深層学習に基づく特許翻訳における数値表現の扱い

A method of translating numerical expressions in neural machine translation

株式会社日立製作所　研究開発グループ
岩山　真

1992 年株式会社日立製作所入社。文書検索、自然言語処理等の研究に従事。また、NTCIR において特許検索用テストコレクションの作成に携わる。2009 年度より特許版産業日本語委員会委員。

1　はじめに

機械翻訳には、二つの手法がある [1]。第一は、規則に基づく手法である。あらかじめ用意しておいた規則や辞書を使って文を翻訳する。規則や辞書は人手で作成することが多い。第二は、統計に基づく手法（SMT：Statistical Machine Translation）である。SMT では、文単位の対訳対から、単語や句単位の対訳テーブルを学習し、このテーブルを用いて確率的に一番尤もらしい訳文を生成する。人手で規則や辞書を作成する必要はないが、大量の対訳対が必要になる。

近年、新たな手法として深層学習に基づく手法 [3][4][8]（NMT：Neural Machine Translation）が提案された。NMT では、再帰ニューラルネットワークを用いて原文から訳文への翻訳モデルを直接学習する。NMT も、学習に大量の対訳対が必要となるが、end-to-end の学習となるため、チューニングを行わなくても SMT と同程度もしくは SMT を上回る精度が得られる [7]。

NMT の問題の一つは、数値表現に代表される低頻度表現の翻訳である。SMT では、対訳テーブルにより単語や句の翻訳が制御できるために、原文の数値をそのまま訳文にコピーすることができるが、NMT ではこのような制御が行いにくい。特許では数値が発明の本質となることもあるため、数値表現は正確に翻訳する必要がある。

本研究では、NMT において、原文の数値を訳文に強制的にコピーする方法を提案し評価した。

2　NMT の概要と問題点

本研究が対象とする NMT は、再帰ニューラルネットワーク（RNN：Recurrent Neural Network）を用いたエンコーダ/デコーダモデル [3][4][8] である。図 1 にその概略を示す。翻訳元の文（原文）を RNN の内部状態にエンコードし、エンコードされた内部状態から翻訳結果の文（訳文）をデコード（生成）する。更に、アテンション機構 [3] を有しており、訳文中の単語の生成時に、原文の情報を使うことができる。効果的に学習が行われると、対訳関係となっている原文中の単語に強くアテンションがあたるようになる。アテンション機構も訓練

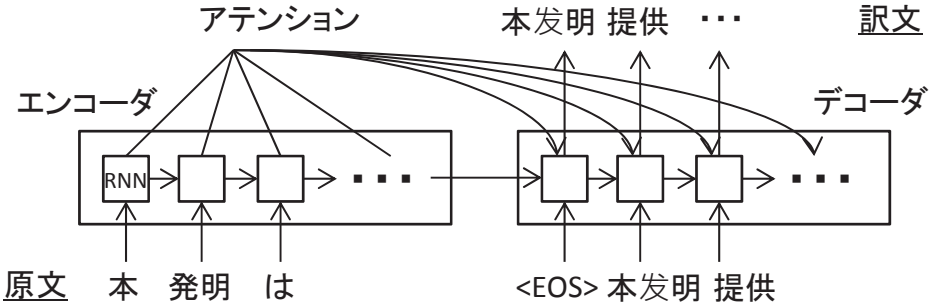

図 1

データから自動で学習する。

NMT では、対訳関係は再帰ニューラルネットワーク内の状態や重みとして学習されている。SMT のように明示的に対訳テーブルを作成するわけではない。そのため、訓練データに出現しない、もしくは数回しか出現しない低頻度表現の制御が困難である。SMT では、対訳テーブルから低頻度表現が同定できるため、低頻度表現をそのまま翻訳先にコピーするといった個別対応が可能である。一方、NMT では低頻度表現も無理やり訳そうとするために、低頻度表現は意味不明な翻訳結果となることが多い。

