

双方向の逆翻訳を利用したニューラル機械翻訳の 教師なし適応の検討

森田 知熙^{1,a)} 秋葉 友良^{1,b)} 塚田 元^{1,c)}

概要: 本論文では、収集が容易な原言語側および目的言語側の単言語コーパスを用い、統計的機械翻訳の性能を向上させる手法を提案し、ニューラル機械翻訳によるドメイン適応実験で有効性を示す。提案法は、原言語側と目的言語側の単言語コーパスから疑似対訳コーパスを作成しこれを学習データに追加することで、双方向の翻訳システムを再学習する。さらにこれを繰り返して、徐々に疑似対訳コーパスの質を向上させることで翻訳システムの性能を向上させる。本手法により、少量の対訳コーパスと両言語の単言語コーパスがあれば精度の高い翻訳システムの作成が可能となる。対訳コーパスと単言語コーパスが別ドメインであっても適用できるため、統計的機械翻訳の教師なしドメイン適応で特に有効である。ニューラル機械翻訳を用いた実験で、逆翻訳を用いる従来手法でドメイン適応を行った場合と比較して、大幅に翻訳精度が向上することを確認した。

1. はじめに

近年、ニューラルネットワークを用いた系列変換をベースとするニューラル機械翻訳 (NMT) が高い翻訳精度を持つことが報告されている。一方、ニューラル機械翻訳の学習には数十万~数百万文の規模の、大量の対訳コーパスを必要とすることが知られている。大量の対訳コーパスを構築することはコストが高く、翻訳システムを構築する際の大きな問題となっている。また、ドメインの異なる（以降、ドメイン外）対訳コーパスから精度の高いモデルを学習することが困難なため、翻訳対象とドメインの等しい（以降、ドメイン内）対訳コーパスを作成しなければならないことも大きな問題となっている。

そのため、ドメイン内の訓練データが少ない場合、ドメイン外で学習したモデルをドメイン内の対訳コーパスを用いて適応させる研究が行われている。しかし、これらはドメイン内の対訳コーパスが必要であり、その作成コストの問題は依然深刻である。

本稿では、この問題を解決するため、対訳になっていない原言語側と目的言語側の2つの単言語コーパスを活用する教師なしドメイン適応手法を提案する。この方法では、翻訳システムに変更を加える必要はなく、対訳コーパスに

翻訳システムにより生成された疑似対訳コーパスを追加するだけで実行可能である。また、両方向の翻訳システムを更新する手法のため、同時に両方向の翻訳システムの性能を改善できる利点もある。

英日翻訳、日英翻訳を対象にドメイン適応の実験を行った結果、適応前に比べて英日方向で+15.87、日英方向で+9.9とBLEU[1]が大きく改善し、翻訳精度が向上していることを確認した。Sennrichらの先行研究[2]との比較でも、英日方向で+7.86、日英方向で+8.04と改善し、提案法の効果を確認した。さらに、ドメイン内の半教師あり学習との比較から、学習を繰り返す提案法はドメイン適応においてより効果的であることが分かった。

2. 関連研究

2.1 ニューラル機械翻訳の半教師あり学習手法

比較的少量の対訳コーパスと共に、入手が容易な単言語コーパスを利用する機械翻訳の半教師あり学習手法が提案されている。Gülçehreら[3]は目的言語側の単言語コーパスを使い学習したRNN言語モデルをニューラル機械翻訳と統合する2つの手法を提案している。Sennrichら[2]は、目的言語側の単言語コーパスを逆翻訳することで得られる疑似対訳コーパスを学習データに追加する手法を提案している。Zhangら[4]は目的言語側と原言語側の両方の単言語コーパスを利用し、両方向の翻訳システムを連結して得られるオートエンコーダとのマルチタスク学習を行う手法を提案している。Xiaら[5]は対訳コーパスから学習する

¹ 豊橋技術科学大学
Toyohashi University of Technology

a) morita@nlp.cs.tut.ac.jp

b) akiba@cs.tut.ac.jp

c) tsukada@brain.tut.ac.jp

翻訳システムの双対性, すなわち対象な2つの方向の翻訳システムが構築できることに注目し, 両モデルを強化学習により同時に学習する方法を提案している. Zhang ら [6] は, 同様に両方向の翻訳モデルを逆翻訳を利用して同時に学習する手法をEMアルゴリズムで定式化している. この手法は, 本論文の提案手法と同様のアイデアに基づくものであるが, 提案法は作成した疑似対訳コーパスの対訳毎に重みを用いておらず, 一般的な対訳コーパスを用いた機械翻訳の学習の枠組みで実現できる点が異なる. また, 本稿では, この手法のドメイン適応での効果を調査した.

