社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[招待講演] 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？

ニュービッグ　グラム†

† 奈良先端科学技術大学院大学 情報科学研究科
E-mail: †neubig@is.naist.jp

あらまし　機械翻訳は計算機の登場以来の長年の夢でありながら、長い間実用的なレベルに達しなかった。しかし、この十数年に著しい進歩を遂げ、ようやく実用に耐えられる翻訳精度になりつつある。本稿では、この精度の向上に貢献している 5 つの要因について解説する。

キーワード　機械翻訳, 統計的機械翻訳, 自動評価, フレーズベース翻訳, 識別学習, 統語ベース翻訳

[Invited Talk] Machine Translation

Why couldn't we do it? Why are we starting to be able to now?

Graham NEUBIG†

† Graduate School of Information Science, Nara Institute of Science and Technology
E-mail: †neubig@is.naist.jp

Abstract　While machine translation has been a long-held dream dating back to the appearance of the first computers, for many years it had not reached a level that was able to stand up to real use. However, over the past ten or so years, there have been huge advances in the technology, and machine translation is finally seeing large-scale use in real situations. This paper describes 5 major elements that have contributed to these enormous steps forward in machine translation systems.

Key words　machine translation, statistical machine translation, automatic evaluation, phrase-based translation, discriminative training, syntax-based translation

1. はじめに

機械翻訳とは、人間の言葉の間を、計算機によって自動的に翻訳する技術であり、言葉の壁を超える可能性があるとして大いに期待されている技術である。計算機が発展し始めた 1950 年代から、計算機の応用例として注目されてきた。このように長年注目されたにも関わらず、未だに人間の翻訳者と同じ精度で翻訳できる自動翻訳手法が存在しない。

しかし、人間の翻訳者と同じレベルの翻訳ができないと言っても、この 10 年機械翻訳の技術は大きな進歩を遂げ、以前と比べて段違いの性能を実現している。この驚くべき進歩のほどを語るのは、Green らの [1] 研究にある翻訳者のコメントである：

> Your machine translations are far better than the ones of Google, Babel and so on.

この翻訳者は Google 翻訳を利用した経験があり、実験に参加した際に Green らの実験で用いられた翻訳システムが、過去に経験した Google 翻訳より性能が高いと述べている。しかし、Green らの研究で用いられたシステムは Google 翻訳自体であった。つまり、プロの翻訳者にとって結果が見違えるぐらい、性能が向上していることが分かる。

本稿では、過去から最近までたどりながら、このような翻訳性能の向上に大きく貢献している技術について解説する。具体的には、高性能な翻訳システムの 5 つの要素について述べる：

（1）　統計的機械翻訳 (2. 節)
（2）　フレーズベース翻訳 (3. 節)
（3）　自動評価 (4. 節)
（4）　識別学習 (5. 節)
（5）　統語ベース翻訳 (6. 節)

なお、本稿は紙面の関係上、大まかな概要となっているが、機械翻訳についての詳細は [2] を参照されたい。

2. ルールベースから統計ベースへ

機械翻訳を試みる際に、大きく分けて 2 つの問題に直面する。1 つ目は「語彙選択」という、入力された単語に対して、いかにして出力する単語を選択するかという問題である。2 つ目は「並び替え」という、出力する単語をいかにして目的言語らしい語順に並べるかという問題である。

最初に考案された翻訳システムはルールベース翻訳 (RBMT) という手法であった。RBMT では、原言語と目的言語両方に精通した翻訳者が翻訳に用いるルールを人手で作成する。ほとんどの場合は、原言語文の解析を行い、この解析結果に基づいて語彙選択と並び替えを行う。例えば、英語の「run」を日本語の「走る」へ翻訳するか、「実行する」へ翻訳するかを、目的語が「陸上競技の一種」なのか「コンピュータプログラムの一緒」なのかで区別するルールを用いて語彙選択を行うことができる。文の構造を解析し、英語で「S(主語) V(動詞) O(目的語)」の順番になっている句を「S が O を V」などと、日本語らしい語順に並び替えることができる。しかし、このようなルールを書き下すのが非常に労力のかかる作業であり、例外も多い。このため、典型的な文に対して翻訳を行える RBMT システムを構築することは可能であるが、例外的な語彙選択（「run up」→「増やす」）や並び替え（「I have no time」→「ない時間を持っている」ではなく「時間がない」）をすべてカバーするのは非常に困難である。

この問題を解決するのは、対訳データから自動的に翻訳ルールを学習する翻訳方法である [3]。特に、現在の最先端の翻訳システムの中で利用されているのが Brown らにより提案された統計的機械翻訳 (SMT) である [4]。SMT では、原言語文を $\boldsymbol{f}$ とし、目的言語文を $\boldsymbol{e}$ とした時に、文の関係を確率的にモデル化し、ある $\boldsymbol{f}$ に対して事後確率が最大となる $\hat{\boldsymbol{e}}$ を翻訳結果とする。

