The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ユーザ辞書を用いたニューラル翻訳のカスタマイゼーション

The customyzation of neural machine translation with user's dictionary

山内 翔大 *1　村上 陽平 *2　中口 孝雄 *3　石田 亨 *4
Shota Yamauchi　Yohei Murakami　Takao Nakaguchi　Toru Ishida

*1京都大学工学部情報学科
Undergraduate School of Informatics and Mathematical Science, Kyoto University

*2京都大学デザイン学ユニット
Unit of Design, Kyoto University

*3京都情報大学院大学ウェブビジネス技術専攻
Department of Web Business Technology, Kyoto College of Graduate Studies for Informatics

*4京都大学社会情報学専攻
Department of Social Informatics, Kyoto University

Neural machine translation (NMT) has significantly improved quality over traditional statistical-based machine translation (SMT). However, it is known that it is difficult for NMT to translate sentences containing rare terms. Therefore, in this research, we propose a method of replacing rare terms with synonyms and translating it, and replacing the translated synonyms with translations of the rare terms in the bilingual dictionary. This approach has two technical issues: acquisition of synonyms from small scale corpus, and selection of the synonyms. The proposed method shows the better result than the result of the existing technics in sign test.

1. はじめに

ニューラルベース機械翻訳 (NMT) は，従来の統計ベース機械翻訳 (SMT) よりも大幅に品質を向上させている．例えば，Google が開発した GNMT は，エラーの少なさや正確性などにおいて，PBMT よりも良い結果を示している [Johnson 17]．また，TOEIC の点数で換算すると 900 点相当の英文作成能力があるニューラル翻訳機 *1 も登場している．SMT では，入力文の各単語をマッピングテーブルを用いて翻訳し，言語モデルを用いて適切に並び替えることによって翻訳結果を生成し出力する．一方，NMT では，原文はエンコーダと呼ばれるニューラルネットワークによってベクトル表現に変換され，その後，デコーダと呼ばれる別のニューラルネットワークが翻訳文を生成する．したがって，SMT が単語を最小単位としたマッピングテーブルを用いて翻訳するのに対し，NMT では，入力文そのものが最小単位になり，入力文に唯一に対応する翻訳結果を出力する．

しかしながら，NMT には，普段使われることが少ないような専門用語を含む文を翻訳することが困難であるということが知られている．また，従来の SMT では，マッピングテーブルに専門用語を追加することで，辞書を用いた翻訳に対応することができた．しかし，NMT にはマッピングテーブルがないため，辞書を用いて専門語を含む文章を翻訳することが難しい．したがって，ユーザが NMT をカスタマイズすることも困難であると言える．

そこで，本研究では，専門用語を類義語で置き換えて翻訳し，翻訳結果の中に含まれる類義語の訳語を対訳辞書から得られる専門用語の訳語でおきかえるという手法を提案する．この手法により，ユーザは任意の NMT に対して外付けでユーザ辞書を連携させることができ，専門用語を含む文の翻訳が可能になる．このアプローチを実現する上で解決するべき技術課題として，以下の二つが挙げられる．

1. 小規模コーパスからの類義語の獲得

   専門用語は大規模コーパスの中にさえほとんど出現しないため，モデルを学習するにはデータが足りないということがある．その場合，適切な類義語を獲得することが非常に困難である．したがって，小規模コーパスから適切な類義語を獲得する手法が必要である．

2. 翻訳機に適用する類義語の選択

   類義語の中のどの単語を使用すれば良い翻訳結果を出力できるかは，使用する翻訳機に依存する．したがって，翻訳機に適用する類義語の選択手法が必要である．

*1 https://www.miraitranslate.com/index.php#txt-t

2. 未知語の翻訳

NMT には，専門用語を含む文を翻訳することが困難であるという欠点がある．計算の複雑さを制御するために，NMT は小さな語彙を使用しなければならず，語彙外の大量の専門用語はすべて単一の記号 (以下，unk 記号) に置き換えられるからである．このアプローチにより，NMT モデルが専門用語の翻訳を学ぶことができないという問題が生じている．したがって，入力文中のある単語や，その単語の翻訳結果がその頻出単語群に含まれていない場合，入力文の単語と出力文の単語の対応関係が取れず，モデルはその単語の適切な翻訳を生成することができない．また，最悪の場合には，一部の語やフレーズが翻訳結果から欠落してしまう可能性がある [Li 16]．この問題は，語彙外の単語という意味の out-of-vocabulary word の頭文字をとって OOV 問題として知られている．この問題に対して，これまで様々なアプローチから研究がなされてきた．これらのアプローチは，大きく分けて 2 種類に分類することができる．一つは，ニューラル翻訳モデル学習時にカスタマイゼーションを施す手法であり，もう一つは，翻訳時にカスタマイゼーションを施す手法である．