2.2 ニューラル機械翻訳のドメイン適応手法

機械翻訳のドメイン適応手法として最も標準的な手法は, ドメイン外の対訳コーパスで事前学習した翻訳モデルを少量のドメイン内対訳コーパスで fine-tuning する方法である. Chu ら [7] は, 原言語側の文にドメインを表すタグを追加するマルチドメイン学習と, fine-tuning を組み合わせるドメイン適応手法を提案している. また, 上記の半教師あり学習 [2][3][4][5] は, ドメイン内の学習データとして単言語コーパスのみを用いる機械翻訳の教師なし適応手法としても利用できる. Sennrich ら [2] は, 目的言語側の単言語コーパスから構築した疑似対訳コーパスがドメイン適応にも有効であることを示している.

3. ニューラル機械翻訳

本稿では, エンコーダ, デコーダ, アテンション機構から構成されるアテンションベース NMT システム [8] を用いる. エンコーダ, デコーダは再帰ニューラルネットワーク (RNN:Recurrent Neural Network) により構成される. このモデルは, 原言語が与えられたとき, 対訳文の尤度が最大となるように学習する.

入力文 $\mathbf{x} = x_1, x_2, \dots, x_n$ とそれまでに出力した単語列 y_1, y_2, \dots, y_{t-1} が与えられたとき, t 番目の単語 y_t の出力確率を以下の式で求め, 翻訳候補を決定する.

$$p(y_t | y_1, y_2, \dots, y_{t-1}, \mathbf{x}) = \text{softmax}(W_s \tilde{h}_t) \quad (1)$$

ここで, W_s は重みパラメータであり, \tilde{h}_t は, 文脈ベクトル c_t と h_t から以下の式で求める.

$$\tilde{h}_t = \tanh(W_c [c_t; h_t]) \quad (2)$$

h_t はデコーダ側の隠れ状態を表す, c_t は全てのエンコーダの隠れ状態にわたる加重平均であり, エンコーダ側の隠れ状態 \bar{h}_s とアライメント重み $a_t(s)$ により以下の式で求める.

$$c_t = \sum_{s=1}^S a_t(s) \bar{h}_s \quad (3)$$

$a_t(s)$ は \tilde{h}_t と \bar{h}_s をスコア関数で比較することで得られ

る. 本稿では, アテンションモジュールに, Luong ら [8] の dot-product attention を採用し, 以下の式で求める.

$$a_t(s) = \frac{\exp(h_t^\top \bar{h}_s)}{\sum_{s'} \exp(h_t^\top \bar{h}_{s'})} \quad (4)$$

4. 提案手法

提案手法は, ドメイン外の対訳コーパスの他に, 原言語側と目的言語側のドメイン内単言語コーパスを用いてドメイン適応を行う. 2つの単言語コーパスは対応付けられたコンパラブルコーパスである必要はない. また, 提案法は両方向の翻訳システムを同時に学習するので, 原言語と目的言語の区別は重要ではなく, 2つの言語は対等である. 以降, 2つの言語をそれぞれ X, Y と記し, 言語 X から Y への翻訳を X-Y, Y から X への翻訳を Y-X と記す.

提案法の手順は以下の通りである. (図 1)

- 1 ドメイン外の対訳コーパス D_X^{out}, D_Y^{out} から X-Y, Y-X の両方向の翻訳システムを学習する. 以降, これをモデル 0 と呼ぶ.
 - 2 X-Y 翻訳システムを以下の手順で再学習する.
 - 2.1 Y のドメイン内単言語コーパス D_Y^{in} から Y-X 翻訳システムのモデル i を用いて翻訳結果 $D_X^{in'}$ を得る.
 - 2.2 $D_X^{in'}$ と D_Y^{in} の組を疑似対訳コーパスとして D_X^{out} , D_Y^{out} と混合し, X-Y 翻訳システムを学習し, モデル $i+1$ とする.
 - 3 Y-X 翻訳システムを以下の手順で再学習する.
 - 3.1 X のドメイン内単言語コーパス D_X^{in} から X-Y 翻訳システムのモデル i を用いて翻訳結果 $D_Y^{in'}$ を得る.
 - 3.2 $D_Y^{in'}$ と D_X^{in} の組を疑似対訳コーパスとして D_Y^{out} , D_X^{out} と混合し, Y-X 翻訳システムを学習し, モデル $i+1$ とする.
 - 4 $i \leftarrow i+1$ としてステップ 2 に戻る.
- 上記の 2~4 の手順を繰り返し, 疑似対訳コーパスの品質を向上させることで, 翻訳システムの性能改善を図る.

