単語属性変換による自然言語推論データの拡張

石橋 陽一 [1]　須藤 克仁 [1,2]　中村 哲 [1]
[1] 奈良先端科学技術大学院大学　[2] 科学技術振興機構さきがけ
{ishibashi.yoichi.ir3, sudoh, s-nakamura}@is.naist.jp

1 はじめに

自然言語推論は前提文と仮説文が与えられ、それらの間にどのような関係が成立するかを予測するタスクである。これまでいくつかのタスクとデータセットが公開されており、例えばその中で代表的なものとして Stanford Natural Language Inference corpus (SNLI) [1] がある。SNLI は前提文と仮説文の間に成り立つ関係として「含意」「矛盾」そしてそのどちらとも判別できない例に対して「中立」の 3 種類のラベルが付与されているコーパスである。SNLI のデータ数は 550k ペアと、自然言語推論の他のコーパス（MultiNLI [2]:433k・SciTail [3]:27k）よりも比較的多いが、現在主流のニューラルネットワークに基づく手法を利用するにあたっては、その性能を向上させるためにより多くのデータで学習することが望ましい。そこで本研究では自然言語推論のデータ拡張に取り組む。

自然言語推論のデータは語の多様性が少ないため文の表層的な情報に過学習してしまう問題点があることが指摘されている [4]。この問題の解決のためには、表層的な情報への過学習を回避するようなデータが必要である。そこで本研究では SNLI 文中の特定の単語に対して、その意味を反転した単語に置換することで、表層的な情報をできる限り維持したまま意味類似性が低い文を生成する手法を提案する。例えば、前提文 "The boy is eating a fruits." と、含意となる文 "The boy is eating an apple." が仮説文として与えられているとき、仮説文に対して性別を変換する単語変換を適用し、"The **girl** is eating an apple." を生成する。このようなデータ拡張は元の仮説文の文法やスタイル等の表層的な部分は変化させずに、意味を矛盾の方向に変化させることができる。したがって本研究のデータ拡張で生成した仮説文をモデルが学習する場合、表層的な情報だけでなく一部の単語の意味の違いに注目する必要があるため、拡張データで学習したモデルの性能が表層的な情報に左右されにくく頑健になると考えられる。

特定の単語の変換を行う手法として鏡映変換による単語属性変換 [5] がある。この手法は次の 2 つの特徴： (1) 変換対象の属性（例：性別）を持つ単語を変換し (2) 変換対象の属性を持たない単語は変換しない という特徴を持つ。鏡映変換に基づく単語属性変換を文中の全ての単語に鏡映変換を適用した場合、文中の一部の単語が変化しそれ以外は元文と同じであるような文を作り出すことができる。これにより元文の表層的な情報をある程度維持したまま意味を変えたデータを生成することが可能となる。そこで本研究では SNLI のデータ拡張に鏡映変換を適用し、その効果を検証する。

2 関連研究

SNLI のデータ拡張は Kang [6] らが提案している。SNLI のデータはクラウドワーカーにアノテーションされているためその語彙が限られており文の多様性が低い。そのためモデルがある種のパターンに過学習することが指摘されている [4]。例えば、前提文 “The dog did **not** eat all of the chickens.” に対して仮説文 “ The dog ate all of the chickens.” が与えられたとき、正しいラベルは「矛盾」であるにもかかわらず、既存のモデルは「含意」と誤分類した例が報告されている。この問題に対して Kang らはルールテンプレートを介して大規模な語彙を文に組み込みデータを拡張することで SNLI と SciTail の精度を向上させた。本研究では別のアプローチとして単語の変換に基づく方法を提案する。

3 手法

3.1 単語属性変換

本研究では文のデータ拡張として、文中の特定の単語を変換する鏡映変換に基づく単語属性変換 [5] を用いる。鏡映変換に基づく単語属性変換は、変換対象の属性を持つ単語を変換し、変換対象の属性を

