教師なし英日ニューラル機械翻訳の検討と文の潜在表現の分析

米田 昌弘 † 鶴岡 慶雅 ‡
† 東京大学 工学部電気電子工学科
‡ 東京大学 大学院情報理工学系研究科
{komeda,tsuruoka}@logos.t.u-tokyo.ac.jp

1 はじめに

近年, ニューラル機械翻訳 (Neural Machine Translation; NMT) [7, 4] が提案され, 従来手法である統計的機械翻訳 (Phrase-Based Statistical Machine Translation: PBSMT) を上回る性能を実現し, 大きな注目を集めている. しかし, NMT が高い性能を発揮するためには翻訳の学習に用いられる対訳コーパスが大量に必要となり, 対訳コーパスを大量に得られない言語対, ドメインにおいて高精度な翻訳システムを獲得することが難しいという問題がある.

このような問題に対応する方法として, 一切の対訳コーパスを用いず, 各言語における単言語コーパスのみを用いて翻訳を実現する教師なしニューラル機械翻訳 (Unsupervised Neural Machine Translation; Unsupervised NMT) [2, 3] が提案された. 教師なしニューラル機械翻訳はソース言語とターゲット言語の文章が, 共有された単語埋め込み表現と, 大部分のパラメータを共有したエンコーダにより同一空間上にエンコードされることを期待し, Denoising AutoEncoder [8] としての学習と逆翻訳 [6] により作成される疑似対訳コーパスを用いた翻訳モデルの学習を組み合わせることにより実現される手法である.

本研究では教師なしニューラル機械翻訳を用いて, 英語, 日本語間における翻訳システムの検討を行い、同様の手法で構成した英語, フランス語間における翻訳システムと比較分析を行う.

2 教師なしニューラル機械翻訳

Artetxe らにより提案された教師なしニューラル機械翻訳 [2] では, 一切の対訳データを用いず, 各言語の対訳関係にない大量の単言語コーパスのみにより翻訳システムを実現する.

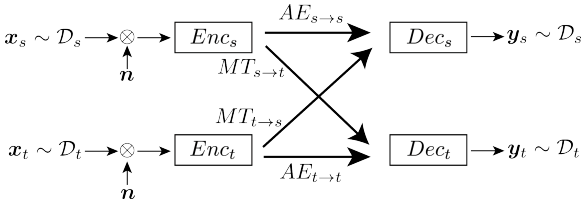


図 1: 教師なしニューラル機械翻訳モデル構造の概要 ($\boldsymbol{x}$ は入力系列, $\boldsymbol{y}$ は出力系列, $\mathcal{D}_s$, $\mathcal{D}_t$ はそれぞれソース, ターゲット言語の系列の分布を示す.)

2.1 多言語単語埋め込み表現の作成

前処理としてソース・ターゲット言語においてそれぞれ単語埋め込みベクトルを作成し, 各言語の埋め込み表現を Artetxe らの手法 [1] により同一空間上に写像し, ソース・ターゲット言語共通の多言語単語埋め込みベクトルを得る.

2.2 モデル構造

教師なし機械翻訳のモデル概要を図 1 に示す. ソース・ターゲット言語それぞれについてエンコーダ Enc_s, Enc_t とデコーダ Dec_s, Dec_t を用意し, それぞれのエンコーダをデコーダを組み合わせることにより, 計 4 種類のアテンション付き系列変換モデルを構成する. Enc_s と Dec_t, Enc_t と Dec_s を組み合わせたモデルはそれぞれ翻訳モデル (以下, それぞれ $MT_{s\to t}$, $MT_{t\to s}$) として振る舞い, Enc_s と Dec_s, Enc_t と Dec_t を組み合わせたモデルは入力を復元するモデル, すなわちオートエンコーダ (以下, $AE_{s\to s}$, $AE_{t\to t}$) として振る舞うことを期待する.