5. 実験

5.1 実験目的

5.1.1 繰り返し学習による教師なし適応

本実験では, 提案手法のドメイン適応での効果を調べるため, 単言語コーパスを用いて提案手法で適応を行い, ドメイン内テストデータに対する翻訳精度の調査を行う. 翻訳精度の評価には BLEU[1] を使用する.

初期の翻訳システムの学習に用いるドメイン外対訳コーパスは, 適応の際にもドメイン内の疑似対訳コーパスと共に学習に利用することができる. しかし, 正確な対訳ペアであるもののドメイン外である対訳コーパスをドメイン内の学習に用いる効果は明らかではない. そこで, ドメイン適応の際にドメイン外対訳コーパスを付加した場合と付加しない場合での実験も行う.

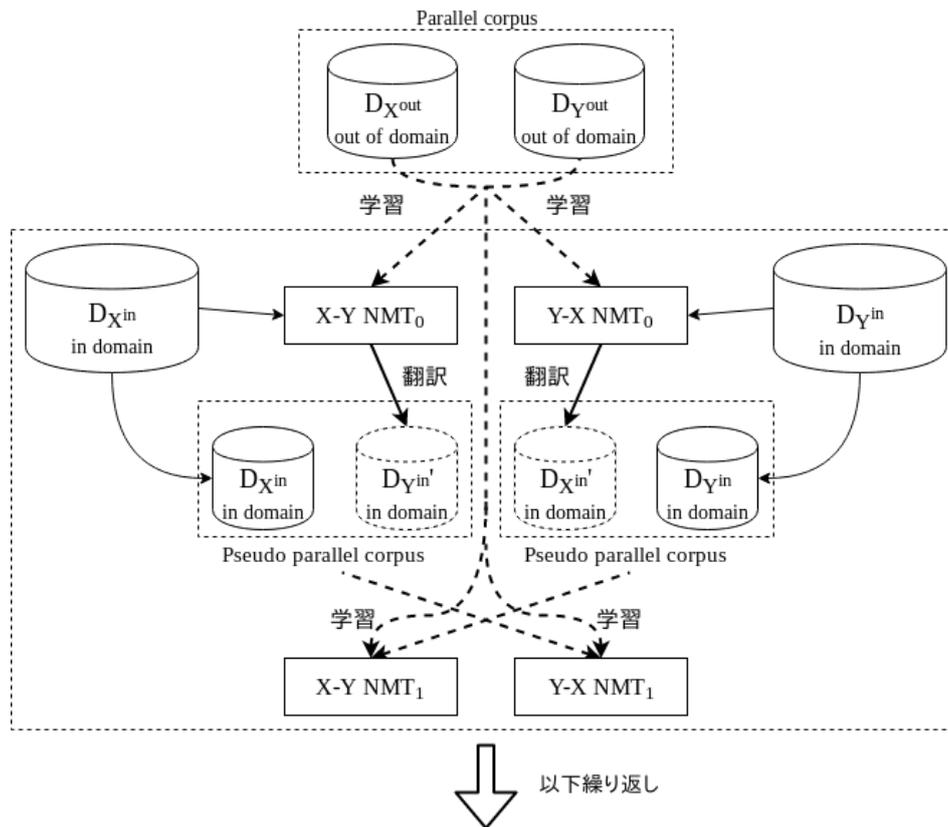


図 1 原言語側と目的言語側の単言語コーパスを用いた同時学習のフロー

5.1.2 同一ドメイン内での半教師あり学習

提案手法は機械翻訳の半教師あり学習手法としても利用可能である。すなわち、初期の教師あり対訳コーパスにもドメイン内コーパスを利用した場合に相当する。そこで、初期対訳コーパスの質の差が最終的な翻訳精度に与える影響を調べるために、初期対訳コーパスをドメイン内コーパスに置き換えた実験も行い、性能の比較を行った。

