教師なし学習による語彙平易化における出力制御

勝田 哲弘　山本 和英
長岡技術科学大学
{katsuta, yamamoto}@jnlp.org

1　はじめに

近年では、難解文と平易文の集合をそれぞれ別の言語とみなし、機械翻訳タスクの枠組みでテキスト平易化を行う研究が盛んに行われている。深層学習の発展による機械翻訳の飛躍的な精度向上も伴って、テキスト平易化の操作をニューラル機械翻訳モデルに学習させる手法が大きく性能を伸ばしている。しかし、テキスト平易化の分野において難解文と平易文がそれぞれ対応付けられているようなコーパスは機械翻訳のタスクと比べると圧倒的に少なく、また、日本語においてそのような言語資源はほとんど存在しない。そのため、日本語テキスト平易化の研究を行うことが困難である。

そこで、容易に大規模なコーパスを用いて学習を行うことのできる教師なし学習による手法を用いる。単言語のコーパスからテキスト平易化のためにそれぞれ独立の難解文コーパスと平易文コーパスを構築し、教師なし機械翻訳の枠組みでテキスト平易化を行う。しかし、教師なし学習では一般的に単語埋め込みによって単語の対応付けを学習するが、必ずしも単語埋め込みから同義語を抽出できるわけではなく、ノイズを多く含んでいる。そのため、WordNet を用いたクリーニングを行い、その効果を調査する。

2　関連研究

2010 年以降、統計的機械翻訳 (SMT) を用いたテキスト平易化の研究は盛んに行われている。英語では平易化コーパスとして、English Wikipedia と Simple English Wikipedia から対応する文対を抽出した単言語パラレルコーパスがよく用いられている。[4, 5, 9]

テキスト平易化の評価においても翻訳タスクと同様の評価尺度がよく用いられている。機械翻訳のための評価尺度である BLEU は参照文と出力文を比較して出力文の意味や文法の正しさを評価する。BLEU の計算は各文間の単語列の一致度を計算している。そのため、平易化タスクなどの単言語間の翻訳では何も書き換えを行わなくてもある程度 BLEU スコアが高く出てしまう問題がある。そのため、より積極的に書き換えを行っていることを評価する SARI[10] が提案された。SARI は入力文と出力文と参照文の 3 つを用いる自動評価尺度であり、平易さの観点では BLEU よりも人手評価との相関が高いことが知られている。

また、教師なし機械翻訳の分野では Artetxe et al. [1] が、事前学習済みのクロスリンガル n-gram 埋め込みを SMT のフレーズテーブルに適用させて訓練し、従来の教師なし NMT システムを大幅に改善させた。そして Phrase-based SMT の構造は、この教師なし翻訳により適していると主張している。

WordNet による SMT の改善としては大山鉄郎 et al. [11] が語義曖昧性の解消に WordNet を素性として利用している。その際に語義間の経路長を求める計算方法が最も精度に貢献していることを示している。

3　方法

3.1　教師なし翻訳モデル

Artetxe et al. [1] が提案した Unsupervised Statistical Machine Translation(USMT) の学習の流れは図 1 に示すとおりである。

まず、各言語で phrase2vec によって n-gram 単語埋め込みを構築し、各単語埋め込みを教師あり、または教師なしで構築されたシード辞書によって共有の空間へのマッピングを行う。今回は各言語間の同一単語をシード辞書とするマッピングを行っている。[2] そして、統合された単語埋め込みをもとにフレーズテーブルを構築し、それと言語モデルを用いて SMT を初期化する。構築され

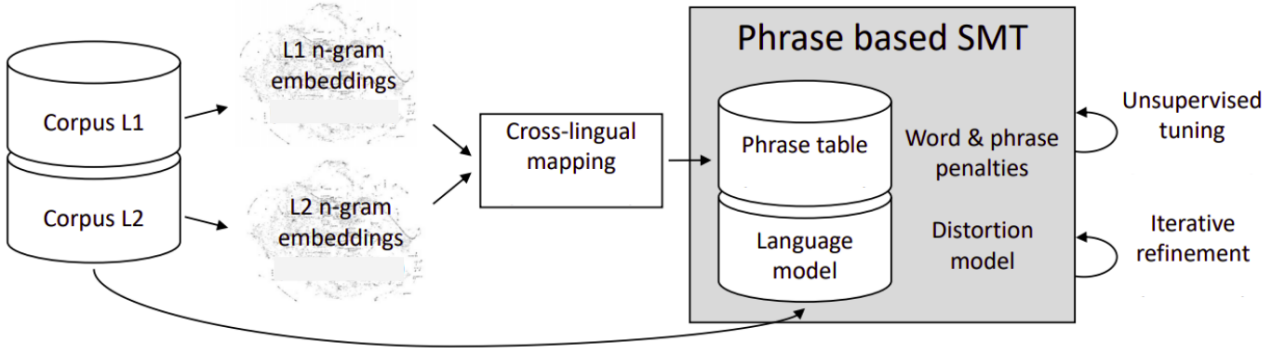

図 1: USMT のモデル構造 ([1] より抜粋)

