相互情報量を用いた質問応答システムのためのクエリ拡張

阿部 裕司(東京農工大学大学院 情報工学専攻)
古宮 嘉那子,小谷 善行(東京農工大学大学院 工学研究院)

本研究では、質問応答タスクのためのクエリ拡張手法を提案する。質問応答におけるクエリ拡張は、入力された質問文に出現しない語彙を含んだキーワード群で文書検索を行なうことで情報源の多様性を増し、システムの回答精度を向上させるものである。FAQ を想定した質問応答において、コーパス内の単語共起頻度と相互情報量を用いたクエリ拡張手法は一定の成果を挙げているが、ドメインに依存しない質問応答を実現する際には、話題がクエリ拡張語の影響でずれてしまうことが問題となる。そこで我々は、質問側の 2 単語に対する回答側 1 単語の共起情報をもとに、話題のずれを抑えつつクエリ拡張を行なう手法を提案する。評価実験として、クエリ拡張を行なう場合と行なわない場合の質問応答システムの性能を比較した結果、正解率が 3 ポイント向上した。

1 はじめに

質問応答(QA)は、自然言語文の質問を入力とし、それに対する回答自体を出力するタスクである。近年は、質問のドメインを限定せず、Web や新聞などの大規模な情報源から適切な回答となる部分を抽出する開領域質問応答の研究が盛んに行なわれている。情報源として Web を利用する際、Web 検索エンジンを質問応答用に独自開発することは困難であるため、既存の Web 検索エンジンを用いて分析対象の文書を得ることが一般的である。

質問応答タスクにおけるクエリ拡張は、入力された質問文に出現しない語彙を含んだキーワード群で文書検索を行なうことで情報源の多様性を増し、システムの回答性能を向上させるものである。クエリ拡張語としては、もとの質問文に含まれる語彙の同義語、類義語、関連語などが用いられ、それらを取得する様々な手法が提案されている。本研究では、コミュニティ Q&A サイトの質問応答事例における語の共起情報をもとに、クエリ拡張語を取得する手法を提案する。

2 関連研究

[Berger00]は、質問側にある語が出現した際に回答側にどのような語が出現しやすいかを FAQ(frequently-asked question)コーパスを用いて学習し、その情報をクエリ拡張に利用する手法を提案している。[Berger00]では、質問語と回答語の依存関係の度合いを測る尺度として相互情報量を利用している。相互情報量の式を以下に示す。

$$I(Wq,Wa) = P(w_q, w_a)\log\frac{P(w_q, w_a)}{P(w_q)P(w_a)} + P(w_q, \overline{w_a})\log\frac{P(w_q, \overline{w_a})}{P(w_q)P(\overline{w_a})} + P(\overline{w_q}, w_a)\log\frac{P(\overline{w_q}, w_a)}{P(\overline{w_q})P(w_a)} + P(\overline{w_q}, \overline{w_a})\log\frac{P(\overline{w_q}, \overline{w_a})}{P(\overline{w_q})P(\overline{w_a})} \quad (1)$$

ここで、W_q は質問側に語 w_q が出現するかどうかを示す確率変数、W_a は回答側に語 w_a が出現するかどうかを示す確率変数である。

$$Wq = \begin{cases} w_q & \text{質問側に}w_q\text{が出現する} \\ \overline{w_q} & \text{質問側に}w_q\text{が出現しない} \end{cases}$$

$$Wa = \begin{cases} w_a & \text{回答側に}w_a\text{が出現する} \\ \overline{w_a} & \text{回答側に}w_a\text{が出現しない} \end{cases}$$

質問側の語 w_q と回答側の語 w_a のコーパス内における共起性が強いほど、それらの語相互情報量は大きくなる。[Berger00]は、新しい質問が与えられた際に、質問文内のそれぞれの語に対し相互情報量を最も大きくする語を一つずつクエリ拡張語として利用する手法を提案している。以降、この手法で行なわれるクエリ拡張について、{“質問側の語”→”回答側の語”}と表記する。

学習データ、テストデータともにドメイン依存の強い文書[1]を利用する場合、この手法で有効なクエリ拡張を行なうことができる。しかし、ドメイン非依存の質問応答する場合、クエリ拡張語による話題のずれが大きな問題となる。予備実験段階で[Berger00]の手法を実装した際の事例を以下に示す。なお、相互情報量計算のための学習データとして『Yahoo!知恵袋データ』の質問回答事例を利用した。