持たない単語は変換しないという性質を持つ。例えば性別の属性変換 f_{gender} によって “man” のベクトル $\mathbf{v}_{man}$ を “woman” のベクトル $\mathbf{v}_{woman}$ に変換する。一方で、性別に対して不変であるような単語、例えば “person” のベクトルが与えられた場合は変換せず入力ベクトルと同じ $\mathbf{v}_{person}$ を出力する。これらの 2 つの性質によって文中の特定の単語（例：性別に関する単語）のみ変換し、それ以外の単語は変換せずにノイズの少ないデータを自動で生成可能となる。ここで単語ベクトルのアナロジーでも変換可能のように思えるが、その場合は変換対象の単語（例：性別単語）に関して事前知識を用意する必要がある。例えばアナロジーで男性を表す単語を女性に変換する場合、男性のベクトルから差ベクトル $d = \mathbf{v}_{woman} - \mathbf{v}_{man}$ を引くことで女性に変換し、逆に女性から男性への変換の場合 d を足すことで変換する。したがってアナロジーで単語を変換する場合は入力単語ベクトル $\mathbf{v}_x$ が男性もしくは女性のどちらに属すかという事前知識が必要となるが、鏡映変換は事前知識を用いないための条件（単語ベクトル空間中の 2 点を同じ写像で反転可能）を満たす写像であるため、そのような事前知識を用いることなくデータ拡張に適用できる。

3.2 鏡映変換

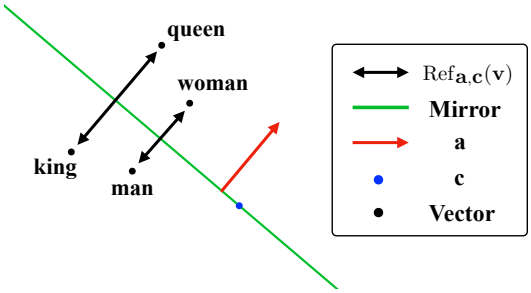

図 1 鏡映変換に基づく単語属性変換

鏡映変換は鏡と呼ばれる超平面によって 2 つのベクトルの位置を反転させる写像である。標準内積が与えられた n 次元実ユークリッド空間 $\mathbb{R}^n$ における鏡映変換は

$$\mathrm{Ref}_{\mathbf{a},\mathbf{c}}(\mathbf{v}) = \mathbf{v} - 2\frac{(\mathbf{v}-\mathbf{c})\cdot\mathbf{a}}{\mathbf{a}\cdot\mathbf{a}}\mathbf{a} \tag{1}$$

と定義される。ここで $\mathbf{a}\cdot\mathbf{a}$ は内積を表す。また $\mathbf{a}$ および $\mathbf{c}$ はそれぞれ鏡（超平面）を決定するパラメタであり、$\mathbf{a}$ は鏡に直交するベクトル、$\mathbf{c}$ は鏡が通る $\mathbb{R}^n$ 上の点である（図 1）。

本研究では全結合の多層パーセプトロン（MLP）によって変換対象の属性ごとに鏡のパラメタである $\mathbf{a}$ と $\mathbf{c}$ を推定することで鏡を学習する:

$$\mathbf{a} = \mathrm{MLP}([\mathbf{z};\mathbf{v}_x]), \tag{2}$$

$$\mathbf{c} = \mathrm{MLP}([\mathbf{z};\mathbf{v}_x]) \tag{3}$$

ここで $[\cdot;\cdot]$ はベクトルの列方向の連結を表す。そして入力単語ベクトル $\mathbf{v}_x$ の属性を反転させたベクトルを鏡映変換し $\mathbf{v}_y$ を得る:

$$\mathbf{v}_y = \mathrm{Ref}_{\mathbf{a},\mathbf{c}}(\mathbf{v}_x) \tag{4}$$

そして予測されたベクトル $\mathbf{v}_y$ が目的の単語ベクトルとなるように平均二乗誤差で最適化する。

3.3 データ拡張

鏡映変換に基づく単語属性変換を利用してデータ拡張を行う。最初に単語属性変換モデルを学習し、その後 SNLI のデータに適用しデータを拡張する。本研究では SNLI データの仮説文のうち含意ラベルが付与されている文に単語属性変換を適用し、文の意味が前提文と矛盾するように単語の意味を反転させることで矛盾ラベルを付与した新たな文を生成する。単語を変換する場合、変換対象外の単語が変換されてしまうことが起きうるが、鏡映変換に基づく単語属性変換では、鏡映写像の特性によって変換対象外の単語が入力された場合はほぼ変換されないため、この問題をある程度回避する事ができる。またこれにより、一部の単語のみが変換されることで、文の表層的な情報をほとんど変えずに矛盾する文を生成できる。そこで本研究では含意ラベルが付与された仮説文の全ての単語を鏡映によって変換する。なお、単語属性変換によって拡張するデータは訓練データのみとし、評価データは拡張を行わない。また、鏡映変換によって変化しなかった文は新たなデータとして追加しない。

4 実験

4.1 単語属性変換の学習と SNLI データの拡張

まずは単語属性変換の学習を行った。[5] の単語属性変換の学習データと評価用データを結合し、学習データとして用いた。変換対象は性別と反意語とし、それら 2 つの単語データを結合し学習データとした。事前学習済み単語埋め込みには GloVe [7] を使用した。単語ベクトルの次元数は 300 次元、5 層の MLP の隠れ層次元数は 300 次元で鏡映変換の学習を行った。最適化には Adam を使用し学習率の設

定は以前の研究で最高精度のモデルの設定を使用した。

表 1　単語属性変換の学習用データセット

変換属性	データ数
性別単語	202
反意語	6286

次に、学習した単語属性変換を SNLI のデータに適用しデータを拡張した。表 2 は追加されたデータ数と拡張前のデータ数の比較である。拡張前と比較して最大で 24 万件データを生成・追加することができている。表 4 に生成されたデータの一例を示す。鏡映変換によって変換対象の特定の単語のみ変換されており、元の仮説文の表層的な類似性を維持したまま意味を矛盾に変えることができている。例えば、性別の鏡映変換によって “An actor is performing in an outdoor park dressed up.” の “actor” を “actress” に変えているが、それ以外の単語は鏡映変換を適用しても変化していない。このため、文の品質を損なうような単語が混在した文を生成することなくノイズの少ないデータを多く生成できたと考えられる。

表 2　データ拡張の結果

Dataset	#Train	#Val	#Test	#Total
拡張前	550,152	10,000	10,000	570,152
+ 性別変換	640,310	10,000	10,000	660,310
+ 反意語変換	699,272	10,000	10,000	719,272
+ 性別 + 反意語	789,430	10,000	10,000	809,430

表 3　単語属性変換で拡張したデータで学習したモデルのスコア

Dataset	Test Accuracy
拡張前	0.797
+ 性別変換	0.794
+ 反意語変換	**0.798**
+ 性別 + 反意語	0.792

4.2　SNLI モデルの学習とデータ拡張による精度の変化

データ拡張を行っていないオリジナルのデータと、単語属性変換による拡張を行ったデータでそれぞれ SNLI のモデルを学習した。本研究では前提文と仮説文それぞれを別の RNN でエンコードし結合したベクトルを MLP に入力し分類するモデル [1] を用いた。モデルの設定はすべての実験で同一とし、RNN には 300 次元の LSTM を用い、バッチサイズは 128 で学習を行った。実験の結果を表 3 に示す。

実験の結果、データ拡張の実施前後でモデルの性能は大きく向上しなかった。この原因としてラベルの偏りによってモデルが過学習したことが挙げられる。本研究では矛盾ラベルは大きく増加しているが含意や中立のラベルは増えていないため、SNLI モデルが矛盾ラベルに過学習し、分類結果が偏るケースが増加したと考えられる。そこで、混同行列を作成しラベルの偏りによって分類結果が異なっているか確認した（図 2）。その結果、矛盾と中立に分類されたケースが 128 から 303 件増加していることがわかった。したがって、ラベルの偏りにモデルが過学習した結果、精度が向上しなかったと思われる。解決策として含意ラベルを付与した文を増加させる方法を検討中である。

