氏名 (本籍)	相場 徹 (秋田県)
学位の種類	博士 (情報科学)
学位記番号	情博第 156 号
学位授与年月日	平成 12 年 3 月 23 日
学位授与の要件	学位規則第 4 条第 1 項該当
研究科、専攻	東北大学大学院情報科学研究科 (博士課程) 人間社会情報科学専攻
学位論文題名	古典サンスクリット語・チベット語の計量的なデータ処理に関する研究
論文審査委員	(主査) 東北大学教授 生出 恭治　東北大学教授 福地 肇 東北大学教授 静谷 啓樹　東北大学教授 磯田 熙文 (文学研究科) 東北大学助教授 ジェレミー・シモンズ

論文内容要旨

1 序論

東洋学、とくにインド学仏教学の学問分野においても、研究に計算機を用いることが不可欠となりつつある。インド学仏教学は文献学を中心とした実証的な学問分野であるが、文献の分量が膨大であることに比して、文献の多様性・言語の難解性などの理由から、いまだ手付かずの状態で放置されている文献が多い。それゆえ、研究に計算機を用いることが非常に効果的な学問分野であると考えられる。ただし現在のところ、これらのテキストを扱うための知識の蓄積は十分ではないため、すでに構築・公開されている電子テキストの用途は、単語検索等の単純なものに限られている。

筆者は、東洋系古典文献を扱う学問分野、とくにインド仏教学における研究目的・研究方法の枠組に従いながら、計算機科学とくに自然言語処理学における方法論、なかでも言語的情報の抽出等の手法を援用した研究を目標としている。ただしインド学仏教学で扱う古典サンスクリット語・チベット語等については言語の特殊性・処理に関する知識の蓄積が十分ではない。そこで本論では、東洋系古典のテキストを計算機で扱う際に生じる問題点をいくつか取り上げる。そして、その問題を解決することによって、東洋系古典の電子テキストを処理するための基礎的な知識の獲得および蓄積を目的とすることとした。本論で取り上げる問題点は以下の 3 つである。

- 転写方式の多様性の問題 (第 2 章)
- 術語辞書構築の問題 (第 3 章)
- 対訳データ構築の問題 (第 4 章)

2 転写方式の多様性の問題

現在すでに構築・公開されている電子データは、構築プロジェクトごとに独自の転写方式が用いられていることが多く、統一的に扱うことが困難という問題がある。そこで第 2 章では、まず実際に構築・公開され

ているテキストの転写方式を可能な限り収集し、それらの転写方式の分類・整理をおこなった。その結果、さまざまな転写方式を大きく以下の2つに分類することとなった。

- 代替表記を用いる方式 (Substitutional Methods)
- 特定の文字コード領域に文字を割り当てる方式 (Allocation Methods)

このうち前者は "ṭ" を ".t" のように表記する方式であり、代替表記への転写に一定の規則性を見出すことができるのが特徴である。そこで、それぞれの転写方式ごとに転写規則を発見し、それらに関する整理をおこなった。後者は、ダイアクリティカルマーク付き文字をサンスクリット語等の転写には用いない文字コード領域、とくに非ASCIIコード領域に割り当てる方式である。この方式については、どの文字をどの文字コードに割り当てるかに関する規則性を見出すことがほぼ不可能であった。

このように転写方式に関する分類・整理をしたのち、任意の電子テキストにおいて用いられている転写方式の機械的な判別に関する実験を、サンスクリット語・チベット語のそれぞれについて行なった。

サンスクリット語　「代替表記を用いる方法」については、転写方式ごとに分類・整理した転写規則に基づく判別をおこなったところ、ほぼ完璧な精度での判別が可能であった。

また「特定の文字コード領域に割り当てる方法」については、サンスクリット語の sandhi という音韻的な制約に注目し、文字の連接傾向に関する情報を電子テキストから抽出した。そして抽出した結果を用いた転写方式の推定をおこなったところ、99% 以上の精度で正しい判別ができることを確認した。ただし、実験のために用いた言語情報の分量、および判別に要する計算量が多すぎることが問題点として残った。それゆえ情報量・計算量を少なくすることが今後の課題である。

チベット語　筆者が収集した転写方式はすべて「代替表記を用いる方法」であったため、それぞれの転写方式ごとの転写規則による判別によって、ほぼ完璧な精度での判別をすることが可能であった。

