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　 ６ 月 １０ 日現在

機関番号：１４６０３
研究種目：若手研究(B)
研究期間：2013～2015
課題番号：２５７３０１３６

研究課題名（和文）訳選択の根拠の自動推定とその機械翻訳における応用

研究課題名（英文）Automatic Prediction of Reasons for Translation, and Use within Machine Translation Systems

研究代表者
Ｎｅｕｂｉｇ　Ｇｒａｈａｍ（Neubig, Graham）

奈良先端科学技術大学院大学・情報科学研究科・助教

研究者番号：７０６３３４２８

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：近年の機械翻訳の進歩に大きく貢献する技術として、対訳データから翻訳ルールを自動的に学習する統計的機械翻訳（SMT）がある。しかしSMTは明示的にある翻訳結果が選ばれる訳の根拠を考慮していないため、人手で構築するルールベース機械翻訳(RBMT)に比べて一部の入力文や言い回しに対して致命的な誤訳を生成することがある。本研究は、人間の翻訳者が用いる根拠を持つ訳選択ルールを自動発見し、SMTと融合する方法の開発を行った。具体的には、言語的情報に基づく翻訳システムの開発、訳選択の根拠の導入、SMTシステムの分析を効率化する枠組みの開発、多言語データにおける訳選択の根拠を発見する手法の開発を行った。

研究成果の概要（英文）：One of the reasons for the recent progress in the field of machine translation is the rise of statistical machine translation (SMT), which automatically learns translation rules from translated data. However, SMT does not directly consider the reason why it chooses a particular translation, and thus has a tendency not made by rule-based machine translation systems. In this work, we developed methods to automatically learn rules about why a machine translation system should choose a particular translation, and introduce them into SMT systems. Specifically, we developed a translation system based on linguistic knowledge, introduced reasons for translation into this system, created a method to efficiently analyze its output, and developed methods to learn the rules for translation from multilingual data.

研究分野：自然言語処理

キーワード：機械翻訳　訳選択　機械学習　構文情報

１．研究開始当初の背景

機械翻訳は人類にとって長年の夢であり、近年の著しい進歩により言葉の壁がなくなる日がようやく見えはじめてきた。この進歩に大きく貢献する技術として、対訳データから翻訳ルールを自動的に学習する統計的機械翻訳（ SMT ）がある。機械翻訳の代表的な存在である Google翻訳は毎月 2 億人に利用されており、 65 カ国語の間の翻訳が可能となっているが、この類を見ない言語数と比較的高い翻訳精度は対訳データから構築可能な SMT の枠組みを採用しているからこそ実現可能である。

しかし SMT は、言語学者が人手で構築するルールベース機械翻訳法 (RBMT) に比べて平均的に高い翻訳精度を実現できる一方、一部の入力文や言い回しに対して致命的な誤訳を生成することがある。この問題は、 SMT が明示的に訳選択の根拠を考慮していないという致命的な欠点に起因する。これにより意味のなさない訳だけでなく、一見自然でありながら原文と全く意味の異なる訳が出力されることが多い。この問題は翻訳の信頼性を大きく損なうものであり、**改善されない限り SMT を安心して利用することができない**。

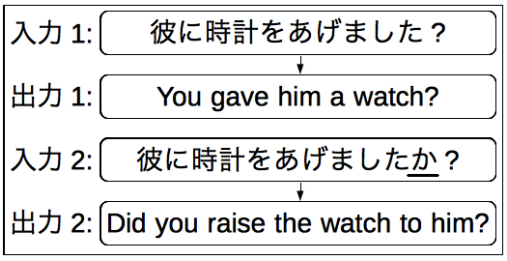

図 1. 入力の微小な差異で大きく異なる統計的機械翻訳の出力

これらの問題の代表的な一例として、意味的曖昧性に基づく誤訳を図 1 に示す。入力文 1 に対して完全な訳が出力されるにも関わらず、原文の意味に影響を及ぼさない「か」を追加した入力文 2 に対して意味不明な訳が生成される。この問題は、「あげました」が「贈った」という意味の「 gave 」、「高めた」という意味の「 raised 」という 2 通りの訳が可能であることに起因する。人間の翻訳者がこの例文を見る時に、「時計」は一般的に「高めるもの」より「贈るもの」であるという知識を用いて曖昧性を解消し「 gave 」という訳語を選択する。これに比べて、 SMT は「ギフトをあげました→ gave a gift 」「幕をあげました→raised a curtain」のような長い単語列からなる翻訳ルールを網羅的に記憶することで曖昧性を解消しようとする。しかし、「時計を上げました→ gave a watch 」がルール集合に存在しない場合、曖昧性が解消されず、ルールの複雑な絡み合いにより出力文 2 のような誤訳を引き起こす。

