日英ニューラル機械翻訳における未知語置き換えの手法とコーパス間の比較

伊部 早紀　　松田 源立　　山口 和紀
東京大学
{ibe, matsuda, yamaguch}@graco.c.u-tokyo.ac.jp

1 はじめに

ニューラル機械翻訳 [2, 5, 16] は，ソース言語の単語列を数値ベクトルによる分散表現で表し，それをニューラルネットワークを用いて変換して求めた数値ベクトルからターゲット言語の単語列を求めることで翻訳を行う手法である．リカレントニューラルネットワークおよびLSTM(Long Short Term Memory) の利用により，長い区間での単語のつながりが考慮されている．そのため，ニューラル機械翻訳を用いると従来の統計的機械翻訳と比べて流暢な文を生成できるが，出力結果に未知語 (UNK) が含まれるという問題が指摘されている [7, 10]．この問題に対処する方法としては，[10, 14] のようにコーパスに前処理を行う手法，[1, 17, 9] のように学習するモデルを変更する手法，[15] のように統計的機械翻訳と組み合わせる手法などがある．

著者らは [18] において，単語間の対応関係に関するヒューリスティックを利用してアテンションから単語アライメント表を推定し，そのアライメントの対応関係から未知語に対応する原言語の単語を探し，コーパスから作成した辞書や外部辞書を利用して未知語を推定する手法を提案した．この手法は未知語を減らし BLEU も向上させられるという点で他の手法より優れているが，口語表現等のコーパスから生成された辞書では上手くいかない場合があった．本論文では，WordNet 等の外部辞書を利用し，提案手法を性質の異なる 3 つのコーパスに適用した結果を記載する．

2 提案手法

提案手法の詳細は [18] を参照されたい．

2.1 単語アライメント表の作成

出力文中の各未知語が原言語におけるどの単語に対応付けられているかを知るために，ニューラルネットが出力するアテンション行列から単語アライメント表を作成する．単語アライメントの作成アルゴリズムとしては intersection(inter)，gdfa-f の 2 種類を用いた．

intersection(inter)． 原言語，目的言語どちらから見てもアテンションが最も強い単語を対応付けする．

gdfa-f． intersection において作成した行列をもとに，対応付けされた単語の隣の単語で，原言語または目的言語のどちらか一方から見てアテンションが最も強い単語があれば対応付けする．さらに，原言語，目的言語それぞれの単語から見て，もし対応付けされている単語がなければ，アテンションが最も強い単語に対応付けする[1]．

2.2 未知語の置き換え

作成したアライメント行列をもとに，出力文における未知語 e_i が入力文のどの単語あるいは単語列 $\boldsymbol{f_i}$ に対応しているかを決める．次に，$\boldsymbol{f_i}$ の訳を決め，未知語 e_i と置き換える．本研究では $\boldsymbol{f_i}$ の訳の決め方として IBM，ChangePhrase(CP)，Dict，IBM+Dict の 4 種類の手法を提案する．

IBM． 対訳コーパスに IBM モデル 4[4] を適用したときの単語翻訳確率 $t(e|f)$ を参照し，$\boldsymbol{f_i} = (f_{i_1}, ..., f_{i_n})$ の要素それぞれにおいて最も確率の高い $e_{best} = \arg\max_e t(e|f_i)$ を訳として選ぶ．

ChangePhrase(CP)． $\boldsymbol{f_i}$ の訳をフレーズテーブル [8] から参照し，コーパスから計算したフレーズ翻訳

[1] すべての単語のアテンションが 0 であれば，どの単語にも対応付けしない．また，アテンションが最も強い単語が複数ある場合，そのすべてと対応付けする．

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 各コーパスのサイズと単語数．train は学習に，dev はパラメータチューニングに，test はテストに用いた．

		train		dev	test
		文	語彙	文	文
ASPEC	日	908.1K	162.3K	1.8K	1.8K
	英		326.8K		
NTCIR-10	日	3.2M	146.5K	2.0K	0.9K
	英		265.0K		
田中コーパス	日	145.9K	31.6K	2.0K	2.0K
	英		21.9K		

確率 $P(\boldsymbol{e}|\boldsymbol{f}) = \frac{c(\boldsymbol{e},\boldsymbol{f})}{c(\boldsymbol{f})}$ が最も高い訳を未知語と置き換える[2]．フレーズで置き換えるため，未知語を複数単語と置き換えることもある．

Dict． 外部の辞書から f_j の各単語の訳を参照し，それを e_i と置き換える．

IBM+Dict． IBM の手法を適用後，Dict を適用する．本論文では IBM と Dict 双方の長所を活かすため，IBM 適用後に Dict を利用する手法も導入した．