3　術語辞書構築の問題

計算機を用いて本格的に電子テキストを扱う研究を行なうためには、電子的な術語辞書・シソーラス類の存在が不可欠となる。しかし、サンスクリット語・チベット語に関する電子的な辞書はほとんど存在していない。それゆえ、すでに存在している電子データから術語間の関係を自動的に抽出する作業をおこない辞書構築の一助とすることができれば、計算機を用いた文献研究の促進が期待される。第3章では、現存する唯一といってよい電子的な術語集である「インド学仏教学論文データベース INBUDS」を利用し、INBUDS 中に記述されている術語を対象とした、術語間関係の抽出に関する実験をおこなった。

INBUDS　INBUDS は論文書誌データベースであり、論文ごとに題目・著者名・掲載誌などの書誌情報が記述されている。また論文ごとに、その内容に関連したキーワードが、地域・時代・文献・人物などといったカテゴリごとに記載されている。

術語の「共起」の利用　「同じ論文データ中に共存している術語は『共起』している」という定義をおこなう。たとえば同じ論文の中で「tarka」と「六句義」とがキーワードとして記載されている場合、これらの術語は共起している、とするのである。そして INBUDS 中におけるこのような術語間の共起の傾向をスコア化することによって、術語間の関係の数値化をおこなった。しかし、抽出することのできた術語間の共

起情報の分量が十分なものではなかったため、抽出した術語間関係のスコアの信頼性が非常に低くなってしまうことが問題となった。

「チェイン」による間接的な術語の共起　術語間の共起に関する情報量の不足の問題を解消するため、「チェイン」という概念を提案することにより、術語間の直接的な共起関係だけでなく、間接的な共起関係に関する情報も抽出するようにした。この「チェイン」について簡単に説明する。たとえば「tarka」と「瑜伽派」という単語があり、両者は直接的な共起関係を持たないとする。このような場合、「tarka」をキーワードに持つ論文Aにおける他の術語、たとえば「(著者) 金倉円照」という項目を利用して、その項目を含む他の論文Bを探し出し、その論文Bに「瑜伽派」という術語がないか調べる。もし論文Bにも「瑜伽派」がないときは、論文Bに含まれる他の項目、たとえば「(人物) パタンジャリ」を元にして、さらに別の論文Cを探し出し、その論文Cに「瑜伽派」が含まれているかを調べる。もし論文Cの中に「瑜伽派」という術語があったときには、「tarka」と「瑜伽派」とは、「(著者) 金倉円照」「(人物) パタンジャリ」をチェインとした間接的な共起関係にある、とするものである。このような共起関係の拡張をおこなうことにより、かなりの分量の術語間関係に関する情報が抽出できるようになった。また、このようにして抽出した情報に関する評価をおこなった結果、ある程度の信頼性を持ったものであることを確認することができた。

4　対訳データ構築の問題

インド学仏教学においては、研究の基礎となるべきサンスクリット語原典の多くが散逸している。それゆえ、原典をチベット語・漢文に翻訳した二次文献がかなり重要視されており、原典研究の際にはこれら翻訳文献との対比は不可欠とされている。しかし計算機上で、これら原典と翻訳文献とを対比させて扱う研究はいまだ行なわれていない。そこで第4章では *Saddharmapuṇḍarīka* のサンスクリット語・チベット語の対訳データを対象とし、これらの文献を扱うための最初の試みとして、対訳データ中における対訳単語を自動的に識別させるための実験をおこなうこととした。

文中の単語語幹の推定　現在のところ、サンスクリット語文中における単語の語幹を計算機に推定させるための技術の蓄積、また解析に必要な電子情報の蓄積が十分ではなく、“bodhisattve”, “bodhisattvān” がともに “bodhisattva” という同じ単語であることを計算機に認識させることさえ困難な状況である。それゆえ、文中における単語の語幹の推定をおこなうことによる、同一単語の同定が重要な課題となる。そこで「活用語尾の長さ」および「語幹ごとの出現確率」という基準を設け、これらの基準による単語語幹の推定を行なうこととした。まず名詞語幹の種類ごとに、サンスクリット語の文中における出現頻度の数えあげをおこない、名詞語幹ごとの出現確率値を用意する。そのうえで、上記の2つの基準に基づく単語語幹の曖昧性の解消を行なう実験をおこなった。実験は名詞のみを対象としたが、動詞等に関する言語情報が現状では不足していることがその主な理由である。実験の結果、実験対象とした名詞語句のうち、語幹の推定をおこなうことができたのは全体の約87%であった。実験結果を分析した結果、これ以上の精度向上のためには辞書等の拡充が不可欠であることが明らかとなったため、この点については今後の課題とせざるを得なかった。

