Vol.2023-NL-256 No.8
2023/5/28

意味と表層の類似度を制御した言い換えによるデータ拡張を用いた事前学習済みモデルの性能改善

小笠 雄也[1,a)]　梶原 智之[2,b)]　荒瀬 由紀[1,c)]

概要：言い換え生成技術は様々な自然言語処理タスクのデータ拡張に応用されてきた．データ拡張においては，意味的な類似度が高くかつ多様な言語表現を提供する言い換えが有益である．しかし原文からの表層の変化が大きくなるにつれて意味を保持することが難しいため，このような言い換え文の生成は困難である．さらにデータ拡張を適用するタスクによっても，望まれる意味・表層の類似度は異なる．そこで本研究では意味類似度が高く表層類似度が低い言い換えを高品質な言い換えと定義し，意味・表層類似度を制御可能な言い換え生成手法を実現する．具体的にはデコーダにサンプリングを適用した折り返し翻訳により，多様な品質の言い換え候補を大量に自動生成する．生成した候補の中から高品質な言い換え文対を抽出し，意味・表層類似度をタグとして付与して事前学習済み系列変換モデルを fine-tuning することで，類似度制御可能な言い換えモデルを構築する．提案手法によるデータ拡張を事前学習済み言語モデルの性能改善タスクに適用し，その有効性を検証した．対照学習を行う手法，fine-tuning 前に中間タスクによる追加訓練を行う手法，それぞれにおいてデータ拡張により既存手法の性能を改善することが明らかとなった．

1. はじめに

言い換え生成 [1] は入力文の意味を保持しながら表現が異なる文を生成するタスクである．生成した言い換え文により疑似的に訓練データを増やすデータ拡張は，事前学習済み言語モデル [2,3]，質問応答 [4]，タスク指向対話システム [5,6]，機械翻訳 [7] の性能改善，学習支援システム構築 [8] などに有効である．原文と意味が近くかつ表層が大きく異なる言い換え文は，原文と異なる表現を多く含むことから，データ拡張において有益であることが多い [9]．本研究では，このような意味類似度が高く表層類似度が低い文を，高品質な言い換え文と呼ぶ．

しかし，原文からの表層の変化が大きくなるにつれて意味を保持することが難しい [10]．**図 1** は，折り返し翻訳で生成した言い換え[*1]と既存の言い換えコーパスである ParaNMT-50M [11] および Paracotta [12] に含まれる言い換えの意味類似度と表層類似度の関係[*2]をヒートマップで表したものである．セルの色の濃さは該当する文対の割合（%）を示し，高品質な言い換えは図の右下のセルに対応する．図 1 から，折り返し翻訳および ParaCotta では，原文と意味は近いが表層も近い文が多くを占めることが分かる．また，ParaNMT-50M では表層は大きく異なる文も多いが，それらの意味類似度は低いものが多くを占めている．さらに，意味類似度や表層類似度が非常に高い（右上のセルに該当），つまり原文と言い換え文がほとんど同一となっているものも多いことも分かる．このように既存の言い換え手法では高品質な言い換えを生成することは難しい．

また，5 章および 6 章で示すとおり，データ拡張に適する意味・表層の類似度はタスク依存である．そのため，類似度を制御できることが最終的なタスクの性能を改善する上で重要であるが，類似度制御機構をそなえた言い換えモデルはほとんど存在しない．

本研究では，高品質かつ意味類似度と表層類似度を制御可能な言い換え生成手法を実現する．まず，折り返し翻訳器のデコーダにサンプリングを適用することで多様な品質の言い換え文対の候補を大量に自動生成する．そして，生成した候補の中から高品質な言い換え文対を抽出するこ

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
[2] 愛媛大学大学院理工学研究科
Graduate School of Science and Engineering, Ehime University
[a)] ogasa.yuya@ist.osaka-u.ac.jp
[b)] kajiwara@cs.ehime-u.ac.jp
[c)] arase@ist.osaka-u.ac.jp

[*1] 5 章で用いる英語版 Wikipedia を折り返し翻訳した結果を集計．

[*2] 記号を除去したのちに 4 語以上からなる文対を 5 万文対ずつランダムサンプリングした．

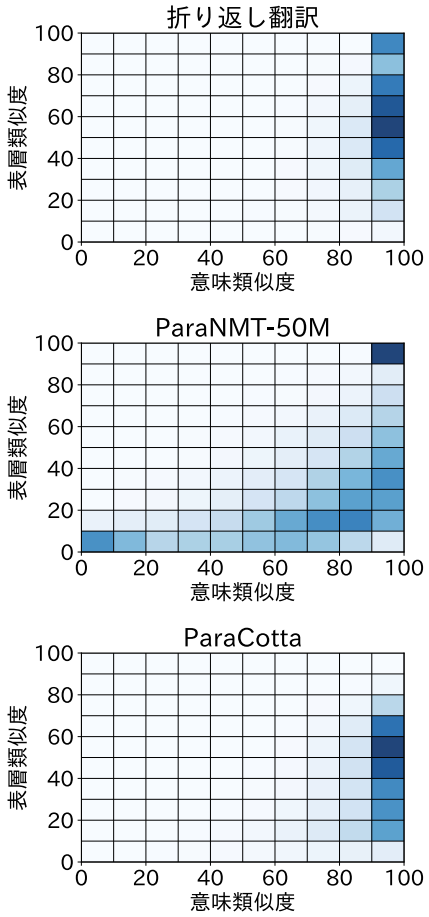

図 1: 既存手法による言い換え文の意味的・表層的類似度の分布（意味類似度は fine-tuning した事前学習済みモデルで，表層類似度は BLEU で測定）

とで，言い換え生成モデルの訓練コーパスを構築する．また，シンプルな機構で意味・表層類似度を制御するために，これらの類似度を表すタグを訓練コーパスに付与し [13]，BART [14] を fine-tuning して言い換え生成モデルを構築する．4.3 節で示す通り，提案手法では所望の意味・表層類似度の高品質な言い換え文を生成できる（**図 3**）．

提案手法の効果を検証するため，事前学習済み言語モデルの性能改善手法にデータ拡張を適用した．実験の結果，対照学習を行う手法 [2, 15] および fine-tuning 前に中間タスクによる訓練を行う手法 [3] の両方において，既存手法を上回る性能を達成した．

2. 関連研究

多様な言い換えの生成は活発に研究されている．Qian ら [9] は複数の言い換え生成モデルを用いる手法を，また Cao and Wan [16] は条件付き GAN を用いる手法を，Park ら [17] や Gupta ら [18] は潜在表現を摂動する手法を提案している．Maddela ら [19] は，単語の削除および文の分割をしてから言い換えを生成することで，生成文の多様性を高めている．

