

能動的質問生成を用いた 対話メディアによる知識の獲得および提供

伊藤 慎吾† 荒木 健治†

†北海道大学 大学院 情報科学研究科

本稿では、Web上の質問文から学習を行い、能動的にユーザに質問することで知識の獲得・提供を行うシステムを提案する。現在知識を共有するシステムとして「Wikipedia」や「Yahoo!知恵袋」「教えて!goo」などが多くのユーザに使用されている。提案手法では、これらの知識共有において能動的に質問を発する対話メディアを用いることで、より多くのユーザに手軽に知識共有に参加してもらうことを目指す。実験の結果、ユーザフィードバックによる質問精度の向上が確認された。また、印象評価実験を行うことでユーザがシステムに対して好印象を持ち、特に「積極性」を感じていることが確認された。さらに、ユーザの質問と獲得された知識を柔軟にマッチングし、知識提供できることを示した。

Acquisition and Offer of Knowledge Using Talk Media with Active Question Generation

Shingo ITO† Kenji ARAKI†

†Graduate School of Information Science and Technology Hokkaido University

We propose the system that acquires and offers knowledge by learning from question on the Web and generating question actively. Recently, many users use "Wikipedia", "Yahoo! chiebukuro" and "Oshiete! goo" as knowledge sharing system. We aim that more users participate in knowledge sharing easily by using talk media with active question generation. As a result of experiments, we confirmed an improvement of the precision of the question generation by the feedback from users. In addition, we confirmed that the users had a good impression for our proposed system and felt aggressiveness in particular by an impression evaluation experiment. Furthermore, we confirmed that our proposed system offered knowledge by matching knowledge for the users' question.

1. はじめに

近年 Web の発達とともに、多くの人間の知識を共有し有効に利用しようという研究が行われるようになってきた。

実際に稼働しているものとしては、Web 上であらゆるものを網羅する百科事典を構築するという目的のもと運用されている「Wikipedia」[1]が有名である。また、ユーザの質問に対してユーザが答えるという形態を取っている「Yahoo!知恵袋」[2]「教えて!goo」[3]なども近年ユーザ数を大きく増やしている。どちらも人手でデータを入力する(質問に答える)ため、精度が高く、共有されている知識を閲覧するだけなら手軽というようなメリットがある。しかしながら、前者は新たな知識を登録する際の手間や百科事典にあるような知識のみ登録されていること、また最近では誰でも書き込めることによる情報の不整合の増加が問題になっている。後者も人間が相手のため気軽に質問できないことや知識が得られるまでのタイムラグなど解決すべき問題点は多い。

一方、人間から全ての知識を得るのではなく、すでに Web 上に存在している大量のデータから知識を獲得しようという試みも数多く行われてきた[4, 5, 6]。しかし、現段階では汎用的な知識獲得は難しいため、評価表現など特定の指標に基づいた知識獲得がほとんどであり、実用段階に至っているものは少ない。

このように、全てを人手で行う、全てをコンピュータが自動で行うという2つの手法はどちらも一長一短であり、2手法の中間位置に属するような半自動の手法が近年少しずつ見られるようになってきた。その一つとしてあげられるのが、エージェントを介させた知識共有システムである。久保田らによって研究された「知識カードを用いた分身エージェント」[7]では、ユーザがQAストーリーを作成し、それらを用いてエージェント同士で知識のやりとりを行う。また、鳥澤の研究[8]では、ユーザの発したキーワードより属性情報、準備方法、利用法などを Web から獲得し、組み合わせで質問生成を試みている。これらは、ユーザから知識を得るため、ユーザが正しい情報を入力するのであれば、従来の知識獲得システムのネックであった知識の精度の問題は解決することができる。特に鳥澤の手法では、対話において能動的に質問することでシステム主導で知識の獲得を行うことができる。知識獲得における対話の意味については小林らによって考察され[9]その重要性が報告されており、対話を活性化させるシステムからの能動的質問は新たな知識を獲得する際にたいへん有用であると考えられる。

