ニューラル機械翻訳における文書トピック情報の利用

高田 凌平　　秋葉 友良　　塚田 元
豊橋技術科学大学
rtakada@nlp.cs.tut.ac.jp, akiba@cs.tut.ac.jp, tsukada@brain.tut.ac.jp

1 はじめに

従来の機械翻訳は、文を単位に処理を行なうように問題設定されてきた。統計的機械翻訳以降の機械学習に基づく機械翻訳では文を単位とした対訳コーパスを学習データとし、原言語の一文を入力として目的言語の一文を出力する翻訳モデルを学習する。このような翻訳モデルでは、文では直接表現されない文脈情報を翻訳に利用できない。例えば、日本語のゼロ代名詞を解消し翻訳するには、対象の文以外から先行詞を特定する必要があるが、文単位の翻訳ではそもそも先行する文を参照することができないため、本質的に不可能である。一方、文書単位で構築された対訳コーパスの整備が進んだことと、多様な情報を柔軟に統合できるニューラル機械翻訳 (NMT) の発展から、文の系列を入力とする機械翻訳の研究が進んできた [6][7][2]。本研究では、首尾一貫した文系列の単位である文書から得られる情報を集約したトピック情報を、ニューラル機械翻訳で利用する手法を提案する。本研究で用いるトピック情報は、翻訳文の前だけではなく後ろを含んだ文脈だと考えることができる。

本論文では、翻訳においてトピック情報を利用する 2 種類の場面を想定し、それぞれの問題設定についてニューラル機械翻訳 (NMT) モデルを提案する。第 1 の問題設定は、翻訳する文にトピックラベルが与えられている場合である。例えば ASPEC コーパスの文には、その文を抽出した論文の分野を表わすラベルが必ず 1 つ付与されている。IWSLT の TED コーパスには、講演の属性を表わすハッシュタグが複数与えられている。これらのトピック情報をニューラル機械翻訳で利用する 2 つの手法を提案する。第 2 の問題設定は、1 文ではなく文書単位で翻訳を行なう場合である。例えば IWSLT の TED コーパスは講演単位で与えられるため、翻訳時に講演全体のテキストが利用可能である。文書からトピック表現を抽出し、ニューラル機械翻訳で利用する手法を提案する。

2 関連研究

文脈情報を利用した NMT に関する研究を紹介する。まず初めに単純な方法で文脈情報を NMT に利用させた研究として Tiedemann ら [6] の研究が挙げられる。Tiedemann らは既存の NMT システムには変更を加えることなく、エンコーダに文脈情報となる以前の文と現在の文を一度に入力することで文脈情報を利用した。

Wang ら [7] は階層的な Recurrent Neural Network(RNN) を使用し、ドキュメントレベル RNN で過去 3 文を 1 つの文脈ベクトルとしてマッピングし、それをセンテンスレベル RNN のエンコーダの初期値とする NMT システムを設計し標準のアテンション機構に基づくエンコーダ-デコーダフレームワークの NMT から BLEU スコアで改善をすることに成功している。Wang らの研究では文脈情報を NMT に利用することで曖昧な単語の訳し分けをすることに成功している。

Maruf ら [2] はセンテンスレベル NMT とメモリネットワークを組合せたドキュメントレベル NMT を提案した。彼らの NMT システムはドキュメントレベルで翻訳を行なうため、より広範な文脈情報を利用することが可能である。そのため、Maruf らの研究では文脈情報を NMT が利用し、語彙の曖昧性の解消や使用語彙の一貫性、代名詞の照応関係などの点でベースラインの NMT から改善することに成功している。

上記の論文から NMT において文脈情報を活用することによって、使用する語彙の一貫性や曖昧な語彙の訳し分け、名詞と代名詞の照応関係、文法上の性がある言語での性の一致など、通常のセンテンスレベル NMT システムでは対処が難しい問題に対して有効であることがわかる。

Tiedemann ら [6]、や Wang ら [7] の研究は翻訳文に先行する文脈を活用するもので、後続する文脈を考慮していない。Maruf ら [2] は学習じは前後の文脈を考慮しいるものの翻訳時には後続する文脈を考慮しきれていない。本研究では文の前後の文脈を考慮するため

