論文検索のための知識の論文の序論からの抽出

佐波 智也　日高 宏紀　渡辺 靖彦　岡田 至弘
龍谷大学 理工学部 情報メディア学科
{t_saba,h_hidaka}@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

1　はじめに

近年，膨大な量の論文が電子化され，インターネットを介して簡単に入手できる環境が整いつつある．ここで重要になるのは，膨大な量の論文の中からユーザが求める情報を含むものを探し出す論文検索の方法である．論文検索の方法として，タイトルや著者名，雑誌名等を対象にしたキーワードによるものが多い．しかし，キーワードで表現できる内容には限界があるので，自然な文によって論文を検索できることがのぞましい．一方，論文検索の結果は論文そのものか，ユーザの質問内容を考慮しない論文の要約であることが多い．検索結果はユーザの質問内容を反映したもの，できれば答えそのものであることがのぞましい．したがって，自然な文での質問をゆるし，答えそのものをユーザに与える論文検索のための質問応答システムを作成することは重要である．そこで本研究では，論文検索のための質問応答システムを実現するために必要な知識を獲得する方法について述べる．

読んだことがない論文に対して，その本論で述べられている内容 [1] について具体的で詳細な質問をするのはむずかしい．そのような場合、ユーザの質問は序論で述べられている内容 [1] を問うものになるのではないかと考えた．そこで、本研究では質問応答のための知識を論文の序論から取り出すことを試みる。論文の序論を調査すると，以下の 5 種類の情報が表現されていた．

目的　論文の目的を含む文
問題点　研究の問題点が記述されている文
背景　研究テーマの背景が記述されている文
手法　研究や実験の手法が記述されている文
必要条件　問題点を解決する方法が記述されている文

さらに，これら情報を表現する文には，その文頭や文末に典型的な表現があった。そこで本研究ではこの 5 種類の情報を表現する文を表層表現を手がかりにして抽出する。本研究の調査および実験には，言語処理学会第 12 回年次大会の予稿集に掲載されている論文を用いた．

2　論文検索に用いる知識

2.1　ユーザの質問に対する調査

先行研究を調べようとしているユーザーにとって、読んだことのない論文についてその内容を具体的に詳しく質問することはむずかしい。したがって、ユーザーの質問は，以下に示すように，論文の序論で述べられている内容を問うものになるのではないかと考えた。

1. 目的
   その研究で明らかにされたことについての質問.
   (例) 自然言語文質問応答システムでどんな知識源を用いたものが報告されていますか
2. 問題点
   既存の技術で不十分なことについての質問.
   (例) 電子化されたさまざまな情報が即時に得られるようになったために，どのような問題がおこりますか
3. 背景
   その研究に関係するこれまでの動向や先行研究についての質問．その研究をしようと思うにいたった経緯，理由についての質問.
   (例) 自然言語処理の分野で何が研究の中心となっていますか
4. 手法
   技術を実現するために用いた手法についての質問.
   (例) 類似文を検索する手法は考案されていますか
5. 必要条件
   解決方法や何かを行うために必要になるようなことについての質問.
   (例) 音声対話システムを実用的に使うにはどのようなことが必要になりますか

2.2　論文序論で表現されている情報

2.1 節で示した 5 種類の質問の答えとなる情報、

目的　目的を含む文
問題点　問題点を述べている文
背景　研究テーマの背景について述べられている文
手法　実験の手法などについて述べられている文
必要条件　必要条件について述べている文

が、論文の序論でどのように表現されているのか調査した。すると、これらの情報を表現する文には、文頭および文末に典型的な表現が含まれていた。以下にそれぞれの情報ごとに例文と典型的な表現を示す.

1. 目的
   - 本稿では〜
     (例) 本稿では，雑談において，どのようなやりとりが対話を継続するのに有効かを明らかにする.
   - 本研究では〜
     (例) 本研究では知識源として web 掲示板を用

いる自然言語文質問応答システムを構築する事を目標とする．

2. 問題点
   - ～困難である
     (例) しかし日本語の場合は，境界を示す明確な手がかりがないため，英語の場合と比べタイプ B の名詞列から境界を識別することは困難である．
   - ～問題となっている
     (例) 近年の計算機性能の向上と，世界規模のネットワークの拡大により，電子化されたさまざまな情報が，即時に得られるようになったが，同時に，ユーザに対する情報過負荷が問題となっている．
3. 背景
   - 近年の～
     (例) 近年の自然言語理解は，意味解析から照応解決や省略処理などの文脈処理が着目されている．
   - 現在の～
     (例) 現在の e-learning システムは選択問題やリスニング問題などある程度多様な問題形式を扱うことができる．
4. 手法
   - ～行っている
     (例) 具体的には，人手で作成した専門分野コーパス・一般コーパスの間での用語の出現頻度の比を用いて用語の専門判定を行っている．
   - ～試みる
     (例) ここでは例として，医学分野の Web 文書集合からそのデータ中の語の出現状況を用いて医学用語間の階層構造を抽出することを試みる．
5. 必要条件
   - ～必要とする
     (例) 多くの自然言語処理システムでは，単語間の意味的な距離を測る処理を，本質的に必要とする．
   - ～不可欠である
     (例) しかし，音声対話システムが実用的に使われるためには，現実の使用条件下でのユーザのふるまいを知ることが不可欠である．

