E-022

質問の関連性を考慮したクエリ生成手法の提案

Proposition of Query Formation by Question Relationship

木村 泰知† 荒木 健治‡
Yasutomo Kimura Kenji Araki

1．まえがき

近年，音声認識技術の発展に伴い，カーナビゲーションシステム，天気情報提供システムに代表される音声対話処理システムが普及している．また，インターネット，新聞記事のような大量データから適切な回答をみつける，質問応答システムもさかんに研究が行われている．音声を利用した対話処理と大規模データによる質問応答では，対話処理として同じカテゴリーにもかかわらず，大きな隔たりがある．音声対話処理システムでは音声認識の問題を克服するために質問範囲が狭くなる．一方，情報検索を用いた対話処理システムでは質問の対象範囲が広いが書き言葉による質問とされている．そのため，話し言葉による対象範囲の広い質問応答システムが望まれる．

また，音声対話の発話は，音声認識精度の影響もあり，システムに対して短い単位で発話する傾向にある．例えば，音声によるカーナビゲーションの場合，行き先の番地を最初から最後まで言い切ってしまうことは稀である．すなわち，発話が多くなることにより文脈処理が必要になる．

さらに，大規模データの質問応答システムにおいても，文脈を考慮した質問が注目されており，NTCIR の Question Answering Challenge においても主要な課題として取り上げられている．特に，大規模データの質問応答において，文脈処理は困難であり，対象の限定が行われていないため，特化した規則を与える方法では，対応できない場面も少なくない．

そこで，我々は話し言葉における対象範囲の広い質問応答システムの第一歩として，文脈を考慮した広範囲の質問応答を行うことを目指し，対象に依存しない文脈処理を目的とする．本稿では，質問応答システムを作成し，文脈処理の有無による比較実験の結果について述べる．

2．処理過程

(1) ベースラインの手法

(ア) 入力文処理

① 入力文から応答タイプを決定する．

1. 予め SVM を用いて学習を行う．
2. ここで利用する学習データは評価データと異なる質問データセットである．
3. 素性は ChaSen により単語分割した結果を用いる．

(イ) クエリ生成処理

① ChasSen による名詞を抽出する．

② 名詞，数詞が連続した場合は結合する

③ クエリの優先順位は各単語の情報量により決定する．情報量により降順にソートしたパターンを生成する．

$$I(x) = -\log_2 P(x) \quad \cdots (1)$$

ここで，x は単語であり，P(x)は単語の生起確率である．

④ 検索結果がない場合もあるため，情報量の小さい単語から削除した，パターンを予め複数生成する．

(ウ) 検索処理

① 新聞記事データ（毎日新聞，読売新聞の各2年分）を予め Namazu [1]でインデキシングを行う．Namazu は TF*IDF によりキーワードに重み付けが行われているため，多くの記事に含まれるキーワードには低いスコアになる．

② 検索結果が出力されるまで，優先順位の高い検索結果から繰り返す．

(エ) 応答抽出処理

① 質問と応答パターンの対から SVM により学習を行った結果を用いる．

② 入力文から応答パターンを判断する．ここでは，人名，組織名，場所，日付，数値，その他の分類を行った．固有表現(Named Entity)の識別は NExT を利用した[3].

(2) 文脈処理

大規模データにおける質問応答は自然言語による入力である．そこで，情報検索のために，入力文からクエリ生成が行われる．クエリは検索結果に大きな影響を与えるため，クエリ生成に文脈情報を含めることは大きな意味がある．ここでは，単語の出現頻度を利用する．例えば，「どこにあるの？」が質問である場合，出現頻度の高い単語で構成されているため，クエリとしてこの質問だけでは処理できないと判断する．さらに，出現頻度が低い単語の場合でも，質問文に含まれる単語から関連ある共起であるか判断することにより，何文前まで扱うか判断する．