2.1 学習時のカスタマイゼーション

[Li 16] では，学習段階においては，まずモノリンガルコーパスから単語の分散表現を学習する．得られたモデルは単語間

の類似度を評価するために用いられる．それと同時に，バイリンガルコーパスを用いて両言語の文章対の中の単語レベルでの対応を学習する．さらに，副産物として，整列されたバイリンガルコーパスから語彙の翻訳テーブルが導出される．このテーブルと類似度モデルを用いて，バイリンガルコーパス内の専門用語の単語対は，類似語に置き換えられる．最終的に，得られた新たなバイリンガルコーパスを用いて，NMT モデルを作成する．

[Jean 15] では，専門用語を含む文章を NMT でうまく翻訳できない問題に対して，NMT モデル作成時に学習させる語彙の量を増やすというアプローチを根幹においた研究を行なっている．そもそも，専門用語が学習できない理由は，語彙サイズに比例して大きくなる計算量を制御するために学習させる語彙数を制限し，それ以外の単語は単一の unk 記号に置き換えて学習させているからである．この研究では，その問題を解決するため，語彙全体をクラスタリングすることで，非常に大きな目標語彙を有するモデルに対してデコードを効率的に行うというアプローチをとっている．

[Luong 15] では，学習時に unk 記号に置き換えられた専門用語がどんな語だったかを追跡することにより，専門用語を含む文章を NMT でうまく翻訳できない問題を解決しようとしてる．具体的には，まず，両言語の対となる文中に含まれる単語の対応関係を情報として含んだデータを用いて訓練し，翻訳後の文中の各 unk 記号に対応する入力文中の語の位置を出力することができるような NMT モデルを作成する．このモデルを用いて翻訳することにより，翻訳後の文に含まれる unk 記号に対応する専門用語の正体を突き止めることができ，最後に対訳辞書等を使って置き換えを行うことで最終的な翻訳結果を出力することができる．

2.2 翻訳時のカスタマイゼーション

[Li 16] は，学習時だけでなく，翻訳時にもカスタマイゼーションが施される．翻訳段階においては，まず，入力文中の専門用語を学習段階で得られた類似度モデルを用いて類義語に置き換える．そして，作成した NMT モデルを用いて翻訳し，翻訳結果の中の類義語の訳語を語彙の翻訳テーブルを用いて専門用語の対訳に置き換えることで，最終的な翻訳結果を獲得する．この手法では，NMT の学習時に使用した類似度モデルや語彙の翻訳テーブルを翻訳の際にも使用しているため，高い翻訳精度を示していると思われる．そして，提案手法は外付けでユーザ辞書を NMT に連携させるアプローチであるため，翻訳時のカスタマイゼーションに分類される．提案手法と [Li 16] との大きな差異は，[Li 16] が自ら作成した NMT モデルを用いて翻訳を行うのに対し，提案手法はユーザが翻訳に使用する任意の NMT に対して適用できるという点である．

3. 専門用語の類義語の獲得

本研究では，word2vec [Lilleberg 15] を用いて獲得された特徴ベクトル間のコサイン類似度が大きいものを類義語と定義する．しかしながら，大規模コーパス中にすら登場回数が少ないような専門用語は，学習を十分に行うことができず，適切な単語ベクトルを生成できないため，類似していない単語が類義語として獲得されてしまう．この問題を解決するために，本研究では word2vec を用いたコーパスの自動生成を提案する．この章では，コーパスの自動生成を用いた専門用語の適切な類義語を獲得する手法について説明する．この手法は，1 回目の類義語学習プロセス，コーパスの自動生成プロセス，2 回目の類義語学習プロセスの 3 つのプロセスから構成される．

