英語空欄補充問題を対象とした質問応答システム

佐々木 稔　遠藤 あゆみ　新納 浩幸
茨城大学工学部

1　はじめに

World Wide Web(WWW) で公開されるデータの数は年々増加し，今では無限に存在していると言える．そのため，文章が多少崩れているなど，データの質がある程度低下することを許せば，WWW 上のデータは非常に有用な情報源として利用することができる．WWW 上のデータから問題解決のための新しい知識を獲得し，蓄積することで，様々な質問に解答する質問応答システムの研究が盛んに進められている [1] [2].

WWW 上のデータには，用語解説などの文を記述するには様々な言語が使われている．このような言語の中で中心となるのが英語で，英語で書かれたページは WWW 上に無限に存在するといっても過言ではない．そのため，英語のページは巨大な用例集であると言える．これらのサイトは，英語の学習者にとって単語，熟語の用例などを調べるための有用な情報であり，それらを検索する研究も盛んである．例えば，入力した英単語列に対して WWW 上でのヒット数を競い，正しく使われていることを調べる Googlefight [3] や，英熟語の用例検索を行う手法も提案されている [4].

本研究では，選択肢のある英文の空欄補充問題を対象とした質問応答システムを構築し，検索エンジンのみを手がかりとした場合での有効性を検証した．本システムでは，WWW で公開されている英文を手がかりとするために，空欄前後の単語と選択肢の語句を組み合わせた単語列に対して検索エンジンを用いて検索を行い，最もよく使われているものを正解とする方法を用いた [5].

2　英語空欄補充問題と知識

本研究では，大学入試や模擬試験などでよく出題される英語の空欄補充問題を用いる．空欄補充問題には過去に記憶した事柄を思い出して，解答を導き出すことを主眼としたタイプがある．このようなタイプの問題には，学習者の記憶を強化することを目的とする以外に，知識を思い出すためにどのような手がかりを探して，利用するのかを訓練させるという目的もある．

英語の空欄補充問題で解答を導き出すために，解答を選ぶまで実際にどのような過程をたどるのかを考える．問題に答える人は，空欄補充問題の選択肢の中から解答となるものを選ぶとき，何が問われているのかを知るために，空欄前後を中心に品詞，構文，語法，意味などのさまざまな手がかりを探す．手がかりが見つかると，それを記憶の中から同じ文構造や文脈をもつ文を探して解答となる選択肢を特定する．このとき，記憶した知識と記憶の中から同じ文を探す手がかりを分けて，知識を WWW 上のデータとし，手がかりを検索式と考えて問題解決を行う．

3　英語空欄補充問題解答システム

本研究で開発した英語空欄補充問題の解答システムの処理を以下に示す．

1. 問題文とそれに対する選択肢を入力する．このとき，問題文が会話文である場合には会話内容を括るダブルクォート「”」を取り除く．
2. 問題文から選択肢とその前後の単語を用いてそれをパターンとし，このパターンを検索式とする．
3. 検索エンジンを用いて，生成された検索式に対するヒット件数を求める．
4. すべての選択肢について同様に検索式作成とヒット件数の抽出を行い，ヒット件数が最大となるものを解答候補として出力する．

検索式に対するヒット件数を求めるために，検索エンジンには最も代表的な Google [6] を用いた．Google では複数の語を含む検索式を与える場合，キーワードの間に “AND” を指定しなくても AND 検索を行うことができる．また，検索式をダブルクォート「”」で括ることでフレーズ検索を行うことができる．これは，”...” で囲まれた単語列が存在するページだけを抽出するものである．

検索エンジンに入力する検索式には，問題文中の選択肢とその前後の単語からなるパターンとした．本実験では，前後にある単語の数の違いにより 12 種類のパターンを設定し，それを検索式としてヒット件数を求める．これらのパターンの詳細を表 1 に示す．表 1 にある検索式の項目で，「(　)」は問題文にある空欄部分で，検索の際にはこの空欄に選択肢がそれぞれ入る．「_」は前後にある 1 単語を表している．

本実験では，これらの中からひとつのパターンに固定した状態で，4 つの選択肢に対する検索結果の比較を行う．これらの選択肢に対する検索結果の中で，最もヒット件数の多い選択肢を解答候補として提示する．このとき，解答候補が最終的な解答であると判定し，出力するために定めた 2 つの条件を以下に示す．

- 4 つの選択肢の中で最も多く検索されたものを解候補とする．
- 検索結果の中で同じ問題の解答，解説をしたページが検索されている可能性がある．そのため，ヒット件数が 10 個以下となる問題については，検索結果を人手により内容を調べて，同じ問題について書かれたページを不正とみなす．このような不正ページを検出して，ヒット件数から削除した．