文脈処理の流れを下記に示す．

(ア) 入力文に含まれる各単語の情報量を求める．

(イ) 求めた情報量の最大値が予め設定した閾値より，低ければ前文をクエリに利用する．

(ウ) 情報量が大きい単語が含まれている場合，前文に含まれる単語との関連性を確認する．

① 自己相互情報量が大きければ，単語の関連が高いと判断し，前文の情報をクエリ生成に利用する．下記に自己相互情報量の式を示す．

† 小樽商科大学
‡ 北海道大学大学院情報科学研究科

$$MI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \quad \cdots (2)$$

P(x,y) は単語 x と単語 y の共起確率である．

② 自己相互情報量が小さければ，文脈情報を利用せずに，クエリを生成する．

3．実験

(1) 実験の目的

(ア) 文脈情報を考慮したクエリ生成の有効性を確認することを目的とする．

(2) 実験方法

比較する手法

- ベースライン
- ベースライン+文脈処理手法

QAC では質問 ID を与えられており，関連ある質問の範囲が ID 番号により識別可能となっている．表 1 に示すように，05-01 から 05-05 は関連ある質問となる．しかしながら，我々の目的は実際の対話システムにおける文脈処理を想定しているため，予め関連ある質問であるという情報を与えない．つまり，関連している質問の範囲を前文と今回の入力の関連性から自動的に判断する．

表１． 質問データセットの例

ID	質問文
04-07	・・・
05-01	"チャールズ皇太子は何歳ですか．"
05-02	"ダイアナ妃とはいつ結婚しましたか．"
05-03	"彼女とはいつ離婚しましたか．"
05-04	"２人の間に子供は何人いますか．"
05-05	・・・

(3) 実験データ

2000 年と 2001 年の毎日新聞と読売新聞の記事を学習データとして利用する．ここでは，記事ごとに分割を行い，858,399 文書として，インデックス化を行った．キーワードは 4,262,281 個存在する．評価データは NTCIR の QAC3 の質問データセットを利用する．問題数は 360 問であり，50 シリーズある．つまり，1 つのトピックに関する質問が 7～8 存在するデータセットである．質問タイプを決定する SVM の学習データは，質問データセット以外の 200 質問を用いた．

(4) 実験結果

ベースラインとベースライン＋文脈処理手法の実験結果を表２に示す．ここで正解とは，システムが出力した 20 の応答候補に正解が含まれていることである．

Baseline の正解応答数は 53，本手法の正解数は 60 であった．文脈を考慮することにより正解数が 7 つ増加した．

表２．比較実験における正解数

	正解数
Baseline	53
本手法	60

(5) 考察

表 3 に正解例を示す．文脈を利用した正解例にある「この宇宙ステーションはいつ廃棄されましたか。」の質問に対しては，Baseline 及び本手法の応答は正解であった．しかしながら，「廃棄によってどこに落下したのですか。」については，本手法のみの正解であった．これは，前の質問文を考慮することにより正解となった例である．関連性のある質問を判断することにより，正解数も向上することが確認された．しかしながら，前の応答をクエリに含めなければ，応答候補をみつけることは困難な質問がある．今回の質問データセットには，このような質問が 50 問存在していた．我々は本稿において質問の関連性のみを扱ったが，前の応答を考慮するしくみも考える必要がある．

表３． 実験における応答例

正解例	
質問	オープン初日のフィルムカット式に出席した俳優は誰でしたか。
応答	アーノルド・シュワルツェネッガー
文脈を利用した正解例	
質問 30-05	この宇宙ステーションはいつ廃棄されましたか。
応答 30-05	２３日
質問 30-06	廃棄によってどこに落下したのですか。
応答	太平洋

5．まとめ

本稿では話し言葉における対象範囲の広い質問応答処理を目的とし，その第一段階として文脈処理手法の提案を行った．質問応答処理では文脈を扱うことが課題とされており，関連ある入力文の範囲を識別することはクエリ生成に大きな意味がある．そこで，関連ある入力文の識別を行い，QAC の課題である 360 問を行った．ベースラインと文脈を考慮したシステムの比較実験では，ベースラインよりも正解数が向上した．しかしながら，前に質問された応答を利用して答えなければならない質問には対処することが困難であるため，今後解決していく予定である．

参考文献

[1]NTCIR http://research.nii.ac.jp/ntcir/
[2]Namazu http://www.namazu.org/
[3] 渡邊一郎，桝井文人，福本淳一: "固有表現抽出ツールＮＥｘＴの精緻化とユーザビリティの向上", 言処理学会第10回年次大会発表論文, 2004.3.