２．研究の目的

本課題は、**人間の翻訳者が用いる少数かつ明示的な根拠を持つ訳選択ルールを自動発見し、 SMT と融合する**方法の開発を行うことを目的とした。これにより、ルールの複雑な絡み合いを少なくし、**SMT 全体の信頼性を向上**させることができると考えられる。

SMT に関する国内外の研究のほとんどはモデルを複雑にすることで翻訳の精度向上を追求しており、モデルの簡素化を目指す本研究とは大きく異なる。モデルの簡素化に関する数少ない関連研究の中では研究代表者が提案したフレーズ抽出法や誤った翻訳ルールの削除法などがある。しかし、これらは主に翻訳の効率化と精度向上を目指したものであり、本研究とは着眼点も手法も大きく異なる。また、翻訳の手がかりを発見するのに用いられる手法として、語義曖昧性の解消法や研究代表者が提案した語彙情報や構文情報の獲得法などがあるが、機械翻訳

のために簡潔かつ人間の翻訳者の直感に沿った知識の獲得を目指す研究は存在しない。

本研究は機械翻訳の安定性向上に向けて、訳選択に役立つ明示的な根拠に着目し、統語情報や文全体の文脈に基づいた翻訳手法の開発に取り組んだ。

３．研究の方法

本研究では、様々な切り口から訳選択の根拠を考慮した翻訳システムの作成に取り組んだ。

(1) **言語的情報を考慮した翻訳システムの開発：**人間の翻訳者が実際に文の翻訳を行う際、原言語の統語的・意味的特徴を理解した上で翻訳を行う。このような情報を簡単に取り入れるために、原言語の言語学的知識を取り入れた統計的機械翻訳システムの開発を行った。［雑誌論文 1, 2, 4］

(2) **ルールや統計モデルにより翻訳システムにおける人間の直感の導入：**上記の翻訳システムを土台に、人間が人手により作成したルールと、統計モデルにより訳選択の根拠を考慮する手法の開発を行った。［雑誌論文 5, 学会発表 13］

(3) **翻訳結果の誤り分析と評価を効率的に行う手法の開発：**訳選択の根拠を効率的に発見するために、翻訳システムが出力した文と正解文から、最も顕著な誤りを発見する方法を開発した。これを用いれば、翻訳システムはどのような誤りを起こしているかを発見し、それに沿ったシステム改良を設けることができるようになる。［雑誌論文 3, 8］

(4) **多言語データにおける訳選択や構文解釈の根拠の発見：**訳選択の根拠を対訳データのみならず、多言語データで発見し、翻訳精度や構文解析精度の向上に焼く立てる手法の開発を行った。［雑誌論文 6, 7］

４．研究成果

「研究方法」で述べた、４つの研究分野において挙げた様々な成果を下記にまとめる：

(1) 翻訳システムの開発

本研究の最も大きな成果の一つは原言語の言語的情報を考慮した翻訳システムの開発である。具体的には、下記の図　2 のように、入力された構文木を一部一部翻訳し、目的言語の単語列へと翻訳する tree-to-string 翻訳手法である。

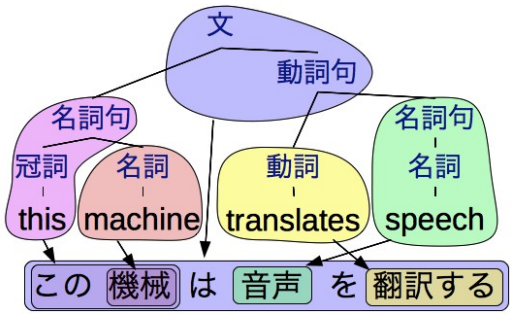

図 1. tree-to-string 翻訳の概念図

この手法の利点として、原言語側の言語的情報を利用するため、本課題の目的である訳選択の根拠や人間の翻訳者の知見の導入を行うことが比較的容易であることが挙げられる。

まず、本研究では、 tree-to-string 翻訳システム「 Travatar」を開発し、オープンソースで公開し［ホームページ 1］、その構成やデザイン理念について研究発表を行っている［雑誌論文 1］。既に、このソフトウェアを利用して研究開発を行っている団体が約 10箇所にのぼる。