既存研究では語彙および構文という，特定の属性の多様性に着目した言い換え生成も研究されている [12, 19–28]．語彙の多様性に着目した手法として，Vijayakumar ら [20] は以前の生成単語との意味空間 [29] 上でのハミング距離が大きくなる単語を生成する手法や，n-gram を多様化させる等の手法をビームサーチに導入することで，多様な言い換え文の生成を行った．また ParaBank [21] は，チェコ語から英語への逆翻訳により言い換えを生成する ParaNMT [11] を拡張し，デコーダに語彙制約 [30, 31] を加えることで多様な言い換えを生成した．語彙制約により，入力文と重複したトークンを出力しないよう制約を与えることで，言い換え文対間での語彙の多様性を高めた．同様に Niu ら [22] は入力文と重複したトークンを連続して出力することを制限するデコード手法により，Zeng ら [23] は原文とキーワードを入力とし，そのキーワードを含むような言い換えを学習することで，語彙的に多様な言い換え生成を行った．ParaCotta [12] は折り返し翻訳を用いて言い換え候補を生成し，その中から BLEU [32] の値が低いものを抽出することで，語彙的に多様な言い換えを収集することを目指した．しかし図 1 からも分かる通り，単純な折り返し翻訳では多様な言い換えは獲得し辛い．

一方，構文の多様性に着目した手法として，Iyyer ら [24] は入力に構文解析木を加え，構文を制御した言い換え生成を行うことで入出力間での構文の多様性を高めた．また，Hosking and Lapata [25]，Chen ら [26]，Bao ら [27] は，構文解析木の代わりに，構文の模範となる文を入力に加えた．さらに，Goyal and Durrett [28] は既存の言い換えコーパスから構文規則を学習し，その規則を基にフレーズを適切に並び替えることによって多様な言い換えを行った．

これらの既存手法には，言い換え生成における多様性を制御できるものは存在しない．Bandel ら [10] は既存の言い換えコーパスを用いて言い換え文対間の典型的な意味・構文・語彙の類似度を推定し，これらを制御ベクトルとして言い換え生成する手法を提案した．ただし，これらの類似度の制御は，入力文に応じてモデルが推定した制御ベクトルに依存するため，ユーザが意図した通りの類似度を持つ言い換えを生成する機構は持たない．また，Bandel らの研究では応用タスクにおける手法の有効性は検証されていなかった．これに対し提案手法では，タグの付与というシンプルな機構でユーザが所望の意味・表層類似度を指定できる．さらに本研究では，提案手法によって生成する言い換えが応用タスクの性能改善に有効であることを示す．

3. 提案手法

提案手法の概要を**図 2** に示す．提案手法は，高品質な言い換え文対からなる言い換えコーパスを自動構築し，意味・表層類似度をタグとして付与して BART を fine-tuning する．これにより高品質な言い換え生成を可能とし，言い換

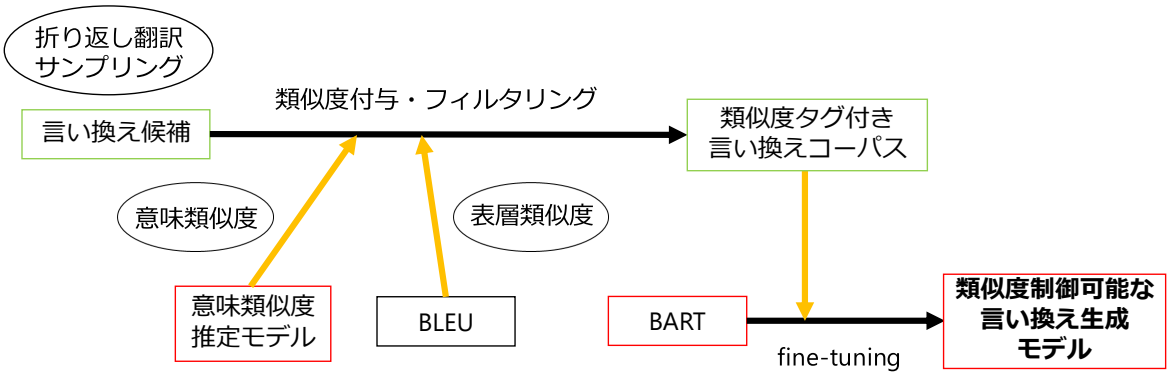

図 2: 提案手法の概要図

え生成における意味・表層類似度の制御を実現する．

3.1 意味・表層類似度推定

意味・表層類似度推定において微細な表記ゆれの影響を避けるため，入力文と言い換え文の双方において事前に記号を除去する．具体的には，アルファベット，数字，空白文字，コンマ，ピリオド以外の記号を除去した．

意味類似度は fine-tuning した事前学習済みモデルで推定する．モデルには Sentence-BERT [33] の Cross-Encoder を用いる．Cross-Encoder は入力された 2 文間の意味的な類似度を $[0, 1]$ の範囲で推定するモデルであり，本研究では推定値を 100 倍した値を用いた．Cross-Encoder の訓練は，Semantic Textual Similarity（STS）タスクのデータセットである STS-B [34] を用いて行う．

表層類似度は 2 文間の Sentence BLEU [32] により推定する．その際，入力文および言い換え文は小文字化し，BLEU 値は 100 倍した値（%）を用いる．

3.2 折り返し翻訳による言い換え文対候補の自動生成

本研究では，折り返し翻訳を用いて言い換え文対の候補を自動生成する．単純なビームサーチを用いた折り返し翻訳では，図 1 に示した通り，言い換え文対の表層類似度が高くなりやすい．そこで，多様な表層類似度の言い換えを生成するため，ビームサーチにサンプリングを適用する．

サンプリングには，温度付き softmax と Top-k サンプリング [35] を組み合わせる．温度付き softmax では温度 T が 1 より大きくなるほど単語の生成確率が均等に近づくため，Top-k サンプリングと組み合わせることで，生成確率が低いトークンも選択されやすくなる．本手法では，表層的に多様な言い換え候補を得られる反面，意味類似度が低い候補も生成されてしまう．本研究では，言い換え文対の候補を大量に生成し，3.1 節で述べた類似度推定に基づくフィルタリングを行うことで，高品質な言い換え文対を確保する．

