様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　５　月　３１　日現在

機関番号：　１３９０１
研究種目：　若手研究(B)
研究期間：　2012～2015
課題番号：　２４７００１３６

研究課題名（和文）プログラム合成・分解による機械翻訳

研究課題名（英文）machine translation through program synthesis and decomposition

研究代表者
松崎　拓也（Takuya, Matsuzaki）

名古屋大学・工学（系）研究科（研究院）・准教授

研究者番号：４０４６３８７２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：組合せ範疇文法を用いた深い構文・意味解析に基づく機械翻訳へむけた基礎研究を行った．具体的には，日本語の基本的な構文を幅広く解析できる日本語文法の半自動獲得，より精緻な意味表示を得るための意味表現・合成枠組みの開発および文法への実装，既存の係り受け解析器を用いた構文解析における曖昧性解消の実現，および翻訳エラーが会話理解におよぼす影響の調査といった項目に対し研究を行った．

研究成果の概要（英文）：Fundamental technical issues were studied toward machine translation through deep syntactic and semantic analysis based on Combinatory Categorial Grammar. Specifically, a Japanese grammar that covers various basic constructions in Japanese was implemented and a parsing and disambiguation technique was developed. The parsing technique is based on chunk-level dependency analyzer trained on shallow annotated corpora and can be adopted to various domain. Furthermore, the effect of translation errors on the comprehension of dialogue was experimentally studied.

研究分野：　自然言語処理

キーワード：　自然言語処理　日本語文法　機械翻訳

１．研究開始当初の背景

(1) 機械翻訳は自然言語処理において最も古くから続く研究課題の一つであり、現在も国内外を問わず活発な研究が続いているが、未だ完全な実用化には至っていない。特に、日本語-英語、日本語-中国語といった異なる性質を持つ言語間での機械翻訳結果は、人間による翻訳に比べはるかに低い品質にとどまっている。

(2) １９９０年代以降、機械翻訳の分野では対訳データからの学習による統計的アプローチが主流となっており、現在先端的な研究では、構文解析によって得た文の統語構造を統計的手法と組み合わせる手法の開発が盛んに行われている。しかし、翻訳手法の工夫による精度の向上は頭打ちに近い状況にあり、WEB から抽出した大量の対訳データを単純な翻訳モデルと組み合わせることで精度をかせぐ、といったその場しのぎのアプローチではない、本質的な技術革新によるブレークスルーが期待されている。

(3) 現在主流の統語に基づく統計的機械翻訳技法では、係り受け解析や単純な句構造文法を用いた構文解析結果を用いるものがほとんどである。これら表層的な構文解析結果からは、並列構造や長距離の依存などを含む文に対し正確な意味を直接得ることは難しい。しかし、並列や長距離依存といった構造は自然な文においても頻出するものである。このように、構文解析のための文法枠組みが単純すぎるために十分に文の意味構造を解析できていないことが、現在の統語に基づく機械翻訳が、単語や単語列といったより表層的な情報に基づく単純な手法に比べ大きな精度向上に至らない本質的な原因であると考えられる。

(4) 主辞駆動句構造文法（HPSG）、組合せ範疇文法（CCG）などの現代的な文法枠組みでは、上記のような複雑な文法構造をもつ文に対してもその意味構造を直接得ることができる。このため、これらの文法枠組みを用いた構文解析を統計的翻訳手法へと融合することで、翻訳精度が大きく向上すると期待できる。しかし、これらの理論的な枠組みに基づき，深い意味解析を可能とする詳細な日本語文法は存在しなかった．また，これら現代的な文法では、意味構造を含む深い解析を行うために複雑な計算・データ構造が用いられているため、文法と整合的に動作する翻訳ルールの獲得や、統計的な翻訳ルール選択モデルとの融合のためにはいくつかの技術的課題を解決する必要があった。

２．研究の目的

精緻な統語解析および形式的な意味表示の出力を可能とする語彙化文法の手法に基づき，統語構造・文意味構造・文脈情報といった多レベルの情報を統一的な枠組みで機械翻訳技術へと導入するための基礎研究を行う．具体的には組合せ範疇文法に基づく日本語文法の開発，曖昧性解消を含む統語・意味解析技術の開発，および現在の統計的機械翻訳技術の限界を探り，多層的な情報による翻訳精度改善へ向けた指針を得るためのユーザスタディを行う．

３．研究の方法

(1) 深い言語解析に基づく機械翻訳のための基礎資源となる日本語文法の開発を行う．特に，分野を問わず必須となる種々の基本構文の正確な解析を可能とするための機能語の統語・意味特性の形式的記述や，機械翻訳への期待が大きい技術文書で重要となる名詞句の意味構造の詳細な解析に注力する．