3 実験

3.1 実験データ・方法

対訳コーパスとしては ASPEC[11]，NTCIR-10 の patentMT[6]，田中コーパス[3]を用いた[4]．コーパスのサイズと単語数を表 1 に示す．モデル学習およびデコーダには nematus[5]を用い，日英および英日の翻訳を行った．英語の構文解析には Stanford Parser[6]を，日本語のトークン化には KyTea[7]を用いた．IBM モデル 4 の実装としては GIZA++[8]を用いた．フレーズテーブルの抽出は mosesdecoder[9]を用いた．未知語の置き換えの手法の Dict としては，[18] では外部辞書として EDict[10] のみを利用したが，今回は WordNet[11]も導入し語彙を大幅に拡張した．動詞は KNP[12]を用いて原形に直して

[2] $c(\boldsymbol{f})$ はコーパス中のフレーズ $\boldsymbol{f}$ の出現回数，$c(\boldsymbol{e},\boldsymbol{f})$ はフレーズ $\boldsymbol{e}$ と $\boldsymbol{f}$ の同時出現回数．

[3] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

[4] 全データ中からランダムに抽出した 2000 文をチューニングに，1983 文をテストに用いた．

[5] https://github.com/EdinburghNLP/nematus

[6] https://stanfordnlp.github.io/CoreNLP/

[7] http://www.phontron.com/kytea/index-ja.html

[8] https://github.com/moses-smt/giza-pp

[9] https://github.com/moses-smt/mosesdecoder

[10] http://www.edrdg.org/jmdict/edict.html

[11] http://compling.hss.ntu.edu.sg/wnja/

[12] http://nlp.ist.i.kyoto-u.ac.jp/?KNP

から検索を行った．コーパスの前処理として [12] を参考に，単語数 51 以上の文は学習データから取り除いた．学習においては語彙数を ASPEC および NTCIR-10 では 4 万語に，田中コーパスでは 1 万語に制限した．

nematus のデフォルトのパラメータでの実験を baseline とした．また，低頻度語を分割し語彙数を減らす手法 (BPE)[14]，未知語に位置情報を付け加えて学習を行う手法 (PosUNK)[10] も実験し，結果を比較した．

翻訳精度の評価指標には BLEU[13]，METEOR[3] の 2 つを用いた．METEOR は日本語の評価には対応していないため日英翻訳の評価のみに用いた．

3.2 実験結果と考察

nematus のデフォルトの設定で学習したモデルをベースラインとし，出力結果に提案手法による修正を加えたものと翻訳精度を比較した．日英での実験結果を表 2 に，英日での実験結果を表 3 に示す．

表 2 より，BLEU は ASPEC では+inter+IBM+Dict が最大で+0.60，NTCIR-10 では+inter+IBM+Dict で +0.06，田中コーパスでは gdfa-f+IBM+Dict で+0.34 の上昇を確認できた．METEOR ではどのコーパスでも gdfa-f+IBM+Dict の場合が最大で，ASPEC では+0.72，NTCIR-10 では+0.15，田中コーパスでは+0.76 の上昇を確認できた．BLEU は完全一致をもとに測る指標であるため，意味は同じだが違う表現に置き換えられている場合でも値は上昇しない．一方 METEOR はそのような表現を考慮している指標である．gdfa-f+IBM+Dict は METEOR の値の方が大きく上昇していることから，未知語を意味的に正しい単語で置き換えられていることが示唆される．

ASPEC，NTCIR-10 はともに学術的な文からなるコーパスのため，英語と日本語の単語の対応が取りやすく，gdfa-f を適用すればアテンションから単語アライメントをほぼ正確に作成できた．学習コーパス中，もしくは WordNet のエントリに存在する単語であれば IBM+Dict を適用し適切な単語に置き換えることができるが，どちらにも存在しない単語は置き換えられない．この 2 つのコーパスでは共に化合物の名称などの専門用語が多く，単語の対応は取れているが IBM+Dict を適用できない例が多数見られた．専門用語を扱う辞書を取り入れることでさらなる BLEU，METEOR の向上が期待できる．

一方田中コーパスでは口語的表現が使われているため，ASPEC や NTCIR-10 に比べ単語の対応が 1 対 1 に取れないことが多く，アテンションが正確にならな

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

表 2: 日英での翻訳結果の精度評価の比較．太字は最も高い値 (UNK の場合は最も低い値).

	ASPEC			NTCIR-10			田中コーパス		
	BLEU	METEOR	UNK [%]	BLEU	METEOR	UNK [%]	BLEU	METEOR	UNK [%]
baseline	24.97	31.50	2.72	35.31	33.88	1.59	27.33	28.29	3.02
BPE	21.47	28.34	**0.00**	32.95	31.12	**0.00**	26.24	27.97	**0.00**
PosUNK	24.25	31.63	0.07	34.88	33.52	0.09	26.48	28.18	0.23
+inter+IBM	25.56	32.06	0.42	**35.37**	33.99	0.21	27.63	28.87	0.97
+inter+CP	25.34	31.97	1.10	35.33	33.92	1.12	27.49	28.67	1.72
+inter+Dict	25.37	31.81	0.42	35.33	33.91	1.16	27.40	28.66	0.97
+inter+IBM+Dict	**25.57**	32.11	0.42	**35.37**	33.99	0.21	27.63	28.92	0.97
+gdfa-f+IBM	25.51	32.16	**0.00**	35.31	**34.03**	**0.00**	**27.67**	29.00	0.05
+gdfa-f+CP	25.37	32.06	0.89	35.33	33.90	1.01	27.52	28.76	1.19
+gdfa-f+Dict	25.26	31.82	0.00	35.32	33.91	1.17	27.40	28.74	0.05
+gdfa-f+IBM+Dict	25.53	**32.22**	**0.00**	35.31	**34.03**	**0.00**	**27.67**	**29.05**	0.05

