様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　3 月　26 日現在

研究種目：基盤研究（C）
研究期間：平成 19 年度～平成 21 年度
課題番号：19500119
研究課題名（和文）　深い言語知識に基づくパターン変換型機械翻訳システムに関する研究

研究課題名（英文）　A Study of Machine Translation System using Pattern-transform method with deep linguistic knowledge

研究代表者：　池田尚志(IKEDA TAKASHI)
岐阜大学・工学部・教授
研究者番号：10232183

研究成果の概要：古典的な方式であるが、パターン変換型の機械翻訳エンジンを構築し中国語を始めベトナム語、シンハラ語、さらに日本の手話への機械翻訳システム jaw を試作した。C++言語のオブジェクトのパラダイムを利用した相手言語の“表現構造”を介する翻訳方式が特徴である。また、jaw において用いる日本語解析システムとして、文節構造解析に基づく解析システム ibukiC を開発した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
平成 19 年度	1,000,000	300,000	1,300,000
平成 20 年度	1,100,000	330,000	1,430,000
平成 21 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,100,000	930,000	4,030,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理

１．研究開始当初の背景

機械翻訳の技術はこれまで長年に渡って多くの研究がなされ発展してきた。今日では使い方を工夫すれば有効に利用できると評価される翻訳ソフトがいくつも市販されている。しかし、たとえば技術文書やビジネス文書などに一般的に使用できるレベルが達成されているわけではなく、完成・成熟した技術と言うにはまだはるかに遠いというのが実情である。一方で翻訳の需要は、インターネット・通信技術の飛躍的な進歩によってますます大きくなってきている。また、英語を中心とするメジャーな言語の間だけでなく、たとえばベトナム語やモンゴル語などいわゆるマイナーとされる言語も含めた言語処理研究、対照言語的研究、機械翻訳研究の意義・重要性も高まってくるのも必然の流れというべきである。IT 技術の進歩、言語処理研究の蓄積を踏まえ、今日の時点での機械翻訳研究の重要性と可能性は大きい。

ところで現在の機械翻訳の研究の主流は、大規模コーパスに基づく統計的手法あるいは用例ベースの手法である。これは電子化された大規模な言語資料が利用できるようになってきたことによるものであり、探求すべき重要な手法である。しかし本課題の申請者らは、統計的な手法は限定された処理の場面では有効であるが、それだけでは言語の全般

を捉えていくのには無理があると予測している。言語の表層のみに注目するいわば力ずくの方法だけでは、言語の全般を捉えていくのは困難であり、深い言語知識に基づいた言語処理の規則、機械翻訳の規則を探求するいわば古典的な方法を併せていくことがやはり必要であると考える。またそもそも大規模な電子化対訳コーパスは未だ存在しない。英語を一方の言語とする対訳コーパスは分野によってはある程度の規模で存在するが、全般的な分野に対して十分な量が存在するとは言えない。ましてベトナム語やシンハラ語などアジアの諸言語については対訳コーパスはほとんど存在せず統計的手法の適用は不可能と言わざるを得ない。

我々は、これまでに日本語からいろいろの言語に翻訳することのできる機械翻訳エンジンを設計・開発し、日本語から中国語、ベトナム語、シンハラ語、ミャンマー語などアジアの言語への機械翻訳システムを試作してきた[2006,S.Thelijagoda, et al.][2006, N.M.Chau, ほか][2005, Z.BU, et al.][2005, ト, ほか][2004, 謝, ほか] {その他, 6 ページの研究業績リスト参照}。また、このエンジンを用いて手話テキストへの翻訳にも挑戦している[2006, 松本, ほか]2005, Tadahiro Matsumoto, et al.]。

