様　式　Ｆ－７－１

# 科学研究費助成事業（学術研究助成基金助成金）実施状況報告書（研究実施状況報告書）（平成２５年度）

1．機関番号　1 4 6 0 3　　2．研究機関名　奈良先端科学技術大学院大学

3．研究種目名　若手研究(B)　　4．補助事業期間　平成２５年度～平成２７年度

5．課題番号　2 5 7 3 0 1 3 6

6．研究課題名　訳選択の根拠の自動推定とその機械翻訳における応用

7．研究代表者

| 研究者番号 | 研究代表者名 | 所属部局名 | 職名 |
|---|---|---|---|
| 70633428 | ニュービッグ　グラム<br>Ｎｅｕｂｉｇ　Ｇｒａｈａｍ | 情報科学研究科 | 助教 |

8．研究分担者

| 研究者番号 | 研究分担者名 | 所属研究機関名・部局名 | 職名 |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

9．研究実績の概要

本年度は、機械翻訳における訳選択の精度向上に向けた調査とシステム構築に取り組み、主に3つの研究成果があった。
１つ目の成果は、実験のベースとなる翻訳システムの構築である。人手により構築されたルールベース機械翻訳(RBMT)の知見を統計的機械翻訳(SMT)に取り入れるために、RBMTと類似した形のSMTシステムが必要となる。これを実現するために、文の構造を利用したSMTシステムを構築し、オープンソースソフトとして公開した。また、システムの実験的評価において、文の構造を英日・日英機械翻訳に直接取り入れることで、既存の翻訳手法を大幅に上回る翻訳精度を実現できた。
２つ目の成果は、SMTに用いる対訳データの小規模化に関する研究である。データを小規模化することにより、本研究の目標であるモデルの小規模化を実現することができるが、単純にデータをランダムに選択すると大幅な精度低下が起こり得る。そこで、大量のデータの中から、頻繁に起こる対訳パターンを特定し、この対訳パターンを確実にカバーするデータを選択する手法を確立した。この対訳データを学習に利用することで、精度の低下を防ぎながらモデルを小規模化できることを、実験的評価により確認した。
３つ目の成果は、訳選択の根拠を自動的に発見するのに欠かせない自動評価尺度の調査である。調査の結果、既存の評価尺度の問題を特定し、訳選択の正確性を正しく評価できる新たな評価尺度の確率の重要性を明らかにした。

10.キーワード

| | | | |
|---|---|---|---|
| (1) 機械翻訳 | (2) 訳選択 | (3) 自然言語処理 | (4) 機械学習 |
| (5) 評価尺度 | (6) | (7) | (8) |

11.現在までの達成度

（区分）（2）おおむね順調に進展している。

（理由）

25年度の目標である「人間の訳選択の根拠に関する調査」に関しては、対外発表に至っていないが、データの収集と初期の分析を完了している。この分析で得られた知見に基づいて、訳選択誤りの分類が確立しつつあり、この分類に基づいてさらにデータを作成してもらう予定である。
25～26年度の目標である「訳選択の根拠の自動発見技術の開発」に関しては、基礎的な学習アルゴリズムを開発し、小規模なデータに対して確認済みである。また、ルールベース翻訳の知見を人手で統計翻訳に取り入れた実験も行っている。
26年度以降の目標である「訳選択の根拠を考慮した翻訳システムの構築」に関しては、ベースとなる構文情報を用いた翻訳システムの構築が完了した。訳選択の根拠を自動的に発見する技術の開発をこの枠組みと同時に開発しており、すぐに適応可能である。
このことから、25年度の目標は未完成な部分がある一方、26年度以降の目標は大幅に前倒しに進んでいることから、研究はおおむね順調に進んでいると言える。

