IT技術者試験を対象とした質問応答システム
– 事典情報に基づく用語問題の解法 –

藤井 敦[†,††]　石川 徹也[†]
[†] 図書館情報大学
[††] 科学技術振興事業団 CREST
fujii@ulis.ac.jp

1 はじめに

情報通信技術の発展とコンピュータ利用者の増加を主な要因として，World Wide Web に存在するページの数は依然として増加の一途をたどっており，今や未曾有の情報源となりつつある．ある程度熟練した利用者ならば，書物やマスメディアなどの既存の媒体に頼らなくても，Web から取得した情報を駆使して日常の様々な問題解決に活用できるほどである．

このような現状は情報処理の研究にも強く影響し，事実，Web を対象にした情報検索や知識発見などの研究が数多くなされている．Web ページの多くはテキスト（書き言葉）情報なので，自然言語処理の研究においても多くの可能性を秘めた対象であることは間違いない．

著者らは Web に基づいて事典情報（用語説明）を抽出する手法を提案し，その結果，既存の事典が網羅していない情報も取得できるようになった [2, 11, 12]．こうして得られた事典情報は，人間が利用するだけでなく，計算機による自然言語処理にも様々な形で応用できる．

特に，自然言語理解に関する研究は知識の整備が高価なため，比較的小規模な対象に限定されていた．しかし，Web を一種の知識ベースとして利用すれば，研究を大規模に展開することが期待できる．

以上の背景を踏まえ，本研究では自然言語理解の一環として，Web から抽出した事典情報に基づく日本語の質問応答システムを実現した．質問応答に関する近年の研究では，統計的な検索手法や浅い言語解析に基づくシステムが主流であり [3, 5, 10]，本研究のような取り組みは稀である．

工学的な観点からシステムの評価や改善を行うためには，対象とする例題の選択が重要である．近年の自然言語処理や情報検索に関する研究は，コーパスやテストコレクションなどのベンチマークを素材として発展した経緯がある．しかし，質問応答に関するベンチマークの整備は比較的遅れている．米国では TREC において質問応答のテストコレクションが整備されたものの，日本語は対象になっていない [7]．また，本研究が目的とする Web に基づく質問応答は，TREC のように新聞記事を対象にした質問応答とは性質が異なる可能性がある．

本研究では，情報処理技術者試験（情報処理技術者試験センター）の「第2種」を評価用のベンチマークとして利用した[1]．第2種は，情報処理技術者を目指す者が初歩の段階で修得すべき基礎知識と応用能力を試す初級レベルの試験として位置付けられている．

実際には，基礎知識として情報処理の専門用語に関する問題が数多く出題される．これらの問題は事典情報に基づく質問応答の対象に適している．また，過去の試験問題と模範解答は各種書籍によって入手できるため，システムの性能評価を客観的かつ安価に行うことができる．

以下，2章で質問応答システムについて概説し，その中核となる事典情報の生成法について3章で詳説する．そして，4章で評価実験について説明する．

2 質問応答システム

2.1 解法の原理

本システムが対象とするのは，第2種試験から収集した用語問題である．全て4択式であり，ある用語に関して適切な記述を選択する問題と，問題文の記述に該当する用語を選択する問題に大別できる．それぞれを「記述選択」「用語選択」と呼ぶ．平成11年度秋の午前問題から抜粋した出題例を以下に示す [9]．なお，この試験では全80問のうち，記述選択22，用語選択18が出題され，用語問題が全体の半数を占めている．

記述選択の例

再帰的プログラムに関する記述として，適切なものはどれか．

ア 自分の中から自分を呼び出すことができる．
イ 主記憶の任意のアドレスに配置して実行できる．
ウ 複数のタスクから同時に呼び出されても正しく処理できる．
エ ロードし直さずに繰り返し実行できる．

用語選択の例

LANのアクセス方式のうち，複数の端末が同時に送信を行い，送信の衝突が起きる可能性のあるものはどれか．

ア ATM，イ CSMA/CD，ウ FDDI，エ トークンリング

[1] 平成13年度から制度が変更され，第2種に相当する試験は「基本情報技術者試験」に移行する．http://www.jitec.jipdec.or.jp/

人間の受験者は，記述選択に対しては，問題文の用語に関する説明を自分の知識から探し（思い出し），それに最も類似する記述を選択する．逆に，用語選択に対しては，まず選択肢ごとの説明を思い出し，問題文の記述に最も類似する説明に対応する用語を選択するだろう．