5.2 データセット

ドメイン外の教師ありデータには Asian Scientific Paper Excerpt Corpus (ASPEC)[9] の英日対訳コーパスを利用した。逆翻訳システムの学習には対訳文の全文 (1,000,000 対) を用いた。

ドメイン内の教師なし単言語コーパスには NTCIR-8 PATMT の英日対訳コーパスを利用した。また、目的言語側と原言語側の単言語コーパスに対訳となるような文ペアが含まれないことを確実にするために、このコーパスの先頭 10 万行を除いた 3,086,284 対を 2 分割し、前半側からは英語のみ、後半側からは日本語のみを抽出し、単言語コーパスとして用いた。テストデータは NTCIR-8 PATMT の学習データ以外のテストセットから 899 対、開発データは 2000 対を用いた。両方のコーパスとも、日本語文は Mecab により分かち書きを行った。

5.3 実験条件

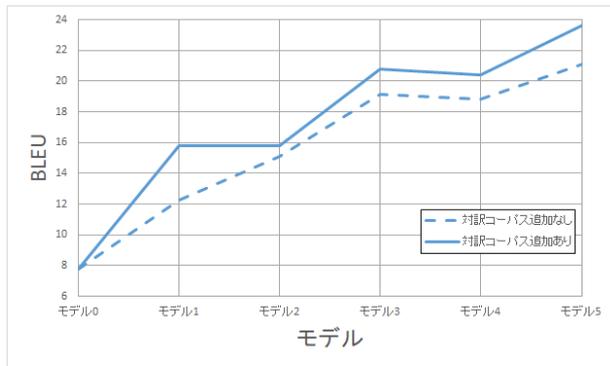
ニューラル機械翻訳システムには OpenNMT[10] を用いた。英日翻訳システム、日英翻訳システム共にエンコーダは 1 レイヤーの BRNN (500 次元)、デコーダは 1 レイヤーの RNN (500 次元) とし、学習アルゴリズムは Adam とし、学習率は 0.001 で 10 エポック学習した。ボキャブラリサイズは 3 万とした。

5.3.1 繰り返し学習による教師なし適応

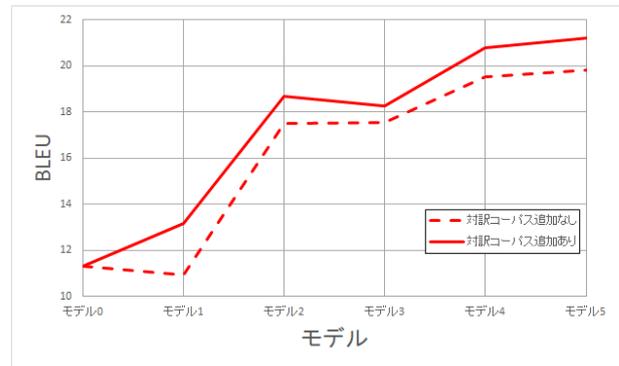
APSEC コーパスの全文から英日、日英の双方の翻訳システムを学習した (以降、これをモデル 0 と呼ぶ)。次に学習した日英翻訳システムで NTCIR8-PATMT コーパスから抽出した日本語の単言語コーパスを翻訳し、英語の翻訳結果を求めた。これを元の単言語コーパスとペアにして疑似対訳コーパスを作成した。この対訳コーパスと元の ASPEC コーパスを混合し (あるいは混合せずに) 英日翻訳システムを再度学習した。この際に翻訳システムのボキャブラリには、疑似対訳コーパスだけを用いて出現頻度の高い 30K を選択した。同様の処理を英語の単言語コーパスに対し適用し、日英翻訳システムを再学習した。この翻訳システムをモデル 1 と呼び、以降翻訳システムを再学習するごとにモデル 2, モデル 3, ... モデル n と呼称する。また、翻訳システムの再学習の際に、疑似対訳コーパスと ASPEC 対訳コーパスを混合して学習するモデルの他に、疑似対訳コーパスのみで学習したモデルも作成し、性能の