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} P(\boldsymbol{e}|\boldsymbol{f}) \tag{1}$$

これを更に、ベイズ則に基づき、以下のように展開する。

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} \frac{P(\boldsymbol{f}|\boldsymbol{e})P(\boldsymbol{e})}{P(\boldsymbol{f})} \tag{2}$$

また、$\boldsymbol{f}$ が与えられているため 2 式に影響を及ぼさないことを利用して、式の分母を省略する。

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} P(\boldsymbol{f}|\boldsymbol{e})P(\boldsymbol{e}) \tag{3}$$

この中で、$P(\boldsymbol{f}|\boldsymbol{e})$ は入力文と出力文の関係を表す翻訳モデル (TM)、$P(\boldsymbol{e})$ は出力文の目的言語文らしさを表す言語モデル (LM) である。LM は一般的に、音声認識など様々な応用で利用される n-gram モデル [5] を利用することが一般的である。TM に関しては、Brown らが統計的機械翻訳を提案した当時は各単語がどの単語に翻訳されるかの翻訳確率、各単語がどのように並べ替えられるかの並び替え確率などからなる。対訳データが与えられた時に、一般的には文と文が対応づいたデータが与えられ、どの単語がどの単語に対応するかが与えられないため、EM アルゴリズムを用いて単語の対応と翻訳確率などを自動的に学習する手法が用いられる。このような自動学習手法は GIZA++[(注1)] [6] というツールキットに実装され、現在でも広く利用されている。

3. 単語ベースからフレーズベースへ

単語に基づく SMT モデルはルールベース翻訳は学習が容易であることから研究の対象となった。しかし、単語に基づくモデルだけでは高い翻訳精度を実現するのが困難である。例えば、「run」が「走る」か「実行する」に翻訳されるかは言語モデル確率 $P(\boldsymbol{e})$ に基づいてある程度正確に判断できるが、「run up」→「増やす」のような複数の単語が 1 つの慣用句になる訳をモデル化することは容易ではない。

このような複数の単語からなる対訳候補を扱う方法として、フレーズベース統計的機械翻訳 (PBMT) [7] が提案された。この手法では、単語ごとに翻訳候補を覚えるのではなく、単語列と単語列の対応を翻訳モデルとして記憶する。これにより、語彙選択や局所的な並び替え（「run a marathon」→「マラソンを走る」）を同時に学習し、翻訳性能の大幅な向上につながった。現在でも PBMT は多くの言語で最高の性能を誇り、Google 翻訳[(注2)]や広く使われているオープンソースソフト Moses[(注3)] [8] の中で利用されている。

4. 人手評価から自動評価へ

機械翻訳の研究で大きな問題となるのが翻訳の評価である。従来のルールベースシステムでは、ルールの構築と同様に、翻訳の評価を人手で行っていた。しかし、翻訳システムの最終評価を人手で行ったとしても、システムの開発中に小さな改良を加えるたびに人手で評価を行うことが非常にコストがかかる。

この評価の効率を大幅に向上させたのは、機械翻訳のための自動評価である [9]。機械翻訳の自動評価では、ある入力文に $\boldsymbol{f}$ に対して、人手で 1 つ以上の正しい訳（参照訳）$\boldsymbol{e}^*$ を用意する。これとシステムが出力した答え $\hat{\boldsymbol{e}}$ を比較し、システム出力と参照訳が近ければ近いほど高いスコアを与える。最初に提案された評価尺度 BLEU [9] では、$\boldsymbol{e}^*$ と $\hat{\boldsymbol{e}}$ の距離を、長さ 4 までの単語列の一致率などに基づいて計算する。BLEU は今でも広く使われているが、日英・英日翻訳で問題となる並び替えに着目した RIBES [10] など、ほかの尺度も数多く提案されている。

5. 生成モデルから識別モデルへ

2. 節で説明した統計的モデルでは、言語モデルと翻訳モデルを両方ベイズ則で組み合わせた形式になっていた。また、それぞれのモデルを、モデルの尤度が高くなるように最尤推定される。しかし、確率計算の式をパラメータ化し、パラメータを 4. 節で説明した評価尺度が高くなるように調整すると更なる性能の向上が実現できる [11]。

(注1) : http://code.google.com/p/giza-pp/
(注2) : http://translate.google.co.jp
(注3) : http://www.statmt.org/moses/

例えば、3 式に、以下のように言語モデル確率と翻訳モデル確率に対してそれぞれ λ_{lm} と λ_{tm} という指数パラメータを導入する。

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} P(\boldsymbol{f}|\boldsymbol{e})^{\lambda_{tm}} P(\boldsymbol{e})^{\lambda_{lm}} \quad (4)$$