【質問】
ソフトバンクとヤフーはどんな関係にありますか？
【クエリ拡張】
{ソフトバンク→ホークス}
{ヤフー→メール}

”ホークス”および”メール”は”ソフトバンク”および”ヤフー”とは関連性の高い語であるが、もともとの質問の内容とは何ら関係の無い語である。これらの語をクエリ拡張語として利用して文書検索を行なった場合、話題が大きくずれてしまうため芳しい結果は得られない。

そこで我々は、質問側の 2 単語に対する回答側 1 単語の共起情報をもとに相互情報量を計算することにより、話題のずれを抑えつつクエリ拡張を行なう手法を提案する。次章で提案手法の詳細を述べる。

3　質問側の 2 語に対するクエリ拡張手法

既存手法のクエリ拡張手法では、質問語の 1 語に対して出現に相関のある語をクエリ拡張語として利用するため、クエリ拡張語により話題がずれるリスクが大きかった。そこで、質問側の 2 語に対して出現に相関のある語をクエリ拡張語として利用する。この手法の実現のために、相互情報量の計算式を以下のように変更した。

[1] [Berger00]は、コンピュータ関連の FAQ および航空会社のコールセンターダイアログを利用している。

$$I(Wq_1, Wq_2, Wa) = P(w_{q1}, w_{q2}, w_a)\log\frac{P(w_{q1}, w_{q2}, w_a)}{P(w_{q1}, w_{q2})P(w_a)}$$
$$+P(w_{q1}, w_{q2}, \overline{w_a})\log\frac{P(w_{q1}, w_{q2}, \overline{w_a})}{P(w_{q1}, w_{q2})P(\overline{w_a})}$$
$$+P(\overline{w_{q1}}, \overline{w_{q2}}, w_a)\log\frac{P(\overline{w_{q1}}, \overline{w_{q2}}, w_a)}{P(\overline{w_{q1}}, \overline{w_{q2}})P(w_a)}$$
$$+P(\overline{w_{q1}}, \overline{w_{q2}}, \overline{w_a})\log\frac{P(\overline{w_{q1}}, \overline{w_{q2}}, \overline{w_a})}{P(\overline{w_{q1}}, \overline{w_{q2}})P(\overline{w_a})} \quad (2)$$

この式は、質問側の 2 語 w_{q1}, w_{q2} と回答側の 1 語 w_{a2} の共起性の強さを示すものである。それらのコーパス内での共起性が高いほど、相互情報量の値は大きくなる。
新しい質問が与えられた際に、質問内の語の二つ組をつくり、それらに対し相互情報量を最も大きくする語を一つずつクエリ拡張語として利用する。これにより、話題のずれを抑えつつクエリ拡張を行なうことができる。例を以下に示す。

【質問】
ソフトバンクとヤフーはどんな関係にありますか？
【クエリ拡張】
{ソフトバンク,ヤフー →子会社}

4 評価実験

クエリ拡張の有効性を確認する評価実験を行なった。既存の質問応答システムを基本システムとし、それにクエリ拡張モジュールを追加したものを利用した。相互情報量を計算するための学習コーパスには『Yahoo!知恵袋データ』を利用し、回答候補の取得には Web を利用した。評価用の質問として NTCIR-8 の ACLIA2 タスク 100 問[Mitamura10]を用い、正解判定は人手で行なった。回答は文単位でスコアの上位 5 件を出力するものとし、ACLIA2 の正解データと同一の内容を示していると思われる場合に正解として扱った。システム性能の評価指標として正解率と MRR を用いた。

4.1 基本となる質問応答システム

基本システムとして、[石下 09]の質問応答システムを再現したものを利用した。このシステムは Web を情報源としており、回答候補の評価指標として「質問の内容との関連性」と「質問の型に応じた記述スタイルを満たす度合い」を計算し、二つの指標をある混合比で掛け合わせて最終評価値とするものである[2]。以下に回答候補の評価式を示す。

$$Score(S_i) = \frac{1}{\ln(1+length(S_i))} \cdot \left\{\sum_{j=1}^{l} T(w_{i,j})\right\}^{\gamma} \cdot \left\{\sum_{k=1}^{m}\sqrt{\chi^2(b_{i,k})}\right\}^{1-\gamma} \quad (3)$$

ここで、l は文 S_i 内の語 $w_{i,j}$ の異なり数、m は文 S_i 内の言語表現(「質問の型に応じた記述スタイルを満たす度合い」の計算時の特徴) $b_{i,k}$ の異なり数、$length(S_i)$ は文 S_i の文字数、T は各語の内容関連度、χ^2 は各言語表現のスコア、γ は評価尺度の混合比を決めるパラメータを表す。実験では γ を 0.5 に固定し、言語表現として 1-gram と 2-gram を使った場合を個別に実験、評価した。