低頻度表現の代表的な例として数値表現がある。

（原文 1）本発明では、2 つの導体の間に、3 個以上の放電部を直列に接続させる。
（原文 2）これにより、正孔注入層 70 の駆動電圧を低下させることができ、有機 E L 装置 1 を長寿命化することができる。

いずれの例も、数値自体は低頻度ではないが、「2 つの導体」「3 個以上の放電部」「正孔注入層 70」など、数値の前後の表現も含めると低頻度になりやすい。

上記の例を NMT で翻訳すると以下のようになる。

（訳文 1）本发明涉及一种在 0 个导体之间串联连接的多个放装置。
（正解 1）在本发明中、在 2 个导体之间串联连接 3 个以上的放电部分。

（訳文 2）由此、能够降低空穴注入层（53）的驱动电压、并且能够延长有机 EL 装置（0）。
（正解 2）由此、可降低空穴注入层（70）的驱动电压、可使有机 EL 装置（1）长寿命化。

いずれも、数値以外はほぼ問題なく訳せているが、数値が正確に訳せていない。訳文 1 では、「3 個以上の放電部」が「多くの放電部」となり、原文の意味が正確に伝わらない。訳文 2 では、「正孔注入層 53」が「正孔注入層 70」となり、場合によっては全く別の発明内容となってしまう。特許では、数や番号が重要な意味を持つことが多く、数値を正確に訳すことが求められている。

3 従来手法

NMT で低頻度表現を扱うために、幾つかの手法が提案されている。[2] では、NMT に対訳辞書を導入している。対訳辞書は、ある単語がある単語に訳される確率として定義されており、SMT と同じ手法で作成する。訳文の各単語を生成する際に、原文のどの単語に注目しているかをアテンション機構で同定し、対訳辞書を用いて翻訳確率を計算する。翻訳確率と、デコーダによる生成確率から訳文の単語を決める。

数値に限れば、原文中の数値をそのまま訳文にコピーすればよい。[6] では、アテンション機構付きのエンコーダ･デコーダモデルにコピー機構を導入した。アテンションが当たっている入力単語をそのまま出力にコピーする確率を導入し、このコピー確率と、デコーダによる生成確率から生成単語を決める。コピー機構は訓練データから自動的に学習される。

NMT の前処理と後処理で対応する手法もある。[9] では、原文中の未知語（低頻度語でもよい）をプレースホルダで置き換えて翻訳を行い、訳文中のプレースホルダを原文の対応する単語に置き変える。学習はプレースホルダ付きの対訳対で行う。例として、以下のような対訳対（下線が未知語）を考える。

（原文 3）The ecotax portico in Pont-de-Buis
（訳文 3）Le portique ecotaxe de Pont-de-Buis

この場合、プレースホルダ付きの対訳対は以下のようになる。

（原文 4）The 〈unk〉 portico in 〈unk〉
（訳文 4）Le 〈unk_1〉 〈unk_－1〉 de 〈unk_0〉

原文中の未知語は全て、〈unk〉というプレースホルダで置き換える。訳文中の未知語には、原文中で対応する単語の相対位置を付記する。例えば、訳文中の未知語“portique”は訳文中で 2 単語目にあり、原文中では 3 単語目の“portico”に対応するため、相対位置は 1 となり、プレースホルダは〈unk_1〉となる。このようにして対訳対を全てプレースホルダで置き換えて、NMT の学習を行う。翻訳時には、原文中の未知語を〈unk〉

と置き変えた文を NMT に入力する。相対位置付きのプレースホルダが出力された場合は、対応する原文中の単語で置き換える。

4 提案手法の概要

プレースホルダを使う方法 [9] は、コピーすべき単語を明示的に指定出来る点が利点である。それに対し、NMT に対訳辞書を導入する方法 [2] や、NMT にコピー機構を導入する方法 [6] では、入力単語がコピーされるかどうかは、NMT の学習がうまく行われるかどうかに依存する。既に述べたように、特許では数字は正確に訳して欲しい（コピーしてほしい）ため、本提案手法でも、プレースホルダを使い数字を強制的にコピーする。

