DAMを用いた文レベルのアライメントによる テキスト平易化コーパス構築手法の提案

Construction of Corpus for Text Simplification by Sentential Alignment based on Decomposable Attention Model

> 永塚 光一^{*1} Koichi Nagatsuka

渥美 雅保^{*1} Masayasu Atsumi

*1 創価大学大学院 工学研究科 Graduate School of Engineering, Soka University

Text Simplification is a task to generate a sentence which is easier to understand than original. Text Simplification helps beginners such as children and foreigners learn languages. Recently, seq2seq models based on large scaled datasets have achieved state-of-the-art results in many areas including Machine Translation, Summarization, and Question Answering, to name a few. Although these model can be applied in text simplification as well, it requires a large number of parallel sentence pairs. Since available sentential corpora for text simplification are inadequate, building new corpus is so critical. In this paper, we suggest the application of neural textual entailment method to detection of simplified sentence pairs so that we are able to automatically construct text simplification dataset. In experiment, we evaluated the performance of identification of simplified sentences by using manually annotated dataset, and our proposed framework outperformed a baseline method.

1. はじめに

テキスト平易化とは、ある文が与えられた時に、読者にとってよ り分かりやすい文に表現を変換する処理のことを指す.テキスト 平易化は、新たに言語を学び始めた子供や、外国人、難読症患 者や、リテラシーの低い人々を支援するために重要な技術であ ると言える.ここで、ある文を対応する簡易な表現の文に変換す るタスクは、単一言語における翻訳処理の一種であるとみなすこ とができる.この観点から、機械翻訳で成功を収めているニュー ラル seq2seq モデルをテキスト平易化へ応用した手法の研究が 行われてきた[Nisioi 17].

こうしたモデルの学習には入力文と出力文が1対1に対応し ている大規模なパラレルコーパスが必要とされる.しかしながら, テキスト平易化にはそのような文間でアライメントが取れたコー パスが構築されていない現状がある. テキスト平易化のための パラレルコーパスが不足する理由として,2つの理由が挙げられ る.まず、一つ目の理由として、ある同一のトピックに対して、複 数のレベルで明確に平易化された文書の数が少ないということ が挙げられる.これは、一般的に平易化という処理が翻訳に比 べてニーズが低いために、 平易化された文書のペアが手に入り にくいということに起因する. また, パラレルコーパスの構築が難 しいもう一つの原因として、平易化された文が、元の文と1対1 の対応関係を常に持つとは限らないということが挙げられる. テ キスト平易化は、元の文章の重要な部分だけを残す処理である という観点から要約の一種であり、1文を短くするだけに留まらず、 元の文章から重要でない文の削除が起きたり、2 文を1 文に統 合したりすることがある.したがって、平易化された文書ペアを取 得することができた場合でも, 文の対応関係を見つけることは容 易ではない. これに対し、人間が手作業で対応関係にアノテー ションを行うことは、質の高いコーパスのために理想的ではある が、かなりのコスト及び時間を要するため、コーパス構築のボトル ネックとなっている.

こうした中,先行研究[Xu 15]では,元文と平易化文のペア(こ れを平易化文ペアと呼ぶ)を自動的にアライメントする手法が提 案されている.しかし,これまでの手法は,表層的な情報のみに 基づいており,平易化の処理におけるパラフレーズや構文構造 の変化を判定することが難しいという課題点が残されている.

そこで、本研究では、より深層的な意味特徴に着目した Decomposable Attention Model を平易化文ペアの検出に応用 することを提案する. Decomposable Attention Model はテキスト 含意認識の分野において使用されているモデルである. テキス ト含意認識は、二つ文が与えられた時に、それらの間で含意関 係が成り立つがどうかを判定するタスクである. ここで、平易化さ れた文は元の文と等価か、少なくとも含意関係にあると言える. すなわち、平易化文ペアの検出は、含意関係認識の問題に置 き換えることができる. 含意関係の推論処理において、従来の表 層的特徴ではなく、セマンティックな特徴を効果的に利用するこ とができれば、パラフレーズが起きた平易化文ペアの検出が可 能になると期待される.

2. 先行研究

2.1 テキスト平易化コーパス

これまで開発されてきたテキスト平易化のためのデータセット として, Simple Wikipedia Corpus[Zhu 10]と Newsela Corpus[Xu 15]がある.