に、文書のトピック情報を活用する手法を提案する。

3 提案手法

3.1 トピックラベルを直接利用する手法

第 1 の問題設定として、文にトピックラベルが付与されている場合を考える。NMT でトピック情報を利用する 2 つの手法を提案する。

3.1.1 タグ付加法

本手法の概略図を図 1 に示す。まず、トピックラベルの種類数だけ新たなシンボルを生成し、原言語側のボキャブラリに追加する。そして原言語文を表現する際、文の先頭、または末尾、あるいは両方に文のトピックラベルに対応するシンボルを挿入し翻訳モデルの学習を行なう。評価の際も同様に原言語側にシンボルを挿入し、翻訳結果を得る。図 1 のように'Die Welt ist alles was der Fall ist.' という文が哲学のトピックの文であるとすると、タグ付加法での入力は文頭のみにシンボルを挿入する場合、'#Philosophy Die Welt ist alles was der Fall ist.' となる。

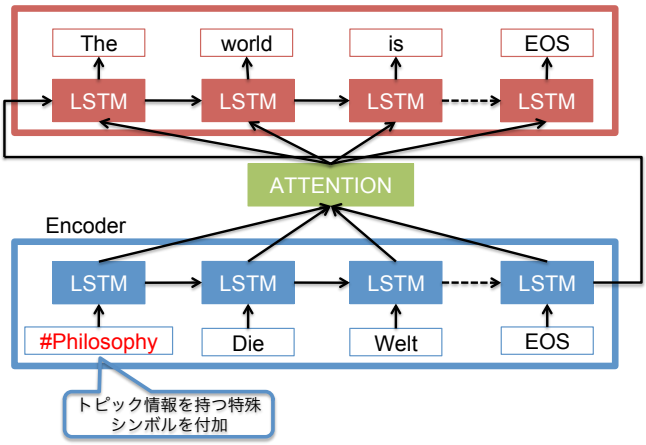

図 1: タグ付加法

3.1.2 タグ変換法

本手法の概略を図 2 に示す。タグ変換法はトピックラベルの情報を RNN 隠れ状態の初期値として利用する手法である。$\mathbf{t}$ を各文に付属するトピックタグと対応する次元だけが 1 でそれ以外の要素が 0 となる N-hot ベクトルとする。以下の式で示す演算を行なうことで、RNN の初期値 $\mathbf{h_{init}}$ とする。

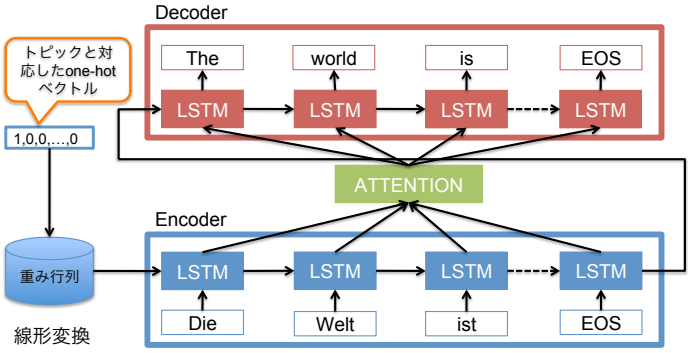

図 2: タグ変換法

$$\mathbf{h}_{init} = \tanh(\mathbf{W}_{topic}\mathbf{t} + \mathbf{b}_{topic}) \quad (1)$$

ここで $\mathbf{W}_{topic}$ と $\mathbf{b}_{topic}$ はそれぞれ重みとバイアスパラメータであり、NMT の他のパラメータと同時に学習する。

3.2 文書からトピック情報を抽出して利用する手法

第 2 の問題設定として、翻訳システムへの入力として文書全体が与えられる場合を考える。文書のトピック情報を単語のワードエンベディングから抽出する。以後この手法を文書平均ベクトル法と呼ぶ。まずベースとなる NMT システムを対象となる対訳コーパスで学習する。訓練されたワードエンベディングを用いて文書に含まれる全ての単語の平均ベクトルを求め文書のトピック情報とし、これを RNN の初期化に利用する。具体的には以下の式で初期化ベクトル $\mathbf{h}_{init}$ は作成される。

$$\mathbf{t} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Emb}_i \quad (2)$$

$$\mathbf{h}_{init} = \tanh(\mathbf{W}_{topic}\mathbf{t} + \mathbf{b}_{topic}) \quad (3)$$

