

係り受け関係に基づくグラフ構造を用いた質問応答システム

倉田 岳人[†] 岡崎 直觀[†] 石塚 満[†]

† 東京大学大学院 情報理工学系研究科 〒113-8656 東京都文京区本郷 7-3-1

E-mail: †{kurata,okazaki,ishizuka}@miv.t.u-tokyo.ac.jp

あらまし 近年、質問応答システムに関する研究は注目を浴びている。質問応答とは、自然言語で与えられた質問文に対して、大規模な電子テキストなどの知識源から適切な解答を探しだし、提示する技術である。質問応答を実現する場合、多くの解答候補に対して、適切な順位付けを行うことが必要となる。しかし、この処理に関して優れたアルゴリズムはいまだ確立されていない。我々は、係り受け解析に基づくグラフ構造を用いることにより、従来手法より高精度のシステムを構築することができたので、ここに報告する。

キーワード 質問応答システム、係り受け解析、グラフ理論、Dijkstra のアルゴリズム、NTCIR

Question Answering System with Graph Structure from Dependency Analysis

Gakuto KURATA[†], Naoaki OKAZAKI[†], and Mitsuru ISHIZUKA[†]

† Graduate School of Information Science and Technology, University of Tokyo Hongo 7-3-1,
Bunkyo-ku, Tokyo, 113-8656 Japan

E-mail: †{kurata,okazaki,ishizuka}@miv.t.u-tokyo.ac.jp

Abstract Question Answering(QA) is the hot research topic. In the task of QA, queries are written in natural language. Then, the system returns the correct answer from newspaper articles. Ranking answers is very difficult in QA task, and there have been no sophisticated algorithm yet. Our method using Graph Structure from Dependency Analysis is superior to former approaches.

Key words Question Answering System, Dependency Analysis, Guraph Theory, Dijkstra's Algorithm, NTCIR

1. はじめに

近年、計算機性能の向上や様々な電子化された文書の整備により、自然言語処理に関する研究が盛んに行われている。質問応答とは、自然言語で与えられた質問文に対して大量文書中から適切な解答を導き出す技術であり、TREC[1] や NTCIR[2] などの評価型ワークショップも開催され、近年注目されている。

本報告では、まず 2 において質問応答システムの概要と将来性に関して述べる。次に、3 において従来から行われている研究に関して簡単にまとめる。その後、今回提案した手法の概要を 4 で述べ、それに基づき構築した質問応答システムの評価を 5 で行う。最後にまとめと今後の課題を 6 に示す。

2. 質問応答とは

質問応答とは、自然言語で表現された質問に対して、電子化された大量の文書から適切な解答を導き出し、提示する技術である。以下に、質問応答の将来性、要素技術などについて述べる。

2.1 背景、将来性

現在、大量の電子化された文書が存在する。これらの中から必要な情報を適切に取り出すことに関しては様々な研究が行われてきた。情報検索はこれらの検索の中でもっと基本的で重要なものであると言える。ここで一般的な情報検索と質問応答システムとの比較を表 1 に示した。

表 1 質問応答と情報検索の比較

	質問応答	情報検索
入力	自然言語で記述された質問文	検索語
出力	質問文に対する解答	検索語に関連する文書集合

Google などで適切な情報を得ようとする場合、検索語の選択にコツがいるが、質問応答ではユーザはより直感的に利用することができる。また出力に関して、情報検索と質問応答で利用場面が異なる場合も考えられるが、明確な解答を求めている場合、関連する文書から解答そのものを探し出す手間を省くことができる。

また、今後訪れるユビキタス社会を考えてみると、一般的な

ユーザが音声入力で様々なシステムに対して質問を行う、という場面が容易に想像できる。このような中で、質問応答システムには多くの利用方法があると考えられる。

2.2 質問応答を実現するための要素技術

質問応答は從来から研究されている様々な自然言語処理の技術の集合と捉えられ、また上記したような実用性、将来性もあることから、近年盛んに研究されている。以下に、必要となる要素技術の代表的なものを列挙する。