対訳データからの対訳語の識別　名詞単語の語幹推定の結果をふまえ、対訳データにおける対応単語の識別に関する実験をおこなった。実験の内容は、*Mahāvyutpatti* というサンスクリット語・チベット語の対訳辞書を用いて対応単語を探すという単純なものである。この実験の結果、対訳語をきちんと発見できた単

語の割合は、サンスクリット語・チベット語ともに全体の3割程度であった。実験の結果がこの程度にとどまってしまったことについては、先に述べた単語語幹の推定の精度の低さ、および今回用いた対訳辞書の語彙数の不足が主な原因として上げられた。

5 結論

本研究は「文献学的な観点に基づく、自然言語処理学の手法を援用した東洋系古典の電子テキストの解析」を目標とし、この目標を実現するための足掛りとして、東洋系古典を扱うための技術・情報の獲得および蓄積を目的としたものであった。

本研究における成果としては、実際に用いられている多様な転写方式を収集して整理・分類したこと、論文書誌データベースからインド学・仏教学関連の術語間の関係を抽出したこと、サンスクリット語の文章中における名詞語幹の推定をおこない一定の成果を得たこと等があげられる。このいずれも前例のない成果であり、今後、東洋系古典を計算機で扱う際の技術的・情報の蓄積となるのではないかと考えている。

しかし、筆者が本来的な目標としている「計算機を用いた古典学」を実現したとは言えない状況であるので、さらなる研究の展開を図っていきたいと考えている。

論文審査の結果の要旨

近年になって、古典サンスクリット語 (Skt.)・古典チベット語 (Tib.) で書かれた文献を扱う研究分野においても、計算機を用いた研究への関心および需要が高まってきている。しかし現状では、電子データの不足、および言語の複雑性などの東洋系古典言語の特殊性・処理上の困難性によって、本格的に計算機解析を用いた研究が困難である。本研究は、Skt., Tib. の電子テキストを対象として、東洋系古典文学を扱う文献学とくにインド仏教学の観点、および計算機科学とくに自然言語処理学の手法を用いることにより、古典文献の計算機処理に関する新たなアプローチを提案した研究であり、全編 5 章から成る。

第 1 章は序論であり、研究の背景および目的を述べている。ここでは Skt., Tib. に対する計算機解析を進めていくためには 3 つの課題を克服する必要があることが示されている。そして以下の章において、これら 3 つの課題についての論考が順に行なわれている。

第 2 章では、広く公開されている電子テキストを統一的に扱う際に問題となる、転写方式の多様性に関して述べている。本章ではまず、多様な表記方法に関する整理および分類を行ない、その整理分類の結果を元にした転写方式の機械的な自動判別について述べている。このような研究は従来行なわれておらず、重要な成果である。

第 3 章では、現状では不足している電子化された術語辞書・シソーラスの自動的な構築について述べている。まず既存の電子的な論文書誌データベースにおける術語の「共起」を利用した術語間関係の推定について述べられている。また、構築するデータの情報量を増やすため「チェイン」という概念による共起の拡張をおこない、その結果に関する評価をおこなっている。この成果は広く応用される可能性をもつものとして評価される。

第 4 章では、Skt., Tib. 文献に対する対訳データの処理の第一歩となる、対訳単語の自動識別について述べている。Skt. においては複雑な単語の活用と sandhi という独特の音韻規則のため、文中の単語から語幹を推定することが大きな課題となるが、本章ではまずこの単語の語幹の推定について述べられ、一定の成果が確認される。そのうえで対訳単語を自動識別する手法に関する実験をおこない、その評価をおこなっている。これは Skt., Tib. の対訳テキストを計算機的に処理しようとした最初の業績であり、当該研究分野に対するきわめて大きな貢献となっている。

第 5 章は結論であり、全体の総括および今後の展望について述べている。

以上要するに本論文は、東洋における古典文献を計算機で扱う際の特殊性・困難性を明らかにし、それに対処するための見通しを立てたうえで、その見通しに沿った一定の成果を示したものであり、情報科学の発展に寄与するところが少なくない。

よって、本論文は博士 (情報科学) の学位論文として合格と認める。