3.3 類似度制御可能な言い換え生成モデルの構築

生成した言い換え候補の意味・表層類似度を推定し，高品質な言い換えであると期待できる文対を抽出する．本研究では，意味類似度が 70 より大きいかつ表層類似度が 45 以下の文対を，高品質な言い換え文対と定義する．これらの言い換え文対を用いて BART [14] を fine-tuning することで，類似度制御可能な言い換え生成モデルを構築する．ここで，入力文の先頭には，意味類似度および表層類似度を示すタグ[*3]を付与する．**表 1** に，構築した訓練コーパスの例を示す．

意味類似度のタグには，類似度の値を 5 単位で区切った <SIM70>，<SIM75>，<SIM80>，<SIM85>，<SIM90>，<SIM95> の 6 種類を使用する．例えば，<SIM70> のタグは，文対の意味類似度が 70-75 の間であることを表す．同様に表層類似度のタグには，<BLEU0_5>，<BLEU10>，<BLEU15>，<BLEU20>，<BLEU25>，<BLEU30>，<BLEU35>，<BLEU40> の 8 種類を使用する．なお，表層類似度が 0 から 10 となる文対が少なかったため，表層類似度が 0-10 の文対をまとめて <BLEU0_5> のタグを付与した．

4. 類似度制御可能な言い換えモデルの構築

本章では，提案手法である類似度制御可能な言い換えモデルを構築し，生成される言い換え文を分析する．

4.1 コーパス作成

Kajiwara ら [36] に従い，英語の言い換え文対を収集するために，英独・独英の折り返し翻訳を行う．翻訳器には，Ng ら [37] によって訓練された Transformer [38] に基づく英独翻訳器[*4]および独英翻訳器[*5]を用いた．デコーダにおける温度付き softmax と Top-k サンプリングの設定のため，$k = \{10, 20, 30, 40\}$，$T = \{1.0, 2.0, 3.0, 4.0\}$ において k, T の組み合わせを試し，言い換え文対候補の意味類似度

[*3] タグは BART の語彙に追加した．
[*4] https://huggingface.co/facebook/wmt19-en-de
[*5] https://huggingface.co/facebook/wmt19-de-en

表 1: 類似度制御可能な言い換え生成モデルの訓練コーパスの例

入力文	言い換え	意味類似度タグ	表層類似度タグ
The bridge's construction date is unknown.	Nothing is known about the date of construction of the bridge.	<SIM90>	<BLEU0_5>
Who was ready for the truth?	Who was prepared for the truth?	<SIM95>	<BLEU35>
There was nobody coming out that door.	No one came out of that apartment door.	<SIM80>	<BLEU10>
This is a moral indictment of the state of our world.	This is an accusation that lies against the state of our world.	<SIM70>	<BLEU40>

と表層類似度の分布を検証した．また，言い換え文対候補が不自然な文になっていないか，各組み合わせについて目視による確認を行った．その結果，意味類似度が高く，表層類似度が低い文対が多く生成されており，不自然な文の割合が小さかった $(k, T) = (20, 3.0), (30, 2.0)$ の 2 つの組み合わせを用いることとした．さらに，原言語 → 目的言語，目的言語 → 原言語，それぞれの方向で 2 種類の設定を用いることで，1 つの入力文から 4 つの言い換え文対の候補を得た．

折り返し翻訳の入力文として，WikiMatrix [39] の英独対訳コーパスおよび NewsCrawl [40] からサンプルした約 3,000 万の英文を用いた．3.2 節で述べた方法により折り返し翻訳を行い，約 1.2 億文対の言い換え候補を生成した．これら言い換え候補について意味・表層類似度を推定し，高品質な言い換え文対を抽出した．意味類似度タグ 6 種類と表層類似度タグ 8 種類の組み合わせ，計 48 種類の文対の数が均等になるように，訓練データ 500 万文対，検証データ 2700 文対，テストデータ 2700 文対をサンプルした．なお，<BLEU0_5> のタグが使われている文対は，表層類似度が 0-10 であり，他のタグと比較し範囲が 2 倍であるため，2 倍の数の文対を抽出した．

4.2 実装の詳細

提案手法は PyTorch[*6] および Hugging Face[*7] を用いて実装した．意味類似度推定モデルで用いる事前学習済み言語モデルとして，DeBERTaV3 [41][*8] を用いる[*9]．意味類似度推定モデルは STS-B を用いて fine-tuning するが，初期シードによって性能が変動する．そこで 10 回シードを変えて訓練を行い，検証データで最も高い性能を示したモデルを意味類似度推定モデルとして使用した．

類似度制御可能な言い換え生成モデルは BART[*10] を 4.1 節で構築したコーパスにより fine-tuning することで構築した．バッチサイズは訓練，検証ともに 128 とした．最適化アルゴリズムには，AdamW [44] を用い，学習率は 1e-5[*11] とした．訓練が 1 エポック終了する度に，検証データで損失を評価し，5 回改善が見られなくなったところで訓練を終了した．言い換え文のデコードにはビーム幅 20 のビームサーチを用い，入力の 0.75 倍から 1.5 倍のトークン数になるように出力長を制限した．

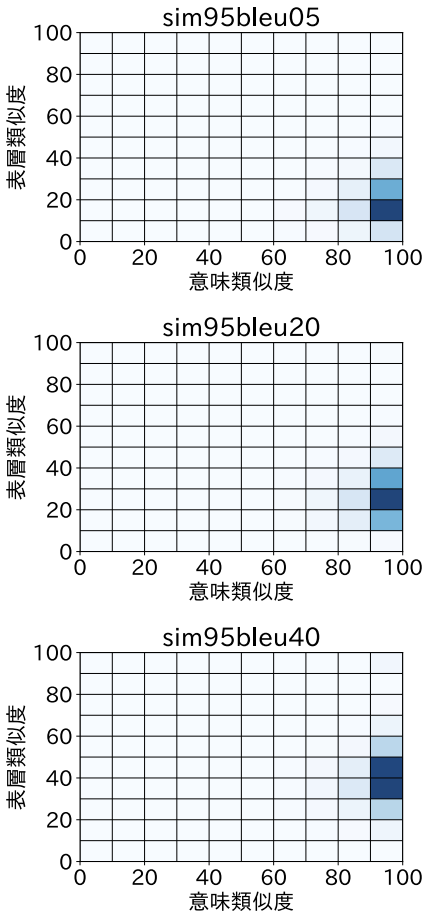

図 3: 類似度制御可能な言い換えモデルによる言い換えの意味・表層類似度の関係

4.3 提案手法による言い換えの性質

構築した類似度制御可能な言い換えモデルにおいて，<SIM95> <BLEU0_5>，<SIM95> <BLEU20>，<SIM95> <BLEU40> のタグを用いた際の言い換え文の意味・表層類似度を分析する．英語版 Wikipedia からサンプルした 5 万文についてそれぞれのタグを用いて言い換えを生成した．記号を除去したのちに 4 語以上の文対に対して[*12]類似度を測定し，図 3 に示すヒートマップを作成し