以上を踏まえて，本システムの解法原理を決定した．すなわち，記述選択，用語選択いずれの場合も事典情報（人間の知識に相当）を検索し，問題文/選択肢を用語説明に変換する．次に，問題文と各選択肢の類似度を計算し，類似度を最大化する選択肢を解答として出力する．

情報検索では，問い合わせ（query）とデータベース中の文書を索引語のベクトルとして表現し，両者の類似度を定量化する手法が提案されている．この手法を用語説明間の類似度計算に応用した．本システムでは，問題文と4つの選択肢が問い合わせと文書データベースにそれぞれ相当する．

2.2 システム構成

本研究で提案する質問応答システムの構成を図1に示す．本システムは大きく分けて，事典情報生成（右上）と問題解決（左下）で構成される．

問題文と選択肢が入力されると，システムは，いずれかに含まれる用語に対する事典情報を生成する．そして，問題解決によって問題文に最も類似する選択肢を出力する．記述/用語選択の区別は人間が明示的に与える．

事典情報生成の原理は著者らが提案した手法[2, 11, 12]に基づいている．要約すると，与えられた用語を含むページをWebから検索し，文章表現やHTMLレイアウトに関する規則に基づいて用語説明を抽出する．ページの検索には「Google[2]」を用いた．

今回は，抽出した用語説明の組織化手法を新たに導入した．抽出された用語説明の多くは互いに類似して冗長であり，また語義も区別されていない．そこで，語義などの一定の規範に基づいて用語説明を整理する．

しかし，用語ごとに質の良い語義分類を定義することは高価である．そこで，語義が分野に強く関連しているという仮説[8]を利用した．例えば「パイプライン」という用語は，コンピュータ分野では「処理方式」，建築分野では「輸送管」の意味で使われる．すなわち，一般的な分野の集合を定義しておけば，用語ごとに語義分類を定義する必要はない．

組織化手法は事典情報の質を高めるだけでなく，質問応答への応用においても重要である．すなわち「輸送管」としてのパイプラインに関する用語説明は，情報処理技術者試験を解く上では不要な情報（ノイズ）である．また，分野を区別できれば，情報処理以外の資格試験問題への応用も容易である．事典情報の生成については3章でさらに詳しく説明する．

問題解決では，問題文や選択肢に対応する用語説明間の類似度を情報検索の手法に基づいて計算する．原理的には任意の類似度計算法が適用できる．今回は，確率的手法の一つ[6]を用いた．そこで，形態素解析システム「茶筌」[15]を用いて，用語説明に含まれる内容語に関する頻度情報を抽出して利用する．

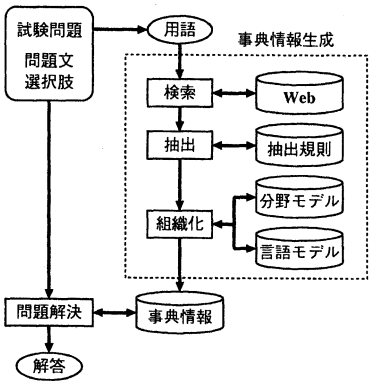

図 1: 質問応答システムの構成

3 事典情報生成

3.1 概要

図1に示すように，事典情報生成は「検索」「抽出」「組織化」の処理で構成される．本章では，以下，今回新たに導入した組織化の手法について説明する．

2.2節で議論したように，我々は用語説明を専門分野に対応付けることで間接的に語義を区別する．そこで，抽出処理によって得られた情報のうち，各分野（語義）に対して最適な用語説明を一つ（あるいは高々数件）選択し，最終的な事典情報を生成する．

ここで，対象となっている用語がどの分野に関連するのかがあらかじめ分かっていると仮定しよう．我々の目的は，それぞれの関連分野 c に対して最適な用語説明 d を選択することである．確率論的な観点からは，各 c に対して $P(d|c)$ を最大化する d を選択することに相当する．ベイズの定理によって式(1)が成り立つ．

$$P(d|c) = \frac{P(c|d) \cdot P(d)}{P(c)} \quad (1)$$

式(1)の右辺において，分母 $P(c)$ は対象となっている分野 c に関する定数なので，分子のみが組織化の中核である．$P(c|d)$ は用語説明 d が分野 c に関連する度合を定量化し，$P(d)$ は d が言語（用語説明）として妥当である度合を定量化する．両者をそれぞれ「分野モデル」