2.3 学習アルゴリズム

オートエンコーダ $AE_{s\to s}$, $AE_{t\to t}$ はノイズ $\boldsymbol{n}$ を付加された文章を入力とし, ノイズが付加される前の文

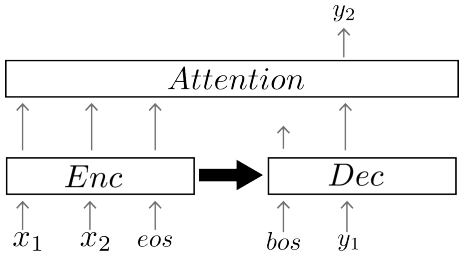


図 2: 文の潜在表現

章を復元するように学習される．翻訳モデル $MT_{s\to t}$ は翻訳モデル $MT_{s\to t}$ により生成された, 疑似対訳データを用いて通常のニューラル機械翻訳モデルと同様の手法により学習される．同様に, $MT_{s\to t}$ は $MT_{s\to t}$ により生成された疑似対訳データを用いて学習される．

オートエンコーダ $AE_{s\to s}$, $AE_{t\to t}$ の学習と翻訳モデル $MT_{s\to t}$, $MT_{t\to s}$ の学習を交互に行うことにより, 教師なしニューラル機械翻訳の学習が行われる．

3 教師なし英日ニューラル機械翻訳

本研究では Artetxe らの教師なしニューラル機械翻訳 [2] により英語と日本語の翻訳システムの検討を行う．Artetxe らの手法ではソース・ターゲット言語において, 多言語単語埋め込み表現をもとに言語間の対訳関係を獲得することが期待される．日英においては語順等の文法的側面における差異が大きいことが翻訳性能にどの程度の影響を与えるか検証する．

4 PCA による文の潜在表現の分析

本研究では教師なしニューラル機械翻訳の文の潜在表現に注目し, これを調べることにより言語をまたいだ文の潜在表現の共通化がどれほど達成されているのかを分析する．ここで文の潜在表現とは, 図 2 中の太い矢印で示される, エンコーダの出力のうちデコーダに直接入力されるベクトルを指す．文の潜在表現に対し, 主成分分析 (PCA) を用いて 2 次元に圧縮したのち, 言語ごとに可視化することにより, それぞれの言語の潜在表現の分布の分析を行う．

教師なしニューラル機械翻訳では, ソース・ターゲット言語についてエンコーダの出力空間が共通になることが期待されている．すなわち, 単一言語内のみならずソース・ターゲット言語間においても, 意味が近い文章の潜在表現は近つき, 対訳関係にある文章同士の潜在表現は十分近くなることが期待される．

表 1: 使用したコーパスの文章数

言語対		train	dev	test
en ↔ fr	en	21,480,041	5000	3003
	fr	12,338,590		
en ↔ ja	en	12,870,853	5000	5000
	ja	9,042,720		

5 実験

5.1 コーパス

本実験では, 英仏, 英日のモノリンガルコーパスとして linguatools の提供する Wikipedia Comparable Corpora[1] より, HTML タグと table 要素を除外して作成した, それぞれの言語対におけるコンパラブルコーパスを用いた．なお、文章中の単語数が 50 を超えるものは除外した．

なお, 本研究は教師なしニューラル機械翻訳の分析をするための開発用データ, 翻訳システムの性能の評価のためのテストデータとして対訳コーパスを用いた．英仏の開発用, テスト用の対訳コーパスとして WMT14[2] より newstest2014 におけるの英仏対訳コーパス, Common Crawl corpus をそれぞれ用いた．英日の対訳コーパスとして Wikipedia 日英京都関連文書対訳コーパス[3] より開発用, テスト用にそれぞれランダムに 5000 文を用いた．

英語, フランス語の単語分割には mosesdecoder[4] を用いた．日本語の単語分割には Kytea[5] を用い, ユニコード正規化の後, mojimoji[6] を用いて英単語と数値はすべて半角に, カタカナはすべて全角に置換した．

最終的に使用したコーパスの文章数はそれぞれ表 1 に示す．

[1] https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/
[2] http://www.statmt.org/wmt14/translation-task.html
[3] https://alaginrc.nict.go.jp/WikiCorpus/
[4] https://github.com/moses-smt/mosesdecoder
[5] http://www.phontron.com/kytea/index-ja.html
[6] https://github.com/studio-ousia/mojimoji

表 2: 各言語対における BLEU スコア

言語対	$s \to t$	$t \to s$
en ↔ fr	11.1	13.50
en ↔ ja	3.86	2.95

表 3: 英 → 日翻訳における翻訳例 (上段は比較的正しい翻訳例, 下段は誤った翻訳例)

入力文	however, it is said that nobunaga declined this offer.
出力文	しかし, 信長はこの申し出を断っている.
参照文	しかし, 信長はこれを辞退したとされる.
入力文	and, dai shogun was placed in every three troops.
出力文	駿や幕府に唯一の一軍兵として入った.
参照文	また, 三軍ごとに大将軍一人を置く.

