\{ ln \frac{P(x, p)}{P(x)P(p)} \right\} \Big/ -lnP(x, p) \tag{4}$$

$$W_{ij} = \begin{cases} NPMI(x_i, p_j) & (NPMI(x_i, p_j) > \theta) \\ 0 & (NPMI(x_i, p_j) \leq \theta) \end{cases}, (\theta \geq 0) \tag{5}$$

これによって，W_{ij} の取りうる値は，$[0, 1]$ に正規化される．

ラプラシアン行列 L は半正定値性を満たす必要があるため，我々は，式（5）に示すように，W から閾値 θ 以下の要素を削除した．これは，ラプラシアン行列 L の半正定値性を満たすと共に，インスタンスとパターンが互いに独立，あるいは負の相関を持つようなノード間のエッジをカットした事を意味する．

5. 言語モデル

本研究では，式（2）におけるクエリの訂正候補の尤度 $P(c)$ の計算のために，検索クエリ

ログから文字 N-gram の言語モデルを作成した．我々が使用した言語モデルを以下に示す．

$$P(c) = \prod_{i=0}^{N-1} P(x_i|x_{i-N+1},\ldots,x_{i-1}) = \prod_{i=0}^{N-1} freq(x_{i-N+1},\ldots,x_i)\big/freq(x_{i-N+1},\ldots,x_{i-1}) \quad (6)$$

ここでは，訂正候補 c を，$c = \{x_0, x_1, \ldots, x_{n-1}\}$ という文字の並びと考え，$P(c) = \prod_{i=0}^{N-1} P(x_i|x_{i-N+1},\ldots,x_{i-1})$ としている．

式（6）最終行の分母の頻度 $freq$ が 0 の場合には，$P(c) = P(x_i)$ として計算した．また，式（6）の分子の $freq$ が 0 の場合には，$freq = 1$ として扱うことで平滑化した．[⋆1]

6. 評価実験

6.1 実験設定

本来であれば，**訂正すべきクエリ集合**と**正しい訂正候補**を定義して精度評価を行うべきであるが，この 2 つを用意するのは容易ではない．**訂正すべきクエリ集合**を用意するためには，事前に検索クエリログ中の誤りを含むクエリを検出し，各クエリに対する全ての正しい訂正クエリを用意する必要がある．また，本手法の各種誤りに対する有効性について評価を行うためには，クエリの誤りのタイプを分類する必要がある．

そこで，本実験では，クエリの訂正として従来の手法では対処が難しく，本手法でカバーすべき問題としている略語の展開について評価を行った．

6.1.1 入力クエリ

本実験では，**訂正すべきクエリ集合**に略語を用いる．略語は，日本語 Wikipedia の略語ページ[⋆2]から，**欧文略語一覧**，**漢字略語一覧**，**カタカナ略語一覧**の 3 つを使用した．上記のページから略語のリストを取得し，重複や 1 文字のみのものを排除した 1,916 件の略語を実験の入力クエリとして用いた．

6.1.2 インスタンス・パターン行列の作成

ヤフー検索の検索クリックスルーログから，クエリをインスタンス，アドレスをパターンとして用いた．検索クリックスルーログは，2009 年 10 月 22 日から 2009 年 11 月 9 日，及び 2010 年 1 月 1 日から 2010 年 1 月 16 日までの期間について集計を行い，これを用いた．また，計算時間の短縮と行列のサイズの圧縮のため，頻度 10 で足切りを行った．インスタンスパターンのエッジ数は，16,988,516 件となった．

インスタンスパターン行列の要素 W_{ij} の閾値 θ は，予備実験の結果を踏まえ 0.1 とし，NPMI の値がこれ以下のエッジはカットした．

Quetchup に与えるパラメータ α は 0.0001 とした．また，ラベル伝播の反復回数は 1 回のみとした．

6.1.3 言語モデルの作成

言語モデルの作成には，ヤフー検索の 2009 年 8 月 1 日から 2010 年 1 月 27 日までの検索クエリログを集計し，これを用いた．こちらについても，計算時間の短縮のために，頻度 10 で足切りを行った．言語モデルの作成に用いた異なりクエリ数は，52,399,621 件であった．

実験では，言語モデルを作成する際に $N = 5$ とした．また，予備実験を行ったところ，正解の部分文字列となっている訂正候補の尤度が，正解の訂正候補の尤度よりも高くなったため，実験では，尤度を訂正候補の表層の文字列長で正規化した．

6.2 評価

出力に対する正解，不正解の評価は，検索エンジンの評価などを専門とする有識者 5 人が行った．評価には式（8）に表す，順位 k での精度（precision at k）と，順位 k での再現率を用いた．ただし，ここでの再現率は，入力クエリ全体中少なくとも 1 件以上正解を提示できた割合を表す．

$$precision = \frac{\text{順位 } k \text{ での正解件数}}{\text{順位 } k \text{ での出力件数}} \quad (7)$$

$$recall = \frac{\text{順位 } k \text{ における正解を 1 件以上提示できたクエリ数}}{\text{入力クエリ全体}} \quad (8)$$