本研究では属性情報ではなく質問文自体に着目し、Web上に存在する実際に多数の人間が書いた質問文から学習することで、よりバリエーションに富んだ質問の生成・知識の獲得を目的としている。また、

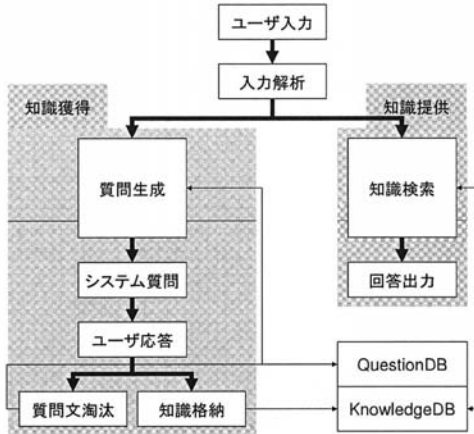


図 1 システム概要図

鳥澤が最初に大量の Web 文書から質問生成を行う方法をとっているのに対し、本システムは、事前に蓄えた質問文の他に、ユーザとの対話によって徐々に質問を生成・淘汰していく方式を取っている。これは、ユーザからのフィードバックを受けることによって、最終的に精度の高い質問をすることが可能になるためである。フィードバックのための評価がユーザの負担となる恐れもあるが、自動生成される質問文全てを管理者側が評価するのは労力・時間の観点からほぼ不可能であるため、実際にシステムを公開するためには必要不可欠な要素であると考えられる。さらに対話メディアとして用いるにあたって、獲得した知識を柔軟に提供できるように知識提供部の開発を行った。

2. システム概要

本研究で提案するシステムを図 1 に示す。システムは、大きく分けて知識獲得・知識提供の 2 つからなり、それぞれ以下の処理部から構築される。システムの入力解析等で用いられている形態素解析には MeCab [10] を使用している。

本システムの動作例を図 2 に示す。ここでシステムは、「熊」に対する上位概念を確認した後、その上位概念に適した質問（「強いせいでしょうか」「小説はありますか」）をしている。しかし、ユーザは「強いせいでしょうか」は「熊」にたいする質問として不適切であると評価し、「小説はありますか」に対してだけ情報を入力している。

2.1 知識獲得

知識獲得は以下の 5 つの処理部から構成される。上位概念・下位概念をそれぞれ取得する際には表 1 のような構文パターンを用いる。DB は MySQL [11] を使用している。

