6ZC-2

コミュニティベース Q&A からの類似質問検索手法

Query Retrieval from Community-based Q&A Databases Based on Language Model

高橋 輝† 高須 淳宏‡ 安達 淳‡

Akira Takahashi Atsuhiro Takasu Jun Adachi

東京大学 大学院 / The University of Tokyo†

国立情報学研究所 / National Institute of Informatics‡

1. はじめに

近年，新たな知識源としてコミュニティベースの Q&A サイト(cQA)に注目が集まっている．cQA はコミュニティのユーザが質問を投稿し，他のユーザが回答する Web サービスである．一日に大量の質問が投稿・回答されるので，そのアーカイブデータは非常に豊かな，かつ自然言語によりアクセス可能な知識源と言える．我々は Q&A サイトのアーカイブデータからの類似質問検索の実現を目指している．類似質問検索[1][2]とは，クエリとして与えられた質問文に類似した内容の質問をアーカイブから検索するタスクである．類似質問検索は通常の文書検索と異なり，対象文書のサイズが小さいため，質問に含まれる語が検索対象となる質問文に現れないことが多い．そこで，単語の意味の類似性を考慮する必要がある．このような問題に対処するため，たとえば文献[1]では，質問文に基づく言語モデルに加え，回答文を用いたモデル，質問文の変換モデルを組み合わせた確率モデルに基づく類似検索手法を提案している．本論文では，この手法をベースに，(i) 回答文情報をより適切に利用する，(ii) 混合パラメータを教師なしで学習することで可搬性を高める手法を提案する．

2. 提案手法

本論文では，言語モデルに基づいた手法を提案する．まず提案手法の基盤として Latent Dirichlet Allocation (LDA) について述べる．次いで手法の特徴について，モデルの推定法及び混合法に分けて述べる．

2.1. Latent Dirichlet Allocation

LDA[3]は文書コーパスのような離散データの集合の生成モデルである．LDA では文書の生成を以下のようにモデル化している．

(i) 文書内の単語の数だけ latent なトピックが生成される．

(ii) １つのトピックから１つの単語が生成される．

トピックの生成は，同一のディリクレ分布より文書ごとに生成された多項分布に従う．

単語の生成は，同一のディリクレ分布よりトピックごとに生成された多項分布に従う．

2.2. モデルの推定

言語モデルによる Q&A 類似検索では，アーカイブ中の各 Q&A ペア($\boldsymbol{q}, \boldsymbol{a}$)と質問 $\boldsymbol{r}$ との適合度を，($\boldsymbol{q}, \boldsymbol{a}$)が $\boldsymbol{r}$ を生成する確率 $P(\boldsymbol{r}|(\boldsymbol{q}, \boldsymbol{a}))$で表現する．これを求めるため本研究では，４つのモデルの混合モデルを考える．第一のモデルは，Q&A ペアの質問文 $\boldsymbol{q}$ から $\boldsymbol{r}$ の各語を生成するモデル M_1 である．このモデルは以下に示すように最尤推定で求める．

$$P(w|(\mathbf{q},\mathbf{a}),M_1) = P_{ml}(w|\mathbf{q}) = \frac{\#(w,\mathbf{q})}{|\mathbf{q}|}$$

ここで#(w, $\boldsymbol{q}$)は $\boldsymbol{q}$ における w の頻度，$|\boldsymbol{q}|$は $\boldsymbol{q}$ に含まれる単語の総数である．

第２のモデルは，質問文 $\boldsymbol{q}$ が $\boldsymbol{r}$ の各語に変換されるモデル M_2 である．このモデルは変換確率を用いて推定する．変換確率 $P(w/t)$は，単語 t が単語 w に変換される確率である．Q&A ペアの集合をパラレルコーパスとみなし，統計的機械翻訳の技術（本研究では IBM Model 1 を用いた）により変換確率を学習する．これにより以下のように推定できる．

$$P(w|(\mathbf{q},\mathbf{a}),M_2) = \sum_{t\in\mathbf{q}} P(w|t)P_{ml}(t|\mathbf{q})$$

第３に，background smoothing のために，背景分布をモデル M_3 として用いる．

$$P(w|(\mathbf{q},\mathbf{a}),M_3) = P(w|M_3) = P_{ml}(w|C)$$

ここで C は Q&A アーカイブ全体を指す．

以上３つのモデルは Xue らが文献[1]で提案したものである．我々はさらに，回答文 $\boldsymbol{a}$ が $\boldsymbol{r}$ の各語に変換されるモデル M_4 を提案する．回答文から質問文への変換を扱うために，特に回答文集合をソース，質問文集合をターゲットとして学習した変換確率 $P_{Q/A}(w/t)$を利用する．

$$P(w|(\mathbf{q},\mathbf{a}),M_4) = \sum_{t\in\mathbf{a}} P_{Q|A}(w|t)P_{ml}(t|\mathbf{a})$$

