局所的トピック情報を利用した論文抄録(ASPEC)の英日機械翻訳

渡邊拓斗[*1]　高田凌平[*2]　佐橋 広也[*2]　山本一公[*1]　秋葉友良[*2]　中川　聖一[*1]

[*1] 中部大学　[*2] 豊橋技術科学大学

[*1] {ep15150-0772@sti. kazumasayamamoto@isc. nakagawa@isc.}chubu.ac.jp [*2]{rtakada　sahashi　akiba}@nlp.cs.tut.ac.jp

1. はじめに

大量のパラレルコーパスを用いたニューラルネットワークによる機械翻訳（ニューラル翻訳、NMT）よって機械翻訳性能は、従来の統計的機械翻訳（SMT）と比べて、飛躍的に向上した。NMT の改善法として、アテンション機構の導入、未知語に頑健なバイトペア単位の利用、SMT と NMT の統合化手法、NMT の翻訳結果のリスコアリング[1,2]、文脈情報の利用[3,4,5,6]などが提案されている。文献[3]は LDA を用いたトピック情報の利用、文献[4]は当該文と直前の文を入力に加える方法、文献[5]は直前 3 文から文ベクトルを抽出し文脈ベクトルとして利用、文献[6]は同一文内のコンテクストの一貫性を利用している。

本稿では、音声認識用の言語モデルのトピック適用法[7]を参考にして、当該翻訳文の前後の文の情報を局所的トピック情報として利用する方法を提案し、論文抄録である ASPEC コーパスで評価した結果を報告する。

2. ニューラル機械翻訳

NMT の主流であるエンコーダ-デコーダモデルについて説明する。原言語 F の入力文を単語レベルの埋め込みベクトルに変換してエンコーダへ入力する。エンコーダから出力される分散表現は入力文の意味や構造を捉えた文ベクトルとなる。文ベクトルをデコーダに入力した場合、最初の目的言語の単語 $e1$ を出力確率によって予測する。次の単語を予測するために、出力された単語を入力として与え、終端記号が予測されるまで単語の予測を繰り返し、最終的に目的言語文 E を出力する (図 1)。単語の予測の際にそれぞれどの原単語に対して注目するかを与えるために、エンコーダから出力される単語ベクトルに重みをかけるアテンション機構によって制御する 。θ をモデルのパラメータとしたとき、デコーダの計算式は、以下のように表すことができる。

$$P(\boldsymbol{E} \mid \boldsymbol{F};\theta) = \prod_{j=1}^{J} P(e_j \mid \boldsymbol{F}; \boldsymbol{E}_{<j}; \theta)$$

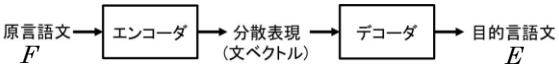

図 1: NMT のブロック図

3. 局所的トピック情報の利用

3.1 単語分散表現の利用

エンコーダ-デコーダモデルの原言語文の入力は、1 単語ずつ、単語列として入力されていくが、この単語表現は 1 hot ベクトル表現である（つまり、当該単語に対応するビットのみ 1 で、他のビットはすべて 0 の語彙サイズの 0,1 からなるベクトル）。この 1 hot ベクトルは、全単語に対して、翻訳上関連する単語が似通った表現になるように、単語分散表現として学習される。通常は 500 次元程度を用いる。単語分散表現は、ベクトルの加減算に対して、近似的に意味表現空間上で閉じている（例：＜王様＞－＜男性＞＋＜女性＞＝＜女王＞；＜＞は分散表現）。つまり、入力文の単語の分散表現の総和または平均値は、大まかな文の意味を表すと考えられる。

本研究では、文単位から局所的なトピック情報を抽出するために、原言語文の文ごとの名詞のみを利用して、名詞の分散表現の総和もしくは平均値を局所的トピック情報として用いる。英語文から名詞を抽出するためには、treetager [†]を使用した。

