Web検索を利用したYes-No型QAシステム

藤原 佑斗　浦谷 則好
東京工芸大学大学院工学研究科

1 はじめに

質問応答システムは従来のWeb検索に比べ，より具体的にユーザが必要とする情報だけを提示するシステムである．

本研究では，factoid 型で回答できる質問文に対応し研究したものであり，non-factoid 型については対応していない．質問文に含まれる新情報と旧情報を利用して Yes-No を判断するシステムを構築する．旧情報に対して新情報が本当に正しいかを判断することで回答を決める．

2 関連研究

金井らの研究[1]のように factoid 型の質問応答システムでは高い精度を得られている．若目田の研究[2]では,ユーザによって与えられた質問文に対して回答を提示するために，質問文に含まれる名詞，動詞，形容詞をキーワードとして抜き出しそれらの共起頻度を利用している．抜き出したキーワードの一つを解仮定のキーワードと定め，残りのキーワードを検索キーワードとして用いる．検索キーワードでWeb検索を行い，取得した情報に対して形態素解析を行う．取得した品詞情報をもとに解仮定のキーワードと同じ品詞となる形態素を抽出して候補とする．候補の出現頻度のトップとなる単語が解仮定のキーワードとなるかを調べる．各々の解仮定のキーワードに対してこの手法を行い，二つ以上トップとなるキーワードが存在したら「Yes」そうでなければ「No」と判別している．

3 研究内容

本研究は,若目田の研究[2]と同様にユーザの提示する質問文に対し，Web情報を用いてYesまたはNoの判断を自動的に行うシステムを構築する．

3.1 質問文の制限

factoid 型の質問文に対応し，必ず回答を導くことのできるものとする．また，他との比較といったような相対的な関係性によって初めて回答を得ることが可能なものや個人の趣意により回答が変わってしまうようなものは質問文として使用しない．

〈質問文の例〉
　海の日は第三月曜日ですか
〈対象としない質問文の例〉
　日本はドイツより面積が大きいですか
　日本人は真面目ですか

3.2 人名辞書の作成

今回は，事前にWikipediaに登録されている人名の一部を「固有名詞-人名-一般」として形態素解析に利用する茶筌[3]の固有名詞辞書に追加登録しておく．これにより茶筌使用後における結果に対する誤りを減らし，その後の形態素に対する補正を軽減することができる．

3.3 旧情報と新情報の定義

ある文をその品詞情報をもとに旧情報と新情報を取得するために，係助詞「は」と格助詞「が」をキーとして利用し，その前後文を主語と述部に分割する．この時，各々の主語と述部から旧情報と新情報を定義する．係助詞「は」の場合は主語が旧情報，述部が新情報となり，格助詞「が」の場合は，主語が新情報となり，述部が旧情報となる．

$$\frac{\text{主語}}{\text{旧情報}}\ \text{は}\ \frac{\text{述部}}{\text{新情報}}$$

$$\frac{\text{主語}}{\text{新情報}}\ \text{が}\ \frac{\text{述部}}{\text{旧情報}}$$

4 システムの構成

本研究で構築したシステムを以下の図1に示す．

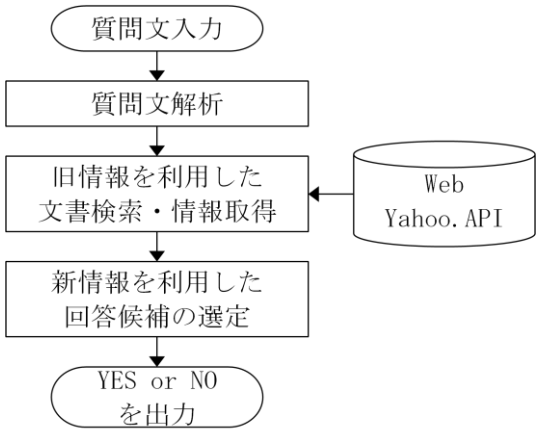

図１　システムの構成

質問文の解析には3.2で作成した人名辞書を追加した茶筌で解析を行う．3.3 で述べた定義を利用して質問文を旧情報と新情報に分割する．このとき，旧情報と新情報に含まれる意味情報からノイズとなる動詞や助詞などを取り除く．

旧情報を検索キーワードとして Yahoo！API を利用してWeb検索を行う．回答候補を抽出するための

スニペット(Summary)は 300 件とする.検索キーワードの例を以下に示す.