[*6] https://pytorch.org/ (ver. 1.11.0+cu113)
[*7] https://github.com/huggingface
[*8] https://huggingface.co/microsoft/deberta-v3-large
[*9] BERT [42] および RoBERTa [43] と比べて，検証データで高い性能を示した．
[*10] https://huggingface.co/facebook/bart-base
[*11] 5e-6，1e-5，2e-5 の中で，検証データにおける損失が最も小さかった．
[*12] 4 語未満の文対では表層類似度の指標となっている BLEU が 0 になってしまうため．

表 2: 類似度制御可能な言い換え生成モデルによる出力例（入力文は CNN-Daily Mail [45] よりサンプル）

入力文 : Sources claimed he'd go if Lib Dems are reduced to fewer than 30 MPs .	
タグ	出力文
<SIM95> <BLEU0_5>	Sources claimed he would leave if the Liberal Democrats were reduced to less than 30.
<SIM95> <BLEU40>	He claimed he would leave if the Liberal Democrats were reduced to fewer than 30 MPs.
<SIM70> <BLEU0_5>	Sources said he would leave the party if the number of MPs was reduced to 30.
<SIM70> <BLEU40>	Sources said he would leave if the party were reduced to fewer than 30 MPs.
入力文 : The tiger sanctuary has been told their 147 cats must be handed over .	
タグ	出力文
<SIM95> <BLEU0_5>	The tiger sanctuary was told to hand over its 147 cats.
<SIM95> <BLEU40>	The tiger sanctuary has been informed that its 147 cats will have to be handed over.
<SIM70> <BLEU0_5>	The animal sanctuary has been informed that its 147 cats will have to be handed over to the animal welfare organisation.
<SIM70> <BLEU40>	The shelter has been told its 147 cats must be handed over to the animal welfare organisation.

た．セルの色が濃いほど，該当する意味・表層類似度を持つ言い換え文の割合が多いことを示している．図 3 より，それぞれのタグで概ね目標とした類似度の文が生成できていることが確認できる．図 1 で示した既存の言い換え手法とは異なり，提案手法は高い意味類似度・低い表層類似度のタグを指定することで高品質な言い換えを生成できる．

表 2 に類似度制御可能な言い換えモデルによる出力例を示す．意味類似度タグが <SIM95> の出力文を見ると，文の意味を維持しながら表層類似度タグに従って多様な表現を生成できていることが分かる．一方，意味類似度タグが <SIM70> の出力文に注目すると，1 番目の例では，「Lib Dems（自由民主党）」という表現が「party」に抽象化されている．また，2 番目の例では，「the animal welfare organisation」という原文には無かった情報が追加されている．これらによって，意味類似度を下げた言い換え文を生成できていることが分かる．

4.4 評価実験設計

類似度制御可能な言い換えモデルによるデータ拡張の有効性を検証するため，事前学習済みモデルの性能改善に関する 2 種類の評価実験を行った．その内，対照学習に適用した結果を 5 章，事前学習済みモデルの中間タスクによる追加学習（transfer fine-tuning）に適用した結果を 6 章で議論する．これらの評価実験のベースラインとして，言い換え自動生成の代表的手法である折り返し翻訳を用いた．折り返し翻訳器は 4.1 節で訓練コーパス作成に用いたものと同じものを用い，貪欲法によるデコードを行った．

5. 事前学習済みモデルの対照学習への適用

本章では対照学習による事前学習済みモデルの文埋め込み改善における提案手法の効果を検証する．当技術のデファクトスタンダードである SimCSE [2] を用いる．

5.1 SimCSE の学習

SimCSE では，入力文と近い意味を持つ文（正例）の埋め込みが近づき，異なる意味を待つ文（負例）の埋め込みが離れるよう事前学習済みモデルの fine-tuning を行う．SimCSE には，生コーパスを用いる設定と自然言語推論タスクのために構築された NLI コーパスを用いて学習する設定が存在する．

生コーパスを用いる場合，事前学習済みモデルに同じ文を 2 回入力し，異なる Dropout を適用した埋め込みのペアを正例とし，負例はミニバッチからサンプルした文を用いる．本実験でも SimCSE の設定に従い，生コーパスとして英語版 Wikipedia からサンプルした 100 万文を用いる．NLI コーパスを用いる SimCSE では，含意関係を持つ文対を正例，矛盾関係を持つ文対を負例とする．MNLI [46] および SNLI [47] を合わせた約 28 万文対からなるコーパスを用いて対照学習を行う場合に最も高い性能を示すことが SimCSE の実験によって明らかとなっている．SimCSE の実装は著者らによって公開されているプログラム[*13]を用い，実験設定は全て SimCSE に従って BERT-base[*14]に対し対照学習を行う．

5.2 評価タスク

対照学習を行った BERT の性能評価は教師なし STS タスクにより行う．STS12-16 [48–52], STS-B, および SICK-R [53] の 7 種類のテストセットにおいて，各文対の意味類似度を文埋め込み間のコサイン類似度により計算する．人手の類似度ラベルとの相関をスピアマンの順位相関係数（ρ）によって評価する．SimCSE の訓練における初期シードによる性能への影響を考慮し，シードを変え訓練・評価した 5 回の結果の平均値を最終的な評価値とする．

5.3 提案手法の適用

本実験では生コーパスおよび NLI コーパスを学習に用いる SimCSE において，提案手法の効果を検証する．生コーパスを用いた SimCSE に提案手法を適用する場合，Dropout によって生成していた正例を言い換え文を用い

[*13] https://github.com/princeton-nlp/SimCSE
[*14] https://huggingface.co/bert-base-uncased