いことが多い．今回提案したアテンションから単語アライメント表を作成する手法ではアテンションが正しいことを前提にしているため，アテンションが間違っていると正確な単語アライメント表を作れず，未知語を適切な単語で置き換えることができない．今回の実験では ASPEC，NTCIR-10 では gdfa-f+IBM+Dict を用いれば未知語を完全になくすことができたが，田中コーパスではすべてをなくすことはできていない．これは，アテンションが間違っており，gdfa-f を適用する際どの単語にも対応付けされていない単語があったためである．ニューラルネットワークのモデルの改善などでアテンションをより正しく出力するようにできればこの問題に対処可能と期待される．

また，表 3 より，英日翻訳でも BLEU が向上することが分かった．

4 おわりに

本論文では，アテンションをもとに単語アライメント表を作成することで，出力結果における未知語が入力文のどの単語に対応しているかを判別し，統計的機械翻訳でのモデルを用いて未知語を適切な単語に置き換える手法を，性質の異なる 3 種類のコーパスに適用した．その結果，提案手法の gdfa-f+IBM+Dict を用いて単語アライメント表を作ることで未知語をなくすことができた．また，学習コーパスから作成した辞書だけでなく外部辞書を利用し，用語が統制されたコーパス，口語主体の表現が多様なコーパスの両方においてBLEU 値などを向上させることができた．

参考文献

[1] P. Arthur, G. Neubig, and S. Nakamura. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on EMNLP*, pp. 1557–1567, 2016.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, Vol. 29, pp. 65–72, 2005.

[4] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[5] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. pp. 1724–1734. ACL, 2014.

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

表 3: 英日での翻訳結果の精度評価の比較．太字は最も高い値 (UNK の場合は最も低い値).

	ASPEC		NTCIR-10		田中コーパス	
	BLEU	UNK [%]	BLEU	UNK [%]	BLEU	UNK [%]
baseline	34.58	2.07	36.73	0.37	21.01	4.01
BPE	29.54	**0.00**	33.26	**0.00**	20.22	**0.00**
PosUNK	34.47	0.17	36.39	0.03	19.82	0.82
+inter+IBM	34.63	0.50	**36.83**	0.05	21.48	2.54
+inter+CP	34.64	0.57	36.76	0.20	21.40	2.88
+inter+Dict	34.59	0.81	36.74	0.28	21.21	2.89
+inter+IBM+Dict	34.64	0.57	**36.83**	0.05	21.48	2.42
+gdfa-f+IBM	34.66	**0.00**	**36.83**	**0.00**	**21.52**	**0.00**
+gdfa-f+CP	**34.68**	0.16	36.77	0.21	21.48	0.16
+gdfa-f+Dict	34.59	1.10	36.73	0.26	21.43	2.32
+gdfa-f+IBM+Dict	34.66	**0.00**	**36.83**	**0.00**	**21.52**	**0.00**

[6] I. Goto, K.-P. Chow, B. Lu, E. Sumita, and B. K. Tsou. Overview of the patent machine translation task at the ntcir-10 workshop. In *NTCIR*, 2013.

[7] S. Jean, K. Cho, R. Memisevic, and Y. Bengio. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on NLP*, Vol. 1, pp. 1–10, 2015.

[8] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL HLT-Volume 1*, pp. 48–54. ACL, 2003.

[9] M.-T. Luong and C. D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of ACL*, Vol. 1, pp. 1054–1063, 2016.

[10] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of ACL and the 7th International Joint Conference on NLP*, Vol. 1, pp. 11–19, 2015.

[11] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. ASPEC: Asian scientific paper excerpt corpus. In *Proc. of LREC 2016*, pp. 2204–2208.

[12] G. Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *ACL System Demonstrations*, pp. 91–96, 2013.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

[14] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of ACL*, Vol. 1, pp. 1715–1725, 2016.

[15] I. Skadina and R. Rozis. Towards hybrid neural machine translation for english-latvian. In *Proceedings of the 7th International Conference Baltic HLT 2016*, Vol. 289, p. 84.

[16] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[17] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of ACL*, Vol. 1, pp. 76–85, 2016.

[18] 伊部早紀, 松田源立, 山口和紀. 日英ニューラル機械翻訳におけるアテンションを用いた未知語置き換えの手法. 自然言語処理, Vol. 25, No. 5, pp. 511–526, 2018.

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.