プレースホルダを使う従来手法の問題点は、訳文中にプレースホルダが複数個存在する場合、それぞれの原文中での位置を相対位置で指定する点にある。英仏のように、構造が似ている言語間では、相対位置の範囲は 0 を中心にあまり広く分布しない。ところが、日中のように構造が大きく異なる言語間では、相対位置の分布は広くなる。更に、日本語は語順が比較的自由であるため、相対位置の分布はより広くなり、訓練データのスパースネスの問題が生じやすい。

提案手法では、翻訳時にプレースホルダが複数個存在する場合、アテンション機構を使って、訳文と原文でプレースホルダを動的に対応付ける。一方、学習時は、対応は曖昧なまま処理する。ここでは、プレースホルダが現れる文脈情報から曖昧性が解消されていることを期待する。

5 学習時の処理

例えば、以下のような対訳対を考える。

（原文 5）本発明では、2 つの導体の間に、3 個以上の放電部を直列に接続させる。
（訳文 5）在本发明中、在 2 个导体之间串联连接 3 个以上的放电部分。

本提案手法では、原文、訳文共に数値の部分を同一のプレースホルダ〈num〉に置換する。上記の例は以下のようになる。

（原文 6）本発明では、〈num〉つの導体の間に、〈num〉個以上の放電部を直列に接続させる。
（訳文 6）在本发明中、在〈num〉个导体之间串联连接〈num〉个以上的放电部分。

置き換え対象の数値は、整数（123 など）および、小数点表記の実数（12.56 など）とする。プレースホルダに置き変えた対訳対を用いてアテンション機構付きの NMT を学習する。

数値を全て同一のプレースホルダに変換するため、例のように複数の数値が存在する場合は、それらの対応は無視して学習されることになる。

6 翻訳時の処理

翻訳時には、前節と同じ方法で、原文中の数値をプレースホルダに置換する。例えば、

（原文 7）L/S が 0.5～0.8 である。

を以下に変換する。

（原文 8）L/S が〈num〉～〈num〉である。

また、デコード時のために、プレースホルダが置換した数値を記録しておく。

次に、変換後の原文を学習した NMT に入力する。NMT による翻訳結果が以下になったとする。

（訳文 8）L/S 是〈num〉～〈num〉。

この場合、訳文にもプレースホルダが 2 個あるため、どちらを 0.5 に戻し、どちらを 0.8 に戻すかを決める必要がある。本手法では、アテンション機構を用いて、訳文と原文間のプレースホルダの対応を計算する。

訳文の各単語を生成する際は、原文中のどの単語に注目しているかがアテンション層に現れる。図 2 は、訳文 8 を生成した際のアテンション層を図示したものである。縦軸が訳文に相当し、訳文中の各単語 w_f に対

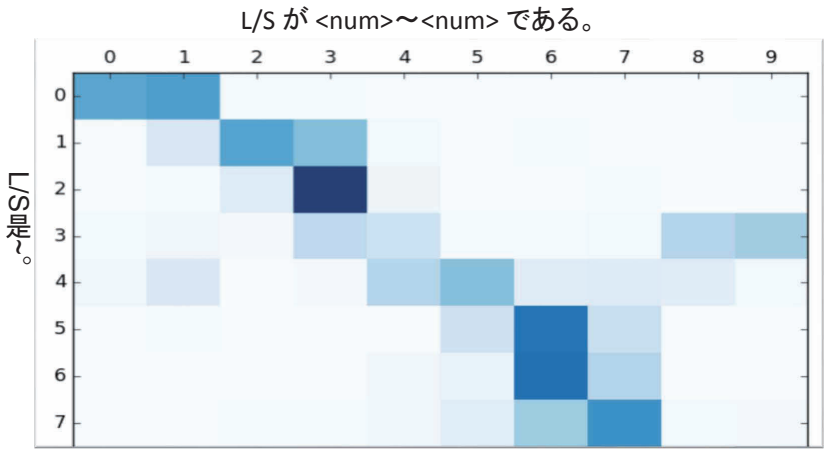

図 2

表 1

		文数	数値を含む文数	プレースホルダ数
訓練データ	日本語	164697	61271	148335
	中国語	164697	38076	89603
テストデータ	日本語	41147	15232	36949
	中国語	41147	9659	22590