表 3: STS テストセットにおけるスピアマンの順位相関係数 $\rho \times 100$（提案手法では意味類似度タグは <SIM95> に固定）

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
生コーパス								
SimCSE	67.19	**81.13**	**73.13**	80.51	77.72	76.08	70.66	75.20
SimCSE + 折り返し翻訳	59.16	71.47	65.66	75.55	72.81	69.21	65.84	68.53
SimCSE + 提案手法 <BLEU0_5>	69.42	77.66	72.07	80.87	**79.54**	**77.99**	72.54	75.73
SimCSE + 提案手法 <BLEU20>	**69.53**	78.95	72.83	**81.07**	79.19	77.98	72.64	**76.02**
SimCSE + 提案手法 <BLEU40>	68.92	78.06	72.18	79.96	78.78	77.87	**72.86**	75.52
NLI コーパス								
SimCSE	75.32	84.81	80.30	85.58	81.05	84.39	80.42	81.70
SimCSE + 折り返し翻訳	76.27	83.82	80.59	**86.02**	81.76	84.70	80.47	81.95
SimCSE + 提案手法 <BLEU0_5>	**76.75**	84.73	80.45	85.95	81.78	84.82	80.56	**82.15**
SimCSE + 提案手法 <BLEU20>	76.37	84.72	80.46	85.83	**81.87**	84.76	80.48	82.07
SimCSE + 提案手法 <BLEU40>	76.23	**84.87**	**80.60**	85.93	81.63	**84.84**	**80.65**	82.11

るよう置き換える．NLI コーパスを用いる SimCSE では，入力文・正例・負例それぞれを言い換えたものをコーパスに追加して学習に利用する．すなわち，提案手法を用いた SimCSE では，訓練データである NLI コーパスが 2 倍に拡張される．

SimCSE では意味的な類似度をよく表現する文埋め込みを生成できるモデルを構築することを目的としているため，正例としては意味類似度が高い言い換えが望ましいと考えられる．そこで提案手法である類似度制御可能な言い換え生成モデルに使用するタグは，意味類似度は <SIM95> に固定し，表層類似度タグは <BLEU0_5>，<BLEU20>，<BLEU40> からそれぞれ一つ組み合わせて入力した．

5.4 実験結果と考察

評価実験の結果を**表 3** に示す．生コーパスを用いる場合，SimCSE + 提案手法 <BLEU20> が，ベースラインであるデータ拡張を行わない SimCSE と比較して平均スコアを 0.82 ポイント改善した．また，その他のタグを用いた SimCSE + 提案手法においてもベースラインの平均スコアを上回っている．一方，比較手法である SimCSE + 折り返し翻訳では，ベースラインよりも平均スコアが 6.67 ポイント悪化した．提案手法でも表層類似度が最も高い <BLEU40> の際の性能改善が最も小さいことから，表層類似度が高い言い換えでは対照学習による事前学習済み言語モデルの改善には寄与しないことが分かる．

NLI コーパスを用いた場合，折り返し翻訳によるデータ拡張でも平均スコアを 0.25 ポイント改善したが，提案手法 <BLEU0_5> では平均スコアを 0.45 ポイント改善している．また NLI コーパスを用いる場合，表層類似度は最も小さい <BLEU0_5> が最高性能となっており，生コーパスを用いる場合とは異なっている．以上より，ベースとするコーパスやそのサイズによってデータ拡張における最適な表層類似度が異なり，SimCSE の性能向上には類似度制御可能な言い換え生成が必要であることが明らかとなった．

5.5 「言い換え」の効果の検証

5.4 節の結果より，NLI コーパスを用いた SimCSE において提案手法によるデータ拡張が有効であることが示された．本節では，この性能向上が言い換えによって表現の多様性が確保されたことによるものであるか，訓練データが増えたことによるものであるのか分析する．ただし組み合わせ数が増大するため，本節の実験では負例はミニバッチからランダムにサンプルするものとし，入力文（文 A）と正例（文 B）のみを提案手法によって言い換える（それぞれの言い換え文を「文 A'」「文 B'」と表記する）ものとする．

実験の結果得られた 7 種類の STS タスクの平均スコアを**表 4** に示す[*15]．表中の「AB」の列はオリジナルの SimCSE を示し，A'B'，A'B，AB' の列はそれぞれ言い換え文によって元の文を置き換えた場合を示す．すなわち，A'B' では文対数はオリジナルの SimCSE と同じだが，入力文・正例それぞれが提案手法によって言い換えられたものである．AB + A'B' の列は，AB のデータと A'B' のデータを合わせ，2 倍に拡張したものである．

表 4 から分かる通り，A'B'，A'B，AB' はすべてオリジナルの SimCSE のスコア（列 AB）を上回っており，提案手法によって表層類似度を制御した言い換え生成が有効であることが示された．さらに興味深いことに，2 倍のデータ拡張を行った AB + A'B' 列と，言い換え文により元の文対を置き換えた A'B' 列の性能が同程度となっている．このことから，提案手法による SimCSE の性能向上は訓練コーパスが大きくなったことによるものではなく，言い換えによって訓練コーパスにおける言語的表現の多様性が増したことによるものであることが分かる．この特性はモデルの訓練における計算コスト低減の観点からも望ましい．

[*15] 負例をランダムサンプルしているため表 3 のスコアより性能が低下していることに注意されたい．

表 4: 言い換えによる対照学習への効果の分析

	A'B'	A'B	AB'	AB + A'B'	AB
SimCSE + 提案手法 <BLEU0_5>	**79.25**	79.01	79.23	79.22	
SimCSE + 提案手法 <BLEU20>	79.11	78.91	79.03	79.18	78.57
SimCSE + 提案手法 <BLEU40>	79.13	78.89	79.02	79.16	

6. Transfer fine-tuning への適用

事前学習済みモデルを所望のタスクのコーパスを用いて fine-tuning する前に，関連するタスクによって追加学習する transfer fine-tuning を行うことで，最終タスクでの性能が向上することが示されている [3, 54]．本節では，Phang らが提案した STILTs [3] について，提案手法を適用しその有効性を検証する．Phang らの実験で最も高い性能が得られた組み合わせとして，事前学習済みモデルに BERT-large を，transfer fine-tuning のタスクに MNLI を用いる．Transfer fine-tuning の訓練は Phang らの設定を用いて行う[*16]．

6.1 評価タスク

Transfer fine-tuning を行った BERT の性能評価は GLUE ベンチマークに含まれる 8 種類のタスク[*17]によって行う．8 種類のタスクには, 文法が正しいかを判定する CoLA [55], 二値の感情分析を行う SST-2 [56], 言い換え認識の MRPC [57] と QQP[*18], STS-B, 含意関係認識の MNLI, QNLI [58][*19], および RTE [59] が含まれる．CoLA はマシューズの相関係数，MRPC と QQP は F1 スコア，STS-B はピアソンの相関係数，その他のタスクでは accuracy を評価指標とする．評価は，GLUE ベンチマークの評価サーバ[*20]で行う．