(1) Simple Wikipedia Corpus

Simple Wikipedia Corpus は最初に開発されたテキスト平易化 のための大規模データセットである.データの構築には Wikipedia にある記事とその記事をより簡単な英語に置き換えた Simple English Wikipedia の記事が文書ペアとして使われている. Simple Wikipedia Corpus は、テキスト平易化の研究において貴 重なベンチマークデータセットであるものの、平易化のための再 編集をしているエディタがボランティアであることや、明確な平易 化の統一基準が不足していることに起因する平易化処理の質 の低さが問題点として指摘されている[Xu 15].

連絡先:永塚光一, 創価大学大学院工学研究科, 東京都八王 子市丹木町1丁目236, e18m5212@soka-u.jp

(2) Newsela Corpus

こうした問題意識に基づき、その後 Newsela Corpus が作成された. Newsela Corpus は、児童のための学習教材からデータを 収集しており、プロのエディタにより複数の学習者レベルに応じ てテキストが編集されていることが特徴である. Newsela Corpus は Simple Wikipedia Corpus に比べて、テキスト平易化の質が高 いコーパスであるものの、どちらのコーパスも依然として文レベ ルのアライメントがなされていないという課題点がある.

2.2 自動アライメント

文レベルのアライメントを自動的に検出する手法がいくつか 提案されている. Newsela Corpus の作成者[Xu 15]は、最もシン プルな文間の類似度計算手法として Jaccard similarity を使用 することを提案している. Jaccard similarity は以下の式で与えら れる.

$$Jaccard(X,Y) = \frac{count(X \cap Y)}{count(X \cup Y)}$$
(1)

また、平易化文ペア間の1対1関係のみではなく、1対Nや N対1の対応関係の検出に特化した Vicinity-driven sentence alignment が提案されている. Vicinity-driven sentence alignment は TF-IDF に基づいた手法であり、実際に Newsela Corpus の 自動アライメントに適用された研究が報告されている[Scarton 18]. しかしながら、パラフレーズや省略が頻繁に起こるテキスト 平易化では、単語頻度により意味的な類似度を計算することに は限界がある.

一方で、より深層の意味関係を推論するタスクにテキスト含意 認識がある.テキスト含意認識では、ある文のペアを含意、矛盾、 中立の3つの関係に分類する.

- Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.
- Bob is awake.
- It is sunny outside.

例えば、最初の文章では、「he cannot sleep」から bob が起き ている状態が推論されるので、二つ目の「bob is awake」という文 意は最初の文意に含まれ、含意関係があると言える.一方で、最 後の文の「it is sunny outside」という内容は最初の文の「the thunder and lightning」という描写に相反するので、矛盾関係に あると結論できる.本研究では、テキスト平易化文ペアの検出に テキスト含意認識モデルを適用する.

3. Decomposable Attention Model(DAM)

Decomposable Attention Model (DAM)[Parikh 17]は、テキスト 含意認識を学習するニューラルネットワークモデルである. DAM は文間の意味推論をある単語レベルのアテンションに分解して、 最終的な含意関係の予測を行う. このモデルではまず、アテン ション機構を用いて 2 つの入力文 $S_{\alpha}[w_{\alpha 1}, \cdots w_{\alpha l}, \cdots w_{\alpha l_{\alpha}}],$ $S_{\beta}[w_{\beta 1}, \cdots w_{\beta j}, \cdots w_{\beta l_{\beta}}]中の単語間の注意重み<math>e_{ij}$ を計算する. その後,注意重みで単語分散表現 α_{i} 及び β_{j} を獲得する.

$$\alpha_i = \sum_{i=1}^{l_a} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_a} \exp(e_{ki})} \bar{a}_i \tag{2}$$

$$\beta_j = \sum_{j=1}^{l_b} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_b} \exp(e_{kj})} \overline{b}_j \tag{3}$$

ここで, l_a , l_b は文 S_a , S_β の単語長である. また, \bar{a}_i 及び \bar{b}_j は, 各入力文中の単語に対する事前学習された単語分散表現であ る. [Parikh17]は, GloVe[Pennington 14]を単語分散表現の事前 学習手法として用いているが, 今回のモデルでは, ELMo[Peters 18]を使用する.

次に、2 つ入力文の単語分散表現 \bar{a}_i 、 $\bar{b}_j \geq a_j$ 、 $\beta_i \geq 0$ 間で意味 的関係をフィードフォワードネットGにより計算して、各単語と文 間の関係を表すベクトル $v_{a,i}$ 、 $v_{b,i}$ を求める.

$$v_{a,i} = G([\bar{a}_i, \beta_i]) \qquad (4)$$

$$v_{b,j} = G([\bar{b}_j, \alpha_j]) \qquad (5)$$

$$v_a = \sum_{\substack{i=1\\l_b}}^{l_a} v_{a,i} \qquad (6)$$

$$v_b = \sum_{\substack{j=1\\j=1}}^{l_b} v_{b,j} \qquad (7)$$

最終的に、フィードフォワードネットHにより、2つの文間の関係 を表すラベルyを出力する.

$$y = H([v_a, v_b])$$
(8)

この手法は、アテンションとフィードフォワードニューラルネット ワークに、単語分散表現の事前学習を組み合わせた単純なモ デルとなっており、従来の複雑なアーキテクチャに比べて少ない パラメータで学習することができる.