質問生成部 上位概念に対する生成質問が QuestionDB に格納されていない場合は、Web 上

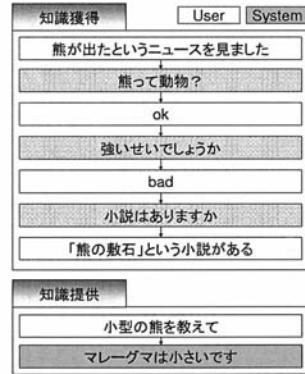


図 2 動作例

の質問文を学習し質問を生成する。

システム質問部 生成質問群より過去の質問状況などを考慮し選択されたものを、システム質問として出力する。

ユーザ応答部 ユーザは質問に対する答え、または質問文の評価を入力する。

質問文淘汰部 ユーザが誤りと判断した質問文を淘汰する。

知識格納部 ユーザ応答より獲得された知識を KnowledgeDB に格納する。

2.2 知識提供

知識提供は以下の 2 つの処理部から構成される。

知識検索部 ユーザの入力した質問文に適した知識を KnowledgeDB に格納されたデータより検索する。

回答出力部 知識検索部で得られた知識をユーザに提示する。

3. 処理過程

3.1 質問生成

知識獲得部は Web 上の質問文から学習を行った能動的質問生成を軸としている。

3.1.1 下位概念取得

ユーザ入力中の下位概念と同じ上位概念を持つ下位概念を自動的に収集した。下位概念の自動取得に関しては、様々な過去研究 [12, 13, 14] が存在するが、今回は更新が容易であり、ユーザが使用しながらも学習し続けることができるという点に着目し、安藤らの「定型表現を利用した下位概念の自動取得 [12]」を用いて行った。安藤らの研究によると、定型表現を利用して、上位概念から下位概念を自動取得できる。定型表現の種類は表 1 の 7 つが最も信頼性が高いと述べられており、実際「動物」などの下位に一般名詞を多く含むものに対しては、6 割以上の正解率で下位概念が取得できたと報告されてい

表 1 定型表現

定型表現
N1などのN2
N1などN2
N1というN2
N1のようなN2
N1に似たN2
N1以外のN2
N1と呼ばれるN2

N1:下位概念
N2:上位概念

る。本研究では、安藤らが用いた新聞記事コーパスの代わりに Web を用いた。

3.1.2 質問文収集

取得した下位概念を用いて Web より質問文の収集を行った。質問文収集には、通常の検索サイトを用いるより高い効率で大量の質問文を集めるために質問サイトである「Yahoo!知恵袋」「教えて!goo」を用いた。下位概念をクエリとして検索したページから、パターンマッチを利用して、質問文を収集する。パターンマッチに使用したのは、「*(下位概念)*(文末表現)」(*はワイルドカード)のような文構造である。文末表現は「ですか」「ますか」「でしょうか」「ませんか」「教えて」「?」を使用した。

3.1.3 下位概念に対する質問テンプレートの生成

収集した下位概念に対する質問文に汎用性を持たせるために構文解析を行い、「文末表現にかかる文節」+「文末表現」のセットを作成し、これを質問テンプレート L とした。構文解析には CaboCha [15] を用いた。さらにこの段階で、下位概念にあたる名詞が含まれていた場合「@」という変数に置き換え、下位概念を入れ替えることで様々な質問を生成できるようにした。

3.1.4 スコア計算

質問テンプレート L に含まれている名詞は、その名詞が質問テンプレート L 全体でよく使われていればいるほど、下位概念と共起していることとなる。下位概念と共起している名詞を含む質問テンプレート L は、よりの確な質問であると考えられる。そこで、どの質問テンプレート L が有用であるかを推定する尺度として、共通の上位概念の質問テンプレート L の集合に対するある質問テンプレート L 全体での名詞の出現頻度を使用した。用いた名詞は「固有名詞」と「普通名詞」で、代名詞などは、上位概念・下位概念に関係なく出現するため除いた。今回は最も単純なものとして、その質問テンプレート L に出てくる名詞の出現頻度の合計をテンプレート中の単語数で割ったものを ScoreT とした。式 (1) に ScoreT の定義式を示す。

$$\text{ScoreT} = \frac{\text{名詞の出現頻度の合計}}{\text{総単語数}} \quad (1)$$

3.1.5 淘汰処理

作成した質問テンプレート L には、「店を教えて」と「お店を教えて」のように、字面上類似した質問が多く存在する。このような質問は、どちらをユーザに尋ねても得られる答えは同じであるため、淘汰する必要がある。そこで、質問テンプレート L 間の類似性を判断するために、類似度 α を定義した。 α の定義式を式 (2) に示す。質問テンプレート L1 と L2 を比較した場合の類似度 α は、L1 の総単語数 < L2 の総単語数 という条件のとき、L1 と L2 の単語で一致した数を L1 の総単語数で割ることで求められる。

この類似度の閾値は、予備実験により、最適値 0.6 と求められた。類似度が閾値を越えた場合、単語数の少ない質問テンプレート L は削除される。質問テンプレートの汎用性を考えると単語数の多い方を削除した方が良いが、その場合文として成立しないほど短い質問文が残されてしまう可能性があるためである。

$$\alpha = \frac{L1 \text{ と } L2 \text{ で一致した単語数}}{\text{総単語数 (L1)}} \quad (2)$$