このように，WWW 上のデータの中で同じ問題に対する解答，解説以外のページを対象として解答候補を決定する．

4　評価実験および考察

本節では，前節に述べた英語空欄補充問題の解答システムに対して評価実験を行い，実験結果とその結果に対する考察を述べる．

表 1: 検索式となる各パターン

	内容	検索式
パターン 1	選択肢のみ	(　)
パターン 2	選択肢と前 1 語	_ (　)
パターン 3	選択肢と後 1 語	(　) _
パターン 4	選択肢と前後 1 語ずつ	_ (　) _
パターン 5	選択肢と前 2 語	_ _ (　)
パターン 6	選択肢と後 2 語	(　) _ _
パターン 7	選択肢と前 2 語と後 1 語	_ _ (　) _
パターン 8	選択肢と前 1 語と後 2 語	_ (　) _ _
パターン 9	選択肢と前後 2 語ずつ	_ _ (　) _ _
パターン 10	選択肢と前 2 語と後 3 語	_ _ (　) _ _ _
パターン 11	選択肢と前 3 語と後 2 語	_ _ _ (　) _ _
パターン 12	選択肢と前後 3 語ずつ	_ _ _ (　) _ _ _

4.1　データ

本研究の評価実験で用いたデータは，1998 年度から 2003 年度までに大学入試センター試験の本試験，追試験における「英語」で出題された空欄補充問題を用いた．この空欄補充問題は各年度の本試験，追試験でそれぞれ 10〜14 問出題され，問題の総数は 138 問である．各問題には選択肢が 4 つ与えられ，その中から 1 つの正解を選択する．

4.2　実験結果

前節において述べた解答システムに対して，評価実験を行った．その結果，各パターンにおける正解候補数，不正数，最終正解数と正解率を表したものを表 2 に示す．

この表の中で，各行の意味は以下のようになる．

正解候補数 4 つの選択肢のうち，最もヒット件数の多い選択肢が正しい解答となっている問題の数

不正 解答候補の中でその問題の解答，解説をした Web ページを参照していた問題数 (ヒット件数が 10 件以下の場合にのみ不正かどうかを人手で判断)

最終正解数 解答候補数から不正を除いた最終的な正解数

正解率 全 138 問中で正解した問題の割合
(最終正解数/138(問)) × 100 (%)

表 2 において，正解候補数はパターン 7 からパターン 9 にかけて最も多くなり，正解率も最も高くなった．しかし，さらに前後の単語を加えると検索される Web ページが限定されてしまったために正解率が伸びなかった．また，不正となる問題に注目

表 2: 各パターンの解答候補数と正解率

検索パターン	1	2	3	4	5	6	7	8	9	10	11	12
正解候補数 (問)	41	56	51	70	64	67	77	77	76	69	64	62
不正数 (問)	0	1	1	2	4	4	6	9	17	23	25	31
最終正解数 (問)	41	55	50	68	60	63	71	68	59	46	39	31
正解率 (%)	30	40	36	49	43	46	51	49	43	33	28	22

すると，結果的にパターン 1 から 12 にかけて順次不正の数が増加した．これは，検索式に空欄前後の単語を追加することにより，同じ問題文が WWW から抽出しやすくなったためだと考えられる．

4.3 考察

検索パターンの正解率を全体的にみると，平均して 40% 程度の精度であった．このパターンの中で最も精度が高かったパターンは，選択肢とその前 2 語と後ろ 1 語から成るパターン 7 で 51% であった．逆に最も精度が低かったのが，選択肢とその前後 3 語ずつから成るパターン 12 で，22% とランダムに選択して正解する確率 25% を下回る結果となった．パターン 12 を用いて検索を行った結果，その問題の解答，解説をした英語学習サイトやメーリングリストのバックナンバーが抽出されたが，これらのページを不正として除くことで同じ用例がなくなったためだと考えられる．

選択肢のみを検索式としたパターン 1 は，1 語の名詞や前置詞など利用頻度の高い単語が多いため，それだけでは解答を特定する有力なパターンとはならなかった．たとえば，

“At first no one in class could find an answer, but finally David came up (　　) one”

1. for　2. on　3. to　4. with

(2001 年度本試験より抜粋)

という問題文について考える．この問題文に対する選択肢はすべて前置詞単語である．このとき，パターン 1 とパターン 7 における Google でのヒット件数を比較したものを表 3 に示す．パターン 1 で検索を行うとヒット件数は 12〜34 億件となり，非常に多くの文書が検索される．すべての選択肢についてこのような膨大なヒット件数が結果として返されたとしても，どの選択肢が解答であるかを判断することは非常に困難である．次に，選択肢だけではなく前後の単語を加えたパターン 7 で検索を行った結果，パターン 1 と比較して大幅にヒット件数が減少している．また，通常使われない単語の並びについてはヒット件数が非常に少なくなり，“came up with” といったよく使われるフレーズについては他の選択肢と比較して多くの文書で使われている．それぞれの選択肢で正解と不正解の間に有意差を見つけることで正解を導き出すことが可能となる．

表 3: 各パターンとヒット件数の比較

検索式	ヒット件数	解答
パターン 1		
1. for	2,050,000,000	
2. on	1,430,000,000	
3. to	3,440,000,000	誤
4. with	1,170,000,000	
パターン 7		
1. came up for one	170	
2. came up on one	1,100	
3. came up to one	876	
4. came up with one	14,500	正

もうひとつの例として，以下の問題文について考える．

“Are John and Mary still living in New York? No, they (　　) to Dallas.”