4. 実験

4.1 データセット

実験では、英語学習者向けのニュースサイト Breaking News English¹から収集したマルチレベルの平易化テキストを用いる. Breaking News English は、easy タイプの記事(level 0, 1, 2, 3) と hard タイプの記事(level 4, 5, 6)から構成されている. 今回の 実験では、収集サンプルから平易化文ペアであるものと平易化 文ペアでないものを 50 ペアずつランダムに抽出し、計 100 事例 を用意した. easy タイプの記事からはレベル 0 と3 を、hard タイ プの記事からはレベル 4 と 6 の文を使用する. これらのペアに 人手でアノテーションを施すことで、モデルの平易化文ペア識 別性能をテストする. ベースラインには、Jaccard Similarity を用 いる. 各モデルの閾値は、[Xu 15]に従い 0.5 とする.

4.2 実験結果

実験結果を表1に示す.表1から, F値において, DAM が ベースラインを大きく上回っていることが分かる. 特に, Recall に おいて, Jaccard Similarity に大きな差をつけていることから, 平

¹ https://breakingnewsenglish.com/

表1平易化文ペア特定の実験結果

	Recall	Precision	F値
Jaccard	0.36	1.00	0.52
DAM	0.76	0.82	0.79

表2特定された平易化文ペアの例

元文

They said our sleeping brain is much more aware of the outside world than we thought.

平易化された文

The researchers said our sleeping brain is active.



図1単語間のアテンション

易化文ペア間の共通単語頻度などの表層的特徴だけでは捉え きれない含意関係の認識に成功していることが分かる.

一方で、Precision においては、Jaccard Similarity が DAM より も頑健な性能を示した.これは、ベースライン手法が文のペア間 で共通する単語の重なりが一定の割合以上観測されない場合 は全て負例に分類する方式であるためと考えられる.

実際に、Jaccard Similarity では検出に失敗したが、DAM によって検出に成功した平易化文ペアを表 2 に示す. この例では、 平易化において、パラフレーズが起きているものの、含意関係が 認識され検出に成功している. また図1に、単語間のアテンションの相関を示す. ここで、平易化された文中の"acitive"が元文 中の"much more aware"というフレーズに強く対応していることが わかる.

5. まとめ

本研究では、テキスト平易化のパラレルコーパス構築のため に、含意関係認識モデルを用いて、自動アライメントを行うことを 提案した.実験結果より、従来の手法よりも含意関係認識モデル がより意味的な関係性を捉え、平易化文ペアの高精度なアライ メントに貢献することがわかった.今後の課題の1つは、この枠組 みを利用して、実際に作成中のコーパスに自動的にアノテーションを付与する機構を構築することが必要である.

参考文献

- [Nisioi 17] : Sergiu Nisioi, Sanja Stajner, Simone Paolo Ponzetto, Liviu P. Dinu: Exploring neural text simplification models, In Proceedings of ACL. 2017.
- [Xu 15] Wei Xu, Chris Callison-Burch, Courtney Napoles: Problems in Current Text Simplification Research: New Data Help, Transactions of the Association for Computational Linguistics, vol. 3, pp. 283–297, 2015.
- [Zhu 10] : Zhu Z, Bernhard D, Gurevych I: A monolingual treebased translation model for sentence simplification, In Proceedings of the 23rd International Conference on Computational Linguistics (COLING). 2010.
- [Scarton 18] Carolina Scarton, Gustavo Henrique Paetzold, Lucia Specia: Text Simplification from Professionally Produced Corpora, In proceedings of Irec, 2018.
- [Parikh 17] Ankur P. Parikh, Oscar Tackstrom, Dipanjan Das, Jakob Uszkoreit: A Decomposable Attention Model for Natural Language Inference, arXiv preprint:1606.01933, 2017.
- [Pennington 17] Jeffrey Pennington, Richard Socher, Christopher D. Manning: GloVe: Global Vectors for Word Representation, Proceedings of EMNLP, 2014.
- [Peters 18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner: Deep contextualized word representations, In NAACL, 2018.