ウェブ文書を知識源とした曖昧な質問に対する質問応答

05T4047H

田中洸一

構成

- １．問題提起
- ２．システム概要
- ３．解答リストを含む表の抽出
- ４．テキスト解析による解答群の作成
- ５．実験とその結果
  5.1表　5.2解答群　5.3組み合わせ
- ６．全体の評価と今後の課題

１．問題提起（１）

質問「ワールドカップの優勝国はどこですか？」

↓

答え・イギリス（ラグビーのワールドカップ）

・ブラジル（サッカーのワールドカップ）

・ノルェー（スキーのワールドカップ）

↓

答えがひとつではない

１．問題提起（２）

・曖昧なキーワードがある質問はキーワードの意味に応じて解答が複数存在してしまう。

（先ほどの例の場合、ワールドカップが曖昧なキーワード）

↓　解決法

↓

◎複数の解答をリスト化して出力する

１．問題提起（３）

- 質問に対する解答を探し出す知識源として
  ウェブを用いる。

↓

◎解答リストとなりうる表を抽出（新しく導入）

＋

◎テキスト解析による解答リストの生成

（従来のテキスト解析手法を、
新聞記事対応からウェブ対応に改変）

2. システム概要(1)

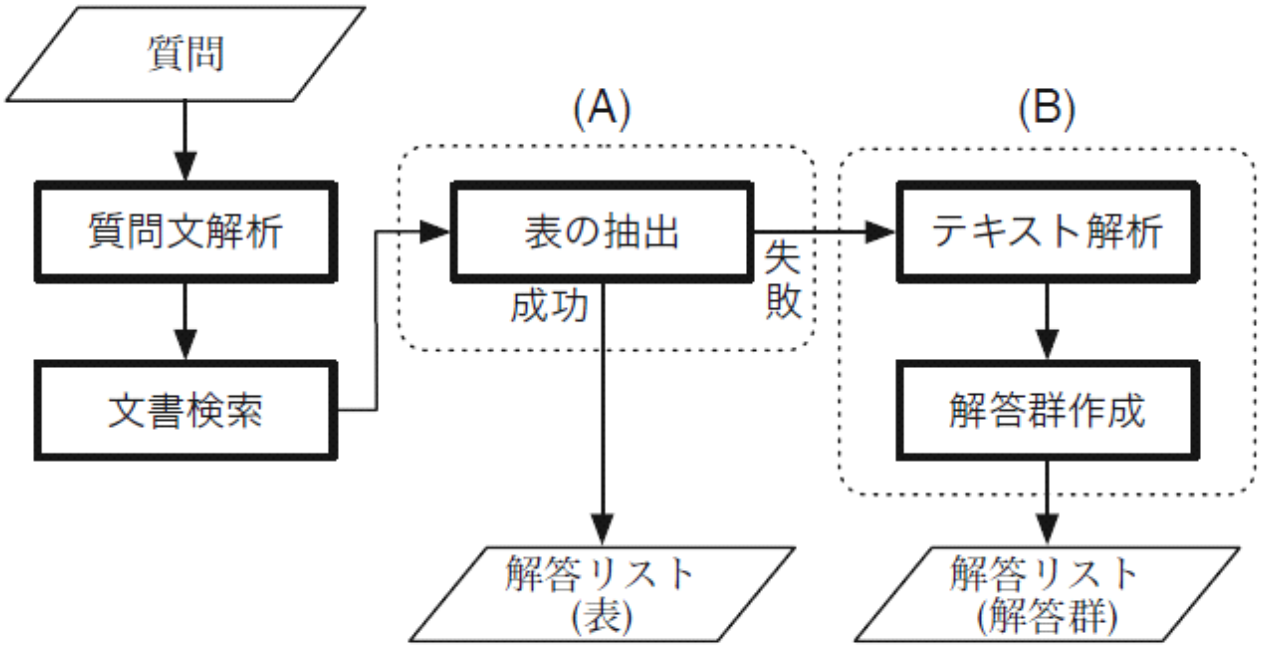

図 1: 提案システムの概要

２．システム概要（２）

質問文から以下に分類し検索

・プライマリーキーワード
　　＝最も関係の深いキーワード1つ

・セカンダリーキーワード
　　＝プライマリーキーワード以外

・使用検索エンジン　→　Tsubaki

・利用範囲→検索結果の上位100件のページ

３．解答リストを含む表の抽出(1)

- 複数の解答を含む表を抽出する手続き

１．tableタグで定義されている表を検出。

２．表の1行目または1列目あるセルがプライマリキーワードk_pと一致する。

（ただし、セル内の文字列が複合名詞や末尾などに一致すればk_pであるとみなす）

３．解答リストを含む表の抽出(2)

３．すべてのセカンダリキーワードが以下のいずれにも存在しない場合は除外。
（１）表のキャプション
（２）表の前にある3つのセグメント
（３）そのページのtitleタグの中

４．２．で検出したセルの同じ行または列に解答が含まれるかチェックする。

3. 解答リストを含む表の抽出(3)