3.1 1 回目の類義語学習プロセス

まずはじめに，大規模コーパスと，専門用語が使用されている小規模コーパスを統合し，1 回目の学習用コーパスを作成する．次に，学習用コーパスに対し word2vec による学習を行うことにより，学習用コーパス内の用語の単語ベクトルを学習したモデルを獲得する．最後に，この学習モデルから学習用コーパス内の用語の単語ベクトルを獲得する．ここで得られた各用語の単語ベクトルの類似度を求めることにより，各用語の類義語を獲得することができる．

3.2 コーパスの自動生成プロセス

図 1 は，コーパス自動生成の手順を表したものである．1 回目の学習で獲得された word2vec モデルを用いて，専門用語が含まれた文の中の専門用語以外の名詞を類義語に置き換えることで多様な文を獲得する．

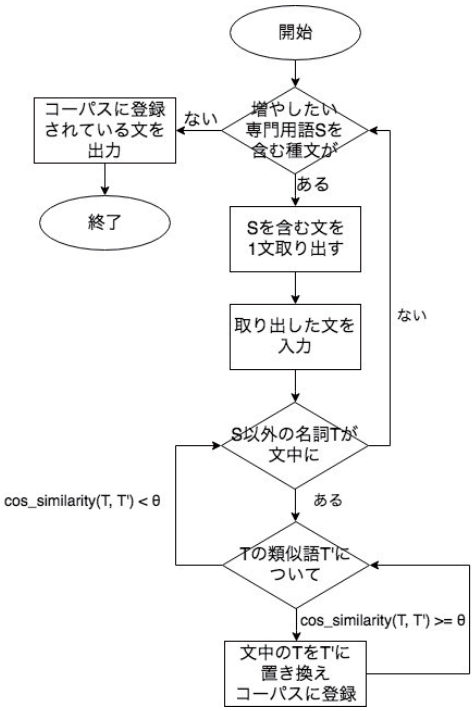

図 1: word2vec を用いたコーパスの自動生成のフローチャート

コーパス自動生成の手順は以下の通りである．まず，増やしたい専門用語 S を含む文（種文とする）を一つ選択する．種文中に S 以外の名詞 T があれば，名詞 T の類義語の中で，T とのコサイン類似度が 1 番高い T’ を選ぶ．名詞 T がなければ，別の種文を選択し直す．T と T’ のコサイン類似度が，閾値 θ 以上ならば種文の中の T を T’ に置き換える．そうでないならば新たな T を選択し直す．T を T’ に置き換えた文を獲得する．その後，T’ を再び選択し同様の処理を繰り返す．（この際，T’ は前回選んだものの次に類似度が高いものを選ぶ．）

3.3 2 回目の類義語学習プロセス

まず，1 回目の学習用コーパスと，自動生成したコーパスを統合し，2 回目の学習用コーパスを作成する．次に，2 回目の学習用コーパスに対し word2vec による学習を行うことにより，コーパス内の用語の単語ベクトルを学習したモデルを獲得する．最後に，この学習モデルから専門用語の単語ベクトルを獲得する．ここで得られた各専門用語の単語ベクトルと，その他の用語との類似度を求めることにより，各専門用語の類義語を獲得することができる．

4. 適切な類義語の選択

この章では，適切な類義語の選択手法として，類似度積を用いた，最終候補となる類義語の選定と，折り返し翻訳及び BLEU 値計算 [Papineni 02] を用いた，最終候補からの類義語の選択を提案する．

4.1 類似度積を用いた類義語選定プロセス

まず，類似度積について説明する．類似度積とは，翻訳対象文中に登場する専門用語の単語ベクトルと，各専門用語に対応する類義語の単語ベクトルとのコサイン類似度の積のことである．まず，提案手法の翻訳を行うための必要十分条件は，類義語で専門用語を置き換えた文の翻訳結果の中に類義語単体の翻訳結果が含まれていることである．したがって，翻訳対象文中に含まれる各専門用語の類義語のうち，この条件を満たし，なおかつ類似度積が高いような類義語の組を最終候補として選定する．