これをすることで、λ_{lm} を大きくすることでより言語モデルを重視し、目的言語において自然な文を生成することができる。逆に λ_{tm} を大きくすると、より翻訳モデルを重視し、原文に対して忠実な文を生成するようになる。更に並び替えモデル、フレーズではなく単語を独立して翻訳する翻訳モデルなど、10-20 通りの生成モデルを組み合わせて翻訳候補の重み付けを行う手法が広く利用されている。

このようなパラメータを調整する手法として誤り率最小化学習 (MERT) が広く利用されている [12]。MERT では、各パラメータごとに線形探索を行い、評価尺度が高くなるように反復を繰り返すことで、任意の評価尺度を最大化できる。さらに、各フレーズ対に対する素性など、より多くの素性を導入することで性能の更なる向上が見込めることが知られており、多くの素性を最適化するアルゴリズムも提案されている [13]。

6. フレーズベース翻訳から統語ベース翻訳

フレーズベース翻訳では単語列を単語列に翻訳する仕組みになっていた。しかし、単純に単語列では表せない現象も多く存在する。例えば、並べ替えの例として挙げた、「S V O」→「S は O を V」という訳は、連続した単語列ではなく、単語（「は」「を」）と変数（「S」「V」「O」）からなるパターンの間を翻訳することができれば自然と表すことができる。これらのパターンをフレーズベース翻訳で正確に捉えるのが困難であり、特に長距離の並び替えを正確に行えないのがフレーズベース翻訳の最も大きな課題とされている。この問題を解決するために、木構造や統語情報に基づく統計的機械翻訳モデルが提案されている。

もっとも単純なモデルでは、「X1 X2 bread」→「X1 はパンを X2」のように、変数を全て「X 数字」で表した階層的フレーズベース翻訳 (Hiero) モデルが提案されている [14]。更に、「NP1 VP2 bread」→「NP1 はパンを VP2」のように、翻訳パターンに「名詞句 (NP)」「動詞句 (VP)」のような統語的な情報を取り入れた手法もある。特に、ルールベース翻訳と同様の形式で、予め文を解析して、この解析結果に基づいて翻訳を行う手法は長距離の並べ替え精度を改善できる [15], [16]。事前に文を解析して、この原言語側の構造に基づき翻訳を行う手法は Travatar(注4) [17] などのオープンソースツールキットで実装されており、英日・日英翻訳などで高い精度を実現している [18]。

7. おわりに

本論文では、現在の機械翻訳システムの軸となっている技術を 5 つ取り上げた。これらの技術を組み合わせて、医療対話文 [19] に対するへ自動翻訳した際の結果を表 1 に示す。結果を見て分かる通り、高い精度で翻訳を行えていることが分かる。しかし、SMT にはまだ様々な課題が残されている。その課題の例を表 2 に示す。

まず、原言語文に明示的に書かれていない情報を目的言語へ再現することは未だに困難である。例えば、日本語において、主語を省略されることが多いが、英語では主語は必須である [20]。表 2 a) の例では、話者は指示を出しているため、主語は明らかに「あなた」であるが、翻訳システムはこれを判別できず、主語を「私」にしてしまった。主語の省略以外にも、活用が比較的少ない言語（英語など）から、活用が比較的豊富な言語（フランス語やロシア語など）へ翻訳する際、正しい活用形を復元することが困難であるとされている [21]。

また、現在の機械翻訳は曖昧な単語の訳し分けは完璧にできるわけではない。例えば、表 2 b) の例では、「lie」は「横になる」、「うそをつく」という 2 つの可能な訳があり、医療現場において発話者が聞き手に「うそをつけてください」と頼むことはめったにないにも関わらず、誤った訳を選択した。この問題は、訳し分けに必要な情報が文章全体のトピックや文脈に基づく場合は特に顕著である [22]。

最後に、音声認識結果 [23] やウェブ上のテキスト [24] など、崩れた入力に対する機械翻訳は難しいとされている。その一例として、表 2 c) に現れるひらがな表現がある。ここで、「ご苦労さま」が翻訳できたとしても「ごくろうさま」が誤って「ごく」と「ろう」に分割されてしまい、翻訳が失敗したことが見受けられる。特に、構文解析を行う手法はテキストを正しく解析できることを前提としているが、崩れた入力に対して解析が正確に行えるとは限らない。また、統計翻訳に必要な学習データは、崩れた表現を含むものは著しく少なく、崩れた表現をまず学習データに含まれる標準的な表現に変更してから翻訳を行うなどの対策が必要となる。

(注4)：http://phontron.com/travatar

文　献

[1] S. Green, J. Heer, and C.D. Manning, “The efficacy of human post-editing for language translation,” Proc. CHI, pp.439–448, 2013.