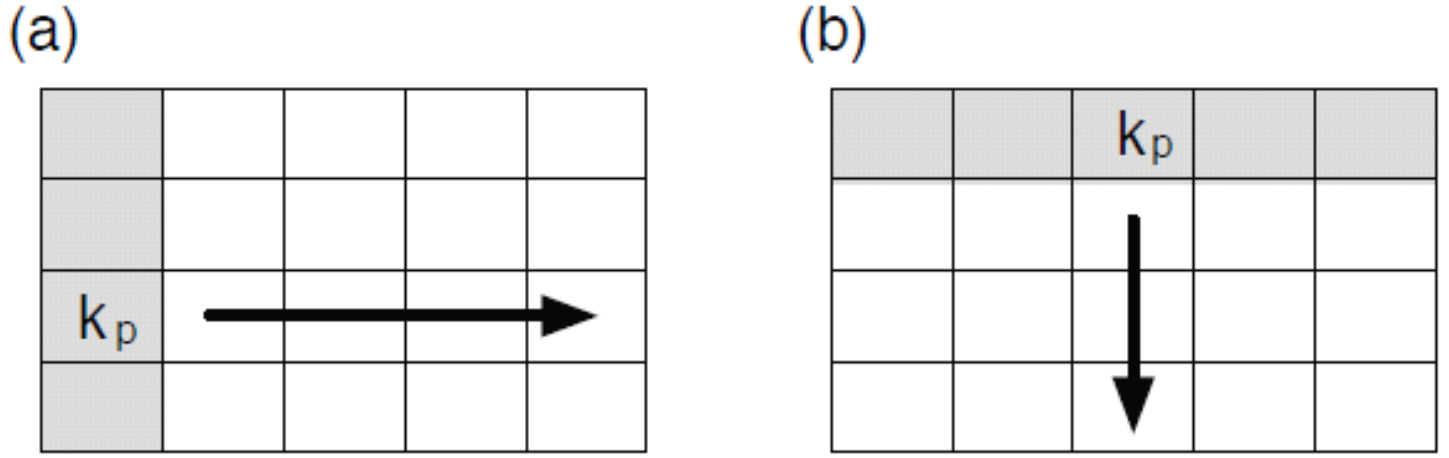

図 2: 複数の解答を含む表の検出

３．解答リストを含む表の抽出(４)

・表に解答を含むかのチェックには、”南瓜”を用いる。

◎南瓜によって解析されたセル内のテキスト

　　｜

　　｜ 一致している割合が0.3以上なら、

　　｜　　その表を抽出し提示する

◎質問の解答タイプ

４．テキスト解析による解答群の作成（１）

- セグメント＝HTMLのタグによって分割された領域

address, blockquote, div, dl, h1, h2, h3, h4, h5, h6, hr, ol, p, pre, table, ul, noframes, noscript, dir, menu

以下の条件を満たすセグメントを抽出

◎全てのキーワードが含まれるセグメント

◎セカンダリキーワードのいずれかがtitleタグに含まれ，かつプライマリキーワードを含む残りのキーワードが含まれるセグメント

４．テキスト解析による解答群の作成（２）

- 得られたセグメントを南瓜で解析し、解答候補を抽出して以下の組を生成。

（a_i,k_j,s_k）　　a_i=解答候補
k_j=質問文内のキーワード
s_k=キーワードk_jの意味を限定する限定表現

※限定表現はあらかじめ用意された抽出パタンを用いる

４．テキスト解析による解答群の作成（3）

・（a_i,k_j,sk） から共通の属性（attr）を発見し、解答群AGを生成。

AG(k,attr)={(a_i,s_i)}　　k=曖昧なキーワード

a_i=解答候補

s_i=キーワードkの限定表現

例　質問「シドニー五輪の柔道の金メダリストは誰?」

AG(金メダリスト,数+キロ級)=

{(井上康生,男子１００キロ級),

(井上康生,五輪１００キロ級),

(田村亮子,女子４８キロ級)}

４．テキスト解析による解答群の作成（4）

複数得られる解答群AGを

(1)AGにおける限定表現や解答候補の異なり数

(2)限定表現の共通属性attrのタイプ

(3)解答候補の信頼度

(4)キーワードと限定表現の関連度

などに応じてスコア付けをおこない、最大のスコアを持つ解答群を解答リストとする。

５．実験とその結果

・テストデータ
  ・３０問の曖昧な質問

・実験結果
  ・解答リストを含む表の抽出による結果（5.1）
  ・テキスト解析による解答群生成の結果（5.2）
  ・上記の２つを組み合わせた結果（5.3）

5.1解答リストを含む表の抽出による結果

表 1: ウェブページから抽出された表の評価

抽出された表の数	24
精度	88%
再現率	46%

精度＝出力された表のうち、正解の割合

再現率＝正解となる表のうち、この手法で抽出できた割合

5.2テキスト解析による解答群生成の結果

表 2: 生成された解答群の評価

スコア 1 位	13(43%)
10 位以内	22(73%)
正解解答群の平均順位	2.1

5.3組み合わせ手法の結果

表 3: 解答リストの手法別の評価

	表	解答群	併用
(A) 出力	10	30	30
(B) 正解を含む	9	22	25
(C) 正解が 1 位	9	13	17

６．全体の評価と今後の課題（１）

- ウェブページの表とテキストを知識源として利用することで、曖昧な質問に対するより多くの解答を抽出できた。

- しかし、得られる解答リストの多くは大会の開催年や開催回数の曖昧性を反映したものであり、多様な質問に対応できているとは言い難い。

- 多様な曖昧性に対応し、適切な解答を作成できるようにすることが今後の課題。

6. 全体の評価と今後の課題(2)

質問「全英オープンで優勝したのは誰ですか?」

全英オープンゴルフ・歴代優勝チャンピオン

年	優勝者	出身国	開催地
2006	タイガー・ウッズ	米	ロイヤルリバプール・ゴルフクラブ
2005	タイガー・ウッズ	米	セント・アンドリュース
2004	トッド・ハミルトン	米	ロイヤルトルーン
2003	ベン・カーティス	米	ロイヤルセントジョージズ
2002	アーニー・エルス	南ア	ミュアフィールド

図 3: 抽出された表の例