表 1 繰り返し学習におけるドメイン適応とドメイン内での半教師あり学習における各モデルの BLEU 値

| 教師あり対訳コーパス | 適応時の初期対訳コーパスの追加 | 英日翻訳 | 日英翻訳 |
|---------------|-----------------|-------|-------|
| ASPEC (モデル 0) | あり | 7.78 | 11.31 |
| ASPEC (モデル 1) | あり | 15.79 | 13.17 |
| ASPEC (モデル 5) | あり | 23.65 | 21.21 |
| ASPEC (モデル 1) | なし | 12.28 | 10.91 |
| ASPEC (モデル 5) | なし | 21.09 | 19.80 |
| NTCIR (モデル 0) | あり | 25.30 | 25.21 |
| NTCIR (モデル 1) | あり | 32.42 | 30.87 |
| NTCIR (モデル 5) | あり | 35.14 | 33.37 |



(1) 英日翻訳モデルの各学習段階における BLEU 値の推移



(2) 日英翻訳モデルの各学習段階における BLEU 値の推移

図 2 繰り返し学習によるドメイン適応における各学習段階での BLEU の推移

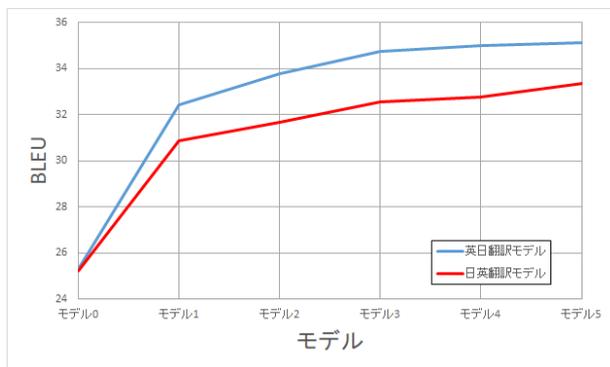


図 3 同一ドメインでの半教師あり学習における各学習段階での BLEU 値の推移

比較を行った。

5.3.2 同一ドメイン内での半教師あり学習

ドメイン内の教師あり対訳コーパスとして、NTCIR8-PATMT コーパスの内、単言語コーパスとして使用しなかった先頭 10 万対を使用し、モデル 0 を学習した。その後は、5.3.1 説で述べた教師なし適応と同じ手順で、NTCIR8-PATMT の単言語コーパスを用いて両方向の翻訳システムの再学習を繰り返した。

5.4 実験結果

図 2, 図 3 は、それぞれの実験条件において、各学習回数における BLEU の推移を表している。表 1 に、各条件

での BLEU 値を示す。表 1 に示すモデルはドメイン適応を行ったモデルとドメイン内での半教師あり学習を行ったモデルで初期の翻訳システムの学習に用いた対訳コーパスが異なる。その後の学習に用いた単言語コーパスは全て同じである。

5.4.1 繰り返し学習による教師なし適応

図 2 に結果を示す。なお、モデル 1 は Senrriich ら [2] の手法に相当する。図 2(a) の英日翻訳、(b) の日英翻訳共にモデルを再学習するごとに BLEU が向上していることがわかる。表 1 の結果では、ドメイン適応ではドメイン外の対訳コーパスも学習データに付加したモデルが最も BLEU 値が改善され、モデル 0 (通常ドメイン適応に相当) に対しモデル 5 では、英日翻訳で BLEU 値が +15.87, 日英翻訳で +9.9 改善された。また、Senrriich ら [2] の手法に相当するモデル 1 と比べても、英日翻訳で BLEU 値が +7.86, 日英翻訳で +8.04 改善されている。これらにより提案手法である繰り返し学習が翻訳精度の大幅な改善に効果的であることがわかる。また、ドメイン外の対訳コーパスを学習データに追加したモデルの方が BLEU 値が改善されていることから、ドメイン外の対訳コーパスであってもドメイン内での翻訳システムの学習に貢献することを示している。また、図 2(a) の英日翻訳モデルの BLEU と (b) の日英翻訳モデルの BLEU を比べると、日英翻訳ではモデル 2 の学習時に BLEU が大きく向上し、英日翻訳ではモデル 3 の学習時に BLEU が向上している。また、日英翻訳システ

ムと英日翻訳システムのBLEUは、同時に上昇するのではなく、交互に上昇していることがわかる。これは、英日翻訳に比べて日英翻訳モデルの精度が高く、最初に作成した疑似対訳コーパスの品質に差があり、その品質差が後のモデルの精度に影響しているためだと考えられる。学習回数を重ねる毎に、2つのモデルの精度の差が縮まり、学習曲線が滑らかになっていることが見て取れる。