た SMT を最終的に逆翻訳を行って相互に繰り返し学習をすることで重みのチューニングを行う。

実験で使用したパラメータは公開されているモデルのデフォルト値を用いている[1]。

平易化のタスクでは平易文と難解文は同じ言語であるため、各コーパスを結合して単語埋め込みを学習させることができるが、今回はそれぞれのコーパスで独立に学習させた後に共有の空間へのマッピングを行っている。

3.1.1 フレーズテーブルの構築

USMT のフレーズテーブルは単語埋め込みを用いて以下の 2 つの値を各言語間の両方向で計算される。

- フレーズ翻訳確率
- 語彙の重み

我々はこのモデルのフレーズテーブルの構築段階で WordNet を用いたフレーズテーブルのクリーニングを行う。

3.1.2 フレーズ翻訳確率

対応するフレーズ翻訳確率を推定するために、それぞれの単語埋め込みのコサイン類似度にソフトマックス関数を適用する。より具体的には、原言語句 $\bar{e}$ に対して翻訳候補 $\bar{f}$ が与えられた場合、フレーズ翻訳確率は次のように計算される。

$$\phi(\bar{f}|\bar{e}) = \frac{\exp(\cos(\bar{e},\bar{f})/\tau)}{\sum_{\bar{f}'}\exp(\cos(\bar{e},\bar{f}')/\tau)} \quad (1)$$

[1]https://github.com/artetxem/monoses

τ は予測の信頼性を制御する温度パラメータである。τ を調整するために、最近隣検索で言語間の埋め込み自体に辞書を作成し、両方向に辞書を誘導し最尤推定を適用する。

$$\min_{\tau} \sum_{\bar{f}} \log \phi(\bar{f}|NN_{\bar{e}}(\bar{f})) + \sum_{\bar{e}} \log \phi(\bar{e}|NN_{\bar{f}}(\bar{e})) \quad (2)$$

3.1.3 語彙の重み

語彙の重みを計算するために、対象言語の句に含まれる各単語を最も可能性の高い原言語側の単語に合わせて、それぞれの翻訳確率の積を取る。

$$\mathrm{lex}(\bar{f}|\bar{e}) = \prod_{i} \max(\epsilon, \max_{j} \phi(\bar{f}_i|\bar{e}_j)) \quad (3)$$

3.2 WordNet を用いたクリーニング

WordNet は synset と呼ばれる同義語の集合が階層構造となって表現されている概念辞書である。日本語 WordNet[3] を用いることで語と語の意味的な類似度を計算できる。現在日本語 WordNet には、57,238 概念 (synset 数)、93,834 語が収録されている。

$$\mathrm{rel}(\bar{f},\bar{e}) = \frac{1}{\min_{i,j}(length(\bar{f}_i,\bar{e}_j)) + 1} \quad (4)$$

式 (4) は各フレーズ間の最小経路長を考慮した類似度の尺度である。各フレーズに含まれる単語間で最も経路の短い synset の対をそのフレーズ間の類似度として計算する。

今回我々はこの類似度を用いてフレーズテーブルの足切りを行う。フレーズ翻訳確率、語彙の重みのスコアを計算する際に、WordNetによる類似度が閾値σを下回る場合に、そのスコアを$1e-9$で上書きする。また、今回はWordNetで類似度を計算できないフレーズ対は何も処理を行っていない。

$$score'(\bar{f}|\bar{e}) = \begin{cases} \text{1e-9}, & \text{rel}(\bar{f},\bar{e}) < \sigma \\ score(\bar{f}|\bar{e}), & otherwise \end{cases} \quad (5)$$

3.3 平易化の学習

学習データには日本語ウェブコーパス 2010 (NWC 2010)[2]から自動的に難解文と平易文に分割した擬似コーパスを使用する。SNOW S13:やさしい日本語チェッカー[3]を用いてNWC2010の文に難解語が含まれているかをチェックし、難解文と平易文をそれぞれ50億文ほど抽出し学習用のコーパスとして使用する。抽出する際に、コーパスのクリーニングのために記号の削除、3回以上連続する単語の削除を行っている。

テストデータとしてSNOW T15[8] とSNOW T23[6] を合わせた平易化コーパス[4] から747文を抽出してBLEU、SARIを用いて評価する。

また、形態素解析器としてUniDic[5]を用いたMeCab[7] を使用している。

NWC2010から平易化をUSMTに学習させる際、閾値を変化させ、そのときの精度への影響を調査する。

4 結果及び考察

表 1: 閾値による平易化の精度変化

Threshold	BLEU	SARI
0.00	49.91	53.53
0.10	49.68	51.94
0.25	50.13	56.57
0.50	**50.58**	**58.17**
入力文	45.78	21.22