- ・日本語形態素解析
- ・日本語係り受け解析
- ・日本語全文検索システム
- ・固有表現抽出

3. 日本語質問応答に関する従来手法

3.1 質問応答の流れ

日本語質問応答を実現するための一般的な流れを図1に示す。

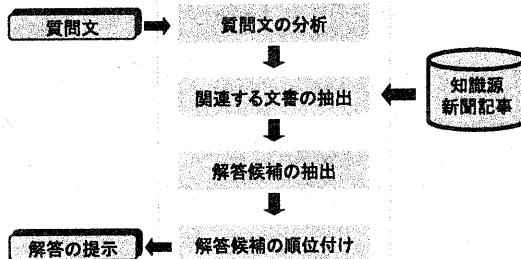


図1 質問応答システムの一般的な流れ

以下に各段階での処理に関して簡単にまとめる。

3.1.1 質問文の分析

自然言語で与えられた質問文を解析する。検索語が与えられる情報検索とは異なり、質問応答では知識源から関連する記事を検索するために適切な検索語を選び出す必要がある。また、質問文から解答に関する情報を出来るだけ多く抽出しないといけない。このように質問文の解析は非常に重要な処理である。

3.1.2 関連する文書の抽出

質問文の解析結果に基づき、知識源から質問文に対する解答を含むと思われる文章を抽出する。ここでは従来の情報検索の技術を用いることができる。

3.1.3 解答候補の抽出

抽出された文章の中から、解答の候補を抜き出す。ここでは、どのような規則で解答候補を抽出するか、ということが非常に重要な要素となる。従来から研究されている固有表現抽出や複合語抽出の技術をここで用いることができる。

3.1.4 解答候補の順位付け

抽出された解答候補の中で、どの解答候補が正しいかを順位付けする。質問応答を実現する上で最も難易度の高い段階はこの順位付けである。

3.2 関連研究

従来行われていた質問応答に関する研究では、主に Answer

Selection の部分を改良することがよく行われていた。その中のいくつかを簡単にまとめる。

3.2.1 木構造の類似度に基づく順位付け

テキストの構文的類似度を求める尺度がいくつか提案されている[3]。高橋等は質問文と解答候補を含む文との類似度に従って質問応答を実現する手法を提案している[4]。しかし、これらの手法は計算量が膨大になり、また高い精度を得るには至っていない。

3.2.2 単語の属性に従ったルールを大量に記述する順位付け

Leeらは、単語の属性などを Lexico-Semantic Pattern という形で大量に記述する手法を提案している[5]。この手法は非常に精度の高い結果を残しているが、質問文などを非常に多くのパターンに分類し手間のかかる手法であると言える。

3.2.3 解答候補と検索語の距離に基づく順位付け

検索語と解答は近い位置に現れる、ということを前提とし、解答候補と検索語の距離に基づいて順位付けを行う手法がいくつか提案されている。福本らは、解答候補と検索語の位置関係に基づいた手法を提案している[6]。また単語距離に着目した手法もいくつか提案されている。ここで前提として与えられている、「検索語と解答の位置は近い」ということは非常に重要と考えられるが、単語間距離に着目した手法で高精度の結果が得られている手法はない。

3.3 従来手法の問題点

前述したように、単語間距離に基づく順位付けは非常に有効であると考えられる。しかし、単語間距離を用いて、高精度な結果を残しているシステムは未だに構築されていない。これは、単純な単語間距離を用いると、余計な文節などが間にに入ったりすることがよく起こり、解答候補と検索語の距離が大きくなる場合があることが原因と考えられる。

4. 提案手法の概要

本報告では、上述したような単純な単語間距離に基づく手法の問題点を克服できるように、グラフ構造を用いた距離尺度を導入する。グラフ構造を導入することにより、単純な単語間距離の問題点以外にも、質問応答を実現するための様々な問題点に対処することができる。これらについては6においてまとめることとする。本節では、本報告で提案するグラフ構造を用いた日本語質問応答システムについて説明する。図2に、提案手法の概要を示した。

以下に、処理の流れを示す。

4.1 新聞記事の索引付け

質問応答システムでは、大量のテキストを知識源として扱う。今回実験で用いた NTCIR-3 の QAC1 の場合、新聞記事2年分を扱った。このような大量の記事を扱うために、事前にどの単語がどの文書に含まれるかを調べ、索引付けを行った。今回は、各文書の各文ごとに索引付けを行った。索引付けには“Namazu”的 mknmz コマンドを用いた。また、索引付けを行う際に必要となる日本語の分かち書きには、辞書の整合性を保つために、“茶筌”を用いた。