[2] http://www.google.com/

「言語モデル」と呼ぶ．言い替えれば，我々の組織化手法では，ある特定の分野との関連度が高く，かつそれ自身が用語説明らしい情報が最終結果として出力される．

実際の処理では，まず全ての専門分野に対して $P(d|c)$ を計算し，$P(d|c)$ の値がある閾値以上の用語説明だけを選択する．その結果，対象用語が関連する分野と適切な用語説明を同時に特定することができる（対象用語が関連する分野をあらかじめ知る必要はない）．

我々が以前提案した手法 [2, 11] では，言語らしくない情報を排除するために，言語モデルを一種のフィルタとして利用した．しかし，今回の枠組は統一的な確率モデルであり，理論的に分かりやすいという特長がある．

3.2 分野モデル

既存の文書分類法 [4] を利用し，$P(c|d)$ を式 (2) によって推定する．

$$P(c|d) = P(c) \cdot \sum_t \frac{P(t|c) \cdot P(t|d)}{P(t)} \qquad (2)$$

ここで，$P(t|d)$，$P(t|c)$，$P(t)$ はそれぞれ，d，c，分野全体における単語 t の出現確率である．$P(c)$ は定数として扱った．実際には，$P(t|d)$ は用語説明における単語の相対出現頻度として計算する．

$P(t|c)$，$P(t)$ を計算するためには，まず分野を定義し，それらに関する語の頻度分布を推定する必要がある．分野対応が付いた大量の文書は高価なので，機械翻訳用に作成された（株）ノヴァの専門用語辞書を利用した[3]．

この辞書は 19 の専門分野で構成され（表 1），合計約 100 万件の日英対訳を定義している．日本語項目からは「茶筌」を用いて単語を抽出した．日本語の用語説明は英単語を含むことがあるので，英語項目も併用して $P(t|c)$，$P(t)$ を計算した．辞書は実世界における頻度情報を含んでいない．しかし，分野固有の単語は複合語項目に繰り返し出現するため，本手法は妥当な近似である．

表 1: ノヴァ専門分野辞書の構成分野

航空・宇宙，バイオテクノロジー，ビジネス，化学，コンピュータ，土木・建築，防衛，地球環境，電気・電子，原子力・エネルギー，金融，法律，数学・物理，機械工学，医療・医学，金属，海洋・船舶，プラント，貿易

3.3 言語モデル

抽出処理がうまく機能しなかった場合や元の Web ページが装飾のための特殊記号，電子メールアドレスなどを含む場合は，言語的に無意味な情報が用語説明として抽出されることがある．言語モデルは，これら非言語的な情報を排除するために重要な役割を果たす．

[3] http://www.nova.co.jp/

多くの統計的言語処理と同じように，単語の Nグラムを用いて言語モデルを作成した．具体的には「茶筌」を用いて「CD-ROM 世界大百科事典」（約 8 万語収録）[14] を単語に分割し，CMU-Cambridge ツールキット [1] を用いてトライグラムを学習した．ここで，対象用語の表層的な違いに左右されないように，世界大百科事典の見出し語はあらかじめ共通の変数に置換した．

通常の Nグラムモデルでは，短い単語列ほど高い確率値が与えられる傾向がある．この傾向は，機械翻訳や音声認識のように，比較対象となる単語列がほぼ同じ長さである場合には問題にならない．しかし，本研究では，用語説明の長さは様々であり，質に拘わらずに短い用語説明が常に選択されやすくなる．この問題を回避するために，用語説明中の単語数によって $P(d)$ を正規化する．

4 評価実験

4.1 方法

平成 11 年度秋の「第 2 種情報処理技術者試験」午前問題から収集した 40 件の用語問題を対象に質問応答システムの性能を評価した．評価尺度として，被覆率（coverage）と正解率（accuracy）を用いた．被覆率は，正解・不正解に拘らず解答した問題の割合であり，正解率は解答した問題のうち正解した割合である．

用語問題の問題文や選択肢に含まれる用語は 96 語あった．それら全てに対して，Google は複数のページを検索し，ページ総数は 18,864,298（用語あたり平均 196,503）であった．しかし，実際には Google は上位 1,000 件のページ内容のみを取得するので，用語説明抽出の対象となったのは，用語あたり最大 1,000 ページであった．抽出処理の結果，96 語中の 85 語に対して少なくとも 1 件の用語説明が抽出された．