[2] 本研究の主張点では無いため説明を割愛する。詳しくは[石下 09]を参照されたい。

4.2　情報源の取得方法

回答候補の取得元となるWeb文書の取得方法の概略を図1に示す。

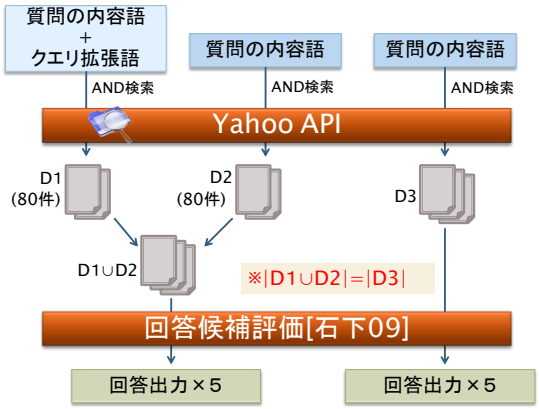

図1　Web文書の習得方法概略

クエリ拡張語の選択方法として、まず質問文内の内容語のすべての二つ組に対し、最も相互情報量を高くする語をそれぞれ1語ずつクエリ拡張語候補として選出した。クエリ拡張語候補が3語以上ある場合は、相互情報量の値上位3語のみをクエリ拡張語として利用した。

図1で*D1∪D2*と表記された文書集合は提案手法を利用して得られた文書集合を表しており、クエリ拡張を用いて検索された文書を含んでいる。一方、*D3*と表記された文書集合はクエリ拡張を用いて検索された文書を含んでいない。*D1∪D2*と*D3*の文書数を同じにし、共通の回答候補評価モジュールを利用することにより、クエリ拡張を用いた文書検索の有用性を確認することができる。

5　実験結果および考察

実験結果を表1,2に示す。

表1　記述スタイルの特徴として1-gramを用いた場合

	クエリ拡張の有無	
	なし	あり
正解率	0.42	0.45
MRR	0.262	0.237

表2　記述スタイルの特徴として2-gramを用いた場合

	クエリ拡張の有無	
	なし	あり
正解率	0.27	0.23
MRR	0.145	0.142

記述スタイルの特徴として1-gramを用いた場合はクエリ拡張を行なう方が、2-gramを用いた場合はクエリ拡張を行なわない方が、高性能であるという結果が得られた。全体として1-gramを用いた場合の性能が高いのは、2-gramを利用した場合には混合比γ=0.5という値は適切ではなく、「質問の内容との関連性」が軽視されすぎたためであると考えられる。混合比γは様々な値を試す必要があるが、本実験ではクエリ拡張を用いた場合に最も高いシステム性能が得られた。

実験時に行なわれたクエリ拡張を調査した所、一般的な語に対するクエリ拡張語が大きく話題をずらしてしまう可能性が高いことが分かった。例を以下に示す。

【クエリ拡張】
{する,年　　→結婚}
{きっかけ,する　→結婚}
{元,年　　→別れる}
{くださる,する}　→クリック}

このような例はシステム性能を低下させ

る原因となる。TF-IDF などを用いて、クエリ拡張を行なう語と行なわない語を選別する必要があると思われる。

6 おわりに

質問応答タスクのためのクエリ拡張手法を提案した。質問側の 2 語に対する回答側の 1 語の共起情報をもとにクエリ拡張を行なう事で、話題のずれを緩和することに成功した。評価実験として、クエリ拡張を行なう場合と行なわない場合の質問応答システムの性能を比較した。結果、記述スタイルの特徴として 1-gram を用いた際に正解率がもとのシステムに比べ 3 ポイント向上した。

謝辞

本研究を行なうにあたり、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』を利用させて頂きました。利用を快諾して下さった各社に感謝いたします。また、評価実験の際に NTCIR-8 の質問応答テストコレクション『ACLIA2』を利用させて頂きました。NTCIR の運営にご尽力をいただいている皆様にも感謝いたします。

参考文献

[石下 09]石下 円香, 佐藤 充, 森 辰則:Web 文書を対象とした質問の型に依らない質問応答手法, 人工知能学会論文誌, pp.339-350(2009).

[Burger00] Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal:Bridging the lexical chasm: Statistical approaches to answer-finding, In Proceedings of SIGIR, pp.192-199(2000).

[Mitamura10]Teruko Mitamura, Hideki Shima Tetsuya Sakai, Noriko Kando, Tatsunori Mori, Koichi Takeda, Chin-Ywe Lin, Ruihua Song, Chuan-Jie Lin and Cheng-Wei Lee: Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access, In Proceedings of 8th NTCIR Workshop Meeting(2010).