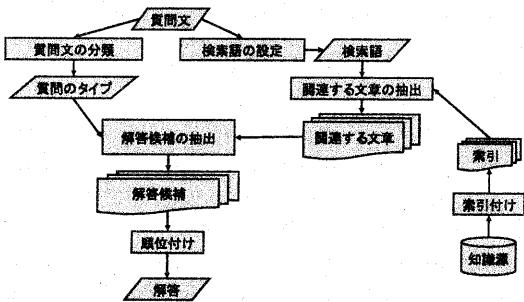


図 2 構築したシステムの概要

なお、新聞一年分のデータを文ごとに切り分け、“茶筌”で解析し、索引付けをするために要した時間は、CPU：Pentium4 2.8GHz、メモリ：1GB のマシンを用いて、およそ一日程度であった。

4.2 質問文の分析

4.2.1 検索語の設定

自然言語で記述された質問文から、索引付けされた新聞記事を検索するためのキーワードを選択する。検索語の選択手順の概要を以下に示す。

- (1) 質問文を“茶筌”で解析する。
- (2) “茶筌”での解析結果から得られる最初の検索語セット $K = (k_1, k_2, \dots, k_n)$ で検索を行う。
- (3) 検索語が多すぎて “Namazu” が検索結果を返さない場合、 K から検索語を減らした検索語セット K' で再び検索を行う。
- (4) 文章が得られるまで、(3) の操作を繰り返す。
- (5) 検索語セット K'' がキーワードを全く含まなくなつた時点で、質問に対する解答を発見できなかったことをユーザに提示する。

ここで、(2) でどのような検索語セットを作成するか、(3) で検索語を減らしていくアルゴリズムなどは、本論文の範疇を逸脱するため、ここでは述べないこととする。

4.2.2 推定される解答の形に基づく質問文の分類

質問文を幾つのタイプに分類するかに関しては様々な手法が提案されている。NTT が構築している質問応答システム SAIQA [7] では 30 種類程度に、SAIQA-2 では 100 種類程度に分類を行っている。しかし、質問文に対する分類の数を多くすることは、それだけ手動でのルールの設定が多く必要になり、また、システムの肥大化も招く。そのため、本報告では、4 種類のタイプに分類することとした。以下に各々の概要を示す。

Type 1 このタイプには以下の様な質問が属する。

- * 木星は何個の衛星を持っていますか。
- * 江戸幕府は何年続きましたか。

このタイプの質問は、「何 + 単位」もしくは「何 + 接尾語」という形での質問を行っている。その結果、解答となる表現の単位、もしくは接尾語が限定されることとなり、解答の候補の数が非常に少なくなる。

Type 2 このタイプには以下の様な質問が属する。

- * 日本人で初めて大リーガーになったのは誰ですか。
- * スペイン村は三重県のどこにありますか。

このタイプの質問文は、疑問詞が解答の属性を指示している。例えば、「誰」という質問に対する解答は「人名」である、ということがわかる。このタイプの質問に関しては上述した固有表現抽出を用いることができる。今回は、以下の表 2 示した疑問視についてこのタイプに分類することとした。

表 2 Interrogative Words in Type 2

疑問詞	解答
だれ、誰	人名
どこ、何処	地名、会社名、機関名
いつ、何時	時間、日付

Type 3 このタイプには以下の様な質問が属する。

- * 東京湾アクアラインの全長はどのくらいですか。
- * リニアモーターカーの走行試験で出た最高速はどのくらいでしたか。

このタイプに属する質問は、「どのくらいですか」のような表現で、解答として数字表現を求めている。つまり解答の候補としては、数字表現を抽出すればよい、ということとなる。ここで、数値表現に対する単位がわからないことが、Type 1 との違いとして挙げられる。村田らは数値表現に対する単位を自動推定する方法を提案している [8]。我々のシステムにおいても、この方法と類似の方法で単位の推定を実装したが、この点に関しては詳しく触れないこととする。