するアテンション層を図示したものが横軸となる。横軸の各要素は原文の単語 w_e に相当する。色の濃さは、アテンションの強さ a（w_e|w_f）に相当している。アテンションの強さは、確率値と解釈できる。

一文中でのプレースホルダの数は、日本語で 2.42 個、中国語で 2.35 個となり、平均して 2 個以上のプレースホルダがある。よって、プレースホルダの対応付けは必須の処理であることがわかる。

7 実験データ

本手法の効果を確認するために、公報から抽出した対訳対を用いて実験を行った。対訳対は、複数の国で出願されているファミリー特許から自動的に抽出している。本実験では、電気分野の特許の要約から対訳対を選んだ。訓練データとテストデータは 4 対 1 にランダムに分割した。表 1 にデータの詳細を示す。なお、日本語の形態素解析には MeCab[1] を、中国語の形態素解析には Jieba[2] を用いた。

表 1 を見ると、特許には数値が多く含まれていることがわかる。日本語の場合は、37%、中国語の場合は、23% の文に数値が含まれている。中国語での割合が日本語より少ない理由は、中国語では数字が漢字で記載されることが多いためである。今回は、漢数字の処理は行わなかった。

1 http://taku910.github.io/mecab/

2 https://github.com/fxsjy/jieba

8 実験方法

日本語から中国語への翻訳で実験した。日本での外国特許の調査という観点では、中国語から日本語へ翻訳する方が現実的だが、今回は、原文を筆者らが理解できることを重視した。

まず、表 1 の訓練データを用いて、アテンション機構付きの NMT を学習した。実装には TensorFlow[3] を用いた。RNN セルには GRU（Gated Recurrent Unit）を用い、256 ユニット 2 層からなる階層的なエンコーダ・デコーダモデルを構築した。語彙数は、日本語、中国語、共に 100,000 とした。最適化アルゴリズムには Adam を用いた。Adam のパラメータはデフォルト値（論文推奨値）のままである。また、RNN への入力は逆順とした。

学習したモデルを使い、テストデータで評価を行った。評価指標は、BLEU を用いた。ベースラインとし

3 https://www.tensorflow.org/

て、プレースホルダを使わない手法で実験を行った。また、比較対象として、プレースホルダを用いた従来手法[9]でも実験を行った。従来手法では、プレースホルダ間の対応を事前に同定しておく必要がある。本実験では、全データに対して GIZA++[4] を適用し、単語単位での対応を自動的に計算した。

9 実験結果と考察

表 2 に実験結果を示す。

表 2

	数値有	数値無	全体
ベースライン	0.1710	0.1928	0.1848
従来法	0.1813	0.1968	0.1911
提案手法（置き換え前）	0.2288	0.1924	0.2059
提案手法（置き換え後）	0.2192	0.1924	0.2023

- 提案手法を用いることで、ベースラインの NMT に比べ、全体として BLEU が 1.7 ポイント、割合で 9% 向上した。提案手法は、数値を含まない文に関してはベースラインと同じ処理を行うため、数値を含む文のみで比較すると、BLEU の向上は 4.8 ポイント、割合で 28% となる。このことから、プレースホルダを用いて数値をコピーすることの有効性がわかる。プレースホルダを用いる従来法も、ベースラインよりも BLEU が高い。
- 提案手法を用いることで、数値を含む文の BLEU が、数値を含まない文の BLEU と同等、もしくは上回るようになった。
- 従来法と比べると、本手法は、全体として 1 ポイント、数値を含む文で、3.8 ポイント BLEU が向上している。
- 本手法において、デコード時にプレースホルダを置き変えない場合と、置き変えた場合との BLEU を比較すると、置き換えにより BLEU が 1 ポイント低下している。これは、アテンションによる対応付けの誤りに起因している。
- 従来手法も提案手法も、数値を含まない文に対する BLEU はほぼ同じとみなせる。よって、プレースホルダが数値以外に及ぼす副作用はほぼ無いと言える。

