複単語表現の意味計算を要する文類似度評価データの構築

神戸 隆志

東北大学 工学部 電気情報物理工学科

1. はじめに

2 つの文の類似度を計算することは，機械翻訳の自動評価尺度 [Papineni 01] や，類似したテキストデータの検索に基づく retrieval-based の手法 [Guu 20] など様々な手法に用いられる，自然言語処理における重要な基盤的技術である．そうした背景から，与えられた 2 つの文の類似度を推定するタスクは，言い換え識別 (Paraphrase Identification; PI), 意味的類似度 (Semantic Textual Similarity; STS), 機械翻訳の自動評価尺度など複数の異なる文脈で取り組まれており，それぞれ多くの評価用データセットが作成されている [Zhang 19, Agirre 12, Mathur 20]．

複単語表現 (Multi-Word Expression; MWE)[Sag 02] とは複数の単語からなる句であり，句を構成する単語から句全体の意味を計算することが難しく，これまでの文類似度計算手法も複単語表現の意味を適切に計算できないことが予想される．特に文類似度計算手法にも広く用いられる単語埋め込みは，複単語表現の意味を理解する能力が低いことが知られている [Shwartz 19]．より多様な言語構造に頑健な文類似度評価手法の開発を促進するため，複単語表現という言語現象を考慮したデータセットが必要である．

そこで本研究では，複単語表現の意味計算を要する文類似度評価データを構築する．主たる指針としては，文類似度計算手法の複単語表現に対する頑健性を評価するために，必ず文ペアの片方のみが興味の対象である複単語表現を含み，もう一方の文はその複単語表現の言い換えもしくは無関係な句を含むように設計する．評価する文類似度計算手法が複単語表現に対する知識を有していなければ，これらの事例間の識別に失敗すると考えられる．具体的には，逆翻訳と制約付き文生成に基づく手法を用いて意味が類似した文ペアを，BERT によるマスクトークンの予測に基づく手法を用いて意味が類似しないペアを生成することを試みる．実験では，複単語表現の一種である句動詞を含む 593 文からなるコーパスに対して提案した 2 つの手法を適用し，さらに生成した文ペアに対してクラウドソーシングを用いて文法的・意味的に問題がないかの確認と類似度スコアの付与を行った．アノテーションの結果，逆翻訳と制約付き文生成に基づく手法は類似した文ペアを生成しやすい傾向があり，BERT によるマスクトークンの予測に基づく手法は類似した文ペアも生成するが，十分な数の類似しない文ペアを生成することが確認できた．

こうして作られた文類似度評価データは，様々な文類似度計算手法の複単語表現に対する頑健性を評価でき，システムによる複単語表現の意味理解を促すという点において自然言語処理分野を前に進めることに繋がると考える．

2. 関連研究

本研究では、類似度が付与された文ペアを生成するための 2 つの手法を提案する．一つは逆翻訳 (2.1 節) と制約付き文生成 (2.2 節) に基づく手法であり，もう一方はマスク言語モデルの一種である BERT(2.3 節) を用いた手法である．以下詳述するが，前者の手法では主に類似した文ペアが生成されやすく，一方で後者の手法は主に類似しない文ペアが生成される傾向があると考え，これら 2 つの手法により生成される文ペアを組み合わせることで，正例と負例のバランスが取れたデータセットが構築できると考えられる．

2.1 逆翻訳に基づく言い換え文生成

逆翻訳 [Sennrich 16] は，文をある特定の言語の文に翻訳し再度逆方向に翻訳することで，元の文の言い換え文を生成する手法である．逆翻訳によって生成された文は，元の文の意味を保存したまま別の表現に言い換えられていることが多い．この特徴から，元の文と逆翻訳後の文をペアにすることで言い換え文のペアを生成する手法が広く用いられている [Zhang 19, Hu 19b]．本研究においても，逆翻訳に基づく手法を用いて類似した文を生成する手法を提案する．