Type 4 このタイプには以下の様な質問が属する。

- * マカオはポルトガル語でどのように表しますか。
- * DVD とは何のことを指しますか。

このタイプに属する質問は、疑問詞や「何 + 接尾語」のような形で解答の属性や表現に関する情報を明示的には示していない。つまり解答の候補の絞り込みを行うことができず、解答の候補の数が非常に大きくなり、正解を提示することが困難となる。

以上の 4.2.1, 4.2.2 の処理を行い、質問文を図 3 に示すようなファイル形式に変更する。

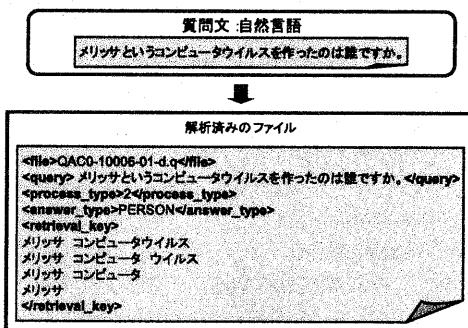


図 3 質問文の解析結果を記述するファイル

作成したシステムでは、上記したようなファイルを入力として受け取り、解答を提示する。

4.3 コーパスからの記事の抽出

提案手法では、質問に対する解答は、コーパス中の検索語を含む文、もしくはその近傍に存在するということを仮定する。そのため、解答を含む可能性のある文として、質問文中の検索語を含む文をまずははじめに抽出した。そして、検索語を含む文に解答の候補が含まれない場合は、その文の前後の文を抽出することとした。ただし、前後の文を抽出する場合、代名詞の問題が生じる。なお、「Namazu」で検索を行った場合、tfidf に従って順位付けを行って出力が行われるため、大量の文が検索結果として返された場合はより順位の高い文を選択した。このコーパスからの文の検索の流れを図 4 に示す。

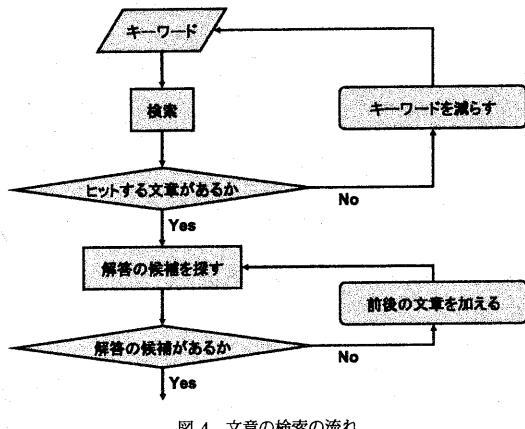


図 4 文章の検索の流れ

4.4 解答候補の抽出

4.2.2 で推定された解答の形に基づき、4.3 で抽出された文から解答の候補を抽出する。以下に、各タイプ毎の解答候補の抽出方法を示す。

Type 1 質問文から解答候補の形は「XXX 時間」などのように推定されている。よって、抽出された文から、「XXX 時間」のような形でパターンマッチングを行い、解答の候補を抽出する。

Type 2 質問文から解答がどのような固有表現であるかがわかっている。よって、4.3 で抽出された文に対して固有表現抽出を行い、質問文から推定された属性を持つ単語を解答候補として抽出する。

Type 3 質問文から解答として数字表現を求めていることがわかっている。よって、4.3 で抽出された文から、数字表現を抽出し、それらを解答候補とする。

Type 4 質問文から解答に関する情報は得られない。そのため、4.3 で抽出された文の中の名詞、未知語のうちで、tfidf 値上位 20 個を解答候補とした。

4.5 解答候補の順位付け

4.4 で抽出された解答候補に対して、どの解答候補が最も解答らしいかという点に関して順位付けを行う。提案手法では以下の様にして順位付けを行った。

(1) 4.4 で抽出された文に対して係り受け解析を行う。今回は係り受け解析に CaboCha [9] を用いた。

(2) 複数の文から得られた文節間の係り受け関係に従い、各文節をノードとするグラフ構造を作成する。

(3) グラフ内で、質問文から抜き出された検索語を含むノードに関しては、検索語とその他に分割する。具体的には、キーワードに「発明」があり、グラフ中に「発明品」というノードがあれば、「発明 → 品」という形にする。