翻訳対象の当該原言語文の直前文、もしくは後続文の名詞による分散表現を局所的トピック情報として、図 2 の単語分散表現の先頭に追加する。

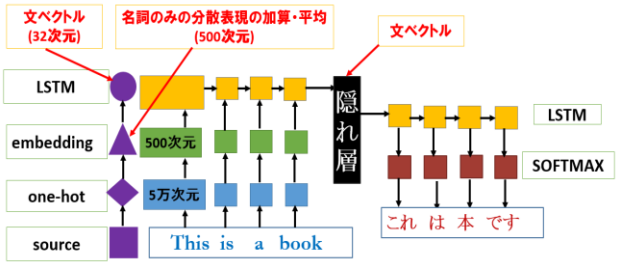

図 2 : NMT の構造と追加する局所的トピック情報

3.2 文ベクトルの利用

エンコーダ-デコーダモデルでは、エンコーダで原言語文を隠れベクトル（本稿では、文ベクトルと呼ぶ）に変換し、デコーダで、この文ベクトルと原言語文、直前までに予測された目的言語文を入力として目的言語の次単語を予測していくモデルである。エンコーダから出力さ

[†] http:/ www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

れる文ベクトルは原言語の入力文の意味や構造を表現した高次元連続空間上の実数値であると考えられる[8,9]。この文ベクトルは、原言語文の情報を保存しているため、局所的トピック情報としては詳細過ぎる。通常、隠れベクトルは 500 次元程度を用いるが、これを8次元とか 16 次元に低次元化して、翻訳モデルを学習すれば、この時の文ベクトルからは目的言語の翻訳が精度よくできず、おおまかな内容の文が生成される。これを局所的文脈情報として利用する。

翻訳対象の当該原言語文の直前文、もしくは後続文の低次元の文ベクトルを局所的トピック情報として、図２の LSTM 分散表現の先頭に追加する。

４．ASPEC の英日翻訳実験

4.1 ASPEC コーパス

論文の抄録である ASPEC コーパスは、英日の 100 万文のパラレルコーパスからなる。研究分野として 23 分野とその他の分野の計 24 分野に分けられており、100 万文をこの分野別にソーティングした。また各論文の抄録は数文から構成されており、3〜5 文程度が大部分であり、この抄録通りの順にソーティングされている(図 3 参照)。100 万文のうち研究分野が変わる文境界は 23 か所であり、無視できる程度である。一方、論文抄録の境界は約 20 万弱箇所あり、無視できないが、このまま学習に利用した。つまり、論文抄録の境界文では、当該文の直前または直後の文が、同じ論文抄録の文ではないことになり、悪影響がある。しかし、テスト文の 1812 文については、このような境界文（399 箇所 399 文）については、評価対象文から除外した。

直前文: it was used for the monitoring in a radiotherapy facility and gave a practically satisfactory result .
当該文: gamma-ray emission probabilities for 187w have been determined with uncertainties less than 1 % from measurement of absolute γ-ray intensities and disintegration rate .
直後文: investigation of the hint for challenge for future was made through review of the history of technology and typical inventions in the twentieth century .

直前文: 放射 線 治療 施設 の モニタ に 使用 し , ほぼ 満足 できる 結果 を 得 た 。
当該文: 絶対 γ 線 強度 と 崩壊 率 を 測定 し , 187w の γ 線 放出 確率 を 1 % 以下 の 不確かさ を 以て 決定 した 。
直後文: 20 世紀 の 技術 史 と 代表 的 な 発明 から , 未来 へ の 挑戦 の ヒント を 探 った 。

図３　抄録文章の例

4.2　ＮＭＴの構造

ＮＭＴのエンコーダは双方向 LSTM を用い、隠れユニット数は 500+500 とした。デコーダは単一の LSTM を用い、隠れ層のユニットは 1000 とした。単語の分散表現は 500 次元にした。単語の語彙サイズは 50000 語である。学習エポック数は 9 回である（文献[2]は 10）。