評価は，検索クリックスルーログから訂正候補を 50 件抽出した後，これに対して，(1) 言語モデルの尤度のみでランキング，(2) ラベル伝播のスコアのみでランキング，(3) 言語モデルの尤度+ラベル伝播のスコアでランキング (提案手法) の 3 つを行い，これらを比較する．

6.2.1 評価ガイドライン

我々の定義した正解の基準を以下で説明する．

我々は略語に対する正しい訂正のパターンとして，(1) **欧文略語**から**欧文正式表記**，(2) **欧文略語**から**日本語表記**，(3) **日本語略語**から**日本語正式表記**，(4) **日本語略語**から**英語表記**を定義した．訂正候補として良いものの基準は，(1) **固有名詞**，(2) **展開したフレーズ**，(3)

⋆1 大規模データではスムージングが結果に大きな影響を与えないことが知られている[4] ため，本研究では実装の手軽な Add-one Smoothing を用いた
⋆2 http://ja.wikipedia.org/wiki/略語

表 1 略語と訂正候補の例

略語	訂正候補（順位の降順）	タイプ	訂正のパターン
adf	asian dub foundation	固有名詞-組織の正式名称-	欧文略語から欧文正式表記
ana	全日空，全日本空輸株式会社	固有名詞-組織の通称/正式名称-	欧文略語から日本語表記
ny	ニューヨーク	固有名詞-地名-	欧文略語から日本語表記
tos	テイルズオブシンフォニア	固有名詞-製品名-	欧文略語から日本語表記
イラレ	illustrator	固有名詞-製品名-	日本語略語から英語表記
ハンスト	ハンガーストライキ	展開したフレーズ	日本語略語から日本語正式表記
阪神	阪神タイガース	固有名詞-組織の正式名称-	日本語略語から日本語正式表記
fyi	for your information	展開したフレーズ	欧文略語から欧文正式表記
fyi	参考までに	略語に対応する日本語の意味	略語に対応する日本語の意味
wktk	ワクテカ，ワクワクテカテカ	略語に対応する日本語の意味	略語に対応する日本語の意味

表 2 順位 k に対する精度と再現率

k	言語モデルのみ		ラベル伝播のスコアのみ		言語モデル+ラベル伝播のスコア	
	precision	recall	precision	recall	presicion	recall
1	0.157	0.157	0.114	0.114	**0.161**	**0.161**
3	0.142	0.278	0.122	0.256	**0.157**	**0.321**
5	0.128	0.346	0.121	0.341	**0.142**	**0.392**
10	0.102	0.425	0.114	0.453	**0.115**	**0.465**
30	0.0776	0.529	**0.0871**	0.536	0.0817	**0.542**
50	0.0731	**0.557**	**0.0733**	**0.557**	0.0732	**0.557**

表 3 入力クエリと出力訂正候補の例

入力クエリ	訂正候補
写植	写真植字，写植屋，写植機，写植 方，漫画
満鉄	満鉄調査部，南満州鉄道株式会，南満州鉄道，満鉄会，満州鉄道
はねトび	はねるのとびら，はねるのトびら，はねるの，はねるのトびら，はねるのトビラ，はねとび
vod	ビデオオンデ，ビデオ・オン・デマンド，ビデオ オンデマンド，ビデオオンデマンド
ilo	日本 ilo，ilo 協会，国際労働機関，国際労働期間，ilo 条約
pr	パブリック・リレーションズ，パブリックリレーションズ，prohoo!マ，pr 会社，プラ

略語に対応する日本語の意味とした．

表 1 に，具体例を示す．

ただし，上記のパターンに属さないものについては，別途有識者と協議した上で，例外として扱った．また，上記のパターンに属するものについても，一部例外として扱った．例えば，欧文略語から日本語表記への訂正については，欧文略語からひらがな表記での訂正は誤りとして扱った．

6.3 実験結果

表 2 に，k を変化させた際の精度と再現率を，表 3 に，入力クエリとそれに対する出力の例を表す．

実験結果では，$k = 50$ とした時に，3 つの手法に特に差は見られなかった．これは，そもそもランキングを行う対象をラベル伝播で抽出したものに絞り込んだ上で，3 つの手法を比較しているためである．

$1 \leq k \leq 10$ の範囲では，言語モデル+ラベル伝播のスコアでランキングを行った方法（提案手法）が，精度，再現率共に高くなった．この理由として，言語モデルによるランキングとラベル伝播のスコアによるランキングの特徴の違いが挙げられる．言語モデルから求めた尤度は訂正候補のクエリらしさを表しており，検索クエリログ中に良く現れる表記の訂正候補ほど尤度が高くなる傾向にある．一方，こうした訂正候補は検索結果のヒット数も多くなり，クリックスルーログのパターンと多く共起するようになると予想できる．NPMI をインスタンス・パターンの要素として用いた我々のラベル伝播手法では，インスタンスが特定のパターンとだけ共起するような場合に，2 つを繋ぐエッジに強い重みが与えられる．すると，少数の特定のパターンでシードインスタンスと繋がるインスタンスほど，上位に抽出されやすくなる．そのため，検索クエリログ中では低頻度なインスタンスが上位に来やすい．我々の提案する手法では，これらを組み合わせることによって，クエリらしさと入力クエリとの類似度の両方を考慮した訂正候補の抽出を行えたと考えられる．