この翻訳エンジンは、原言語（＝日本語）をパターン翻訳規則と照合し相手言語の“表現構造”に変換する。“表現構造”は、相手言語での表現を生成するために必要な言語表現の部品や生成のために必要な各種の情報を集積しているもので、実装には C++のオブジェクトをそのまま用いている。助詞･助動詞あるいは複合機能語などの機能語は日本語で重要な役割を果たしているが、その翻訳についてはパターン翻訳とは別の規則で“表現構造”上に必要な情報を書き込む。相手言語の文は、このオブジェクト上に集積されている情報を線状に組み起こして作成する。実装には C++のクラスメソッドを用いている（線状化関数と称している)。“表現構造”への変換の過程に“深い言語知識”から導かれる翻訳規則が組み込まれ、線状化関数に相手言語の言語知識が組み込まれる。このように“表現構造”を介しての翻訳が我々のシステムの特徴である。

２．研究の目的

これまでに、以上述べた翻訳エンジンと翻訳規則を記述するためのエディター等を開発し、前述したアジアのいくつかの言語への翻訳システムを、それぞれの国からの留学生との協同で試作してきており、本手法の有効性についての見通しは得られている。本研究では、中国語への翻訳システムや手話への翻訳、また機械翻訳に必要な日本語文の解析システムなど、これまでに取り組んできた研究の成果を発展させることを目的とした。

日中翻訳に関しては、これまでに連体埋め込み表現、取立て詞による表現、テンス・アスペクトに関する表現などについての翻訳規則について研究してきた。この研究期間である３年の間に、これらをさらに発展させて以下の内容を目標として、大学院博士課程留学生や日本人修士学生らとともに研究を進めることとした。

(1) これまでに設計してきたこれらの表現の翻訳規則についての見直し・整理を行い、またシステムへの実装が不十分であった部分についての実装を行う。これによってさらに問題点の発掘、翻訳規則の精緻化を進める。
(2) 存在表現、使役表現、授受表現等について分析し、翻訳規則の設計･実装を行う。
(3) 接続表現、埋め込み表現などを含むさまざまの文種の表現 1000 例文程度を収集し、システム上に翻訳規則を実装して、問題点の発掘･分析を行い、試作システムを拡張する。

日本語の言語表現において、機能語（助詞、助動詞、複合機能語など）の果たす役割は本質的に大きく、機械翻訳においてもその扱いは重要である。従来、これら機能表現に関わる部分の翻訳手法は十分には研究されてきていない。我々は、これまでに長単位の機能語約２万語を収集し、それに基づく文節構造解析システムを開発してきた。本研究では、この２万語の分析を深化し、解析システムの整備を行う。またその中国語への翻訳規則について分析する。

(4) 機能語と本動詞の識別が必要となる表現{たとえば「を通して」} の収集およびそれを識別するための共起表現の収集
(5) 長単位機能語辞書の整備 {時制、判断、接続など機能要素への分割、標準表現への換言など}
(6) 中国語へ翻訳するための規則の設計･実装

以上のほか手話に関す自然言語処理の研究も行う。手話も日本語や中国語と同じく“言語”である。本研究では、手話に関する自然言語処理の研究の一環として、日本語から手話への機械翻訳の研究も行う。

(7) 我々が提案している手話の表記法(jjs-notation)を用いて､日本語から手話テキストへの機械翻訳について研究する。これまで手話文字システム(JSPad)を含む手話に関する研究を行ってきている松本助教と学生達との共同で研究を進める。

３．研究の方法

図１は我々が開発している翻訳エンジンの概要である。日本語文の分析、表現パターン翻訳規則との照合による変換と、変換によって作り出された目的言語の内部表現（我々は目的言語の“表現構造”と呼んでいる）からの目的言語の生成（我々は表現構造からの線状化と呼んでいる）というトランスファー方式の枠組みである。