(2) 語彙化文法を用いた構文・意味の並行的な解析技術の研究開発を行う．深い統語・意味解析における膨大な曖昧性を効率的に解決するための手法を開発する．詳細な文法に基づく解析済みコーパス作成のコストは高いため，ごく少量の深い解析結果を含む訓練データおよび大規模だが浅い解析のみを含む訓練データを有効に用いた学習手法についても研究を行う．

(3) 現在の機械翻訳の出力において，文書理解の阻害原因となる翻訳エラーのタイプを大規模なユーザスタディによって調査し，技術的改善点を見定めるとともに，機械翻訳技術全体の限界を見定める．

４．研究成果

(1) 浅い解析のみを施したコーパスから日本語の大規模な語彙化文法を半自動的に得る試みを行った．これは，文節係り受け構造のような浅い解析のみを人手で施したコーパスに対して，語彙化文法に基づく再解釈を行う種々の変換ルールを適用することで，精緻な文法による解析結果を大量に生成し，そこから辞書を抽出するアプローチによるものである．具体的には，まず京都大学テキストコーパスで与えられる文節係り受け情報および NAIST テキストコーパスで与えられる述語項構造の情報を統合し，文全体の句構造および各述語の項となる句を同定した解析結果を中間的に生成した．この句構造木の各節点に対して，組合せ範疇文法の規則ラベルを指定し，各単語に対して統語的特性を部分的に指定することで，組合せ範疇文法による文の導出が得られる．これにより，数万文に対する深い解析結果を得ることに成功した．この半自動生成されたコーパスを文法抽出

および構文解析器の訓練に用いることで，被覆率の高い日本語語彙化文法および解析器を得ることができた．

(2) コーパスから半自動獲得した文法・辞書では，述語の種々の格フレームに対応した語彙項目がコーパスでの出現例をもとに一挙に得られるという利点がある．しかし，表層的な情報からは判別不能な深い意味構造を半自動的にコーパスに付与することは極めて困難であり，結果として，半自動獲得した文法ではそのような詳細な意味構造の差を捉えることはできない．特に，機械翻訳の応用分野として期待が大きい技術文書の翻訳で必要となるような詳細な意味構造の差を捉えるためには，名詞句が表すエンティティの単数・複数の区別や，普通名詞および固有名詞の区別，述語に対する項の分配および集合的な読みの区別といった様々な意味的差異を捉えることが可能な文法を用いる必要がある．さらに，上記のような名詞句に関連した意味構造に関連する一連の統語・意味現象として，エンティティ間の関係やエンティティの属性を表す，名詞句を中心とした構文にまつわる現象がある．これらの，いわゆる不飽和名詞を中心とした種々の構文の正確な解析は，例えばソフトウェアの仕様など，数学的な側面が強い技術文書の解析・翻訳において極めて重要である．上記のような種々の統語・意味現象を捉える精緻な日本語文法の記述を行った．

また，従来の文法における統語・意味インタフェースでは捉えにくい現象として，数学テキストなど，変数名が名詞の一部として表れる文における変数名の意味解釈の問題がある．これは，変数を含む名詞句の意味内容が文脈に深く依存し，かつ，変数の意味解釈として「変数の名前」および「変数の値」の 2 通りの解釈が存在するために，いわゆる構成性原理に従わないように見える意味合成が必要となるという問題である．これに対し，文脈を表すオブジェクトを単語の意味記述における操作対象とすることで，上記のような現象を含む精緻な意味解析を可能とする意味表示および意味合成の枠組みを開発した．

(3) 上記の，人手によって記述した精緻な文法に基づく解析では，種々の意味構造の差を捉えるために多種の文法規則が用いられ，その中間結果には多様な統語構造が現れる．このため，現実的な時間で解析を行うためにはこれまでとは異なる解析アルゴリズムが必要となる．また，従来の教師付き学習に基づく解析アルゴリズムでは，学習データとして大量の解析結果が必要となる．しかし，精緻な文法による解析結果を大量に作成することはコストが高いという問題がある．

これらの問題を解決するため，文節単位の係り受け構造をあらかじめ定めた後，その構造と整合する範囲内で組合せ範疇文法による導出を探索する手法を開発した．この手法は，文節係り受け構造の制約のもとで探索を行うことで解析の計算コストを削減し，かつ，文節係り受け構造のみを付与したコーパスのみから学習した係り受け解析器があれば曖昧性解消ができるという利点がある．