表1に各閾値による結果を示す。また、比較として入力文をそのまま出力した場合のBLEUとSARIを計算し、併記する。入力文をそのまま出力した場合のスコアはBLEU=45.78、SARI=21.22であるため、閾値σを0.00としている結果と比較するとUSMTがBLEUを4ポイント、SARIを32ポイント近く改善させていることになる。そして、WordNetのによるフレーズテーブルのクリーニングを行うことでより精度が改善していることがわかる。

4.1 WordNetの貢献について

WordNetによるクリーニングの閾値を大きくするほど精度が高くなっているため、クリーニングによってより意味の近い出力をモデルにさせることができることがわかる。しかし、WordNetによる精度の改善が小さい原因として、意味が近い出力になっているが参照文と異なる場合があることや、WordNetを適用できる語彙のカバー率がそれほど大きくないことが考えられる。表2にいくつかモデルの出力例を載せており、閾値が高いほど意味の近い出力になっていることがわかる。しかし、参照文と一致しているわけではない。そのため、WordNetによるBLEUの改善が小さくなっている。また、今回使用した50億文ずつの学習コーパスに含まれる語彙数は出現数が10以上の語彙に限ると99277である。それに対して、実際にWordNetで類似度を計算することができる語彙はその内27708であった。そのため、単語単位で見たときにWordNetは全体の約28%の語彙間の類似度しか計算できないためクリーニングを適用できる割合が少ないという問題がある。

5 おわりに

本研究では、言語のコーパスからテキスト平易化のためにそれぞれ独立の難解文コーパスと平易文コーパスを擬似的に作成し、教師なし機械翻訳の枠組みでテキスト平易化を行った。また、翻訳モデルとしてUSMTを使用し、そのフレーズテーブルに対してWordNetを用いたクリーニングを行い、その効果を調査した。結果としてWordNetを用いたクリーニングを行うことでUSMTの精度を向上させることができたが、貢献はそれほど大きくはない。

今後としては、対義語辞書を加えたモデルの改善やよりより良いWordNetの適用方法を模索していきたい。また、WordNetで全ての語彙を考慮できるわけではないのでより多くの語彙を考慮

[2]https://www.s-yata.jp/corpus/nwc2010/
[3]http://www.jnlp.org/SNOW/S13
[4]http://www.jnlp.org/SNOW
[5]https://unidic.ninjal.ac.jp/

表 2: WordNet の閾値ごとの出力例

原文	私 は 風邪 を ひか ない よう に 外出 し なかっ た
0.00	私 は 風邪 を 引か ない よう に **デート** し なかっ た
0.10	私 は 風邪 を 引か ない よう に **留守** し なかっ た
0.25	私 は 風邪 を 引か ない よう に **お 出かけ** し なかっ た
0.50	私 は 風邪 を 引か ない よう に **お 出かけ** し なかっ た
参照文	私 は 風邪 を 引く こと が ない よう に 外 に 出 なかっ た
原文	なん と 思いやり の ある あなた でしょう
0.00	なん と **誠実** の ある あなた でしょう
0.10	なん と **誠実** の ある あなた でしょう
0.25	なん と **責任 感** の ある あなた でしょう
0.50	なん と **優し さ** の ある あなた でしょう
参照文	あなた は なん と 心 が 優しい の でしょう

できるように retrofitting のような単語埋め込みを改善する手法に WordNet を用いるといったことに取り組んでいきたい。

謝辞

本研究は、平成 29-31 年学術研究助成基金助成金 挑戦的研究 (萌芽) 課題番号 17K18481 の助成を受けています。

参考文献

[1] M. Artetxe, G. Labaka, and E. Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[2] M. Artetxe, G. Labaka, and E. Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019, 2018.

[3] F. Bond, T. Baldwin, R. Fothergill, and K. Uchimoto. Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the 6th global WordNet conference (GWC 2012)*, pages 56–63. Citeseer, 2012.

[4] W. Coster and D. Kauchak. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669, 2011.

[5] T. Kajiwara and M. Komachi. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016.

[6] A. Katsuta and K. Yamamoto. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 461–465, 2018.

[7] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[8] T. Maruyama and K. Yamamoto. Simplified corpus with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1153–1160.

[9] S. Wubben, A. van den Bosch, and E. Krahmer. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012.

[10] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

[11] 大山鉄郎, 関洋平, et al. 統計的機械翻訳における wordnet を用いたフレーズ意味曖昧性解消手法の提案. **研究報告自然言語処理** *(NL)*, 2013(11):1–6, 2013.