(4) 係り受け関係から作成されたグラフは有向グラフであるが、これらをすべて無向グラフにする。このようにして作成されたグラフ構造の例を次頁の図 5 に示した。

(5) ノード間のリンク数に従って、隣接するノード間のコストを定める。ここで隣接するノード A, B 間のコスト $Cost(A, B)$ は式 1 に従って定めた。

$$Cost(A, B) = 1/(N_{link(A, B)})^2 \quad (1)$$

ただし、 $N_{link(A, B)}$ はノード A, B 間のリンク数とする。

(6) Dijkstra のアルゴリズム^(注1)に従い、解答候補と検索語の最短距離を算出する。そして、ある解答候補とすべての検索語との距離の和を、その解答候補のスコアとし、そのスコアに従って順位付けを行った。

$$Score(Candidate) =$$

$$\sum_{All\ keywords} Distance(Candidate, Keyword) \quad (2)$$

ここで $Candidate$ は特定の解答候補、 $Keyword$ は検索語を表し、ノード X, Y の最短距離 $Distance(X, Y)$ はダイクストラのアルゴリズムにより、式 3 の様に定められる。

$$Distance(X, Y) = \min \sum Cost \quad (3)$$

5. 評価実験とその評価

我々の研究室では、NTCIR-4 QAC2 に参加している。しかし、まだ NTCIR-4 QAC2 のデータは利用不可能であるため^(注2)、昨年度行われた NTCIR-3 QAC1 のデータセットを用いて、提案したシステムに対する評価を行った。以下の表 3 に実験の条件を示した。

5.1 評価方法

表 3 に示した様に、今回は Task 1 の条件に従って評価を行った。ここで、Task 1 の評価方法について簡単に述べる。

表 3 評価実験の条件

知識源	毎日新聞 98 年、99 年
質問セット	NTCIR3 QAC1 Task1 200 問

(注1) : Dijkstra のアルゴリズムはグラフ理論の基本的なアルゴリズムであり、閉じたグラフ内のある始点から他の点までの最短距離を求めることができる。

(注2) : QAC のタスクでは、結果報告の一週間程度前に質問文のデータが公開される。

Q. テニスの全仏オープン女子シングルスで3年ぶりの優勝を果たしたのは誰ですか。

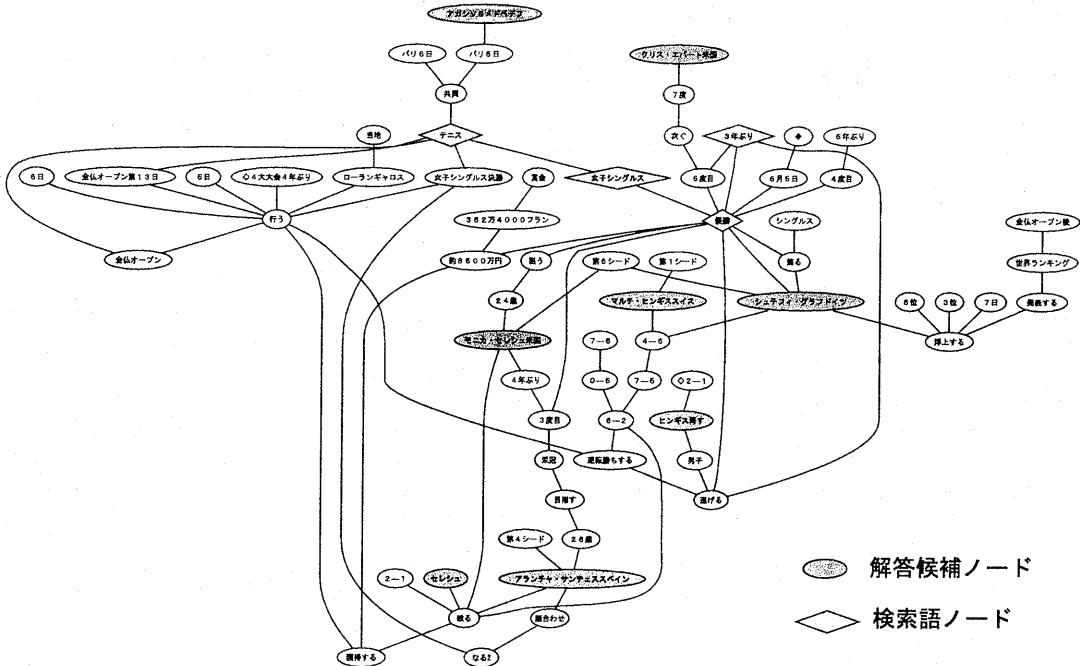


図 5 作成されたグラフ構造の例

Task 1 では、システムは一つの質問に対して、順位を付けて 5 個の解答を返す。ここで、正解を返した最も上位の順位の逆数 RR をその設問の得点とする。そしてその平均値 MRR をシステムの評価とする。