3.1.6 上位概念に対する質問テンプレートの生成

10 の質問テンプレート L の集合から、上位概念に対する質問テンプレート H の集合を生成する。この際、異なる下位概念における質問テンプレート L には類似した質問が存在している。そこで、ここでも類似度 α と同様に類似度 β を定義する。 β の定義式を式 (3) に示す。この際、類似度 α と分母の総単語数の対象が変更されているのは、質問テンプレート L を生成した際に極端に短い質問文は削除されているので、今回は汎用性を重視し類似度の定義を行う必要があるためである。

$$\beta = \frac{L1 \text{ と } L2 \text{ で一致した単語数}}{\text{総単語数 (L2)}} \quad (3)$$

3.1.7 検索エンジン等による淘汰処理

最後に質問テンプレート H を、検索エンジンの Hit 数、質問文の長さ、質問のユニーク度、質問に格助詞が含まれているか、質問に名詞が含まれているか等の観点から評価した。評価するためのスコアを ScoreS とし、この定義式を式 (4) に示す。

$$\text{ScoreS} = a_x \cdot (1 + b_x) \cdot (c_x + (d_x + e_x) \cdot f_x) \quad (4)$$

ここで、 a_x は検索エンジン (Yahoo [16]) での Hit 数、 b_x は正規化された質問文の長さ、 c_x は質問のユニーク度、 d_x は格助詞の有無 (1 か 0)、 e_x は名詞の有無 (1 か 0)、 f_x は重みを表している。質問のユニーク度は、QuestionDB の全質問文に同一質問が

どれだけ含まれているかによって算出している。また、重みは予備実験より 0.4 とした。

ScoreS が閾値を下回った場合、質問テンプレート H は QuestionDB より削除される。閾値は予備実験より、1.5 とした。

3.2 知識検索

知識検索部は知識獲得部でリレーショナル DB に知識とともに格納された質問とユーザ質問を比較して、字面上で類似した質問と組になっている知識をユーザに提供する。知識が格納されていない場合は質問応答システム [17] に受け渡す。

3.2.1 上位概念取得部

ユーザの入力文から抽出した単語の上位概念 H を Web を用いて調べ、確認する。

3.2.2 質問マッチング部

取得した上位概念に対する質問文の中で、ユーザの入力文に類似した文を探す。

3.2.3 類似度チェック

KnowledgeDB に類似した質問文があった場合は、その質問文に対応して格納されている知識を回答とする。この際、MySQL に搭載された全文検索機能を利用して字面上の類似質問の検索を行っている。これは、速度向上のための措置である。しかしながら、MySQL ではたとえ類似度が低くても強制的に一番近い質問文を出すために、全く関係ない知識がユーザに提供されてしまう可能性が高い。そこで、MySQL で一番類似しているかと判断された質問とユーザ質問との類似度をシステムがもう一度計算し、閾値を超えた場合のみ出力するようにした。

3.2.4 質問応答システムへの受け渡し

格納されている知識がなかった場合は質問文を質問応答システム [17] に受け渡し、その出力結果を回答とする。なお、質問応答システムの知識源は Web とした。

3.3 対話メディア

本システムは能動的に作られた質問を対話メディア上で利用することによって、ユーザから知識を獲得、さらに獲得された知識を必要としているユーザに提供することを目的としている。そこで、実際に生成された質問文を利用できるような対話メディアを構築した。対話メディアは CGI を用いて作成した。

4. 知識獲得評価実験

本稿で提案した対話メディアの有効性を確認するための実験を行った。この際、ユーザ評価による生成質問の種類と収束速度をあげるため、明示的にテーマを設定した。11名の被験者 (A~K) に順番にシステムを使用してもらい、ユーザフィードバ

クによる知識の獲得率の変化等について評価した。システムは前の被験者の対話からの学習結果を引継ぎ、次の被験者との対話を行う。被験者の内訳は、理系男子大学院生 10 名 (B 以外)・理系女子大学院生 1 名 (B) である。