訓練コーパスに含まれる文（対）数が 10,000 より多いタスク（SST, QQP, MNLI, QNLI）に関しては 3 エポックの訓練を行う．ただし transfer fine-tuning で MNLI コーパスを用いるため，これを行ったモデルについては fine-tuning は行わず，transfer fine-tuning 後のモデルを評価するものとする．一方，BERT-large では訓練コーパスのサイズが小さい場合に学習が不安定になることが知られている [42]．そこで CoLA，MRPC，STS-B，RTE については，訓練を 10 エポックとし，初期シードを変えながら 5 回実験を行い，検証セットでのスコアが中央値であったモデルを用いて評価する[*21]．Fine-tuning の学習率は，2e-5，バッチサイズは 32 とする．

6.2 提案手法の適用

提案手法である類似度制御可能な言い換え生成モデルにより，transfer fine-tuning に用いる MNLI コーパスのデータ拡張を行う．具体的には MNLI コーパスに含まれる各文対についてそれぞれ言い換えたものをペアとし，コーパスに追加することで，2 倍に拡張する．含意関係ラベルは言い換え元と同じものを付与した．GLUE ベンチマークに含まれる各タスクについて，適切な意味・表層類似度の設定は自明でない．そこで意味類似度タグは <SIM70>，<SIM80>，<SIM95> の 3 種類，表層類似度タグ <BLEU0_5>，<BLEU20>，<BLEU40> の 3 種類を組み合わせた $3 \times 3 = 9$ 通りの設定でそれぞれ言い換えを生成し，データ拡張を行った．言い換え生成ではビーム幅 20 でビームサーチを行い，入力の 0.75 倍から 1.5 倍のトークン数になるように制限を加える．

6.3 実験結果と考察

図 4 に，MNLI コーパスによる transfer fine-tuning を行った STILTs と提案手法の性能比較を，タスクごとにヒートマップで示す．縦軸が意味類似度，横軸が表層類似度を表しており，中間色が STILTs の性能に該当する．青が濃いセルほど提案手法が STILTs を上回っており，赤が濃いものは STILTs が上回っているものである．SST-2, MRPC, MNLI-mm, RTE には濃い青色のセルが存在し, 提案手法によるデータ拡張で STILTs を改善したことが分かる．**表 5** に BERT を直接 fine-tuning した場合，STILTs, MNLI を折り返し翻訳によりデータ拡張した「STILTs + 折り返し翻訳」と提案手法の性能を示す．提案手法については最も高い性能となった意味・表層類似度タグとその結果を示している（最高性能のタグが複数ある場合はその内の一つを掲載）．表 5 より，提案手法は 7 種類のタスクで折り返し翻訳を上回っている．またタスクごとに最適な意味類似度と表層類似度が異なっていることからも，言い換え生成における類似度制御が重要であることが分かる．

提案手法によるデータ拡張が効果的であったタスクは，MRPC や RTE のように，タスク自体の訓練コーパスが小さく，transfer fine-tuning で行う MNLI と同種のタスクであった．これは Arase and Tsujii [54] の実験と一致

[*16] ただし訓練時間を短縮するため，バッチサイズのみ 24 から 32 に変更した．
[*17] WNLI タスクには問題が指摘されているため評価から除外する（https://gluebenchmark.com/faq）．
[*18] https://www.quora.com/q/quoradata/
[*19] Phang らは古いバージョンである QNLIv1 を用いていたが，本実験ではより新しい QNLIv2 を用いた．
[*20] https://gluebenchmark.com/leaderboard
[*21] Phang らの設定では，20 個の初期シードを試し，最高性能のモデルを評価している．しかしこの設定は再現性に懸念があるため，本実験では中央値のモデルを評価に用いることとした．

