【374】

氏名（本籍）	黄　海湘（中　国）
学位の種類	博　士（情報学）
学位記番号	博　乙　第　2601　号
学位授与年月日	平成24年4月30日
学位授与の要件	学位規則第4条第2項該当
審査研究科	図書館情報メディア研究科
学位論文題目	**中国語を対象とした意味訳型翻字手法**

主　査	筑波大学教授	理学博士	石塚英弘
副　査	筑波大学教授	工学博士	田中和世
副　査	筑波大学教授	工学博士	杉本重雄
副　査	筑波大学教授	博士（工学）	佐藤哲司
副　査	筑波大学教授	博士（工学）	山本幹雄
副　査	東京工業大学准教授	博士（工学）	藤井　敦

論文の内容の要旨

科学技術や文化の発展に伴って、新しい技術や概念を表す用語が次々に造られている。ある言語で造られた用語を他の言語へ移入する方法には、大きく分けて「意味訳」と「翻字」がある。意味訳は、原言語の意味を移入先言語の既存語もしくは新語で表記する方法である。翻字は、原言語の発音を移入先言語における音韻体系に基づいて表記する方法である。固有名詞や専門用語は翻字されることが多い。

多くの言語において、翻字には表音文字が用いられる。例えば、日本語ではカタカナを用いて外国語を翻字する。しかし、中国語では表意文字である漢字を用いて翻字するため、翻字結果は原言語の発音を再現するだけでなく、表記に使用された漢字によって一定の意味を持つようになる。その上、一つの発音に対して複数の異なる漢字が対応する。そこで、対象語の意味を考慮した翻字、すなわち「意味訳型翻字」では漢字の選択が重要である。

例えば、飲料の名称である「Coca-Cola」に対して、中国語では「可口可楽」や「口卡口拉」などの様々な漢字列で発音を表記することができる。公式の表記は「可口可楽」であり、「可口」と「可楽」にはそれぞれ「美味しい」と「楽しい」という意味がある。すなわち、適切な漢字を使用することで、「Coca-Cola」の発音と意味をうまく表現できている。他方において、「口卡口拉」は、「口卡」に「喉に詰まる」という意味があるため、飲料の名称として不適切である。このように、使用される漢字によって翻字結果の適否が変わる場合がある。

ここで、翻字結果の適否に関する基準が対象語の種類によって異なる点に注意を要する。例えば、音楽家のChopinは中国語で「肖邦」と表記する。「肖」は中国人の姓によく使われる漢字である。「消」は「肖」と発音が同じであるにも拘らず、「消滅する」という意味があるため人名に使用されることは稀である。しかし、消毒薬や殺虫剤のように何かを消し去る効果を持つ商品名を翻字する際には、「消」という漢字が好んで使われる可能性がある。

以上より、中国語における意味訳型翻字においては、対象語の発音だけでなく意味と種類も考慮する必要

がある。意味訳型翻字の自動手法に関する先行研究では、対象語の発音と意味を考慮した手法が提案されている。しかし、対象語の種類は考慮されていない。さらに、先行研究では対象語の意味を反映した漢字を選択するために、対象語に関連する語句を人手で与える必要があり高価である。

本論文は、意味訳型翻字において、対象語の発音、意味、種類を統合した確率的な手法を提案する。また、対象語の関連語を自動抽出することで人手の負担を削減する手法を提案する。さらに、提案した手法をコンピュータ上のシステムとして実現し、実験によって有効性を評価する。

本システムは、対象語とその種類を入力として、複数の翻字候補を順位付きリストの形式で出力する。現在、原言語として日本語を前提としている。しかし、原理的には、ローマ字表記できる言語は全て原言語として利用可能である。本システムは、対象語の発音、意味、種類を考慮するために、「発音モデル」、「意味モデル」、「カテゴリモデル」を構築し、ある漢字もしくは漢字列が選択される確率を計算する。この中で、本論文の貢献であるカテゴリモデルは、人名や企業名といった用語の種類ごとに中国語のテキストデータを収集し、各漢字が使用される確率を事前に計算して構築する。

本システムの動作は、まず、発音モデルによって対象語の発音に近い複数の漢字列を翻字候補として導出し、漢字列の確率に基づいて翻字候補の順位付きリストを生成する。次に、意味モデルとカテゴリモデルによって、対象語の意味と種類を反映した漢字が含まれる翻字候補を優先して、翻字候補の再順位付けを行う。ここで、対象語の種類に対応したカテゴリモデルを選択的に使用することで、対象語と同じ種類の語に使われやすい漢字を含む翻字候補を優先することができる。また、意味モデルで必要とされる対象語の関連語をWorld Wide Webから自動抽出して利用する。具体的には、対象語に関する日本語Wikipediaの記事を検索し、形態素解析によって名詞と形容詞を抽出する。抽出した名詞と形容詞に対して、対象語との関連度をWeb上の単語分布に基づいて計算する。さらに、関連度が高い上位の語を中国語に機械翻訳し、最終的な関連語の集合とする。

評価実験では、日本語の単語210件を評価用の対象語として利用し、提案手法の有効性を複数の観点から評価した。その結果、対象語の種類に応じてカテゴリモデルを使い分けることで、翻字精度が向上することを確認した。さらに、対象語の関連語を人手で与えた場合と自動抽出した場合の翻字精度を比較した結果、翻字精度をそれほど落とすことなく、関連語を人手で与える負担を削減できることを確認した。