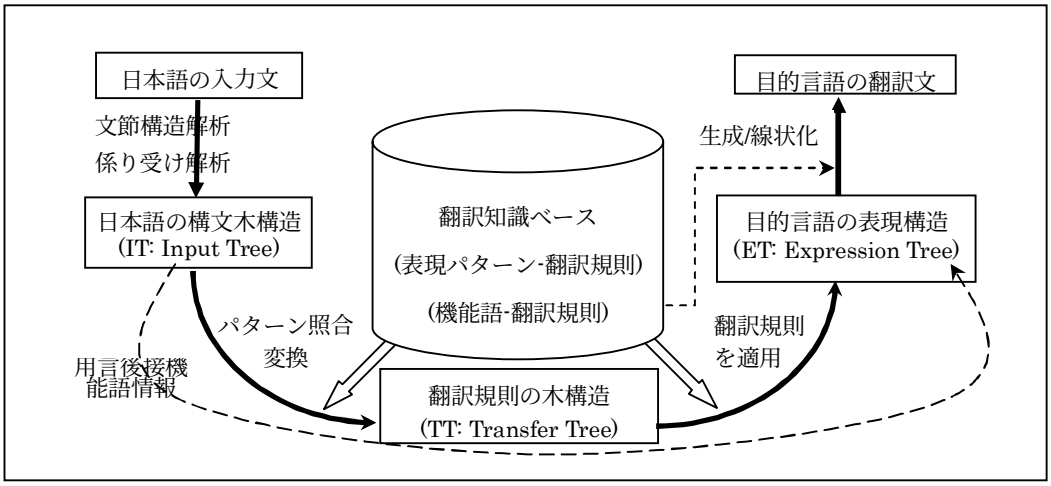

図 1．翻訳エンジンの構成

(1) 入力日本語文の解析システムの整備を進める。文節構造解析システムの長単位機能語辞書を整備し精度の向上を図る。機能語と本動詞の曖昧性除去のための共起表現情報を収集し、システムに実装する。
(2) 構文解析システムについては、変換規則のパターン照合と整合するように､解析システムの細部の設計･プログラミングを行い、規則辞書の整備を行う。
(3) 変換・生成部分のシステム、および翻訳規則エディターのブラッシュアップを進める。
(4) 日中翻訳におけるテンス･アスペクト､取立て詞の翻訳規則を整理･見直し。実装する。
中国語の文型を分析し、線状化関数の整理･見直しを行う。
(5) 存在文の日中対応に関する言語的分析を行い、翻訳規則を設計し実装する。
(6) 副詞、接続詞、機能語の日中対応に関する言語的分析を行い、翻訳規則を設計し実装する。
(7) 入手できる日英対訳コーパスなどから、接続表現、埋め込み表現などを含むさまざまの文種の表現300例文程度を収集し、日中翻訳の実験を行う。問題点を発掘し分析する。

日本語の解析に関わる部分は日本人大学院生の協力を得て行う。変換部分は日本人大学院生と留学生（修士および博士）との協同で行う。相手言語生成の部分は留学生（博士）が中心になって行う。研究代表者（池田）は全体の設計、統括、言語知識の分析を行う。研究分担者（松本）は、パターン翻訳規則、生成関数の実装に関わるシステム面の整備を分担する。

４．研究成果

(1) 我々が開発してきたパターン変換型機械翻訳システム jaw で用いている日本語解析システム ibukiC の側で、機能語部分の解析を機能文節の概念に基づく方法に改版した。 ibukiC については、辞書などこの他の点についても整備を進め、研究室のホームページで公開した。
また機械翻訳エンジン jaw の側でもこれに対応する形に改版し､jaw/Chinese において機能表現に関する翻訳規則を実装して翻訳実験を行い、改版による翻訳規則を柔軟に記述出来ることに関する効果を確認した。

(2) 日中機械翻訳システムについては、存在表現軽動詞構造表現について分析し、翻訳手法・翻訳規則を提案した。また、使役・授受表現に関して機械翻訳の観点からの日中間の日中間の表現の対応について分析を深め、機械翻訳規則としてまとめた。