次に、この翻訳システムに基づいて分析を行い、人間の直感に合わず、さらに翻訳結果の揺れの原因となる誤り要因を分析した。その結果、翻訳モデルを学習する際の単語の対応付け性能、構文解析器の解析性能、探索時の探索性能を特定した［雑誌論文 2］。これらの要因に対して改良を行い、大幅な性能向上がされ、従来の翻訳システムを大幅に上回っていることが分かる。

また、この翻訳システムを用いて、アジア言語における翻訳性能を競う翻訳コ

ンペ Workshop on Asian Translation 2014 に出馬し、その結果を報告した ［雑誌論文 4］。コンペの結果、本研究で開発したシステムが対象となっていた全 4 言語対（英日・日英・中日・日中 ）で最高性能を記録した。

(2) 人間の知見の導入・モデル化

次に、上記の翻訳システムに基づいて、人間の知見で訳選択に重要であると思われる情報を取り入れる手法を開発した。具体的には、ルールや統計モデルという 2 通りの手法で取り入れる手法を検討した。

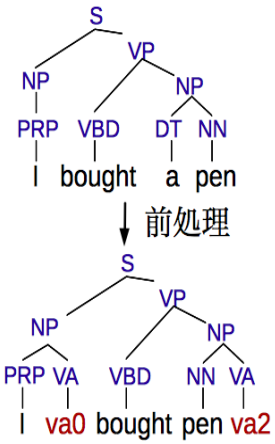

図 4: ルールに基づく前処理

まず、ルールに基づく手法では、原言語文の構造と目的言語の差に着目し、図 4. にあるような前処理でその構造を近づかせる手法を提案した ［雑誌論文 5］。例えば、英語では「I bought a pen」は日本語で「私は ペン を 買った」と翻訳されるが、英語には日本語に存在しない「 a」が含まれ、日本語には英語に含まれない「は」と「が」が含まれる。このような情報の正確な復元は翻訳にとって難しいため、人間の知見を用いて、予め前処理で構造を近づける方法を取り入れた。これによって、実際に人間の持っている知識にもとづいて訳選択の問題を簡単にし、性能の向上を実現することができた。

また、訳選択の根拠として重要でありながら、従来の翻訳システムとして十分考慮されていなかった情報として、文全体に渡る長距離の依存性が挙げられる。このような訳選択の根拠を統計モデルに取り入れる手法として、リカレントニューラルんっとワーク （RNN）を用いた翻訳モデルによる翻訳結果のスコア付けと選択を行った［学会発表 13］。 RNN は長距離の依存性を考慮することのできる統計モデルであり、用いた結果下記表 2 の通り、長距離の依存性を必要とする様々な言語的現象における改善が確認された。

現象	改善	改悪
句の並べ替え	26	4
助動詞の挿入・削除	15	0
並列句の扱い	13	2
名詞・動詞の一致	6	0

表 2: RNN による種類別の翻訳性能向上数

(3) 分析枠組みの開発

訳選択の根拠を翻訳システムに導入することに当たって、まず利用しているシステムの現状と問題点を把握する必要がある。このような現状の把握の効率化を行うために、翻訳システムの出力と人手によって構築された参照訳を用いて、翻訳システムが誤った代表的な箇所を探しだす手法を提案した。具体的に、短い単語列 (n-gram) に誤りの可能性に当たるスコアをつけて、スコアの低い n-gramから分析を行っていく手法である。この手法の概念図を下記図 5 に示す。