5.2 学習アルゴリズム

本実験では教師なしニューラル機械翻訳の学習アルゴリズムとして Artetxe らの手法 [2] と同様の設定を採用する.

ソース言語, ターゲット言語それぞれについて fasttext[7]を用いて 300 次元の単語埋め込み表現を獲得した後, Artetxe らの手法 [1] により, 同様に 300 次元の多言語単語埋め込みベクトルを作成した. 翻訳モデル, オートエンコーダの構造は Luong ら [4] のモデルに従う. 各言語のエンコーダには 2 層の双方向 LSTM を用い, すべてのパラメータを言語間で共有した. 各言語のデコーダには 2 層の LSTM を用い, これらのパラメータは言語ごとに共有をしなかった. なお, すべての隠れ層の次元は 600, 語彙数はソース言語, ターゲット言語ともに 50,000 とした. $MT_{s\to t}$, $MT_{t\to s}$, $AE_{s\to s}$, $AE_{t\to t}$ について 1 度ずつ学習することを 1 イテレーションとし, バッチサイズを 50 として 300,000 イテレーション学習を行った時点で, 損失の値に大きな変化が見られなくなったため学習を終了した.

5.3 結果・考察

本実験では翻訳性能の定量的な評価手法として BLEU[5] を用い, 実際のスコアの計算には mosedecoder の実装を用いた. また, 定性的な評価として翻訳例を確認した.

英仏, 英日それぞれの翻訳システムの結果を表 2 に示す. 表 2 より分かるように, 英仏と比較して英日における翻訳性能は著しく低かった. 英 → 日翻訳において, 比較的正しい翻訳例と誤ったの翻訳例を表 3 に示す.

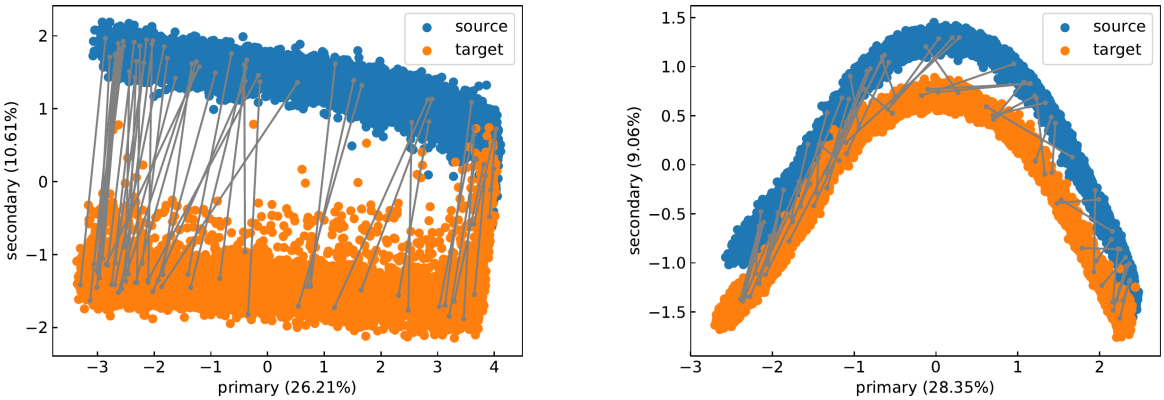
開発用対訳コーパス中の文の潜在表現について、イテレーションごとのの分布と対訳関係を英仏、英日ついてそれぞれ図 3, 4 に示す. 各分布図において, 横軸, 縦軸がそれぞれ第一成分, 第二成分を示しており, 軸ラベルの括弧中の値はその軸の成分が占める全体の分散への寄与率を示す. 図 3, 4 中の灰色の線は潜在表現の対訳関係を示す.