局所的トピック情報を表す文ベクトルは、LSTM 層の隠れユニット数を双方向で 16+16 または 32+32 として学習し、全学習データ文と全テストデータ文を、文ごとに 16+16 次元または 32+32 次元の文ベクトルに変換した。なお、比較の為に直後の文の 500 次元を加えることも行った。この文ベクトルを 500 ユニットの隠れ層の先頭に追加する場合は、前向き方向と後ろ向き方向のベクトルの先頭それぞれに追加し、残りのユニットにはゼロを詰めた。単語分散表現は、固定した場合と、再学習した場合を行った。NMT の学習には、OpenNMT ‡ツールを使用した。

4.2　翻訳結果

(a) 文ベクトルの利用

まず、当該翻訳対象文の直後の文の文ベクトルを局所的トピック情報として用いた BLEU の結果を表１に示す。図３の各文に対する 16 次元の文ベクトルと 32 次元による文ベクトルから日本語に翻訳した結果を図 4 に示す。図からわかりように、文ベクトルは、トピックをとらえているとは言い難い。そのため、この方法では、抄録文章境界文で悪い影響を受けることなどでベースラインを上回ることができなかった。

(16次元)
直前文: その 結果 , 検討 した 。
当該文: 2 . 0 % である 。
直後文: 今後 について , 今後 の 概要 を 紹介 した 。
(32次元)
直前文: 放射 線 施設 として , 放射 線 施設 として , 実用 的 な 結果 を 示 した 。
当該文: x 線 吸収 確率 を 測定 した 。
直後文: 今後 の 研究 について 解説 した 。

図４　図３の文ベクトルからの日本語翻訳結果

(b)名詞の分散表現の利用

当該翻訳対象文の直前の文および直後の文に現われる名詞の分散表現の平均ベクトルを局所的トピック情報として用いた。図 3 の各文から抽出された名詞を図 5 に示す。なお、名詞の分散表現の総和の利用は平均値の利用よりも翻訳性能は悪くなった。

直前文: monitoring/radiotherapy/facility/result
当該文: gamma-ray/emission/probabilities/uncertainties/%/measurement/γ-ray/intensities/disintegration/rate
直後文: investigation/hin/tchallenge/future/review/history/technology/inventions/century

図５　図３の各文から抽出された名詞

この局所トピック情報を利用して翻訳した BLEU の結果を表 2 に示す。全テスト文を使用した場合は、テスト文のうち、抄録文章の先頭の文は直前の文が利用できなく、最後の文は直後の文が利用できない。そこでこれらの文については、ベースラインの結果を用いたのが表 2(a)である。一方、これらの文については評価から除外したのが表 2(b)である。表から、直前の文および直後の文の効果には大きな差はなかった。単語の分散表現は最初に学習した翻訳モデルの値に固定した方が、改めて学習し直すよりも良かった。全テスト文で評価した場合は直後の文を利用することにより BLEU でベースラインの 35.1 から

‡ http://github.com/OpenNMT/OpenNMT-py

35.3 とわずかではあるが改善することができた。一方、抄録文書の先頭文もしくは末尾文を除外して評価した場合は約 0.3 の BLEU の上昇が得られた。図 6 の日本語への翻訳結果例を示す。

なお、4.1 節で述べたように、異なった論文抄録の文を局所トピック情報の抽出に用いる文が、100 万文の学習データ中 20 万文弱あり、これによって、局所トピック情報が十分に学習されていない可能性が大きい。この様な文を除外して学習すれば、局所的トピック情報の利用効果はもっと大きくなると考えられる。

表 1　局所的文ベクトルを利用した翻訳結果
(全テスト文による評価, 直後の文利用)