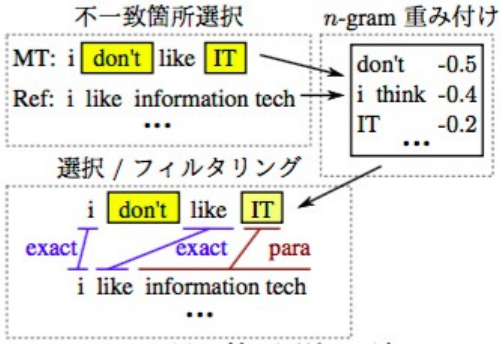

図 5: 誤り分析枠組みの概念図

この手法に基づいて、誤り分析を行った結果、従来の１文ごとの分析に比べて効率的に誤り箇所を探しだすことができることが分かった。

(4) 多言語データ・多モダリティーデータの利用

最後に、複数の言語、もしくは複数のモダリティにおける情報を用いて訳選択の精度を向上させる手法について取り組んだ。

まず、多言語情報に基づく手法で、目的言語で訳選択が曖昧な語彙について、第３の言語でその訳語を確認し、尤もらしさを判定する手法を提案した［雑誌論文6］。例えば、下記の図 6 のような例でアラビア語から中国語へ翻訳する際に、中国語における訳語を英語で確認する具合である。

図 6 ：第３の言語を用いた訳語選択

本研究の結果、第３の言語を用いることによって、前置詞などの機能語の選択をより正確に行うことができることを確認した。

最後に、多言語データを用いて、原言語の構文的な解釈の曖昧性を解消し、構文解析器の学習を行うことで、 4.1 節で紹介したような tree-to-string 翻訳システムの性能向上を図る手法も提案している［雑誌論文7］。具体的には、複数の構文解析結果を用いて翻訳を行い、その中で最も人手で作成された正解訳に近いものに用いられた構文木を用いて構文解析器の適応を行う手法である。この手法を用いて適応された構文解析器を用いた翻訳実験では翻訳性能の向上が実現された。

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 8 件）

1. 赤部晃一 , Graham Neubig, Sakriani Sakti, 戸田智基 , 中村哲 . 機械翻訳システムの誤り分析のための誤り箇所選択手法 . 自然言語処理 26-1. 2016 年 . pp. 87-118. 査読有 . DOI 10.5715/jnlp.23.87.

2. Makoto Morishita, Koichi Akabe, Yuto Hatakoshi, Graham Neubig, Koichiro Yoshino, Satoshi Nakamura. Parser Self-Training for Syntax-based Machine Translation. Proceedings of the 12th International Workshop on Spoken Language Translation. 2015. pp. 147-154. 査読有 . http://workshop2015.iwslt.org/downloads/IWSLT_2015_RP_7.pdf

3. Graham Neubig, Philip Arthur, Kevin Duh. Multi-target Machine Translation with Multi-synchronous Context Free Grammars. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics. 2015. 11. pp. 293-302. 査読有 . http://aclweb.org/anthology/N/N15/N15-1033.pdf

4. Yuto Hatakoshi, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura. Rule-based Syntactic Preprocessing for Syntax-based Machine Translation. Proceedings of the 8th Workshop on Syntax, Semantics, and Structure in Statistical Machine Translation. 2014. 8. pp. 34-42. 査読有 .

5. Graham Neubig. Forest-to-String SMT for Asian Language Translation: NAIST at WAT 2014. Proceedings of the 2014 Workshop on Asian Translation. pp. 1. 20-25. 査読無 . http://aclweb.org/anthology/W/W14/W14-4004.pdf

6. Koichi Akabe, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura. Discriminative Language Models as a Tool for Machine Translation Error Analysis. Proceedings of the 25th International Conference on Computational Linguistics. 2014. 25. pp. 1124-1132. 査読有 . http://aclweb.org/anthology/C/C14/C14-1106.pdf

7. Graham Neubig, Kevin Duh. On the Elements of an Accurate Tree-to-string Machine Translation System. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. 52. pp. 143-149. 査読有 . http://aclweb.org/anthology/P/P14/P14-2024.pdf

8. Graham Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. Proceedings of

the 51st Annual Meeting of the Association for Computational Linguistics. 2013. 51.pp. 91-96. 査読有 .
http://aclweb.org/anthology/P/P13/P13-4016.pdf

〔学会発表〕（計 14 件）

1. 森下睦 , 赤部晃一 , 波多腰優斗 , Graham Neubig, 吉野幸一郎 , 中村哲 . 対訳コーパスを利用した構文解析器の自己学習 . 言語処理学会第 21 回年次大会 . 2016 年 3 月 8 日 . 東北大学 （ 宮城県仙台市 ）.