3　典型的な表現を用いた知識抽出

図 1 に、論文の序論から取り出した知識を用いた質問応答システムの概要を示す。本研究で論文の序論から取り出した情報を、質問応答システムの知識として利用する。

論文の序論から取り出す情報は、2 章で示した 5 種類 (目的，問題点，背景，手法，必要条件) である。2 章で述べたように、これらの情報を表現する文には典型的な表現がある。そこでそれらの表現を手がかりにして 5 種類 (目的，問題点，背景，手法，必要条件) の情報を表現する文

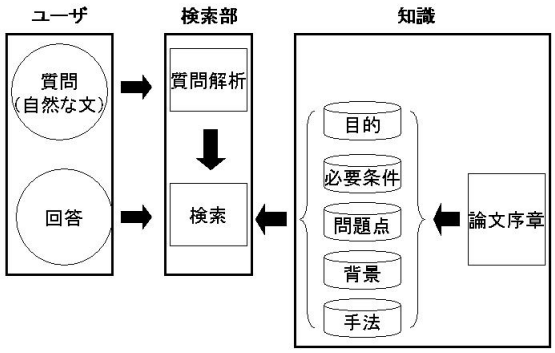

図 1　論文の序論から抽出した知識を用いた質問応答システム

表 1　抽出結果

	正解	再現率	適合率
目的	264	78%	43%
問題点	108	80%	45%
背景	101	45%	23%
手法	296	30%	39%
必要条件	173	63%	68%
合計	1131	57%	43%

を論文の序論から取り出す。この抽出方法を以下に示す。

step 1　論文の序論を形態素解析する。形態素解析には JUMAN[2] を用いた。

step 2　文抽出の手がかりにする表現 (手がかり表現) を用いて、step 1 の結果から 5 種類 (目的，問題点，背景，手法，必要条件) の情報を表現する文を取り出す。本研究で抽出に用いる手がかり表現を図 2 に示す。

step 3　抽出した文に、抽出に利用した手がかり表現に与えられている意味ラベルを与える。1 つの文に対し、複数の異なる意味ラベルが与えられることもある。

4　実験結果と検討

4.1　論文の序論からの情報抽出

人手によって作成した正解と，システムの抽出結果を表 1 に示す．適合率を低下させる原因の 1 つが指示詞である。本研究では、抽出された文が指示詞を含む場合、その指示対象が取り出した文だけでわからなければ、誤って抽出したものと判定した。また、背景と手法は、問題点や必要条件に比べ，再現率も適合率も良くない．この原因として考えられることは、

1. 背景と手法は、問題点や必要条件に比べ、さまざまな表現を用いて説明されている。

1 目的
- 本稿では〜
- 本実験では〜
- 本論文では〜
- 本報告では〜
- 本調査では〜
- 本手法では〜
- 本研究では〜
- 本研究は〜
- そこで〜
- 目的は〜
- 我々は〜
- 〜について報告する。
- 〜検討する。
- 〜調べる。
- 〜検証する。
- 〜観察する。
- 〜取り上げる。
- 〜まとめる。
- 〜考察する。
- 〜推定する。
- 〜目指す。
- 〜目的である。
- 〜目的としている。
- 〜目的とする。

2 問題点
- 既存の〜
- 従来の〜
- 欠点として〜
- 〜困難である
- 〜絶望的である
- 〜容易ではない
- 〜難しい
- 〜問題となっている
- 〜場合がある
- 〜課題がある
- 〜問題点がある
- 〜欠点がある

3 背景
- 近年〜
- 既存の〜
- 現在〜
- 〜重要視されている
- 〜期待されている。
- 〜研究されている
- 〜研究が行われている
- 〜研究されてきた
- 〜活発になっている
- 〜活発になってきている
- 〜注目されている
- 〜注目を集めている

4 手法
- そこで〜
- 本手法では〜
- 本システムは〜
- 具体的には〜
- 〜提案する
- 〜比較した。
- 〜作成した。
- 〜調査した
- 〜考案した
- 〜考案されている
- 〜行った。
- 〜行う。
- 〜試みる。
- 〜アプローチをとった