1. are just moved　2. had just moved

3. have just moved　4. will just move

(1998 年度追試験より抜粋)

この問題文に対する選択肢には “just move(d)” が存在するが，それが受動態や完了形などになっている．このとき，パターン 1 とパターン 12 における Google でのヒット件数を比較したものを表 4 に示す．この表において，パターン 12 の検索式に着目すると，8 単語を並べた単語列は単語数が非常に多いので，それにマッチする文書を検索することが非常

表 4: 各パターンとヒット件数の比較

検索式	ヒット件数
パターン 1	
1. are just moved	493
2. had just moved	57,500
3. have just moved	39,700
4. will just move	6,600
パターン 12	
1. York? No, they are just moved to Dallas	0
2. York? No, they had just moved to Dallas	0
3. York? No, they have just moved to Dallas	0
4. York? No, they will just move to Dallas	0

に難しくなっている．しかも，検索式の中に “York” や “Dallas” といった地名 (の一部) が含まれているが，これらの単語が Web データ中の一文に含まれることは非常に稀である．このような要因から，パターン 12 ではすべての選択肢についてヒット件数が 0 となり,「解答なし」という結果になってしまったと考えられる．

最初の例のように，一般的によく使われる前置詞などの機能語や動詞といった単語はさまざまな用例を使った数多くのページが検索されてしまう．そのため，その単語のみを用いて検索式としても，それにマッチするヒット件数がすべての選択肢で膨大な数となり，選択肢間での有意差は現れない．しかし，空欄前後の単語を多く含めてマッチングする単語数を増加させると，マッチする文書が WWW 上に見つけられなくなる．また，地名や名前 (“John” や “Mary” など) の固有名詞を含んだ検索式も Web ページを狭い範囲に限定させてしまい，用例を検索することが困難となる．これより，問題文から正解と不正解との間に有意差が現れる検索式のパターンを考慮することが非常に重要であることが分かる．

今回実験に用いた空欄の前後を含めたパターンは，イディオムや慣用句のように空欄の前後にヒントとなる単語が存在している場合には正しい結果を得ることができた．しかし，これらのパターンでは，接続詞を選択させる問題のような文と文，あるいは節と節の論理的な接続関係や時制の一致などの問題には対応できていない．そのため，これらの問題に対処する検索式を構成することが今後の課題となっている．

5 おわりに

本研究では，WWW 上の情報を手がかりとして英語の空欄補充問題を解くシステムの提案を行い，検索エンジンのみを手がかりとした場合での有効性を検証した．過去の大学入試センター試験において出題された問題に対して実験を行った結果，すべてのパターンに対する平均正解率は約 40% 程度であった．個々のパターンで見ると，最も正解率の高かったパターンは選択肢とその前 2 語と後ろ 1 語から成るパターンで 51% となった．空欄の前後 1〜2 単語を利用することにより，空欄のみや前後 3 語を加えたパターンと比較して，WWW から情報をより効率的に引き出せることが分かった．

しかし，本システムで生成する検索式では，イディオムや慣用句などには効果的であるが，接続関係や時制などの問題文には対応できないという問題点がある．そのため，空欄前後の語を加えるだけではなく，文節間の意味的なつながりや時制の一致などといった関係を考慮した検索式の表現方法を提案することが今後の課題となる．

参考文献

[1] 藤井 敦:“IT 技術者試験を対象とした質問応答システム －辞典情報に基づく用語問題の解法－”, 言語処理学会第 7 会年次大会発表論文集, pp. 514–517, 2001.

[2] 山田 一郎, 柴田 正啓, 金 淵培: “Ｗｅｂを情報源としたＱ＆Ａシステムの検討”, 言語処理学会第 9 会年次大会発表論文集, pp. 633–636, 2003.

[3] Googlefight: http://www.googlefight.com

[4] 山本 真人, 田中 久美子, 中川 裕志: “検索エンジンに基づく多言語用例指南ツール:Kiwi”, 言語処理学会第 9 会年次大会発表論文集, pp. 15–18, 2003.

[5] 外池 昌嗣, 佐藤 理史: “ウェブを用いて 4 択クイズを解く”, 言語処理学会第 9 会年次大会発表論文集, pp. 641–644, 2003.

[6] Google: http://www.google.com