この手法には文節係り受け構造の決定の段階で誤った場合，組合せ範疇文法による正しい導出が得られないという問題がある．これに対しては，係り受け解析の際に複数の解析を出力（n-best 解析）し，スコアの高い係り受け構造から順に 2 段階目の解析の入力として組合せ範疇文法による解析を行い，最初に得られた完全な導出を出力とするという手法によってほぼ解決した．

(3) 統計的機械翻訳の出力における翻訳誤りには種々のものがある．これらの誤りタイプのうち，翻訳結果の理解において大きな影響を与えるものを同定し，そのうち，語彙化文法に基づく翻訳によって解決可能なものを見定めることで適切な手法設計を行う必要がある．また，語彙化文法の利用を含め，現在可能な機械翻訳の手段では到底解決できないタイプの翻訳エラーに対しては，機械翻訳の利用形態そのものを含め，実践的な利用フローを想定し，それに応じた機械翻訳手法を設計する必要がある．

このような目的のため，現在の機械翻訳システムおよび人手による翻訳がテキスト理解に与える影響を比較調査する実験を行い，その結果を分析した．具体的には，大学入試センター試験「英語」から対話の理解に関する多肢選択式問題を収集し，複数の機械翻訳システムおよび人手によって日本語訳した問題を被験者に解かせ，翻訳システムの違いによる正答率への影響を調査した．

この結果，統計的手法に基づく翻訳システムに比べ，ルールベースの手法による翻訳システムの結果に対する正答率が有意に高いこと，また，文脈を考慮できない設定で行った人手による翻訳と，ルールベース翻訳による結果に対する正解率は同程度であることが明らかになった．

さらに，各翻訳システムの出力における翻訳エラーをタイプ分類し，正答率，回答に対する自信度，および翻訳システムの主観的評価に対する個々のエラータイプの影響を定量的に調査した．この結果，統語・意味解析に基づく翻訳が有効と思われる訳語選択レベルの翻訳誤りは多数存在し，かつ正答率に大きく影響すること，述語に対する項の表層格の選択は多数あるが正答率には大きく影響しないこと，文脈の深い理解が関係する翻訳エラーは正答率に大きく影響するが，実際に観測されるこのタイプのエラーを防ぐのは手法を問わず困難とみられること，などが明らかになった．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① Sumire Uematsu, Takuya Matsuzaki, Hiroki Hanaoka, Yusuke Miyao, Hideki Mima, Integrating Multiple Dependency Corpora for Inducing Wide-Coverage Japanese CCG Resources, ACM Transactions on Asian and Low-Resource Language Information Processing, 査読あり, 14(1), 2015, DOI: 10.1145/2658997.

〔学会発表〕（計 5 件）

① Takuya Matsuzaki, Akira Fujita, Naoya Todo and Noriko H. Arai, EVALUATING MACHINE TRANSLATION SYSTEMS WITH SECOND LANGUAGE PROFICIENCY TESTS, the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2015), 2015 年 7 月 26 日, 北京（中国）

② Takuya Matsuzaki, Hidenao Iwane, Hirokazu Anai, Noriko H. Arai, The Most Uncreative Examinee: A First Step toward Wide Coverage Natural Language Math Problem Solving, The 28th Conference on Artificial Intellegence (AAAI 2014), 2014 年 7 月 27 日, ケベック（カナダ）

③ Takuya Matsuzaki, Hidenao Iwane, Hirokazu Anai and Noriko Arai, The Complexity of Math Problems -- Linguistic, or Computational?, The 6th International Joint Conference on Natural Language Processing, 2013 年 10 月 15 日, 名古屋国際会議場（愛知県・名古屋市）

④ Sumire Uematsu, Takuya Matsuzaki, Hiroki Hanaoka, Yusuke Miyao, Hideki Mima, Integrating Multiple Dependency Corpora for Inducing Wide-coverage Japanese CCG Resources, The 51st Annual Meeting of the Association for Computational Linguistics, 2013 年 8 月 4 日, ソフィア（ブルガリア）

⑤ Takuya MATSUZAKI, Toward Wide-Coverage Natural Language Mathematical Problem Solving, The 9th International Workshop on Logic and Engineering of Natural Language Semantics, 2012 年 11 月 30 日, アミューズメントゾーン宮崎（宮崎県・宮崎市）

〔図書〕（計 1 件）

① 畠山雄二（編著）、本田 謙介 (著), 今仁 生美 (著), 松崎 拓也 (著), 宮尾 祐介 (著), 但馬 康宏 (著), 田中 江扶 (著), 数理言語学事典, 2013 年, pp. 88-97.

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
http://researchmap.jp/mtzk/

６．研究組織
(1)研究代表者
松崎 拓也（MATSUZAKI, Takuya）
名古屋大学・大学院電子情報システム専攻・准教授
研究者番号：40463872