4.1 実験方法

被験者はそれぞれ、対話メディアと「魚」のテーマで 20 ターン、「花」のテーマで 20 ターン、計 40 ターンの対話を行った。ここで 1 ターンは「システム質問」「ユーザ応答」を組としたもの、あるいは「ユーザ質問」「システム回答」を組としたものである。対話は被験者の発話に対してシステムが上位概念を尋ね、上位概念に適した質問をする。その際、質問が文法的・意味的に誤っていた場合、被験者は質問を不正解と評価し、それ以外は質問に対する答えを入力してもらった。

4.2 予備実験

初期状態ではユーザの入力がないため、構文パターン (表 1) を用いた手法により Web 上で上位概念として出現する単語を抽出したものをキーとして、事前に質問文を生成しておく。これにより、対話メディアとして用いたときの反応が素早くなる。実際に 1 つの構文パターンに対して 1,000 ページの Snippet を使い、合計 7,000 ページから頻度の大きい順に 3,000 件の上位概念を取得した。しかしながら、出現頻度が小さい上位概念はユーザとの対話で使用される確率も低い。そこで IRC チャットログの 100 文を入力とし、文中に含まれる単語からシステムが選択したもの 100 個に対して、事前学習する上位概念数との関係を調べた。結果を図 3 に示す。質問可能率は、全てのユーザ入力に対してシステムが質問可能かどうかを判断したものである。質問可能率の定義式を式 (5) に示す。

$$\text{質問可能率 [\%]} = \frac{\text{システム質問可能数}}{\text{ユーザ入力数}} \times 100 \quad (5)$$

なお、ログは IRC サーバ (irc.huie.hokudai.ac.jp) チャンネル (baka) における雑談を 2004 年より収集した 57.7MB のテキストファイルから抽出した。上位概念 1,000 個学習時までは質問可能率が指数関数的に増加しているが、それ以降は 6 割程度で飽和状態になってきている。つまり、Web 上から収集した上位概念は、対話中の質問で使用する上位概念のおよそ 6 割をカバーできることを示している。この結果より、事前に 1,000 件の上位概念に対する質問文を生成し QuestionDB に格納した。

4.3 実験結果

対話の結果、表 2 のような知識が得られた。知識獲得効率の遷移を図 4 に示す。これは被験者 A から被験者 K まで順に実験してもらった際、ユーザフィードバックによる質問精度の向上により、ど