2. Graham Neubig, Makoto Morishita, Satoshi Nakamura. Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015. 2nd Workshop on Asian Translation. 2015 年 10 月 16 日 . Campus Plaza Kyoto (Kyoto, Japan).

3. 森下睦 , 赤部晃一 , Graham Neubig, 吉野幸一郎 , 中村哲 . 機械翻訳の精度を考慮した構文解析器の自己学習 . 情報処理学会第 223 回自然言語処理研究会 . 2015 年 9 月 28 日 . 広島経済大学 （ 広島県廿日市市 ）.

4. 赤部晃一 , Graham Neubig, 工藤拓 , John Richardson, 中澤敏明 , 星野翔 . Project Nextにおける機械翻訳の誤り分析 . エラー分析ワークショップ . 2015 年 3 月 20 日 . 京都大学 （ 京都府京都市 ）.

5. 赤部晃一 , Graham Neubig, Sakriani Sakti, 戸田智基 , 中村哲 . 機械翻訳の誤り箇所選択法における誤選択箇所の調査 . 言語処理学会第 21 回年次大会 . 2015 年 3 月 19 日 . 京都大学 （ 京都府京都市 ）.

6. Graham Neubig, Philip Arthur, Kevin Duh. 複数の目的言語の同時生成による統計的機械翻訳 . 言語処理学会第 21 回年次大会 . 2015 年 3 月 19 日 . 京都大学 （ 京都府京都市 ）.

7. 波多腰優斗 , Graham Neubig, Sakriani Sakti, 戸田智基 , 中村哲 . Tree-to-String 翻訳における構文解析器の自己学習の効果 . 言語処理学会第 21 回年次大会 . 2015 年 3 月 18 日 . 京都大学 （ 京都府京都市 ）.

8. Graham Neubig. 日本語を対象とした統計的機械翻訳の展望 . 第 6 回産業日本語研究会・シンポジウム （ 招待講演). 2015 年 2 月 24 日 . 東京大学 （ 東京都文京区 ）.

9. 赤部晃一 , Graham Neubig, Sakriani Sakti, 戸田智基 , 中村哲 . パラフレーズを考慮した機械翻訳の誤り箇所選択. 情報処理学会第 219 回自然言語処理研究会 . 2014 年 12 月 17 日 . 東京工業大学（神奈川県横浜市） .

10. 波多腰優斗 , Graham Neubig, Sakriani Sakti, 戸田智基 , 中村哲 . 統語ベース翻訳に対する統語的前処理の適用 . 情報処理学会第 217 回自然言語処理研究会 . 2014 年 7 月 3 日 . オホーツク文化交流センター（北海道網走市） .

11. 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？～ . 音楽シンポジウム 2014(招待講演). 2014 年 5 月 25 日 . 日本大学 （ 東京都世田谷区).

12. 赤部晃一 , Graham Neubig, Sakriani Sakti, 戸田智基 , 中村哲 . 機械翻訳システムの詳細な誤り分析のための誤り順位付け手法 . 情報処理学会第 216 回自然言語処理研究会 .2014 年 5 月 22 日 . 東京工業大学 （ 東京都目黒区).

13. 丹生伊佐夫 , 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 . 言語処理学会第 20 回年次大会 . 2014 年 3 月 19 日 . 北海道大学 （ 北海道札幌市).

14. Graham Neubig. 文レベルの機械翻訳評価尺度に関する調査 . 情報処理学会第 212 回自然言語処理研究会 . 2013 年 7 月 18 日 . はこだて未来大学 （ 北海道函館市).

〔図書〕（計 1 件）

1. 奥野陽 , Graham Neubig, 萩原正人 . 自然言語処理の基本と技術 . 翔泳社 . 2016年 .

〔その他〕

ホームページ等

1. Travatar: Forest-to-string Machine Translation.
http://phontron.com/travatar

2. Lamtram: Neural Machine Translation.
https://github.com/neubig/lamtram

６．研究組織

(1) 研究代表者

Graham Neubig
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号： 70633428