[2] 渡辺太郎，今村賢治，賀沢秀人，G. Neubig，中澤敏明，機械翻訳，コロナ社，2014.

[3] M. Nagao, “A framework of a mechanical translation between japanese and english by analogy principle,” Proc. International NATO Symposium on Artificial and Human Intelligence, pp.173–180, 1984.

[4] P.F. Brown, V.J. Pietra, S.A.D. Pietra, and R.L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” Computational Linguistics, vol.19, pp.263–312, 1993.

[5] J.T. Goodman, “A bit of progress in language modeling,” Computer Speech & Language, vol.15, no.4, pp.403–434, 2001.

[6] F.J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” Computational Linguistics, vol.29, no.1, pp.19–51, 2003.

[7] P. Koehn, F.J. Och, and D. Marcu, “Statistical phrase-based translation,” Proc. HLT, pp.48–54, 2003.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation,” Proc. ACL, pp.177–180, 2007.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU:

表 1　統計的機械翻訳システムの英日・日英翻訳例

Table 1　An example of machine translation results.

原文	While taking out food processor blades from the dishwasher, I cut my right hand between my thumb and index finger.
参照訳	食器洗浄器からフードプロセッサーの刃を取り出そうとして、右手の親指と人差し指の間を怪我してしまいました。
システム	食器洗い機からフードプロセッサーの刃を出しながら、私の親指と人差し指の間に、右の手を切ってしまいました。
原文	最近、朝目を覚ますと、左目がドライアイになっていて、瞬きがうまくできません。
参照訳	Recently, when I wake up in the morning, my left eye is really dry, which makes it difficult to blink.
システム	Recently, my left eye is dry and can't blink when I wake up in the morning,.

表 2　翻訳システムの誤り例

Table 2　An example of machine translation errors.

a)	原文	あ、結構です書かなくていいです。
主語の省略	参照訳	Oh, that's fine. You don't have to write.
	システム	I don't write in, thank you.
b)	原文	Please keep lying.
語彙選択	参照訳	このまま休んでいてください。
	システム	うそをつき続けてください。
c)	原文	はい、ごくろうさまでした。
崩れた表現	参照訳	You're all set.
	システム	Yes, it was very good wax.

a method for automatic evaluation of machine translation," Proc. ACL, pp.311–318, 2002.

[10] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, "Automatic evaluation of translation quality for distant language pairs," Proc. EMNLP, pp.944–952, 2010.

[11] F.J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," Proc. ACL, pp.295–302, 2002.

[12] F.J. Och, "Minimum error rate training in statistical machine translation," Proc. ACL, pp.160–167, 2003.

[13] T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki, "Online large-margin training for statistical machine translation," Proc. EMNLP, pp.764–773, 2007.

[14] D. Chiang, "Hierarchical phrase-based translation," Computational Linguistics, vol.33, no.2, pp.201–228, 2007.

[15] J. Graehl and K. Knight, "Training tree transducers," Proc. HLT, pp.105–112, 2004.

[16] M. Collins, P. Koehn, and I. Kucerova, "Clause restructuring for statistical machine translation," Proc. ACL, pp.531–540, 2005.

[17] G. Neubig, "Travatar: A forest-to-string machine translation engine based on tree transducers," Proc. ACL Demo Track, pp.91–96, 2013.

[18] G. Neubig and K. Duh, "On the elements of an accurate tree-to-string machine translation system," Proc. ACL, 2014.

[19] G. Neubig, S. Sakti, T. Toda, S. Nakamura, Y. Matsumoto, R. Isotani, and Y. Ikeda, "Towards high-reliability speech translation in the medical domain," Proc. MedNLP, pp.22–29, 2013.

[20] T. Chung and D. Gildea, "Effects of empty categories on machine translation," Proc. EMNLP, pp.636–645, 2010.

[21] V. Chahuneau, E. Schlinger, N.A. Smith, and C. Dyer, "Translating into morphologically rich languages with synthetic phrases," Proc. EMNLP, pp.1677–1687, 2013.

[22] Z. Gong, M. Zhang, and G. Zhou, "Cache-based document-level statistical machine translation," Proc. EMNLP, pp.909–919, 2011.

[23] M. Ohgushi, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "An empirical comparison of joint optimization techniques for speech translation," Proc. 14th InterSpeech, pp.2619–2622, 2013.

[24] 工藤　拓，市川　宙，D. Talbot，賀沢秀人，"Web 上のひらがな交じり文に頑健な形態素解析，" 言語処理学会第 18 回年次大会発表論文集，pp.1272–1275，2012.