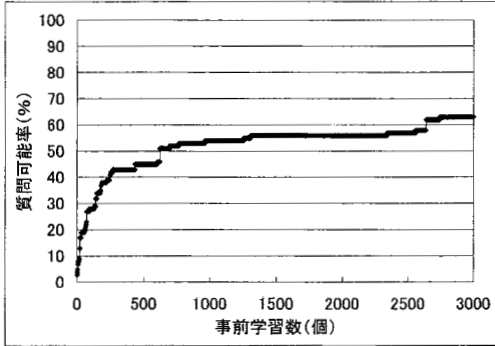


図3 事前学習による質問可能率の推移

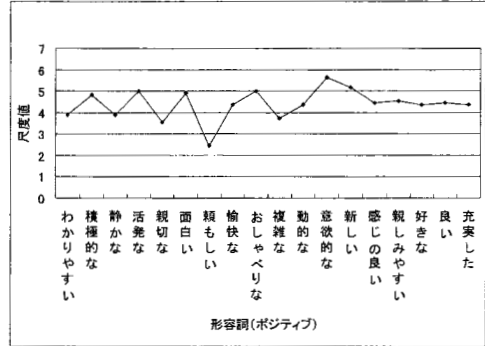


図5 イメージプロフィール

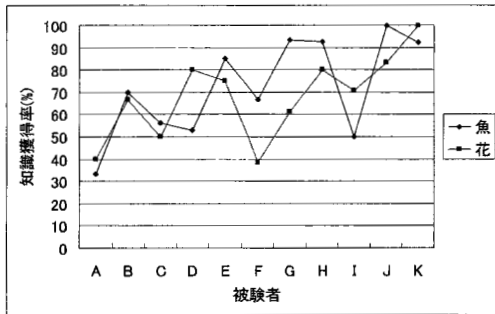


図4 知識獲得率の推移

れだけ知識獲得率が上がっているかを示している。なお、知識獲得率は式(6)のように定義した。

$$\text{知識獲得率} [\%] = \frac{\text{知識獲得数}}{\text{システム質問数}} \times 100 \quad (6)$$

5. 印象評価実験

5.1 実験方法

SD法を用いて印象評価実験を行い、このシステムへの興味やシステム自体の積極性などを評価した。被験者のアンケートは、よく使用される形容詞対18組[18][19]から構成した。

5.2 実験結果

各形容詞対の平均値をイメージプロフィールとしたものを図5に示す。

6. 考察

まず、獲得された知識について考察する。獲得知識の中には、「お店を知りませんか」などの客観的な知識の他、「お勧めを教えてください」等のユーザの嗜好のような主観的な知識も得られている。主観的な知識は

利用が難しいが、統計的に判断することで「ユーザの商品に対する評価」など有用な知識にもなり得る。また、Yes-No質問は鳥澤の手法では生成されていないため、当初の目的であったバリエーションに富んだ質問の生成・知識の獲得の一端が実現できた。

次に、ユーザ評価による質問文の淘汰、及びそれによる知識獲得率の推移について考察する。図4のように、最初の被験者は4割程度のシステム質問に対して知識を入力し、残り6割については質問文が不相当であると判断している。また、質問の正誤判定に個人差があるため、被験者によっては知識獲得率が急に下がってしまう現象も起こっている。しかしながら、実験を進めていくうちにフィードバックによって質問文の精度が向上し、最後の被験者は9割以上のシステム質問に対して知識を入力した。つまり、より多くのユーザが使用することで質問文は洗練され、知識を効率的に獲得できるようになったと評価できる。

さらに、システムの印象評価について考察する。18組の形容詞対の中で特に高得点であるのが「積極的な」「面白い」「活発な」「おしゃべりな」「意欲的な」「新しい」で、全て5.0点程度である。このうち、「積極的な」「活発な」「おしゃべりな」はシステムがユーザに対し能動的にはたらきかけていることが評価されたものであると考えられる。これら以外の形容詞対もほぼ中間値である4.0点を超えており概ね本システムに対してユーザは好印象を持ったことが確認された。「親切的な」「頼もしい」が低評価であることについては、本システムのGUIが未熟であったことが原因である可能性が高い。したがって、知識共有を目的として能動的に質問をする対話メディアはユーザの興味をひき、生成質問の学習元をWeb上の質問文とすることでユーザの満足度を高めることが期待できる。また、システムがユーザに好印象を与えたことから、ユーザ評価による質問文淘汰は非常に有効な手段であると考えられる。

最後に、知識の提供について考察する。実験が小規模なため、獲得された知識がただちに使用されることはなかった。しかし、実際に「鮭のお勧めはありますか」「鮭でお勧めはありますか」等の質問

表 2 獲得知識例

獲得知識	質問	下位概念
時知らずがおいしいらしいよ	お勧めはありますか	鮭
北海道でもたくさん水揚げされます	どこの地域ですか	秋刀魚
網焼きや刺身	調理法を教えてください	鮭
富良野にもありましたよ	スポットってありますか	ユリ
ホームセンターに売ってるよ	お店をしりませんか	朝顔
もう散りました	もう咲いていますか	桜

