A Study on Natural Language Processing for Sinhalese

シンハラ語の自然言語処理に関する研究

සිංහල බස පරිගණක සැකසුම සඳහා අධ්‍යයනයක්

by

Ajantha Herath, M.Eng

January 1997

氏　名(本　籍)	AJANTHA HERATH (スリランカ)
学位の種類	博　士 (工学)
学位記番号	甲第 75 号
学位授与年月日	平成 9 年 3 月 25 日
専　攻	電子情報システム工学専攻
学位論文題目	A Study on Natural Language Processing for Sinhalese (シンハラ語の自然言語処理に関する研究)
学位論文審査委員	(主査) 教　授　池　田　尚　志 (副査) 教　授　後　藤　宗　弘　教　授　藤　田　廣　志

論文内容の要旨

本論文は、スリランカの国語であるシンハラ語の計算機処理に関する研究について述べたものであり、特に日本語からシンハラ語への機械翻訳についての研究が中心となっている。全体は6章からなっている。

第1章の緒論では、機械翻訳を中心とした自然言語処理研究の歴史および機械翻訳の手法について概観し、本研究の目的と意義について述べている。

第2章では、計算機処理をするための前提として、シンハラ語の言語学的側面についてまとめている。すなわち、シンハラ語の言語的な系譜、現代シンハラ語の成立、シンハラ語の構文要素、語順、構文構造などについてまとめている。シンハラ語には性、数、人称等があって主語と述語の間で属性の一致が必要であり、書記法においては語と語の間には空白が置かれるなど英語と類似の性質を持っているが、同時に、語順が自由である、構文要素を随時に省略することが出来るなど日本語と類似の性質も持っている。この章では、このような日本語との対比におけるシンハラ語の特質を明らかにしている。

第3章では、シンハラ語の計算機処理に関していくつかの研究をまとめている。シンハラ語の書記法では語と語の間に空白が置かれるが、実際にはこの一つのセグメントは一つの単語に対応するわけではなく、一般には単語と接尾語から成っている。人間用の従来の文法ではこの接尾語の部分はそれ以上に分析されることはなされていなかったが、本論文では計算機処理の立場からこれをさらに細かく分析し短単位の接尾語に分解して、性、数、格などの属性との対応関係を明らかにしている。本論文ではこのようなセグメントの詳細構造にL-unitという新しい呼び名を与えている。

また逆にこの短単位の接尾語を複合していく際、単純に並べていけばよいというわけではなく、綴りを変化させる必要が生じてくるが、これをLinking ruleとして規則化している。

次に、シンハラ語の主語あるいは目的語を決定するアルゴリズムについて述べている。シンハラ語では、主語と目的語は一般に接尾語によって表現される格によって決められるが、主格と目的格が同形である場合も多く、語順によってそれが決まる場合もあり、また単語の意味や、文脈的情報に依らなければ決まらない場合もある。このような種々の場合を考慮して、主語と目的語を決定するアルゴリズムを与えている。

第4章では、日本語からシンハラ語への機械翻訳に関する研究についてまとめ

ている。一般に機械翻訳は原言語の構文構造と目的言語の構文構造の間の変換を介して行われるが、本論文では日本語とシンハラ語の間の類似性を利用して、形態素レベルの構造の変換で翻訳出来ることを述べている。これは日本語の文節に対応するシンハラ語の単位があることを見出したことに基づいており、この単位をP-unitと名づけている。P-unitは一つあるいは複数のL-unitからなる。日本語の機能語(助詞、助動詞)からP-unitへの対応規則を具体的に作成して、この対応規則を設定することが可能なことを実証している。シンハラ語は日本語と同じように、あるいはそれ以上に語順は自由であり、また同じように構文要素の省略が行われるが、日本語の形態素解析の結果として得られる文節を、そのままの順序でこの対応規則によってP-unitに変換し、そこからシンハラ語の表層構造を生成すれば、複合文、埋め込み文の場合を含めて、十分に理解し得る正しいシンハラ語が得られることを、多くの例文を通じて実証している。また、計算機上に簡単な翻訳システムの構築も行っている。