4.2 折り返し翻訳及び BLEU 値計算を用いた類義語の選択プロセス

図 2 は，類義語選択の手順を表したものである．

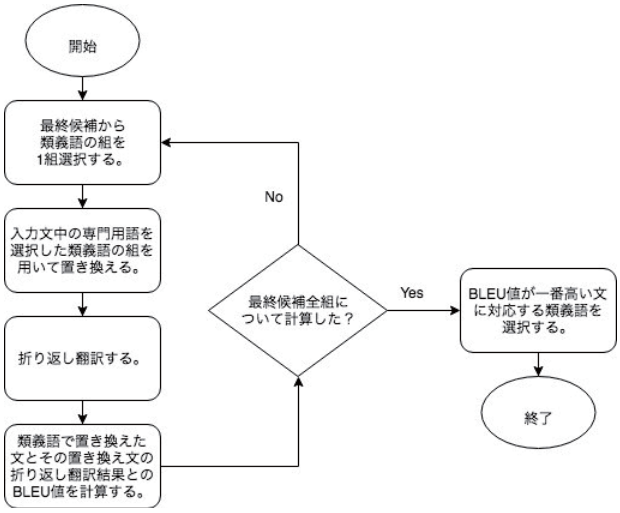

図 2: 折り返し BLEU 値を用いた類義語の選択プロセス

まず，類似度積を用いた類義語選定プロセスによって得られた各類義語の n-組で，入力文中の専門用語を置き換え，置き換え文を作成する．次に，置き換え文を折り返し翻訳する．最後に，置き換え文と，折り返し翻訳結果との BLEU 値を計算し，一番高い BLEU 値を示した際に用いた類義語を選択する．このプロセスにより，最終的に翻訳に使用する類義語が獲得される．

5. 提案手法の評価

ここでは，100 文を対象にして，提案手法の翻訳結果と，専門用語の置き換え語として類義語ではなくランダムな文字列を使用する既存手法 [Murakami 11] を用いた場合の翻訳結果を評価，比較する．今回の実験には，大規模コーパスとして wikipedia コーパス [*1] を，word2vec ライブラリとして gensim[*2] を使用した．翻訳対象の 100 文及び辞書は，京都の観光分野のものを作成した．

∗1 https://dumps.wikimedia.org/jawiki/latest/
∗2 https://radimrehurek.com/gensim/

5.1 平均点による比較

評価方法は，翻訳結果の 5 段階絶対評価であり，評価者は 3 人である．表 1 は，3 人の評価者 A, B, C が二つの手法で実施した 5 段階評価の 100 文の平均点である．

表 1: 3 人の 100 文に対する 5 段階絶対評価の平均値

	A	B	C
提案手法	3.98	3.82	4.10
既存手法	3.86	3.63	4.09

実験の結果，3 人とも提案手法の評価が既存手法の評価を若干上回ったが，t 検定によって統計的に有意な差が得られなかった．

5.2 相対評価による多数決を用いた符号検定結果の比較

次に，100 文に対して，3 人の相対評価による多数決を用いた片側符号検定を行う．この相対評価では，5 段階評価で同点だったものについても強いていえばどちらが良い結果を示したかを答えてもらい，両者に優劣をつけてもらった．どうしても優劣がつけれないものに関しては同点としてもらった．帰無仮説として，「翻訳結果に差がない」という仮説を設定する．また，有意水準 α は 5%とした．

本実験においては，100 文の，2 種類の翻訳手法における 3 人の相対評価の多数決は，表 2 のように分布した．

表 2: 3 人の 100 文に対する相対評価の多数決

提案手法	既存手法	同点
43 文	25 文	32 文

結果，$p = 0.019 < 0.05$ となり，帰無仮説を棄却され，提案手法の統計的優位性が示された．

5.3 翻訳結果の考察

次に，提案手法が，定性的にどのような文に対して効果的な翻訳を行えたのかを考察していく．表 3 の上 3 行は，「仁和寺は平安時代に建てられました.」という文章の二つの手法での翻訳結果である．なお，この文章では専門用語は「仁和寺」「平安時代」であり，提案手法において類義語として選ばれたのはそれぞれ「円城寺」「中世」である．

表 3: 翻訳結果例

翻訳対象	仁和寺は平安時代に建てられました.
提案手法	Ninnaji Temple was built in Heian Period.
既存手法	Ninnaji Temple was built on Heian Period.
翻訳対象	私の父は白雲寺の住職を知っています.
提案手法	My father knows the priesthood of Hakuunji Temple.
既存手法	My father knows the job title of Hakuunji Temple.