質問文：今年の干支は兎ですか
↓
検索キーワード：「今年の干支」

Web 検索は基本的に完全一致検索で行い，情報が得られなかった場合のみ通常検索で情報を取得する.

回答候補抽出のために新情報の品詞情報を利用する． Web から取得したスニペット情報を形態素解析にかけ，新情報から得られた品詞列と同様な品詞列の形態素列を抽出する．これにより新情報に二つ以上の形態素が含まれている場合，単一なものよりも候補となり得る情報の制限が厳しくなるため，新情報が回答候補となる確率が高まると考えられる．しかしながら，単一の表現で正答が存在したとしても除外してしまうデメリットも予測される.

Web から取得した情報には検索キーワードとなる旧情報の形態素が多数存在することは当然である．今回のシステムでは回答候補抽出に品詞情報を利用しているため，旧情報と新情報の品詞が一致した時に間違った候補として旧情報が抽出されてしまう．このとき，旧情報の形態素を含むものを回答候補から除外する.

情報提供者による書き方の違いによって同意の回答候補であっても別の候補として扱われる．そこで今回は，新情報を含む回答候補となる形態素列を統一する処理を加えた．新情報と回答候補において一方が他方を包含している場合，回答候補を新情報に統一する．統一する形態素列の例を以下に示す.

新情報：アメリカ
回答候補：アメリカ
　　アメリカ合衆国
　　北アメリカ
統一：アメリカ

これにより，重複をカウントして降順にランキングした時の順位が入れ替わる．この順位の変動によって正解の語が異表記で書かれていた場合において精度が高まる.

1 位：ワシントン	100
2 位：アメリカ	70
3 位：アメリカ合衆国	50
4 位：ニューヨーク	30
↓	
1 位：アメリカ	120
2 位：ワシントン	100
3 位：ニューヨーク	30

結果，順位が第一位となる回答候補(複数の場合も)に新情報が含まれていれば Yes，そうでなければ No を回答としてユーザに提示する.

5 実験結果

3.1 で制限した質問文で回答が Yes となる質問文 34 問を，No となる質問文 25 問を実験に用いた．また，若目田の研究[2]の結果を評価基準(以降，Base Line)として,本研究の結果と共に表 1 に実験結果を示す．また，不正解であった回答となる新情報を回答候補と共に例として以下に示す.

本手法で行った結果正答率が 9 割を超え，Base Line よりも高い精度となった．これは旧情報に対し新情報の共起頻度を求めたことと新情報の品詞列を利用したことにより，回答候補として出現する形態素列が限定できたためであると考えられる.

正解とならなかったものは，例 1 のように「今」と「現在」を文字列だけでマッチングをとっているため，同一に扱えない．また例 2 のように「ワシントン」は「アメリカ」の一部であるという含意関係にあるものも間違えて回答してしまった.

例 1　質問文：安芸は今の広島県ですか
　　新情報：今の広島県
　　回答候補：現在の広島県

例 2　質問文：ホワイトハウスがあるのは
　　ワシントンですか
　　新情報：ワシントン
　　回答候補：アメリカ

表 1．実験結果

	Yes	No
Base Line	62%	78%
本研究	91%	96%

6 おわりに

本実験結果では Base Line を上回る精度が Yes, No 共に得られた．このことから，質問文を旧情報と新情報に分割して旧情報に対する新情報の共起頻度を用いて Yes,No を提示する手法の有効性を示せた．つまり，通常の factoid 型質問応答システムに対する Yes-No 型の強みを利用することができた.

今後は，日本語語彙体系などを利用し，語意の統一や上位・下位の含意関係に着目してさらに完成度を高めていきたい．また，回答が non-factoid 型の質問文に対応できるよう拡大したい.

参考文献

[1] 金井明, 佐藤充, 石下円香, 森辰則:複数の Web 検索エンジンを用いた factoid 型質問応答, 自然言語処理研究会報告, pp.101-108, 2007
[2] 若目田亜依: Yes-No 型 Q&A システムの構築, 東京工芸大学卒業論文要旨集, pp. 49, 2010
[3] http://chasen.naist.jp/hiki/ChaSen/