E-032

語用論に基づく「論語」検索システムの構築
Analects of Confucius Retrieval System Based on Pragmatics Information

楊 曄† 姜 沛林† 土屋 誠司‡ 任 福継‡
Ye YANG Peilin JIANG Seiji TSUCHIYA Fuji REN

1. まえがき

近年，古典文献を電子化し研究に活用しようとする動きが盛んになりつつある．しかし，古典文献を対象とした情報検索システムを実現するとき，全文検索ではうまくできないことがある．特に，古典と現代文は言葉使いが違うため，一般の利用者はうまく古典の単語を用いて操作することができない．さらに，古典には典故また反語，対偶，比喩など多様な表現法が用いられ，単純なキーワードマッチング技法のみでは検索効率が低下するといった問題がある．

過去の研究において，古典における意味検索を目的とした試みは少なく，新たな意味検索システムの開発が必要である．そこで，本研究では，古典の言葉遣いに着目し，語用情報に基づくアプローチを提案する．

古典意味検索の手法として，オントロジーに基づく手法がある[1]．文献[1] では，中国の唐詩を処理対象として，オントロジーの構築と意味的注釈付与（semantic annotation）の方法を提案している．しかし，単語数が同義語を含めると数万語にも達し，概念分類や精査は手作業のため，相当な時間とコストが必要である．

我々は中国の儒学の代表作である『論語』をドメインとする中国語質問応答システムを構築している．本稿では論語の特徴に基づき対象とする文書から語用情報とカテゴリを抽出し，その結果を用いて検索を行い，適した文書をユーザーに返す検索方法を提案する．

2. 『論語』について

『論語』は二千年余にわたって中国社会，周辺の国々や地域，及び世界各地の人々に大きな影響を与えてきた．この影響は政治，経済，思想，文化，教育など多方面に及んでいる．

『論語』は 20 篇により構成されており，全 20 篇の篇名の大部分は 2 字か 3 字であり，おおむね各篇の第一章の文の最初の 2 文字を取っている．そのため，篇名をみて，直ちにその内容を推測することができない．また，『論語』の各篇には篇ごとに定まったテーマは存在しない[2]．

3. システム構成

本稿で提案する検索手法は『論語』の一般愛読者を対象とし，中国語の自然言語による『論語』に関する問合せに応答するシステムの一部である．本システムの構成を図 1 に示す．

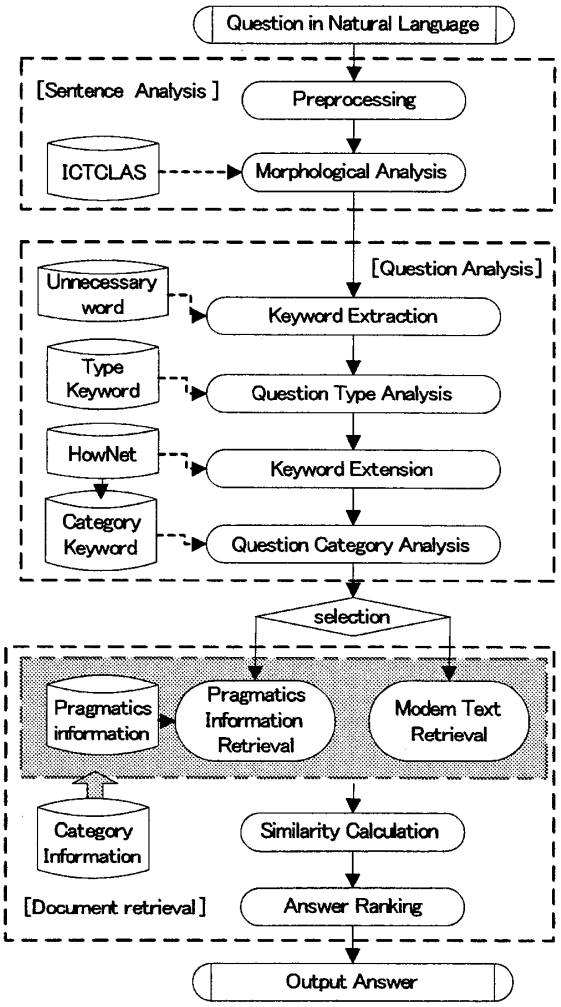

図 1．システムの構成図

4. 語用情報の抽出

発話理解には，語用論の知見が有効である．語用論における関連性理論では，発話の解釈には（1）表意の確定，（2）高次表意の確定，（3）推意の確定の三つの側面がある．表意の確定には（I）文の曖昧性の除去，（II）省略の補完の二つの処理を行わなければならない[3]．本研究では，これらの 3 つの処理により得られる意味を語用情報と定義する．

既存の『論語』関連書籍[4, 5, 6, 7]を参考にし，人手で『論語』各文章の語用情報を抽出した．語用情報の付与に対する恣意性の排除のため，30 人の被験者によって付与した語用情報を読ませ，賛成する人数が半分以上に達する語用情報をシステムに登録した．語用情報の抽出例を表 1 に示す．

† 徳島大学大学院先端技術科学教育部，Graduate School of Advanced Science Technology Education, The University of Tokushima