図 3 より読み取れるように, 50,000 イテレーション時においては, 英仏翻訳モデルのソース・ターゲット言語の文の潜在表現の分布の分離が見られた. しかし, 300,000 イテレーション時において, 50,000 イテレーション時と比較して言語間のが近づく様子が観測された. これは, 言語の違いによる分散への寄与率が小さくなっていることからわかる. 一方で図 4 よりわかるように, 英日のモデルにおいては英仏に見られた文の潜在表現の分布が近づく様子は見られなかった. また, 英仏翻訳と英日翻訳の違いとして, 英仏翻訳においては PCA の結果, 得られた第一成分が言語に依存していないのに対し, 英日翻訳においては明らかに第一成分が言語に依存していることが挙げられる. 以上の事実と, 英仏, 英日の翻訳性能の結果を考慮し, 文の潜在表現の分布が近づくことと、翻訳性能には関連性があると推測できる. また, 高い性能を得るための一つの指標として, 文の潜在表現について注目する価値があるのではないかと思われる.

6 おわりに

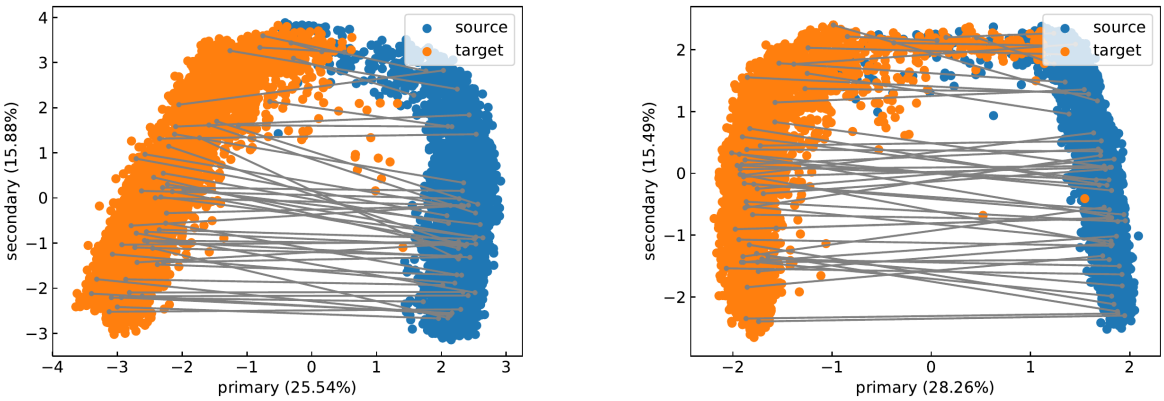
本研究では Artetxe らの教師なしニューラル機械翻訳を用いて英日翻訳システムの検討をしたが, 同様のアルゴリズムを用いた英仏翻訳システムと比較すると, 十分な性能は得られなかった. また, 教師なしニューラル機械翻訳により得られた英仏, 英日それぞれの翻訳システムについて, エンコーダからデコーダに渡される文の潜在表現について分析し, 比較的翻訳性能の高い英仏翻訳において, ソース言語とターゲット言語の分布がイテレーションごとに近づく様子を観測することができた.

[7] https://github.com/facebookresearch/fastText



(a) 50,000 イテレーション時　　(b) 300,000 イテレーション時

図 3: 英仏翻訳における文潜在表現の分布（ソース言語: 英語, ターゲット言語: フランス語）



(a) 50,000 イテレーション時　　(b) 300,000 イテレーション時

図 4: 英日翻訳における文潜在表現の分布（ソース言語: 英語, ターゲット言語: 日本語）

今後の課題として, 教師なしニューラル機械翻訳において大幅な性能の向上が報告された, 複数言語間におけるサブワードの共有が潜在空間の共有にどのように作用するかの分析などが挙げられる.

参考文献

[1] Artetxe et al. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, 2018.

[2] Artetxe et al. Unsupervised neural machine translation. In *ICLR*, 2018.

[3] Lample et al. Unsupervised machine translation using monolingual corpora only. In *ICLR*, 2018.

[4] Luong et al. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[5] Papineni et al. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[6] Rico Sennrich et al. Improving neural machine translation models with monolingual data. In *ACL*, 2016.

[7] Sutskever et al. Sequence to sequence learning with neural networks. In *NIPS*. Curran Associates, Inc., 2014.

[8] Vincent et al. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.