5 必用条件
- 〜必要とされている
- 〜必要になる
- 〜必要とする
- 〜必要がある。
- 〜必要となる
- 〜必要である。
- 〜重要である
- 〜不可欠である
- 〜不可欠となる
- 〜欠かせない
- 〜求められる。
- 〜求められてきている
- 〜しなければならない
- 〜できれば〜
- 〜望まれている。
- 〜要求されている。
- 〜要求が高まっている
- 〜要求が高まってきている

図 2　5 つの情報 (目的，問題点，背景，手法，必要条件) を表現する文を取り出すための手がかり表現

2. 手法を表現する文が複数ある場合が多い。例えば，「まず」「次に」「最後に」などではじまる複数の文から手法の内容が表現されていることがよくある。そのような場合は本来，それらの文すべてを取り出さなくてはいけない。

4.2　種類別の抽出結果についての検討

4.2.1　目的を表現する文の抽出結果についての検討

目的を表現する文の抽出は、再現率は 78% とかなり高いが、適合率は 43% とかなり低い。これは、図 2 で示した手がかり表現によって、目的を表現する文はかなりよく取り出せたが、

- 目的を表現する文でも、指示詞の指示対象が取り出した文だけではわからない文
- 目的以外の情報を表現する文

も数多く取り出していることを示している。特に、方法を表現する文を誤って取り出すことが多かった。例えば以下の例は「本稿では〜」という手がかり表現を含んでいるが、目的を表現する文ではない。この文は目的を表現する文として取り出されるが、誤りと判定される。

(例) 本稿では，これらをまとめて「分類スコア」と

呼ぶ.

一方、以下の例は「〜目的として〜」という表現を含む、目的を表現する文である。しかし、図 2 で示した手がかり表現では、この例文を取り出せない。文末にくる「〜目的とする。」という表現は手がかり表現にしたが、文中の「〜目的として〜」という表現は手がかり表現にはしなかった。これは、文末にくる「〜目的とする。」に比べて、文中の「〜目的として〜」は、その文が目的を表現していると判断することの手がかりとして弱いと考えたからである。

(例) そこで今回国立がんセンター (NCC-CIS) の Web データを元に提供されているすべてのがん（計 54 種類）について用語辞書を作成し，がん情報を必要とする患者のために「がん」に関する文章で用いられる言語的特徴を明らかにすることを 目的として 検討した.

4.2.2 問題点を表現する文の抽出結果と検討

問題点を含む文の抽出も、目的の時と同様に，再現率は 80% とかなり高いが、適合率は 45% とかなり低い。これは、図 2 で示した手がかり表現によって、問題点を表現する文はかなりよく取り出せたが、

- 問題点を表現する文でも、指示詞の指示対象が取り出した文だけではわからない文
- 問題点以外の情報を表現する文

も数多く取り出していることを示している。例えば以下の例では、「下記」が指示する対象がこの文からではわからない。このため、誤りと判定される。

(例) しかし，人が判断する際にも下記のように難しい事例が多くある.

4.2.3 背景を表現する文の抽出結果と検討

背景の情報を表現している文には、図 2 で示した手がかり表現を含まないものが多かった。例えば、以下の文は背景の情報を表現しているが、図 2 で示した手がかり表現を含まない。

(例) 実テキストに対しても，頑健に統語解析を行う技術が発展し，情報抽出や情報検索，機械翻訳等に応用されている.

また、図 2 で示した手がかり表現だけで背景の情報を表現している文であるかどうか、判定するのがむずかしい場合もあった。例えば以下の例では「既存の〜」という手がかり表現を含んでいるが、表現している情報は背景ではなく問題点である。

(例) しかし，既存の 解析系における複合辞の取り扱いは不十分である.

したがって、手がかり表現以外の手がかりを利用して取り出す方法を検討する必要がある。例えば，研究の背景について説明している文は序論の最初に記述されていることが多い。そこで、手がかり表現と文の位置を組み合わせて背景の情報を表現している文を取り出すことが考えられる。

4.2.4 手法を表現する文の抽出結果と検討

図 2 で示した手がかり表現だけでは手法の情報を表現している文をとりだせないことがあった。特に、手法の情報を表現している文が 1 つの序論に複数含まれている場合、それらを取り出すのに失敗することが多かった。

4.2.5 必要条件を表現する文の抽出結果と検討

必要条件の情報が記述されている文は、図 2 で示した手がかり表現でかなり取り出すことができた。取り出すのに失敗した例の多くは、文中の指示詞の指示対象が取り出した文だけではわからない場合であった。

(例) ただし，これらの手法はドメイン依存であったり，膨大な 3 つ組データが 必要になる.

参考文献

[1] 木下: “理科系の作文技術.”, 中公新書 624，(1981).

[2] 黒橋, 河原: 日本語形態素解析システム JUMAN version 5.1 使用説明書, 京都大学, (2005)