5.4.2 同一ドメイン内での半教師あり学習

表1に示す半教師あり学習のモデルでは、初期の翻訳システムであるモデル0に対し、BLEUが英日翻訳で+9.84、日英翻訳で+8.16改善されている。また、モデル1に対しても英日翻訳で+2.72、日英翻訳で+2.50改善されているため、同一ドメイン内での半教師あり学習においても繰り返し学習が翻訳精度を改善できることがわかる。このモデルでは、NTCIR8-PATMTコーパスの先頭10万文のみを対訳コーパスとして使用したが、ASPECコーパスを100万文対訳コーパスに使用したドメイン適応のモデルのよりも最終的なBLEUが高く(+11.49~12.16)なっていることから、対訳コーパスのデータサイズよりも単言語コーパスとのドメインの近さが学習により大きな影響を与えていると言える。学習を繰り返した後もBLEUには明確な差が現れているため、初期の対訳コーパスの選択がその後の学習にも大きく影響を与えていることがわかる。

図2と図3を比較すると、図3の同一ドメインでの半教師あり学習ではモデル3以降は、BLEU値の改善が横ばいであるのに対し、図2のドメイン適応では依然としてBLEU値の改善が進んでいることが見て取れる。このことから、ドメイン適応では提案法の特徴である学習を繰り返すことの効果が大きいことがわかった。

6. まとめ

本稿では、対になる2つのモデルを作成し、単言語コーパスの逆翻訳と学習を繰り返すことで翻訳モデルを逐次改善する手法を提案し、ドメイン適応の有効性について検証した。単言語コーパスを用いて学習を繰り返すことの効果は大きく、その効果はドメイン適応においてより発揮されることが確認できた。また、対訳コーパスの量よりも単言語コーパスとのドメインの近さの方が翻訳システムの品質に与える影響が大きく、学習を繰り返しても影響が残ることを確認した。

本研究では、対訳コーパスと単言語コーパスは別ドメインのものを使用したが、単言語コーパスについては、同一ドメインの対訳コーパスを分割することで作成した。そのため、今後の研究としては、原言語側と目的言語側の単言語コーパスが別ドメインである場合も調査したい。また、提案法にマルチドメイン学習の手法[11]を適用することも検討したい。

参考文献

- [1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. *BLEU: a method for automatic evaluation of machine translation*. In Proceedings of ACL, pp.311–318, 2002
- [2] Rico Sennrich, Barry Haddow, and Alexandra Birch. *Improving neural machine translation models with monolingual data*. In Proceedings of ACL, pp.86–96, 2016.
- [3] Caglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Hui-Chi Lin, Fethi Bougares, Holger Schwenk and Yoshua Bengio. *On using monolingual corpora in neural machine translation*. In Proceedings of CoRR, abs/1503.03535, 2015.
- [4] Jiajun Zhang, and Chengqing Zong. *Exploiting source-side monolingual data in neural machine translation*. In Proceedings of EMNLP, pp.1535–1545, 2016.
- [5] Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu and Wei-Ying Ma. *Dual learning for machine translation*. Advances in Neural Information Processing Systems 29, pp.820–828, 2016.
- [6] Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou and Enhong Chen. *Joint Training for Neural Machine Translation Models with Monolingual Data*. arXiv:1803.00353, 2018.
- [7] Chenhui Chu, Raj Dabre and Sadao Kurohashi. *An Empirical Comparison of Simple Domain Adaptation Methods for Neural Machine Translation*. In Proceedings of ACL, pp.385–391, 2017.
- [8] Minh-Thang Luong, Hieu Pham and Christopher D. Manning. *Effective Approaches to Attention-based Neural Machine Translation*. In Proceedings of EMNLP, pp.1412–1421, 2015.
- [9] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi and Hitoshi Isahara. *ASPEC: Asian Scientific Paper Excerpt Corpus*. In Proceedings of LREC, 2016
- [10] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart and Alexander M. Rush. *OpenNMT: Open-source toolkit for neural machine translation*. arXiv:1701.02810, 2017.
- [11] Catherine Kobus, Josep Crego and Jean Senellart. *Domain control for neural machine translation*. arXiv:1612.06140, 2016.