‡ 徳島大学大学院ｿｼｵﾃｸﾉｻｲｴﾝｽ研究部，Institute of Technology and Science, The University of Tokushima

表1. 語用情報の抽出例

語用情報別カテゴリ	論語の本文	語用情報
推意	先生が言われた，「立派な人間はけっして単なる専門家ではいけないものだ.	君子は博識であるよう.
省略の補完	先生が言われた，「魯と衛の政治は兄弟みたいなものだ.」	魯の国と衛の国の政治は兄弟のように似たものだ.」
高次表意	先生がいわれた，「私は回と一日中，語りあっても，全く逆らわず（異説も反対もなく）まるで愚かもののように見える. だが退出してから，その私生活を見ていると，大切な点はちゃんと身に付けていて人にはっと思いあたらせるところが十分にある. 回は愚かものではない.」	孔子は顔回のことを感心した.

5. カテゴリの抽出

抽出した概念カテゴリは，大分類と小分類とした．大分類は「学習と教育」，「処世」，「政治」，「教養」「人物」，「葬祭」，「天命」，「孔子」，「よそから見る孔子」，「文芸」，「人物評」など12 種類とした．小分類は例えば大分類「学習と教育」の下に「学習法」や「学習内容」，「教育方針」などを定義した．小分類は計40 種類とした．

6. 提案手法の評価実験と考察

『論語』の独特性のため，既存のテストコレクションは利用できない．そこで，我々は独自に『論語』のテストコレクションを作成した．『論語』の文章を被験者に提示し，そこに込められた孔子の知恵や思想を質問文の形式で表現するように指示した．質問作成の際には，一つの文章について複数の質問を作成することを許した．50名の被験者に一人20文章の質問を作成させた．作成した質問を他の10名の被験者でチェックし，テストコレクションを作成した．作成したテストコレクションは199文章である．

語用情報を用いた検索手法の有効性を検証するため，語用情報を利用しない手法と提案手法との精度比較をマクロ平均F値を用いて行った．図2に提案手法と語用情報を利用しない方法の各カテゴリの検索精度を示す．図中のPは語用情報を利用する手法による結果であり，Cは語用情報を利用しない方法の結果である．

カテゴリごとに検索精度を比較した結果，提案手法を用いることで，7カテゴリ全てで検索精度の向上が見られた．これより，語用情報の利用が有効であることがわかる．さらに，各分野のF値の平均を求めたマクロ平均F値，および分野を区別せず全7分野に対してRecallとPrecisionを計算しF値を求めたマイクロ平均F値を評価した[8]．その結果を表2に示す．

語用情報に基づく手法はマイクロ平均F値もマクロ平均F値も50%前後であり，語用情報を利用しない手法より

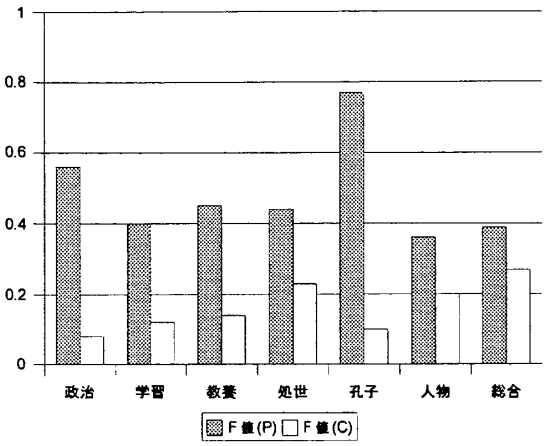

図2. 検索精度の結果

高く，優れた検索精度が得られている．語用情報には語の出現頻度は低いが検索に寄与する重要な語が多いためであると考えられる．

表2. マイクロ平均F値とマクロ平均F値の結果

手法	マイクロ平均F値（%）	マクロ平均F値（%）
語用情報を用いる手法	**52.4**	**48.1**
語用情報を用いない手法	**20.2**	**16.2**

7. おわりに

本研究は，論語に関する質問応答システムの構築であり，本稿では語用情報に基づく検索手法を提案した．

実験の結果，語用情報を用いた提案手法の方が語用情報を用いない手法に比べて約32%検索精度が向上することを示した．

参考文献

[1] 蘇豊文，傅怡停，陳書磊，楊世堯，羅鳳珠，“漢語詩的本体知識與語意検索”，第一文学與資訊科技国際会議，Dec. 2003.
[2] 狩野直禎，“『論語』のすすめ”，月刊しにか，Vol.13, No.6, pp.14-17, 2002.
[3] 田窪行則，西山佑司，三藤博，亀山恵，片桐恭弘，“談話と文脈”，岩波書店，1999.
[4] 南懐瑾，“論語別裁”，上海復旦大学出版社，1990.
[5] 楊伯峻，“論語譯注”，中華書局，1980.
[6] 楊樹達，“論語疏証”，上海古籍出版社，1982.
[7] 松川健二，“論語の思想史”，汲古書院，1994.
[8] 福本文代，鈴木良弥，“WordNetの同義語クラスとその上位関係を利用した文書の自動分類”，情処学論，Vol.43, No.6, pp.1852-1865, 2002.