(3) 機械翻訳エンジン jaw については照合アルゴリズム、多階層の係り受けを含むパターンの処理、機能文節に関する翻訳規則の処理などについての開発を行った。

(4) 日本語文の解析に関して、これまでに主として整備してきた硬い下記言葉を中心とする辞書項目に加え、日常語文を解析するための辞書の整備を行った。
表記ゆれを含む文を解析するための方式の開発、誤り箇所推定処理方式の開発、係り受け解析の曖昧さを絞り込む方式の開発なども行った。
ibukiC の係り受け解析では、係り受けの曖昧さを許す（複数の係りを許す）形での解析を行っているが、本年度は受けの側の格構造を利用して複数の係りを絞り込むことに関する研究を行い成果を得た。
表記のゆれを含む文を解析するための方策について考察し ibukiC に実験的に実装した。表記ゆれを含む文の解析は出来るようにはなったが、処理時間のさらなる短縮が今後の課題である。

５．主な発表論文等
〔雑誌論文〕（計 4 件）
池田尚志, 日本語からアジア諸言語への機械翻訳システムの構築　奮闘記　―中国語へ, ベトナム語へ,シンハラ語へ,日本の手話へ―, 日本語学, 査読無し, 28 巻 12 号, pp62-71, 2009

Tadahiro Matsumoto, Mihoko Kato, Takashi Ikeda, JSPad―A sign Language Writing Tool Using SignWriting, 3rd International University Communication Symposium(IUCS-2009), 査読あり pp.363～367, 2009

王軼諞, 池田尚志, 日中機械翻訳における存在表現の翻訳処理について, 査読あり, 自然言語処理, 14-5, pp65-105, 2007

Samantha Thelijjagoda,Yoshimasa Imai and Takashi Ikeda, Japanese-Shinhalese machine translation system Jaw/Shinhalese, Journal of the National Science Foundation of SriLanka, 査読あり, 35-2, pp81-96, 2007

〔学会発表〕（計 16 件）
薛明恵, 池田尚志, 機械翻訳システム jaw における機能語独立文節に関する処理―jaw/Chinese における事例―, 言語処理学会第 16 回年次大会, PB2-4, 2010(東京大学)

黄暁兵, 池田尚志, 日中機械翻訳における「なる」構文の翻訳処理について, 言語処理学会第 16 回年次大会発表論文集, PB2-2, 2010（東京大学）

脇田貴之,安藤健二郎,太田哲也,池田尚志, 日本語文解析システム ibukiC と文節解析の曖昧さ解消および日常語テキストの解析, 言語処理学会第 15 回年次大会, pp.809-812, 2009
（鳥取大学）

玉置健二,角田慶太,松本忠博,池田尚志, 機械翻訳エンジン jaw について, 言語処理学会第 15 回年次大会, 216-219, 2009（鳥取大学）
吉村康寛,松本忠博,池田尚志, 日本語-手話機械翻訳システム jaw/SL による翻訳実験, 言語処理学会第 15 回年次大会, pp.516-519, 2009（鳥取大学）

Yiou Wang, Takashi Ikeda, Translation of the Light Verb Constructions in Japanese-Chinese Machine Translation, 9 th International Conference on Intelligent Text Processing and Computational Linguistics, Advances in Natural Language Processing and Applications Research in Computing Science33,pp.139-150, 2008(University of Haifa)

Yiou Wang, Takashi Ikeda, Japanese-Chinese Machine YTranslation for *Suru* Expressions, Proceedings of the international Conference on Kinguistic Evidence: Empirivcal,theoretical and Computational Perspectives(LingEvid2008), pp204-207, 2008(University of Tubingen)

池田尚志,脇田貴之,大口智也, 機能文節を導入した文節構造解析システム ibukiC(v2.0)について,言語処理学会大 14 回年次大会,pp. 221-224, 2008（東京大学）

黄暁兵,王軼諞,薛明恵,池田尚志, 日中機械翻訳における使役等の翻訳処理について, 言語処理学会第 14 回年次大会, pp.877-880, 2008（東京大学）

６．研究組織
(1)研究代表者
池田尚志(TAKASHI IKEDA)
岐阜大学・工学部・教授
研究者番号：10232183

(2)研究分担者
なし

(3)連携研究者
なし