N は文書に含まれる単語の総数、 $\mathbf{Emb}_i$ は文書の中の i 番目の単語のワードエンベディングをす。$\mathbf{t}$ は文書毎に生成する。また $\mathbf{W}_{topic}$ と $\mathbf{b}_{topic}$ は重みとバイアスパラメータであり、NMT の他のパラメータと同時に学習する。(2) 式を計算する際文書の全単語を用い

るのではなく名詞のみを用いる手法も検討した。コーパスから名詞を抽出する際には treetagger[4] を用い、POS タグが'NN' の単語のみを使用した。提案手法の概略図を図 3 に示す。

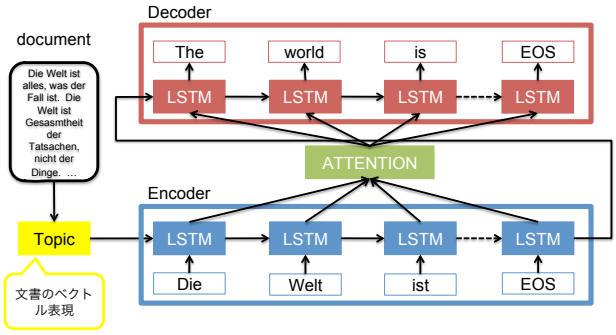

図 3: 文書平均化法

4 実験

4.1 データセット

対訳コーパスには ASPEC 日英対訳コーパスと IWSLT2016 独英対訳コーパスを使用する。ASPEC コーパスは科学技術論文抄録のコーパスで、文単位で様々な抄録か対訳を抽出した対訳コーパスである。各対訳にはそれを抽出した論文が属する研究分野のラベルが 1 つ付与されている。ラベルの種類数 24 である。実験では第 1 の問題設定として、この研究分野ラベルをトピックラベルとして 4.1 節の提案手法の評価に利用する。

IWSLT コーパスは TED の講演書き起こしの話し言葉コーパスで講演単位で対訳を抽出した対訳コーパスである。したがって、各文についてそれが属する文書 (講演) 全体のテキストを参照可能である。実験では、第 2 の問題設定である、翻訳システムへ文書全体を入力する場合として、4.2 節の提案手法の評価に利用する。さらに、各講演には、講演の属性を表すハッシュタグが複数付与されている。ハッシュタグの総数は 472 で各講演平均で 7 個のタグが付与されている。このタグを使い、問題設定 1 の評価にも利用する。

2 つのコーパスの詳細な情報を表 4.1 と表 2 に示す。

ASPEC コーパスでは日本語から英語方向への翻訳を行ない IWSLT コーパスではドイツ語から英語方向への翻訳を行なう。ASPEC では前処理として英語側

表 1: ASPEC の詳細

データセット	文数	分野数
訓練	1,000,000	24
開発	1,790	24
テスト	1,812	24

表 2: IWSLT コーパスの詳細

データセット	文数	文書数
訓練	196,891	1,611
開発	887	8
テスト	1,565	11

で SMT ツールキット Moses[1] に付属したトークナイザによりトークナイズを行ない、小文字化を行なった。日本語側では MeCab[5] を使用して形態素解析を行ない、アルファベットは小文字化した。IWSLT コーパスではドイツ語英語ともに Moses 付属の'tokenizer.perl' を使用してトークナイズした後、小文字化を行なった。

4.2 モデルパラメータ

本研究では、NMT のエンコーダは 1 層の双方向 LSMT を、デコーダは 1 層の単方向 LSTM を使用した。LSTM の隠れ層の次元数は 1,000 とし、ワードエンベディングの次元数は 500 とした。オプティマイザは Adam を使用し、学習率は 0.001 とした。ミニバッチサイズは 256 とし、使用する語彙の数はソース・ターゲット双方の言語で 30,000 語を使用する。