$$MRR = \frac{\sum_{i=1}^n RR_i}{n}$$

$$RR_i = \frac{1}{Rank}$$

5.2 実験結果

次頁の図 4 に、4.2.2 で分類したタイプ別の MRR 、およびすべての設問での MRR を示す。また、比較のために、単純な頻度に基づく順位付けを行った結果も示す。

5.3 実験結果に対する考察

表 4 から、Type 1 に属する設問に対する MRR が非常に高いことがわかる。これは 5.2.2 で説明したように、解答の候補を抽出する時点では、設問が求める形の解答候補が少なかったことが原因と考えられる。

それに対して、Type 4 に関しては、“Namazu”による検索で抽出された文中のすべての名詞、複合名詞を解答の候補としたため、非常に低い MRR しか得ることが出来なかった。また、非常に多くの一般語が解答として提示される、という結果となった。

システム全体の評価としては、表 4 に示したように、実際に提案手法を用いることにより、単純な頻度や tfidf を用いて順位をつけた場合よりも高い MRR を得ることが出来たので、提案手法の有効性を示すことができたと考えている。さらに、

NTCIR-3 QAC1 において MRR が 0.4 を越えるようなシステムはあまりなく、我々の提案するシステムの優位性を示すことができたと考えている。

また、頻度のみの場合でも NTCIR-3 QAC1 において提案されている多くの手法と比肩する MRR が得られている。これは、質問文の分類手法が適切であることを示唆していると言えることができる。

5.4 情報要求の観点からの考察

情報要求とはユーザがある目的を達成するために現在持っている知識では不十分であると感じている状態のことを指す。情報要求に関して Taylor は以下に示す 4 段階を提案している [10]。

Q_1 : Visceral Need 現状に満足していないことは認識しているが、それを具体的に言語化してうまく説明できない状態

Q_2 : Conscious Need 頭の中では問題を意識できるが、あいまいな表現やまとまりのない表現でしか言語化できない状態

Q_3 : Formalized Need 問題を具体的な言語表現で言語化することができる状態

Q_4 : Compromised Need 問題を解決するために必要な情報の情報源が同定できるくらい問題が具体化された状態

現在の情報検索システムは Q_4 の段階にあるユーザを対象としている。それに対して「質問応答システムに質問文を入力する」ということは、より高次の Q_5 などという表現で表される行動に相当すると考えられる。しかし、Type 4 での処理では質問文からキーワードを抜き出して利用することしかせず、情報検索と同じような処理と言わざるを得ない。質問応答システ

表4 評価実験の結果

	Type 1	Type 2	Type 3	Type 4	Total
分類された質問の数	22	94	16	68	200
提案手法で得られた MRR	0.635	0.526	0.625	0.205	0.427
頻度のみを用いて得られた MRR	0.543	0.350	0.250	0.223	0.305

ムを利用するユーザはより高次の情報要求を持っている、という観点からも、より多くの情報を質問文から得られるようなシステム構成にすることが必要である、ということができる。

6.まとめと今後の課題

本報告では、現在構築している質問応答システムの概要について説明し、実際に評価実験を行い、我々のシステムの優位性を示した。以下に、グラフ構造を用いるための前提と、グラフ構造を導入することにより得られた利点を示し、最後に今後の課題を示す。

6.1 グラフ構造を導入するための前提

質問応答を実現するために、語と意味のマッピングについて考えた場合、以下の二つの場合が問題となる。

(1) 一つの意味に多数の単語がマッピングされている場合

(2) 多くの意味が一つの単語にマッピングされている場合

(1) は類義語、同義語の問題。(2) はストップワード、多義語の問題、といふことができる。

(1) に関してはシソーラスを用いることが解決策として挙げられる。しかし、シソーラスを検索で用いることは容易であるが、グラフ構造の中で用いることは難しい。

(2) に関して、ストップワードは一つの単語に多くの意味がマッピングされている典型的例といふことができる^(注3)。そのため、グラフのなかで各々のストップワードが一つのノードにまとまってしまわないようにしなければならない。その他の多義語に関しては、検索を行った時点で一つの閉じたドメインが構成されている、と考えると、この閉じたドメインでは、多義語の多義の中の一つの意味に、その多義語がマッピングされている、と考えることができる。この前提によりグラフ構造を導入することができる、といふことができる。