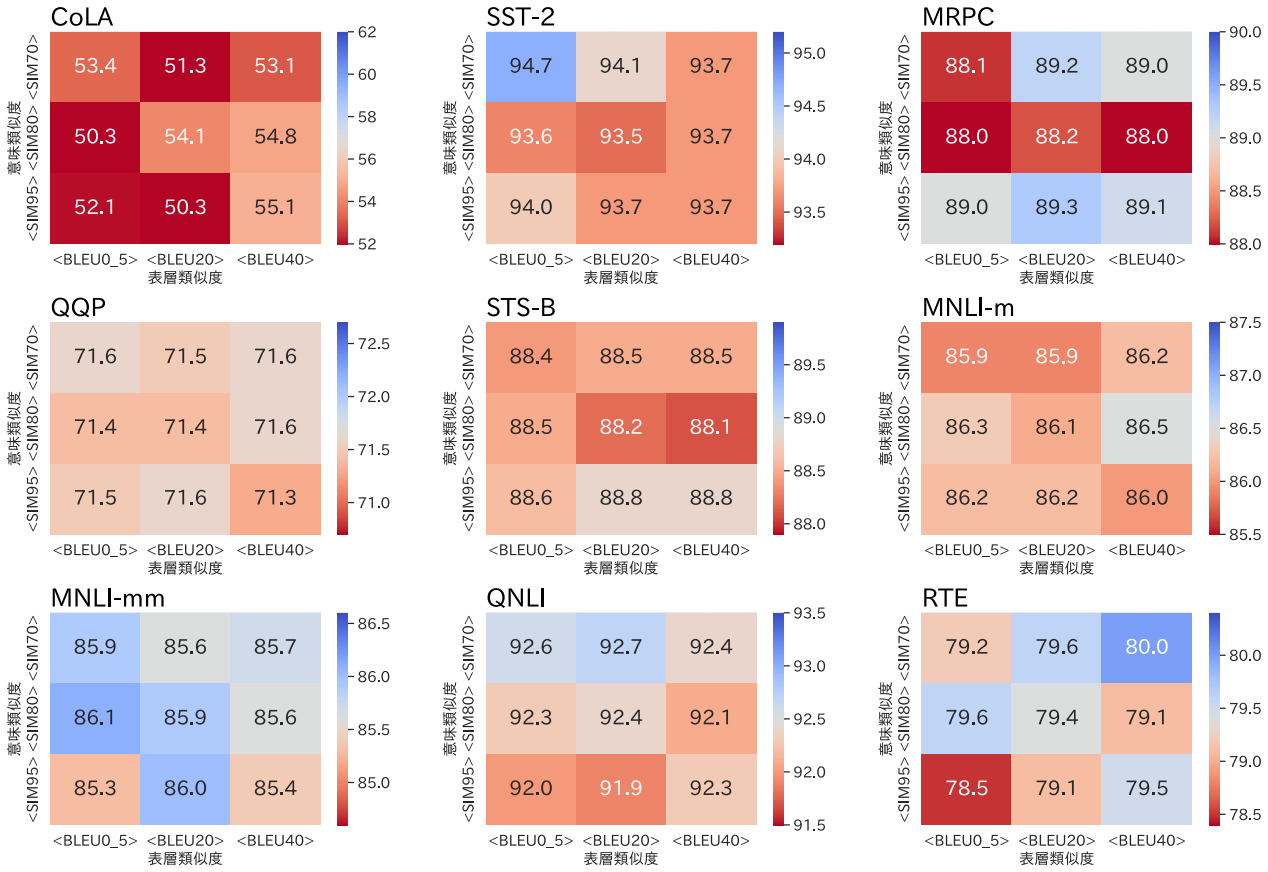

図 4: STILTs を基準とした STILTs + 提案手法の性能のヒートマップ

表 5: GLUE ベンチマークによる実験結果

	CoLA	SST-2	MRPC	QQP	STS-B
fine-tuning	**58.5**	94.3	88.3	**72.4**	86.8
STILTs	57.0	94.2	89.0	71.7	**88.9**
STILTs + 折り返し翻訳	56.7	94.5	88.9	71.7	88.4
STILTs + 提案手法	55.1	**94.7**	**89.3**	71.7	88.8
最高性能のタグ	<SIM95> <BLEU40>	<SIM70> <BLEU0_5>	<SIM95> <BLEU20>	<SIM70> <BLEU0_5>	<SIM95> <BLEU20>
	MNLI-m	MNLI-mm	QNLI	RTE	
fine-tuning	**86.5**	85.6	**92.7**	69.0	
STILTs	–	–	92.5	79.4	
STILTs + 折り返し翻訳	86.1	85.8	92.0	79.2	
STILTs + 提案手法	**86.5**	**86.1**	**92.7**	**80.0**	
最高性能のタグ	<SIM80> <BLEU40>	<SIM80> <BLEU0_5>	<SIM70> <BLEU20>	<SIM70> <BLEU40>	

している．また提案手法は MNLI-mm においても STILTs（fine-tuning に一致）の性能を上回っている．MNLI-mm は fine-tuning 時のコーパスとは類似しない文をテストセットとして用いるものである．提案手法において最良であった表層類似度が最も小さい <BLEU0_5> であったことから，言語表現の多様性を増した MNLI コーパスによる STILTs が BERT の頑健性を向上し，MNLI-mm の改善につながったと考えられる．同様の効果が SST-2 にもあったと推察される．感情分析タスクである SST-2 は MNLI とは関連が薄いが，言語表現の多様性を増した MNLI コーパスによる BERT の頑健性の向上は，SST-2 の改善に資するものであったと考える．

提案手法によるデータ拡張の効果が無かったタスクは，MNLI-m，QQP，QNLI のように fine-tuning のコーパスが大きいものである．これらは fine-tuning のみで十分な転移学習が可能だと考えられる [54]．STS-B については，提案手法は fine-tuning より高い性能を示したが，STILTs を上回ることはなかった．STS-B については MNLI コーパス

の学習で十分であったと考えられる．また文の文法的正しさを判定する CoLA では，すべての手法で fine-tuning より性能が低下しており，MNLI コーパスを用いるアプローチでは CoLA タスクの改善は難しい可能性が示唆された．

7. おわりに

本研究では，意味類似度が高く表層類似度の低い高品質な言い換えの生成モデルを構築し，さらに生成における意味・表層類似度の制御を実現した．提案手法が対照学習および transfer fine-tuning による事前学習済みモデルの性能改善に貢献することを実験的に示し，言い換え生成における類似度制御の重要性を明らかにした．

今後は，テキスト平易化 [60] やスタイル変換 [36] などの言い換え生成タスクへの応用を検討している．高品質な訓練コーパスが少量しかないこれらのタスクにおいて，データ拡張を行うことで性能改善を目指す．さらに提案手法をテキスト平易化やスタイル変換に適するよう訓練するアプローチも考えられる．これにより類似度制御をしながら，タスクに適した言い換えを行うことが可能になると考える．

謝辞

本研究は LINE 株式会社の研究助成を受けたものです．

参考文献

[1] Zhou, J. and Bhat, S.: Paraphrase Generation: A Survey of the State of the Art, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5075–5086 (2021).

[2] Gao, T., Yao, X. and Chen, D.: SimCSE: Simple Contrastive Learning of Sentence Embeddings, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6894–6910 (2021).

[3] Phang, J., Févry, T. and Bowman, S. R.: Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks, *arXiv* (2018).

[4] Yu, A. W., Dohan, D., Luong, M., Zhao, R., Chen, K., Norouzi, M. and Le, Q. V.: QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension, *arXiv* (2018).

[5] Jolly, S., Falke, T., Tirkaz, C. and Sorokin, D.: Data-Efficient Paraphrase Generation to Bootstrap Intent Classification and Slot Labeling for New Features in Task-Oriented Dialog Systems, *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 10–20 (2020).

[6] Gao, S., Zhang, Y., Ou, Z. and Yu, Z.: Paraphrase Augmented Task-Oriented Dialog Generation, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 639–649 (2020).

[7] Effendi, J., Sakti, S., Sudoh, K. and Nakamura, S.: Multi-paraphrase Augmentation to Leverage Neural Caption Translation, *Proceedings of the International Conference on Spoken Language Translation (IWSLT)*, pp. 181–188 (2018).

[8] Okur, Eda and Sahay, Saurav and Nachman, Lama: Data Augmentation with Paraphrase Generation and Entity Extraction for Multimodal Dialogue System, *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 4114–4125 (2022).

[9] Qian, L., Qiu, L., Zhang, W., Jiang, X. and Yu, Y.: Exploring Diverse Expressions for Paraphrase Generation, *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3173–3182 (2019).

[10] Bandel, E., Aharonov, R., Shmueli-Scheuer, M., Shnayderman, I., Slonim, N. and Ein-Dor, L.: Quality Controlled Paraphrase Generation, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 596–609 (2022).

[11] Wieting, J. and Gimpel, K.: ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 451–462 (2018).

[12] Aji, A. F., Tirana Noor Fatyanosa, R. E. P., Arthur, P., Fitriany, S., Qonitah, S., Zulfa, N., Santoso, T. and Data, M.: ParaCotta: Synthetic Multilingual Paraphrase Corpora from the Most Diverse Translation Sample Pair, *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pp. 533–542 (2021).