4.3 トピックラベルを利用する手法の実験結果

評価尺度は機械翻訳で広く使用されている BLEU スコア [3] とする。

ASPEC コーパスでのトピックラベルを利用する手法の実験結果を表 3 に示す。提案手法を用いることで精度が改善され、翻訳にトピックラベルを利用することの効果を確認した。提案法の中ではタグ変換法の性能が最も高い。

IWSLT コーパスでの実験結果を表 4 に示す。IWSLT コーパスでは各文に複数のトピックラベル (ハッシュタグ) が付与されているため、ラベルの有無を N-hot ベクトルで表わしてタグ変換法を適用し

表 3: ASPEC でのタグ利用実験結果

手法	BLEU
ベースライン	31.77
タグ付加法・文頭	33.12
タグ付加法・文末	32.71
タグ付加法・両方	33.29
タグ変換法	33.36

表 4: IWSLT コーパスでのタグ変換法実験結果

手法	BLEU
ベースライン	27.60
タグ変換法	27.93

た。IWSLT コーパスを対象とした実験でも、タグ変換実験で翻訳精度が改善され、トピックラベルを利用することの効果が確認できた。しかし、ASPEC コーパスと比較して効果は限定的である。IWSLT コーパスに付与されたトピックラベルは、利用者が任意に選んだキーワードでハッシュタグを構成しているため、ラベルが多様でかつ一貫性が低いためであると考えられる。

4.4 文書のトピック情報を利用する手法の実験結果

次に文書からトピック情報を持つベクトルを作成する文書平均ベクトル法の実験結果を表 5 に示す。

表 5: IWSLT コーパスでの文書平均ベクトル法の実験結果

手法	BLEU
ベースライン	27.60
提案法	28.16
提案法 (名詞のみ)	28.03

提案法はベースラインの精度を改善しており、文書全体の情報を用いることの効果を確認できた。また、表 4 の結果を比べると、トピックラベルを利用する手法よりも精度改善が大きい。低品質の人手作成ラベルよりも、文書全体を用いる方が頑健にトピック情報を抽出できる可能性が示された。一方、トピック情報の抽出に名詞だけを用いることの効果は確認できなかった。今後、種々のトピック情報抽出法を検討してみたい。

5 おわりに

本研究では文に付与されたトピック情報を直接利用利用する 2 つの手法と文書からトピック情報を生成する 1 つの手法を提案し、ASPEC と IWSLT という特性が異なる 2 つのコーパスで実験を行い有効性を確認した。

どちらのコーパスでもトピックラベルの利用により翻訳精度の改善が見られたものの、効果は ASPEC コーパスの方が大きい。ASPEC は学術論文に基づいており、研究分野をす表トピックラベルが厳密かつ排他的に定義されているのに対し、IWSLT のラベルは利用者が任意に選んだ複数のハッシュタグで構成されており品質が低いことが原因であると考えられる。また、文書からトピック情報を抽出する手法は、IWSLT コーパスにおいてトピックラベルを使う手法を改善した。しかしその効果は大きくない。本研究では、トピックの表現として文書のワードエンベディングを平均する単純な手法を用いたが、今後の課題として、より効果的なトピック表現の構築方法を検討したい。

謝辞

本研究は JSPS 科研費 18H01062 の助成を受けた。

参考文献

[1] P.Koehn et al. Moses: Open source toolkit for statistical machine translation. In Proc. of ACL, 2007.

[2] Sameen Maruf and Gholamreza Haffari. Document context neural machine translation with memory networks. In Proc. of ACL, 2018.

[3] K Papineni, S Roukos, T Ward, and W. J. Zhu. Bleu: a method for automatic evaluation of machine translation. In Proc. of ACL, 2002.

[4] Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop. Dublin*. In Proc. of ACL, 1995.

[5] Yuji Matsumoto Taku Kudo, Kaoru Yamamoto. Applying conditional random fields to japanese morphological analysis. In Proc. of EMNLP, 2004.

[6] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In Proc. of ACL, 2017.

[7] Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploiting cross-sentence context for neural machine translation. In Proc. of ACL, 2017.