二つの翻訳結果の大きな違いは，提案手法では「in Heian Period」と訳されており，辞書連携翻訳では，「on Muromachi period」と訳されている．これは，前者において，「平安時代」が「中世」という時代を司る語に置き換わった結果，適切な前置詞が選択されたと思われる．一方，後者では，「平安時代」が無意味な文字列に置き換えられた結果，その文字列の前に適切な前置詞を配置できなかったものだと思われる．このように，

適切な前置詞を求められる翻訳に対しては，提案手法は有効に働くことが考察される．

次に，表 3 の下 3 行は，「私の父は白雲寺の住職を知っています.」という文章の二つの手法による翻訳結果である．なお，この文章では専門用語は「白雲寺」であり，提案手法において類義語として選ばれたのはそれぞれ「高千穂峰」である．

この翻訳結果においては，提案手法では適切に翻訳されているのに対し，既存手法では無意味な文字列の影響が文全体に伝播し，「住職」の翻訳が不適切なものになってしまったのだと思われる．このように，提案手法においては，既存手法を用いる際に発生した誤訳の伝播を抑える効果が期待される．

6. おわりに

本研究では，専門用語を含んだ文の新たな翻訳手法として，入力文中の専門用語を類義語で置き換え，翻訳した後に翻訳結果に含まれる類義語の翻訳結果を専門用語の対訳と置き換えるという手法を考案した．本研究における技術課題は，以下の二つであった．

1. 専門用語の適切な類義語群の獲得

   出現頻度の低い専門用語の類義語を獲得することは困難であった．したがって，少ないデータから適切な類義語群を獲得する手法が必要であった．解決策として，本研究では word2vec を用いたコーパスの自動生成を考案した．

2. 類義語群からの適切な類義語の選択

   類義語群の中から，正しい翻訳結果を与える類義語を選択する手法が必要であった．解決策として，本研究では，類似度積を用いた類義語選定と，折り返し翻訳及び BLEU 値計算を用いた類義語選択を考案した．

また，提案手法を京都の観光分野のテキストに用いた結果，翻訳精度の向上が示され，既存手法に対する提案手法の優位性が示された．

謝辞

本研究は，日本学術振興会科学研究費基盤研究 (A) (17H00759, 平成 29 年度～32 年度) および日本学術振興会科学研究費若手研究 (A)(17H04706, 平成 29 年度～32 年度) の支援を受けた．

参考文献

[Jean 15] Jean, S., Cho, K., Memisevic, R., and Bengio, Y.: On Using Very Large Target Vocabulary for Neural Machine Translation, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pp. 1–10 (2015)

[Johnson 17] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, *TACL*, Vol. 5, pp. 339–351 (2017)

[Li 16] Li, X., Zhang, J., and Zong, C.: Towards Zero Unknown Word in Neural Machine Translation, in Kambhampati, S. ed., *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pp. 2852–2858, IJCAI/AAAI Press (2016)

[Lilleberg 15] Lilleberg, J., Zhu, Y., and Zhang, Y.: Support vector machines and Word2vec for text classification with semantic features, in Ge, N., Lu, J., Wang, Y., Howard, N., Chen, P., Tao, X., Zhang, B., and Zadeh, L. A. eds., *14th IEEE International Conference on Cognitive Informatics & Cognitive Computing, ICCI*CC 2015, Beijing, China, July 6-8, 2015*, pp. 136–140, IEEE Computer Society (2015)

[Luong 15] Luong, T., Sutskever, I., Le, Q. V., Vinyals, O., and Zaremba, W.: Addressing the Rare Word Problem in Neural Machine Translation, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pp. 11–19 (2015)

[Murakami 11] Murakami, Y., Lin, D., Tanaka, M., Nakaguchi, T., and Ishida, T.: Service Grid Architecture, in Ishida, T. ed., *The Language Grid - Service-Oriented Collective Intelligence for Language Resource Interoperability*, Cognitive Technologies, pp. 19–34, Springer (2011)

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318Association for Computational Linguistics (2002)