[13] Johnson, Melvin and Schuster, Mike and Le, Quoc V. and Krikun, Maxim and Wu, Yonghui and Chen, Zhifeng and Thorat, Nikhil and Viégas, Fernanda and Wattenberg, Martin and Corrado, Greg and Hughes, Macduff and Dean, Jeffrey: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, *Transactions of the Association of Computational Linguistics (TACL)*, Vol. 5, pp. 339–351 (2017).

[14] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871–7880 (2020).

[15] Liu, F., Vulić, I., Korhonen, A. and Collier, N.: Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1442–1459 (2021).

[16] Cao, Y. and Wan, X.: DivGAN: Towards Diverse Paraphrase Generation via Diversified Generative Adversarial Network, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2411–2421 (2020).

[17] Park, S., Hwang, S.-w., Chen, F., Choo, J., Ha, J.-W., Kim, S. and Yim, J.: Paraphrase Diversification Using Counterfactual Debiasing, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33, No. 01, pp. 6883–6891 (2019).

[18] Gupta, A., Agarwal, A., Singh, P. and Rai, P.: A Deep Generative Framework for Paraphrase Generation, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 32, No. 1 (2018).

[19] Maddela, Mounica and Alva-Manchego, Fernando and Xu, Wei: Controllable Text Simplification with Explicit Paraphrasing", *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3536–3553 (2021).

[20] Vijayakumar, A., Cogswell, M., Selvaraju, R., Sun, Q., Lee, S., Crandall, D. and Batra, D.: Diverse Beam Search for Improved Description of Complex Scenes, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1 (2018).

[21] Hu, J. E., Rudinger, R., Post, M. and Van Durme, B.: PARABANK: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-Constrained Neural Machine Translation, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33, No. 01, pp. 6521–6528 (2019).

[22] Niu, Tong and Yavuz, Semih and Zhou, Yingbo and Keskar, Nitish Shirish and Wang, Huan and Xiong, Caiming: Unsupervised Paraphrasing with Pretrained Language Models, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5136–5150 (2021).

[23] Zeng, D., Zhang, H., Xiang, L., Wang, J. and Ji, G.: User-Oriented Paraphrase Generation With Keywords Controlled Network, *IEEE Access*, Vol. 7, pp. 80542–80551 (2019).

[24] Iyyer, M., Wieting, J., Gimpel, K. and Zettlemoyer, L.: Adversarial Example Generation with Syntactically Controlled Paraphrase Networks, *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1875–1885 (2018).

[25] Hosking, T. and Lapata, M.: Factorising Meaning and Form for Intent-Preserving Paraphrasing, *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 1405–1418 (2021).

[26] Chen, Mingda and Tang, Qingming and Wiseman, Sam and Gimpel, Kevin: Controllable Paraphrase Generation with a Syntactic Exemplar, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5972–5984 (2019).

[27] Bao, Yu and Zhou, Hao and Huang, Shujian and Li, Lei and Mou, Lili and Vechtomova, Olga and Dai, Xin-yu and Chen, Jiajun: Generating Sentences from Disentangled Syntactic and Semantic Spaces, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6008–6019 (2019).

[28] Goyal, T. and Durrett, G.: Neural Syntactic Preordering for Controlled Paraphrase Generation, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 238–252 (2020).

[29] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems*, Vol. 26 (2013).

[30] Li, J., Galley, M., Brockett, C., Gao, J. and Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models, *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 110–119 (2016).

[31] Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B. and Callison-Burch, C.: PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification, *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 425–430 (2015).

[32] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318 (2002).

[33] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2019).

[34] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. and Specia, L.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 1–14 (2017).

[35] Fan, Angela and Lewis, Mike and Dauphin, Yann: Hierarchical Neural Story Generation, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 889–898 (2018).

[36] Kajiwara, T., Miura, B. and Arase, Y.: Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34, No. 05, pp. 8042–8049 (2020).

[37] Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M. and Edunov, S.: Facebook FAIR's WMT19 News Translation Task Submission, *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, pp. 314–319 (2019).

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is All you Need., *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008 (2017).

[39] Schwenk, H., Chaudhary, V., Sun, S., Gong, H. and Guzmán, F.: WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia, *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 1351–1361 (2021).

[40] Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V. and Zampieri, M.: Findings of the 2021 Conference on Machine Translation (WMT21), *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, pp. 1–88 (2021).

[41] He, P., Gao, J. and Chen, W.: DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, *arXiv* (2021).

[42] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186 (2019).

[43] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov,

V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv* (2019).

[44] Ilya Loshchilov and Frank Hutter: Decoupled Weight Decay Regularization, *Proceedings of the International Conference on Learning Representations (ICLR)* (2019).

[45] See, Abigail and Liu, Peter J. and Manning, Christopher D.: Get To The Point: Summarization with Pointer-Generator Networks, *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1073–1083 (2017).

[46] Williams, A., Nangia, N. and Bowman, S. R.: A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1112–1122 (2018).

[47] Bowman, S. R., Angeli, G., Potts, C. and Manning, C. D.: A large annotated corpus for learning natural language inference, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 632–642 (2015).

[48] Agirre, E., Cer, D., Diab, M. and Gonzalez-Agirre, A.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 385–393 (2012).

[49] Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. and Guo, W.: *SEM 2013 shared task: Semantic Textual Similarity, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pp. 32–43 (2013).

[50] Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G. and Wiebe, J.: SemEval-2014 Task 10: Multilingual Semantic Textual Similarity, *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 81–91 (2014).

[51] Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L. and Wiebe, J.: SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability, *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 252–263 (2015).

[52] Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G. and Wiebe, J.: SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation, *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 497–511 (2016).

[53] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R. and Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models, *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 216–223 (2014).

[54] Arase, Yuki and Tsujii, Jun'ichi: Transfer Fine-Tuning: A BERT Case Study, *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5393–5404 (2019).

[55] Warstadt, A., Singh, A. and Bowman, S. R.: Neural Network Acceptability Judgments, *arXiv* (2018).

[56] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. and Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642 (2013).

[57] Dolan, W. B. and Brockett, C.: Automatically constructing a corpus of sentential paraphrases, *Proceedings of the International Workshop on Paraphrasing (IWP)* (2005).

[58] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392 (2016).

[59] Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D. and Magnini, B.: The Fifth PASCAL Recognizing Textual Entailment Challenge, *Proceedings of the Text Analysis Conference (TAC)* (2009).

[60] Sun, R., Yang, Z. and Wan, X.: Exploiting Summarization Data to Help Text Simplification, *arXiv* (2023).