# 語用情報に基づく『論語』の質問応答システムに関する研究

楊曄[1]　松本和幸[1]　劉松[1,3]　任福継[2,3]　黒岩眞吾[2]

[1]徳島大学大学院　工学研究科

[2]徳島大学大学院ソシオテクノサイエンス研究部

[3]北京郵電大学　情報工程学院

# Constructing a Question Answering System of Confucian Analects Based on Pragmatics Information

あらまし　我々は，『論語』に関する単なる全文検索システムではなく，『論語』の内容検索を支援するようなシステムを開発している．本論文では，『論語』の言語表現の特徴に応じて，語用情報とカテゴリに基づく検索手法を提案した．抽出したカテゴリは，大分類と小分類がある．大分類は学習，教育，友，治国，教養など 17 種類ある．各大分類の下に小分類がある。また、既存の『論語』関連サイトや著作を参考し，語用情報を抽出した．本システムを用いると，ユーザの質問に対し，『論語』からの適切な文章が返される．

## 1.　まえがき

孔子の『論語』は儒学の古典的著作として，二千年余にわたって中国社会や周辺各国及び世界各地の中国人に大きな影響を与えていた．この影響は中国民族の政治，経済，思想，文化，教育など各方面に及んでいる．近年，人文科学分野でも，古典文献がパソコンを利用し電子テキストとして，格納，蓄積されてきている[1]．このような状況下で,『論語』の研究者は多様な『論語』のデータベース構築を試みている．これらのデータベースは分かりやすい現代語で,『論語』の文章を一つずつ解釈しており,『論語』への理解に役立つと思われる．

しかし，全文検索では用が足りないところも有る．理由として，(A)『論語』は内容が体系的ではないので，一つの概念についての論述が集まっておらず，何箇所にも散在している．よく読んでも分からないという人もいる．(B)目次がないので，演説や文章を書く時，『論語』の名句を引用しようとしても，なかなか難しい、などがある．

人間の物事の認知は概念から始めるように思われる．『論語』の中の概念を抽出し，分類データベースを作成するのは『論語』への理解に役立つと思われる．そこで，本研究は研究者ではなく，一般利用者の立場から，新たな『論語』システムの構築を検討している．『論語』の言語表現の特徴に応じ，語用情報とカテゴリーに基づく検索手法を提案した．

以下，２で『論語』の特徴を紹介し，３で語用情報とカテゴリーに基づくアプローチ，４でシステムの概要について述べ，最後に今後の課題を述べる．

## ２．『論語』の特徴

『論語』原文は 20 巻，500 余の文章からなっている．各巻には文章数が異なり，50 弱の文章（例えば第 14,15 巻）がある巻もあれば,3 文章しか存在しない巻(例えば第 20 章)もある．各文章の長さも異なり，７字しかない文章もあれば，300 字強の文章もある．以下に論語の主な特徴を示す．

(1) 『論語』は内容から言うと，格言に近く意味深い．

(2) 多様な表現法を用いている．対偶や比喩や反語などが挙げられる．

(3) 文字通りの意味ではない情報が含まれている文章もある．

(4) 現代語のシソーラスに含まれていない言葉も少なくない．時代の変遷にしたがって，意味が変わる言葉もある．

## ３．語用情報とカテゴリーに基づくアプローチ

まず，情報検索において問題となるのは，「ことば」と情報の複雑な対応関係である．「ことば」は，文脈に応じてそれが意味する情報が異なるとともに，ある情報を「ことば」によって表現する方法は一般に多数考えられる．[2]『論語』の内容を分析するためには，文字通りの意味だけではなく，情報の抽出が重要である．また，前書きで述べたように，『論語』では，一つの概念についての論述が何箇所にも散在している．この散在した論述をまとめてみると，多角度から一つの概念を解釈しており，かなりの論理性を持つことがわかった．具体的に，例えば「仁」という言葉がよく出てくる．「仁」をめぐって，仁徳の実践方法や仁者の特徴や，仁徳と貧富，仁徳と才能など，多角度から解釈している．

そこで，我々はまず『論語』の主な概念を抽出した．これで，概念による検索が実現され，孔子の思想を把握するのに役立つと思われる．次に，『論語』は人生の道理を説いたもので，現代社会においても道を開くヒントとなるものも含まれている．『論語』を実生活に活かしたいという考えをもとに，既存の『論語』関連サイトや著作を参考し，語用情報を抽出した．

例として，次のような二つの文が挙げられる．

(1)　子の曰わく，これを知る者はこれを好む者に如かず．これを好む者はこれを楽しむ者に如かず．

(2)　宰予，昼寝ぬ．子の曰わく，朽木は雕るべからず，糞土の牆は杇るべからず．予に於てか何ぞ誅めん．

我々は，文(1)から，「学問や仕事はそれを楽しめばこそ上手になる」という語用情報を，文(2)から「勤勉に勉強すべし」という語用情報を抽出した．文(1)と文(2)は共通のカテゴリである「学習」を持っており，細かく分類すると文(1)は学習法で，文(2)は学習の効用である．

抽出したカテゴリは，大分類と小分類がある．大分類は学習，教育，友，治国，教養など 17 種類ある．大分類には例えば「学習」の大分類の下に「学習法」や「学習の効用」，「態度」などのような小分類もある．各大分類の下の小分類数が異なる．語用情報も人手により抽出された．語用情報と文章の対応関係は一対一ではなく，多対一の関係，つまり一つの文章に対して，複数の語用情報を抽出する場合もある．同様に，カテゴリの抽出においても，多対一の場合もある．表 1 と表 2 はカテゴリと語用情報の抽出例を示す．