2.3. モデルの混合

2.2 節で述べた各モデルを混合して最終的なモ

デルを得る．Xue らの手法ではパラメータを用いて混合し，最適な値を実験的に探している．我々は混合パラメータを教師なしで学習する方法を提案する．各モデル M_i は単語を生成する確率分布であるから，LDA におけるトピック M_i と考えることができる．すると，混合モデルは以下のように与えられる．

$$P(w|(\mathbf{q},\mathbf{a})) = \sum_{i=1}^{4} P(w|(\mathbf{q},\mathbf{a}),M_i)P(M_i)$$

ここで $P(M_i)$はトピック M_i の生成確率である．本論文ではこの値は文書（Q&A ペア）によらず同一であると仮定する．トピックは latent であるためこの値は何らかの手法で推定する必要がある．本研究ではアーカイブ内の Q&A を対象とした Gibbs sampling により学習する．用いる完全条件付き確率は「アーカイブ内の全文書の単語 W と，文書 j の k 番目の単語 w_{jk} を生成したトピック t_{jk} 以外のトピック T_{-jk} がわかっているとき，t_{jk} が M_i である確率」で，以下のようになる．

$$P(t_{jk} = M_i | W, T_{-jk}; \alpha)$$

$$\propto \frac{\#M_i + \alpha_i}{|C| + \sum_i \alpha_i - 1} P(w_{jk}|(\mathbf{q},\mathbf{a})_j, M_i)$$

ここで $\boldsymbol{\alpha}=(\alpha_1, \ldots, \alpha_4)$ はハイパーパラメータ，$\#M_i$ はアーカイブ内での M_i の出現頻度である．

最終的に，クエリ質問 $\boldsymbol{r}$ に対する Q&A ペア($\boldsymbol{q}$, $\boldsymbol{a}$)のスコアは以下のように与えられる．

$$P(\mathbf{r}|(\mathbf{q},\mathbf{a})) = \prod_{w \in \mathbf{r}} P(w|(\mathbf{q},\mathbf{a}))$$

3. 評価実験

提案手法の性能を評価するため，質問文をクエリとして与え，クエリに類似した質問を Q&A アーカイブから検索する実験を行った．アーカイブとして Yahoo！知恵袋の研究機関提供用データのうち「インターネット」カテゴリを用いた．１つの質問に対し複数の回答が存在するとき，それぞれの回答と当該質問とを対にした別個の Q&A ペアとした．Q&A ペアの総数は 171,816 件であった．同カテゴリからランダムに選んだ４０件の質問をクエリとした．提案手法，Xue らの手法，言語モデル，Okapi/BM25 の４つの手法の各検索結果の上位２０件ずつを pooling し，クエリと類似しているかどうかを人手で判定して正解セットとした．

本アーカイブにおいて Gibbs sampling により学習された $P(M_i)$の値を表 1 に示す．なおハイパーパラメータはすべての i に対して α_i=12.5 とした．また比較対象として，Xue らの手法（transLM+QL）および，ベースラインとして言語モデルによる手法（LM）についても実験を行った．評価指標として Mean Average Precision (MAP)及び上位１０件の適合率（P@10）を用いた．結果を表 2 に示す．どちらの評価指標においても提案手法が優れていることがわかる．次に有意性を調べるため，対応のある両側 t 検定を行った．その結果，ベースラインである言語モデルに対しては有意水準 95%で有意性が示されたが，既存手法に対しては十分有意でなく，有意性を示すのは今後の課題である．

4. おわりに

変換確率を利用した類似質問検索の手法において，回答文情報の適切な利用と推定値の理論的・自動的な結合を提案した．提案手法では回答文を有効に利用でき，また教師なしで混合比率を決定できる．性能はベースラインに対して有意に優れていることが示された．

参考文献

[1] X. Xue, et al. *Retrieval Models for Question and Answer Archives.* In *SIGIR,* 2008.

[2] K. Wang, et al. *A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services.* In *SIGIR,* 2009.

[3] D. M. Blei, et al. *Latent Dirichlet Allocation.* Journal of Machine Learning Research 3, 2003.

謝辞

本研究の実施にあたっては，ヤフー株式会社が国立情報学研究所に提供した Yahoo！知恵袋データを利用しました．

表 1　$P(M_i)$の値

トピック（モデル）M_i	M_1	M_2	M_3	M_4
$P(M_i)$	0.515	0.104	0.001	0.380

表 2　実験結果

手法	MAP	P@10
LM	0.2432	0.2275
TransLM+QL	0.3028	0.28
提案手法	0.3313	0.325