6.2 グラフ構造を用いることの利点

質問応答を行うにあたり、問題となる点がいくつかあるが、提案手法においてグラフ構造を用いることにより、これらの問題をどのように吸収しているかを以下に示す。

6.2.1 言い換え表現の吸収

質問文と知識源となる文書で同じことを異なる表現で行っている場合がある。提案手法では、グラフ構造の縮退により、以下の様な言い換え表現を吸収できる。

- エジソンはフィラメントランプを発明した。
- フィラメントランプはエジソンの発明品だ。

また有向グラフを無向グラフにすることで、以下のような能動態、受動態の言い換えを吸収できる。

(注3)：厳密に言うとストップワードに与えられているのは意味ではなく機能である

- エジソンはフィラメントランプを発明した。
- フィラメントランプはエジソンによって発明された。

6.2.2 指示代名詞

知識源となる文書がいくつかに別れている場合、指示代名詞が問題となる。グラフ構造でリンクを張ることにより、これらの問題を解消できる場合がある。ただし、厳密な対応を行なう場合には他の処理を加える必要がある。

6.3 今後の課題

今後、NTCIR-4 QAC2 の Formal Run に向けて改良を行おうと考えている点、および今後の課題を列挙する。

知識源からの文の抽出 今回は“Namazu”的仕様となっている tfidif に従って文章を抽出したが、今後はベクトル空間モデルを利用した文章の抽出を考えている。これによりキーワードの選択を行う必要がなくなり、より効率のよい検索を行うことができると考えている。

抽出の単位 今回の報告では、知識源を文単位で索引付け、検索を行った。しかし、グラフ構造の利点を考慮し、指示代名詞に対して頑健なシステムの構築を考えた場合、パラグラフ単位、もしくは文書単位での検索を用いた方がよいと考えられる。今後、これらのシステムを構築し、比較検討を行う予定である。

質問文のさらなる解析 実験の結果からも明らかであり、5.4 でも述べたが、より多くの情報を質問文から得ることが質問応答の精度向上につながることは明らかである。今後は、より多くの情報を質問文から得られるような質問の分類方法、分析方法を検討する予定である。

文 献

- [1] TREC. <http://trec.nist.gov/>.
- [2] NTCIR. <http://research.nii.ac.jp/ntcir/>.
- [3] M.Collins N.Duffy. “Convolution kernels for natural language”. *Neural Information Processing Systems*, 2001.
- [4] INUI Kentaro TAKAHASHI Tetsuro, NAWATA Kozo. “Applying Structural Matching and Paraphrasing”. *Proceedings of the Third NTCIR Workshop*, 2003.
- [5] Gary Guenbae Lee Seungwoo Lee. “SiteQ/J: A Question Answering System for Japanese”. *Proceedings of the Third NTCIR Workshop*, 2003.
- [6] NIWA Tatsuhiro FUKUMOTO Jun'ich, ENDO Tetsuya. “RitsQA: Ritsumeikan question answering system used for QAC-1”. *Proceedings of the Third NTCIR Workshop*, 2003.
- [7] 佐々木裕、磯崎秀樹、平博順、平尾努、質沢秀人、鈴木潤、国領弘治、前田英作. “SAIQA: 大量文書に基づく質問応答システム”. 情報学基礎研究会, No. No.064-12, 2001.
- [8] ISAHARA Hitoshi MURATA Masaki, UTIYAMA Masao. “A Question-Answering System Using Unit Estimation and Probabilistic Near-Terms IR”. *Proceedings of the Third NTCIR Workshop*, 2003.
- [9] “Cabocha”. <http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/>.
- [10] Robert S. Taylor. “question-negotiation and information seeking in libraries”. *College & Research Libraries 1968*, pp. 178–194, 1968.