誤った出力を見たところ，(1) 正解表記の部分文字列である，(2) 正解表記と属性語の組である，(3) クリックスルーログ中に入力クエリが存在しない，(4) 入力クエリに対する関

連語である，などがあった．

(1) に対しては，言語モデルで求めた尤度が正解表記よりもその部分文字列の方が高くなることが原因と考えられる．我々はこの問題に対処するために，文字列長による尤度の正規化を行っているが，これのみでは十分に対処できなかったと考えている．

(2) の例としては，検索クエリ中に良く使われる属性語「とは」,「意味」,「使い方」などと正解表記が共起する場合が挙げられる．誤った出力のうち，上記の 3 つの属性語と共起しているものは 857 件であった．こうした訂正候補は，入力クエリとの類似度が高くなりがちであり，かつ正解表記と共起する単語も良く現れる属性語であるため，言語モデルから得られる尤度も高くなる．この問題については，検索クエリログに含まれる空白区切りのクエリについては，空白で区切った上で，先頭の単語のみを用いて言語モデルを作成することで，ある程度対処可能と考えられる．ただし，必ずしも属性語は空白区切りで入れられるとは限らないため，言語モデル以外に属性語を含むような訂正候補の尤度を下げるような尺度の導入が必要と考えている．

(3) 入力クエリ 1,916 件のうち，クリックスルーログに存在しないために正解できなかったものは 280 件であった．これについては，使用する検索ログの量を増やすことで対処可能と考えている．

(4) は，入力クエリとなる略語が他の一般名詞と同一の表層を持っている場合が多かった．例えば,「dog」などが挙げられる．こうした略語について正しく展開を行うのは現状では難しいと言える．

7. ま と め

本研究では，Noisy Channel Model とラベル伝播を用いた検索ログからのクエリ訂正手法を提案した．

本研究の特徴は，Noisy Channel Model の変換モデルにラベル伝播のスコアを用いていることである．実験では，検索クリックスルーログから抽出した同義語をクエリ言語モデルだけでランキングするよりも，ラベル伝播と組み合わせた方が精度・再現率共に高くなることを確認した．

今後は，実験に使用する検索ログの量をより大規模なものとし，かつ今回の実験で明らかとなった言語モデルだけでは対処できない問題について取り組んでいきたい．

参 考 文 献

1) F.Ahmad and G.Kondrak. Learning a spelling error model from search query logs. In *EMNLP*, pp. 955–962, 2005.
2) S.Bergsma and Q.Wang. Learning noun phrase query segmentation. In *EMNLP-CoNLL*, 2007.
3) G.Bouma. Normalized (pointwise) mutual information in collocation extraction. *GSCL*, pp. 31–40, 2009.
4) T.Brants, A.Popat, P.Xu, F.Och, and J.Dean. Large language models in machine translation. In *In EMNLP-CoNLL*, pp. 858–867. Citeseer, 2007.
5) H.Cao, D.Jiang, J.Pei, Q.He, Z.Liao, E.Chen, and H.Li. Context-aware query suggestion by mining click-through and session data. In *SIGKDD*, pp. 875–883, 2008.
6) Q.Chen, M.Li, and M.Zhou. Improving query spelling correction using web search results. In *EMNLP-CoNLL*, pp. 181–189, 2007.
7) S.Cucerzan and E.Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *EMNLP*, pp. 293–300, 2004.
8) J.Gao, X.Li, D.Micol, C.Quirk, and X.Sun. A large scale ranker-based system for search query spelling correction. *COLING*, 2010.
9) J.Guo, G.Xu, H.Li, and X.Cheng. A unified and discriminative model for query refinement. In *SIGIR*, pp. 379–386, 2008.
10) M.Komachi, S.Makimoto, K.Uchiumi, and M.Sassano. Learning semantic categories from clickthrough logs. In *ACL-IJCNLP*, pp. 189–192, 2009.
11) J.Lafferty, A.McCallum, and F.Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pp. 282–289, 2001.
12) M.Li, Y.Zhang, M.Zhu, and M.Zhou. Exploring distributional similarity based models for query spelling correction. In *ACL*, pp. 1025–1032, 2006.
13) Q.Mei, D.Zhou, and K.Church. Query suggestion using hitting time. In *CIKM*, pp. 469–478, 2008.
14) F.Peng, N.Ahmed, X.Li, and Y.Lu. Context sensitive stemming for web search. In *SIGIR*, pp. 639–646, 2007.
15) K.Risvik, T.Mikolajewski, and P.Boros. Query segmentation for web search. In *WWW, Poster Session*, 2003.