ベースライン	16次元追加モデル	32次元追加モデル	500次元追加モデル
35.10	34.58	34.81	35.02

表 2　局所的名詞の分散表現を利用した翻訳結果
(a)全テスト文による評価

ベースライン	直前1文（再学習）	直前1文（固定）	直後1文（再学習）	直後1文（固定）
35.10	35.12	35.27	35.02	35.32

(b)抄録文章の境界文を評価から外した場合

ベースライン（境界の直前1文除去の1413文）	直前1文（再学習）	直前1文（固定）	ベースライン（境界の直後1文除去の1413文）	直後1文（再学習）	直後1文（固定）
34.39	34.42	34.61	35.59	35.53	35.87

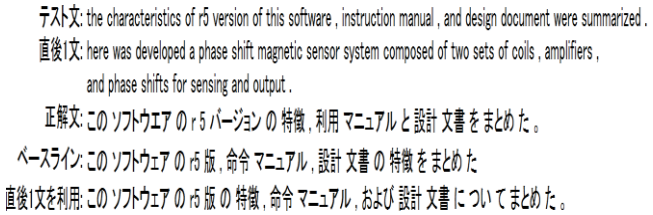

図 6　日本語への翻訳結果例

５．むすび

本稿では、局所的トピック情報を用いたニューラル機械翻訳システムの改善について述べた。局所的トピック情報として、当該翻訳対象文の直前の文および直後の文の低次の隠れベクトル（文ベクトル）表現と名詞の分散表現の平均値を使用する方法を提案し、直後の文の名詞分散表現の平均を利用することにより、BLEU で 0.3 の向上を得ることができた。

なお、高田らは、ASPEC の日英ニューラル翻訳において、24 種類の研究分野のタグを単語の先頭に加えることで BLEU が 1.6 の向上、同じ研究分野の全文の名詞情報を加えることで、1.2 の向上を得ている[10]。

今後は抄録文章の境界文を除いたパラレルコーパスでの学習、 文ベクトルのデコーダへの入力、前後の文にわたる局所トピック情報の利用、などを実装していきたい。

謝辞

本研究は JSPS 科研費 25280062 及び 18H01062 の助成を受けた。

参考文献

[1]今村、隅田、他：双方向リランキングとアンサンブルを併用したニューラル機械翻訳における複数モデルの利用法、情報処理学会、自然言語処理(NL),Ｖｏｌ.2017, No.9, 2017

[2]佐橋、秋葉、中川：科学技術論文抄録と講義音声の英日機械翻訳のリスコアリングの検討、言語処理学会、発表予定、2019.3

[3]J. Zhang, L. Li, A. Way, Q. Liu, “Topic-informed neural machine translation”, Proc. Coling, pp.1807-1817, 2016

[4] J. Tiedemann, Y. Scherrer, “Neural machine translation with extended context”, Proc. EMNLP, pp.82-92, 2017

[5]L. Wang, Z. Tu, A. Way, Q. Liu, “Exploiting cross-sentence context for neural machine translation”, arXiv:1704.04347v3, 2017

[6]B. Zhang, D. Xiong, J. Su, H. Duan, “A context-aware recurrent encoder for neural machine translation”, IEEE/ACM, Trans. ASL, Vol.25, No.12, pp.2424-2432, 2017

[7]W. Naptali, M. Tsuchiya, S. Nakagawa,”Topic-dependent class based n-gram language model”, IEEE Trans. ASL, Vol.20, No.5, pp.1513-1525, 2012

[8]I. Sutskever, O. Vinyals, Q. Le,”Sequence to sequence learning with neural networks”, Advanced in neural information processing systems, pp.3104-3112, 2014

[9] B. Wang, K. Liu, J. Zhao, “Inner attention based recurrent neural networks for answer selection”, Proc. ACL, pp.1288-1297, 2016

[10]高田、秋葉、塚田：ニューラル機械翻訳における文書トピック情報の利用、言語処理学会、発表予定、2019.3