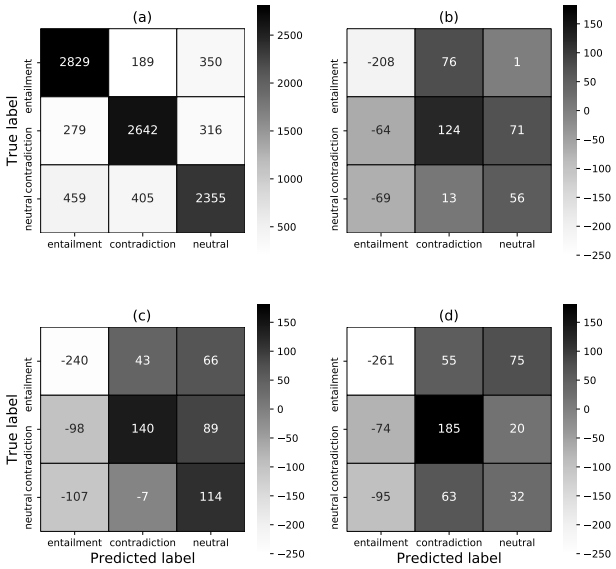

図 2　混同行列の可視化。(a)：データ拡張なし (b)：データ拡張（性別） (c)：データ拡張（反意語） (d)：データ拡張（性別＋反意語）(b)-(d) は (a) からの差分（増減数）を元に混同行列を作成している。

5　まとめと今後の課題

本研究では自然言語推論のデータ拡張のために文の表層的な情報をできる限り変化させず意味を変化させたデータを追加するデータ拡張に取り組んだ。文の表層的な類似性を保ち意味を変換するため文中の特定単語のみを変換する単語属性変換を適用しその効果を検証した。実験の結果、鏡映変換に基づく手法は文中の特定の単語のみを変換させ、ノイズの少ないデータを多く生成することができた。実験では SNLI モデルを学習しデータ拡張による性能を比

表 4 生成されたデータの例。ラベルは元の仮説文に対しては含意、生成された仮説文に対しては矛盾を付与している。

変換属性	前提文	元の仮説文（含意）	生成された仮説文（矛盾）
性別	A woman in a green jacket and hood over her head looking towards a valley.	The woman is wearing green.	The **man** is wearing green.
性別	An actor dressed as a pirate performs in an outdoor park.	An actor is performing in an outdoor park dressed up.	An **actress** is performing in an outdoor park dressed up.
性別	A woman in costume is marching with a large drum.	She plays in a band.	**He** plays in a band.
反意語	Two people climbing up a snowy mountain.	The mountain is cold.	The **valley** is **warm**.
反意語	A woman wearing a blue skirt, high heels, a white shirt, green jacket and headband is walking out of a tunnel.	A woman has a white skirt.	A woman has a **dark** skirt.
反意語	One boy in a striped shirt and one girl in a green shirt each riding on bicycles with training wheels .	Two kids are outside on their bikes .	Two kids are **inside** on their bikes.

較したが、データ拡張によって特定のラベルが増加し過学習が起きた結果、モデルの性能に顕著な差は見られなかった。今後はラベルスムージングや前提文にもデータ拡張を適用し含意ラベルを増やすことでラベルの偏りを調整することを検討している。

謝辞

本研究は JST さきがけ (JPMJPR1856) の支援を受けたものである.

参考文献

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 632–642. The Association for Computational Linguistics, 2015.

[2] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018.

[3] Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5189–5197. AAAI Press, 2018.

[4] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Mitigating Annotation Artifacts in Natural Language Inference Datasets to Improve Cross-dataset Generalization Ability. *CoRR*, Vol. abs/1909.04242, , 2019.

[5] Yoichi Ishibashi, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. Reflection-based Word Attribute Transfer. In Shruti Rijhwani, Jiangming Liu, Yizhong Wang, and Rotem Dror, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2020, Online, July 5-10, 2020*, pp. 51–58. Association for Computational Linguistics, 2020.

[6] Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard H. Hovy. AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2418–2428. Association for Computational Linguistics, 2018.

[7] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1532–1543, 2014.