表 1 カテゴリの抽出例

| 大分類 | 小分類 |
|---|---|
| 友 | 方法，態度，対象，… |
| 学習 | 方法，内容，態度，効用，… |
| … | … |

表 2　カテゴリと語用情報の抽出例

| 大分類 | 小分類 | 語用情報 | 「論語」日本語訳 |
|---|---|---|---|
| 友 | 方法 | 仁徳がある人になるよう | 子の曰わく，徳は孤ならず．必らず隣あり． |
| | | … | … |
| | 態度 | 君主や友には，うるさくせぬよう | 子游が曰わく，君に事うるに数すれば，斯に辱しめられ，朋友に数すれば，斯に疎んぜらる． |
| | | … | … |
| | 対象 | 中庸の人，積極的に進んだ人，正直な人と友達になるよう | 子の曰わく，中行を得てこれに与せずんば，必ずや狂狷か．狂者は進みて取り，狷者は為さざる所あり． |
| | | … | … |
| | その他 | … | … |

## ４．システム

ここでは，以上の考えを元に実現された『論語』の質問応答システムについて述べる．本論文で紹介するシステムは Web で公開するのを目標としたため，評価プロジェクトのように文法的に正しく質問対象が明確な質問ばかりとは限らない．特に，「教育について」のような普通には質問とは考えられないような曖昧な質問も入力される．また，『論語』内容の独特性のため，[3]と[4]のような，既存のテストコレクションが利用できない．我々はアンケート調査を行い，『論語』のテストコレクションを作成した．このテストコレクションから，『論語』に対する質問タイプの種類がそれほど多くないことが分かった．主に，Why 型質問(理由を尋ねるもの)，Symptom 型質問(定義，説明，記述を尋ねるもの)，How 型質問（手順や手法や見方を尋ねるもの）である．[4]そのほか、「○○についてどう考えべきか」のように、物事に対する態度や考え方についての質問も少なくない。図 1 はシステム構成図を示す。

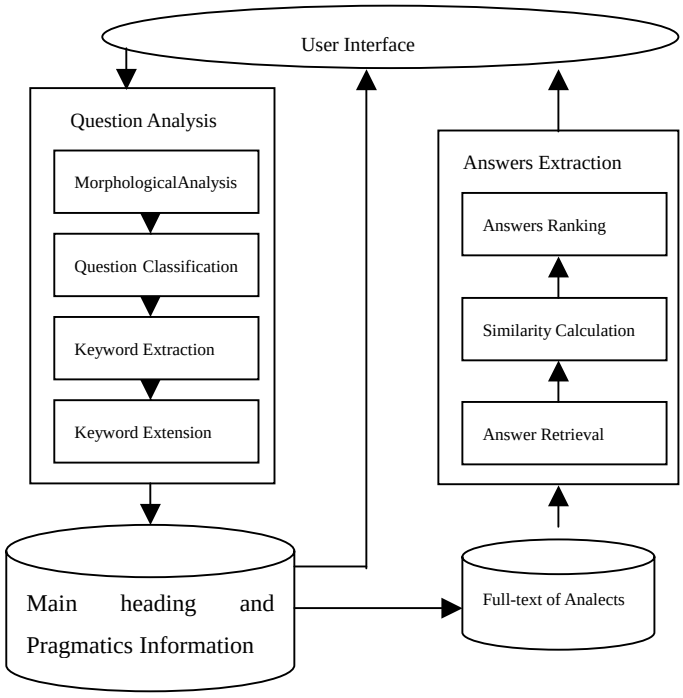


図 1 システム構成図


図 2 システムの検索インターフェース

試作中の『論語』質問応答システムの検索インターフェースのスクリーンショットを図 2 に示す．質問文入力欄にユーザーが質問文を入力し，「検索」のボタンを押すと，質問文に対応した『論語』の内容が検索結果の候補として結果表示画面に複数出力される．図 2 は，「どうやって学習すればいいですか」という質問文を入力し，結果が出力された様子を表している．

## ５．今後の課題

本稿では，語用情報とカテゴリに基づく『論語』内容検索手法を提案し，日本語による『論語』質問応答システムを試作した．今後は質問タイプによる回答抽出規則や，シソーラスに含まれていない単語からの概念抽出について検討し，テストコレクションを用いた評価実験を行いたいと考えている．

**参考文献：**


[1] 桶谷猪久夫，Delmer Brown，大久保裕子　山尾正之：『XML を利用した日本古典史料の英日全文連携検索システムの構築―日米共同研究について』，大阪国際大学紀要「国際研究論叢」第 19 巻第 1 号，pp.87-100,(2005).

[2] 長尾眞，黒橋禎夫，佐藤理史，池原悟，中野洋編：『言語情報処理』岩波書店,(1998).

[3] TREC QA Homepage　http://trec.nist.gov/date/qa.htm

[4] NTCIR QA Homepage http://www.nlp.cs.ritsumei.ac.jp/qac/

[5] 加藤恒昭，福本淳一，桝井文人，神門典子：『質問応答技術は情報アクセス対話を実現できるか』，情報処理学会研究報告 2004-NL-162,pp145-150,(2004).