文を入力すると「時知らずがおいしいらしいよ」という回答を出力したため、ユーザ数が増えれば提供部がより活用されるようになると考えられる。

7. まとめ

本研究では、知識共有を目的とした対話メディアの提案を行った。能動的質問生成をシステムの軸とした知識獲得部と、質問とともに格納された知識に対して柔軟にマッチングし情報の提供を行う知識提供部からなるシステムを構築した。知識提供部は当研究室で開発された質問応答システム [17] と組み合わせることで、獲得された知識が少ない状況でも知識提供を行うことを可能にした。実験では、被験者に順にシステムを使用してもらうことでユーザフィードバックによる質問精度の向上とそれに伴う知識獲得の効率化を確認した。さらに、印象評価実験を行うことでシステムの能動的質問と質問文からの学習が有効であることが示された。知識提供部においても動作が確認された。

今後は、実験規模を拡大することで知識提供部がどの程度使用されるのか確かめる予定である。また、スコア計算部、淘汰処理部などの個々の精度向上やGUIの改善を行うことで、生成質問の充実とユーザ負担の軽減を図りたい。

参考文献

[1] Wikipedia, <http://ja.wikipedia.org/wiki/>
 [2] Yahoo! 知恵袋, <http://chiebukuro.yahoo.co.jp/>
 [3] 教えて! goo, <http://oshiete.goo.ne.jp/>
 [4] 奥村 学, 南野 朋之, 藤木 稔明, 鈴木 泰裕: “blog ページの自動収集と監視に基づくテキストマイニング”, 第6回人工知能学会セミナー Web とオントロジー研究会, 2004-A401-01, pp.1-8 (2004).
 [5] 工藤 拓, 山本 薫, 坪井 祐太, 松本 裕治: “言語情報を利用したテキストマイニング”, 情報処理学会研究報告, 2000-NL-148, pp.65-72 (2000).

[6] 有村 博紀: “データマイニング: ウェブデータからの知識発見を目指して”, 電子情報通信学会技術研究報告, IT2000-32, pp.1-6 (2000).
 [7] 久保田 秀和, 黒橋 禎夫, 西田 豊明: “知識カードを用いた分身エージェント”, 電子情報通信学会論文誌, pp.600-607 (2003).
 [8] 鳥澤 健太郎: “一般ユーザにインタビューする対話エージェント”, 情報処理学会研究報告, 2007-NL-180, pp.25-30 (2007).
 [9] 小林 宏教, 河崎 善司郎, 徐 天晟, 瀧田 啓司, 田島 正登: “知識獲得のための対話方式に関する一考察”, 電子情報通信学会論文誌, pp.901-905 (2000).
 [10] MeCab, <http://mecab.sourceforge.net/>
 [11] MySQL, <http://www.mysql.com/>
 [12] 安藤 まや, 関根 聡: “上位語下位語を含む連体修飾表現の言語的分析”, 言語処理学会第10回年次大会発表論文集, pp.205-208 (2004).
 [13] 新里 圭司, 鳥澤 健太郎: “HTML 文書からの単語間の上位下位関係の自動獲得”, 情報処理学会研究報告, 2003-NL-158, pp.95-102 (2003).
 [14] 芳鐘 冬樹, 野澤 孝之, 辻 慶太, 影浦 映: “ウェブからの関連語・下位語の収集手法の検討と検索システムへの応用”, 第52回日本図書館情報学会研究大会発表要綱, pp.113-116 (2004).
 [15] CaboCha, <http://chasen.org/taku/software/cabocha/>
 [16] Yahoo, <http://yahoo.co.jp/>
 [17] 亀山 恵祐, 荒木 健治: “質問応答システムにおける知識源選択規則の自動獲得の有効性について”, 情報処理学会研究報告, 2007-NL-178, pp.85-90 (2007).
 [18] 末永 俊郎: “社会心理学研究入門”, 東京大学出版会 (1987).
 [19] 井上 正明, 小林 利宣: “日本におけるSD法による研究分野とその形容詞対尺度構成の概観”, 教育心理学研究, Vol.33, No.3, pp.253-260 (1985).