審 査 の 結 果 の 要 旨

本論文の著者は、中国語では表意文字である漢字を用いるため、他言語で造られた用語を中国語に「翻字」する場合は、対象語の意味を考慮した翻字、すなわち「意味訳型翻字」を行う必要があること、それを人間が行うには多くの手間が必要で負担が大きいことに着目し、コンピュータによる意味訳型翻字の支援を目的として、意味訳型翻字に関する計算モデルを提案し、その有効性を評価実験によって検証した。本論文における著者の記述は具体的で説得力があり、本研究の背景と目的、関連研究、本論文で提案した意味訳型翻字の計算モデルと手法、評価実験の方法と結果などが明確に記述されている。そのため、本研究の貢献、新規性、有用性が何れに在るかも明確である。

本論文は５つの章で構成されているため、各章ごとに述べる。

「第１章　序論」では、中国語への翻字に関する歴史的変遷について解説している。外来語の増加に伴い中国では意味訳型翻訳の標準化が必要とされ、新華通信社の世界人名翻訳大辞典には中国語への意味訳型翻字に関する基準が定められている。しかし、この基準は人名と地名以外の用語には対応していないという問題がある。また、人名や地名の翻字においてさえ使われない場合があるという問題がある。その結果、意味訳型翻字は翻訳者の知的活動に依存する度合いが高く、負担が大きいのが実状である。近年急速に増え続け

る大量の新語に対応するためには、この負担がさらに大きくなる。こうした背景を踏まえて、本論文は、コンピュータによる意味訳型翻訳の自動化もしくは部分的な支援が必要である点を指摘している。指摘は具体的で、本研究の目的である意味訳型翻字の必要性が明確に理解できる。

「第2章　関連研究」では、自動翻字に関する先行研究について対象言語を問わず幅広く調査し、それらとの比較を通して本研究の位置付けを明確にしている。とりわけ、中国語における意味訳型翻字に関する先行研究には2つの事例がある。1つ目の研究では、人名を対象として、対象語の起源、性別、姓名を考慮する翻字手法が提案されている。しかし、対象語が人名に制限されているため拡張性に乏しいという問題がある。2つ目の研究では、対象語の発音と意味を考慮した翻字手法が提案されている。この手法は対象語の種類に依存しないという利点がある一方で、対象語の種類に関する特徴を捉えることができないという欠点がある。例えば、音楽家のChopinは中国語で「肖邦」と表記する。「肖」は中国人の姓によく使われる漢字である。「消」は「肖」と発音が同じであるにも拘らず、「消滅する」という意味があるため人名に使用されることは稀である。しかし、消毒薬や殺虫剤のように何かを消し去る効果を持つ商品名を翻字する際には、「消」という漢字が好んで使われる可能性がある。さらに、この手法では、対象語の意味を考慮するために、対象語に関連する語句を人手で与える必要があるという問題もある。以上の検討から、本論文は、対象語の発音と意味に加えて、対象語の種類も考慮する必要があることと、対象語の関連語を自動的に特定することの必要性を指摘している。

「第3章　意味訳型翻字手法」では、2章での議論を踏まえて、対象語の発音、意味、種類を統合した意味訳型翻字手法を提案している。また、対象語の関連語を自動抽出することで人手の負担を削減する手法を提案している。本論文の提案手法を実装したシステムは、対象語とその種類を入力として、複数の翻字候補を順位付きリストの形式で出力する。現在、原言語として日本語を前提としている。しかし、原理的には、ローマ字表記できる言語は全て原言語として利用可能である。本システムは、対象語の発音、意味、種類を考慮するために、「発音モデル」、「意味モデル」、「カテゴリモデル」を構築し、ある漢字もしくは漢字列が選択される確率を計算する。この中で、本論文の貢献であるカテゴリモデルは、人名や企業名といった用語の種類ごとに中国語のテキストデータを収集し、各漢字が使用される確率を事前に計算して構築する。対象語の種類に対応したカテゴリモデルを選択的に使用することで、対象語と同じ種類の語に使われやすい漢字を含む翻字候補を優先することができる。意味モデルを用いるためには対象語の関連語が必要であるため、World Wide Webから関連語を自動抽出して利用する。本章で示されたカテゴリモデルには新規性が認められる。また、World Wide Webから関連語を自動抽出する方法は人手の負担を削減する点で有用性がある。

「第4章　評価実験」では、日本語の単語210件を評価用の対象語として利用し、提案手法の有効性を複数の観点から評価した。その結果、対象語の種類に応じてカテゴリモデルを使い分けることで、翻字精度が向上することを確認した。さらに、対象語の関連語を人手で与えた場合と自動抽出した場合の翻字精度を比較した結果、翻字精度をそれほど落とすことなく、関連語を人手で与える負担を削減できることを確認した。評価実験は工学的な観点から周到に行われており、信頼できる内容である。従って、本論文で提案された手法は有効と言える。

「第5章　結論」では、本論文の貢献を総括するとともに残された課題についても述べている。残された課題はあるものの、カテゴリモデルを用いることによって翻字精度を向上させたこと、対象語の関連語を自動抽出した場合にも翻訳精度をそれほど落とすことなく、人手で関連語を与える負担を削減できることを明らかにしたことは、新規性、有用性の両面で高く評価できる。

以上、本論文の提案手法は意味訳型翻字の研究領域において新規性があり、また有用性もあることから、本論文は博士論文として十分な水準にあると評価できる。

平成24年3月1日、図書館情報メディア研究科学位論文審査委員会において審査委員全員出席のもと、

本論文について著者に説明を求めた後、関連事項について質疑応答を行った。引き続き、「図書館情報メディア研究科博士後期課程の学位論文審査に関する内規」第12項第3号に基づく学力の確認を行い、審議の結果、審査委員全員一致で合格と判定された。

よって、著者は博士（情報学）の学位を受けるに十分な資格を有するものと認める。