組織化処理において，$P(d|c)$ の値が 0.05 以上の用語説明を選択し，各分野ごとに $P(d|c)$ の値が大きい順に上位 3 件を最終的な用語説明として出力した．我々の手法は，高々上位 3 件まで取れば正しい用語説明が含まれることが経験的に分かっているためである．その結果，326 件の用語説明が残った．コンピュータ分野として出力された用語説明が最も多く，200 件（全体の 61%）あった．

さらに，システム性能の目安（ベースライン）を調べるために「英和コンピュータ用語大辞典」[13] を比較対象として利用した．この辞典はコンピュータ分野の専門用語を約 3 万語収録しており，用語問題に含まれる 96 語のうち，42 語に関する用語説明を記述していた．

用語説明の欠落によって解答できない問題に対しては，何も解答しない場合と，選択肢から無作為に 1 つを解答して被覆率を一律 100%にした場合の両方を評価した．

4.2 結果と考察

システム性能の評価結果を表2に示す．326件の用語説明を全て用いた場合も，コンピュータ分野に分類された200件だけを用いた場合も結果は同じだった．コンピュータ以外の分野に分類された用語説明は少数であり，システムの性能に影響しなかったと考えられる．

まず，無作為選択をしない場合，コンピュータ用語大辞典では被覆率が低く，正解率が比較的高かった．それに対して，Webに基づく用語説明を用いると，正解率が低下したものの，被覆率を大幅に向上させることができた．さらに両方の情報を併用することで，コンピュータ用語大辞典を用いた場合の正解率をほとんど保持したまま，被覆率を大幅に改善できた．

次に，解答できない問題に対して無作為選択を行った場合，コンピュータ用語大辞典とWebに基づく用語情報を用いた場合の正解率はほぼ同じであった．しかし，両方の情報を併用することで正解率が顕著に改善した．

以上の結果は，知識源に関する制約から拡張性に乏しかった従来の自然言語理解の研究をWebの利用によって大規模に展開できる可能性を示唆している．

表2: 質問応答システムの被覆率と正解率（%）

	無作為選択なし		無作為選択あり	
事典情報	被覆率	正解率	被覆率	正解率
コンピュータ用語大辞典	50.0	65.0	100	45.0
Webに基づく用語説明	92.5	48.6	100	46.9
併用	95.0	63.2	100	61.3

5 おわりに

本研究では，World Wide Webに基づいて自動生成した事典情報を用いた質問応答システムを提案した．また，事典情報の生成において，統計的な組織化モデルを提案した．情報処理技術者試験を用いた評価実験の結果，既存の事典情報を単独で用いた場合に比べてシステムの性能を向上させることができた．

本研究の成果はWebに基づく自然言語理解の研究における着実な一歩であり，今後も様々な観点から研究を発展させる予定である．

謝辞

専門用語辞書は（株）ノヴァの許諾を得て使用させて頂きました．電総研の伊藤克亘氏には本研究に対して貴重な御意見を頂きました．深謝致します．

参考文献

[1] Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of EuroSpeech'97*, pp. 2707–2710, 1997.

[2] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 488–495, 2000.

[3] Sanda M. Harabagiu, Marius A. Paşca, and Steven J. Maiorano. Experiments with open-domain textual question answering. In *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 292–298, 2000.

[4] Makoto Iwayama and Takenobu Tokunaga. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 162–167, 1994.

[5] Dan Moldovan and Sanda Harabagiu. The structure and performance of an open-domain question answering system. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 563–570, 2000.

[6] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[7] Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 200–207, 2000.

[8] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, 1995.

[9] 藤本喜弘（編）．第二種情報処理技術者過去問題&分析．経林書房，2000.

[10] 賀沢秀人，加藤恒昭．意味制約を用いた日本語質問応答システム．情報処理学会研究報告 2000-NL-140, pp. 173–180, 2000.

[11] 藤井敦，石川徹也．World Wide Webを利用した百科事典的知識の収集法．人工知能学会第48回知識ベースシステム研究会資料 SIG-KBS-A001, pp. 31–36, 2000.

[12] 藤井敦，石川徹也．用語説明抽出に基づくWeb文書の事典的利用．言語処理学会第6回年次大会発表論文集，pp. 296–299, 2000.

[13] 日外アソシエーツ．英和コンピュータ用語大辞典 第2版，1996.

[14] 日立デジタル平凡社．CD-ROM世界大百科事典プロフェッショナル版，1998.

[15] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，浅原正幸．日本語形態素解析システム『茶筌』version 2.0 使用説明書．Technical Report NAIST-IS-TR99012, 奈良先端科学技術大学院大学，1999.