第5章では、この機械翻訳法の問題点について考察している。

日本語の機能語からP-unitへの対応規則は実際には1対多となる場合も多く、その場合どの対応規則を選べばよいか分からないということになる。これへの対処法として、格構造を介して対応をとることを提案している。多くの例文の格構造について検討し、格構造を介することで対応規則の曖昧さを解決できることを実証している。このほか、強調構文の問題、数の問題などについて考察している。

第6章では、本研究の結果を要約し、今後への課題についてまとめている。

論文審査の結果の要旨

本論文は、スリランカの国語であるシンハラ語の計算機処理に関する研究について述べたものであり、特に日本語からシンハラ語への機械翻訳についての研究を中心としたものである。本論文により得られた成果は以下のとおりである。

(1) シンハラ語では英語のように語と語の間に空白が置かれるが、実際にはこの一つのセグメントは一つの単語に対応するわけではなく一般には単語と接尾語から成っている。従来のシンハラ語の文法ではこの接尾語の部分はそれ以上に分析されることはなされていなかったが、本論文では計算機処理の立場からこれをさらに細かく分析し短単位の接尾語に分解して、性、数、格などの属性との対応関係を明らかにしている。本論文ではこのようなセグメントの詳細構造にL-unitという新しい呼び名を与えている。また逆にこの短単位の接尾語を複合していく際、単純に並べていけばよいというわけではなく、綴りを変化させる必要が生じてくるが、これをLinking ruleとして規則化している。

(2) シンハラ語では、主語と目的語は一般に接尾語によって表現される格によって決められるが、主格と目的格が同形である場合も多く、語順によってそれが決まる場合もあり、また単語の意味や、文脈的情報に依らなければ決まらない場合もある。このような種々の場合を考慮して、主語と目的語を決定するアルゴリズムを与えている。

(3) 日本語からシンハラ語への機械翻訳に関する手法を提案し、その有効性を実証している。一般に機械翻訳は原言語の構文構造と目的言語の構文構造の間の変換を介して行われるが、本論文では日本語とシンハラ語の間の類似性を利用して、形態素レベルの構造

の変換で翻訳出来ることを述べている。これは日本語の文節に対応するシンハラ語の単位があることを見出したことに基づいており、この単位をP-unitと名づけている。P-unitは一つあるいは複数のL-unitからなる。日本語の機能語(助詞、助動詞)からP-unitへの対応規則を具体的に作成して、この対応規則を設定することが可能なことを実証している。シンハラ語は日本語と同じように、あるいはそれ以上に語順は自由であり、また同じように構文要素の省略が行われるが、日本語の形態素解析の結果として得られる文節を、そのままの順序でこの対応規則によってP-unitに変換し、そこからシンハラ語の表層構造を生成することで、複合文、埋め込み文の場合を含めて、十分に理解し得る正しいシンハラ語が得られることを、多くの例文を通じて実証している。また、計算機上に簡単な翻訳システムの構築も行っている。

(4) 日本語からシンハラ語のP-unitへの対応規則は実際には1対多となる場合も多く、その場合どの対応規則を選べばよいか分からないということになる。これへの対処法として、格構造を介して対応をとることを提案している。多くの例文の格構造について検討し、格構造を介することで対応規則の曖昧さを解決できることを実証している。

以上、本論文は、日本語からシンハラ語への機械翻訳の手法について提案し多くの例文を通じてその有効性を実証している。シンハラ語の計算機処理に関しての研究はまだ少なく、本論文はパイオニア的な研究の一つに位置づけられる貴重なものであり、学術上、実際上の価値は極めて高い。よって、本論文は博士(工学)の学術論文として価値あるものと認める。