12.今後の研究の推進方策　等

（今後の推進方策）

26年度の予定として主に3つの課題に取り組む予定である。
まず、「人間の訳選択の根拠に関する調査」に関しては、25年度に考案した誤りの分類に基づいて、大規模なデータを作成する予定である。このデータ作成が終了してから、分析を行い、翻訳誤りと見なされる条件について考察を行う。また、誤りと見なされない翻訳の揺れを許しながら、誤りと見なされる翻訳の揺れを許さない翻訳ルール獲得枠組みを考案する。
また、「訳選択の根拠の自動発見技術の開発」に関しては、25年度に開発した学習アルゴリズムを大規模データで利用できるように拡張するとともに、前述の考察の結果を取り入れたルールを学習する枠組みを考案する。特に、RBMTシステムに利用されている情報（例えば、動詞の訳出を選択する時の項のカテゴリー）に着目する。
最後に、「訳選択の根拠を考慮した翻訳システムの構築」に関しては、25年度にベースとなるシステムが完成されたため、訳選択の根拠の自動発見技術が完成すれば、すぐに実環境の翻訳実験も実行可能である。

（次年度使用額が生じた理由と使用計画）

（理由）

平成25年度の目標の1つであった誤り情報付きコーパスの作成が主な原因である。その理由として(1)小規模コーパスに同様のアノテーションを行った際、アノテーション基準の詳細を検討する必要があることが発覚したこと（2)一部平成26年度に予定していたシステムの作成が思ったより早く平成25年度に行うことになったことが挙げられる。

（使用計画）

現在、アノテーション基準は完成しており、コーパスの作成を業者に発注しているため、これに当たる作業にかかる費用を繰越分で満てる予定である。

13.研究発表（平成25年度の研究成果）

〔雑誌論文〕 計( 2 )件　うち査読付論文 計( 2 )件

| 著者名 | 論文標題 |
|---|---|
| Graham Neubig | Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers |

| 雑誌名 | 査読の有無 | 巻 | 発行年 | 最初と最後の頁 |
|---|---|---|---|---|
| Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL) | 有 | - | 2013 | 91-96 |

| 掲載論文のDOI（デジタルオブジェクト識別子） |
|---|
| なし |

| 著者名 | 論文標題 【掲載確定】 |
|---|---|
| Graham Neubig, Kevin Duh | On the Elements of an Accurate Tree-to-String Machine Translation System |

| 雑誌名 | 査読の有無 | 巻 | 発行年 | 最初と最後の頁 |
|---|---|---|---|---|
| Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL) | 有 | - | 2014 | 未定 |

| 掲載論文のDOI（デジタルオブジェクト識別子） |
|---|
| なし |

〔学会発表〕 計( 3 )件　うち招待講演 計( 1 )件

| 発表者名 | 発表標題 |
|---|---|
| Graham Neubig | 文レベルの機械翻訳評価尺度に関する調査 |

| 学会等名 | 発表年月日 | 発表場所 |
|---|---|---|
| 情報処理学会 第212回自然言語処理研究会 | 2013年07月18日～2013年07月19日 | 北海道 函館 |

| 発表者名 | 発表標題 | |
|---|---|---|
| 丹生 伊左夫, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲 | 構文情報を利用した対訳データ選択手法 | |
| 学会等名 | 発表年月日 | 発表場所 |
| 言語処理学会第20回年次大会 | 2014年03月18日～2014年03月20日 | 北海道 札幌 |

| 発表者名 | 発表標題 【発表確定】 | |
|---|---|---|
| Graham Neubig | 機械翻訳 ～なぜできなかったのか？なぜできるようになりつつあるのか？～ | |
| 学会等名 | 発表年月日 | 発表場所 |
| 音学シンポジウム2014 (招待講演) | 2014年05月25日～2015年05月26日 | 東京 |

〔図書〕 計( 0 )件

| 著者名 | 出版社 | |
|---|---|---|
| | | |
| 書名 | 発行年 | 総ページ数 |
| | | |

14.研究成果による産業財産権の出願･取得状況

〔出願〕 計( 0 )件

| 産業財産権の名称 | 発明者 | 権利者 | 産業財産権の種類、番号 | 出願年月日 | 国内･外国の別 |
|---|---|---|---|---|---|
| | | | | | |

〔取得〕 計（ 0 ）件

| 産業財産権の名称 | 発明者 | 権利者 | 産業財産権の種類、番号 | 取得年月日<br>出願年月日 | 国内･外国の別 |
|---|---|---|---|---|---|
| | | | | | |

15.備考

Travatar: A Tree-to-String Translation Toolkit
http://www.phontron.com/travatar/
上記のページは本研究のために開発したソフトの公開ページである。