誤りの原因は大きく以下の 2 点である。

- 訳文にプレースホルダが現れない。主に中国語が漢数字の場合に相当する。

（原文 11）第 1 の移動局は第 1 の通信アプリケーションおよび第 1 の PIN を有し、第 2 の移動局は第 2 の通信アプリケーションおよび第 2 の PIN を有する。
（正解 11）第一移台具有第一通信用和第一 PIN、第二移动台具有第二通信应用和第二 PIN。
（訳文 11）第一移台具有第一通信用和第二 PIN、第二移动台具有第二通信应用和第二 PIN。

この例では、訳文にプレースホルダが全く現れなかった。よって、通常の NMT により翻訳が行われ、数字部分で微妙に間違えている。中国語における漢数字の割合が予想以上に多かったため、今後は漢数字も数値として扱う必要がある。

- 原文に数値が多く出現する。NMT では同じ表現が必要以上に繰り返し生成される傾向があるため、原文と同数のプレースホルダを生成することが難しい。

（原文 12）前記式で、n, m, q, R〈num〉, R〈num〉, X〈num〉, X〈num〉, X〈num〉, X〈num〉 及び A〈num〉は、発明の詳細な説明で定義された通りである。
（訳文 12）在一种使用、nm, q, R, X〈num〉, X〈num〉, X〈num〉, X〈num〉, X〈num〉, X〈num〉, X〈num〉, X〈num〉, X〈num〉, X〈num〉

10 おわりに

深層学習による特許翻訳において、数値を含む表現を正確に翻訳する技術を提案した。約 20 万文対を使った評価実験により、数値を含む文の BLEU が従来手法から 28% 向上することを確認した。

今回は、電気分野の要約のみを対象に実験を行ったが、実用化に向けて、他分野および要約以外の文でも提案手法の有効性を検証するする必要がある。

4 http://www.statmt.org/moses/giza/GIZA++.html

また、数値以外にも、化学式、固有名詞など、翻訳誤りが致命的になる表現は多い。今後、これらの表現へも提案手法を適用していきたい。

謝辞

本研究の一部は、東京工業大学、奥村・高村研究室の田中空さんにインターンとして取り組んで頂きました。

参考文献

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

[2] Arthur, P., Neubig, G., Nakamura, S., “Incorporating Discrete Translation Lexicons into Neural Machine Translation”, EMNLP 2016, 2016.

[3] Bahdanau, D., Cho, K., Bengio, Y., “Neural Machine Translation by Jointly Learning to Align and Translate”, CoRR 2014, 2014.

[4] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”, EMNLP 2014, 2014.

[5] Goto, I., Chow, K. P., Lu, B. Sumita, E., Tsou, B. K., “Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop”, Proceedings of the NTCIR-10 Workshop, 2013.

[6] Gu, J., Lu, Z., Li, H., Li, V. O. K., “Incorporating Copying Mechanism in Sequence-to-Sequence Learning”, pp.1631-1640, ACL, 2016.

[7] Nakazawa, T., Ding, C., Mino, H., Goto, I., Neubig, G., Kurohashi, S., “Proceedings of the 3rd Workshop on Asian Translation（WAT2016）”, pp.1-46, 2016.

[8] Sutskever, I., Vinyals, O., Le, Q. V., “Sequence to Sequence Learning with Neural Networks”, Proc. NIPS, 2014.

[9] Sutskever, I., Le, Q. V., Vinyals, O., Zaremba, W., “Addressing the Rare Word Problem in Neural Machine Translateion”, pp.11-19, ACL, 2015.

[10] Utiyama, M., Isahara, H., “Reliable measures for aligning Japanese-English news articles and sentences”, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pp.72-79, 2003.