2.2 制約付き文生成

制約付き文生成は，出力する文に対して指定した単語やフレーズを必ず出現させる，または出現させない様に制約を付与する手法である．[Hu 19b] らは逆翻訳と制約付き文生成を組み合わせた言い換え文生成手法を提案している．入力文を翻訳した中間言語の文から元の言語の文に翻訳する際に，入力文中の単語をランダムに選択しそれらが出力文中に出現しない様に制限してやることで，多様な言い換え文を生成するという手法である．本研究でも [Hu 19b] らに倣い，逆翻訳と制約付き文生成を組み合わせることによって，複単語表現の箇所が必ず別の表現に言い換えられるような文ペアの生成手法を提案する．

2.3 BERT

BERT[Devlin 19] は自然言語処理分野において広く用られている言語モデルの一種である．BERT は 2 種類の教師無し学習によって事前学習され，事前学習によって得られたモデルに対して出力層を追加し，fine-tuning を行うことで，様々なタスクにおいて良い性能を示すことが知られている．BERT の事前学習の 1 つは Masked Language Modeling と呼ばれるタスクであり，文のサブワード列のうちの一部を “[MASK]” トークンに置き換えてモデルに入力し，元の単語を予測させるというタスクとなっている．本研究ではこの Masked Language Modeling を用いて文生成を行う手法を提案する．

3. データセット構築手法

複単語表現の意味計算を要する文類似度評価データの構築のためには，片方の文中の複単語表現の言い換えを含む類似した文ペアと，無関係の句で置き換えた類似しない文ペアを生成する必要がある．そこで我々は複単語表現を含む 1 つの文を出発点とし，そこから上記の 2 種類の文ペアを生成するための手法を提案する．複単語表現の言い換えを含む類似した文ペアを生成するための手法は逆翻訳と制約付き文生成を組み合わ

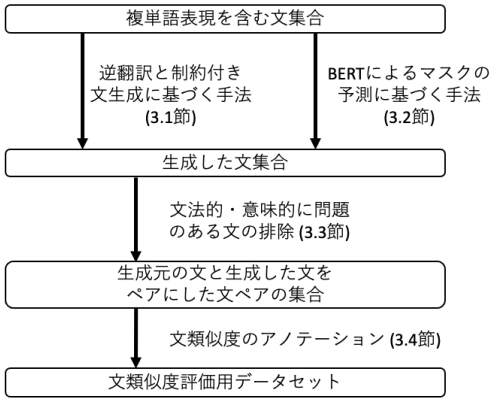

図 1: 文類似度評価データセット構築のフローチャート

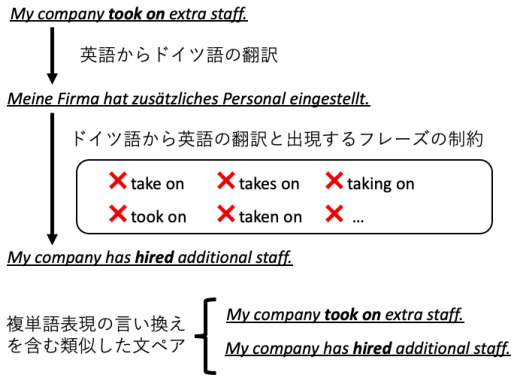

図 2: 逆翻訳と制約付き文生成に基づく類似した文の生成．制約により “take on” とその活用形を出力に含めない様に制限する．

せた手法であり，複単語表現を無関係の句で置き換えた類似しない文ペアを生成するための手法は BERT によるマスクトークンの予測に基づく手法である．さらに生成した文の内，文法的・意味的に問題のある文をクラウドソーシングを用いて排除する．最後に元の文と生成した文をペアにし，文ペアの類似度をクラウドソーシングを用いて判定し，文類似度評価データとする．図 1 はデータセット構築手法のフローチャートを表す．

3.1 逆翻訳と制約付き文生成に基づく文生成

逆翻訳と制約付き文生成を用いた，複単語表現の言い換えを含む類似した文の生成手法を提案する．基本的には逆翻訳を用いて英語文のペアの生成を行うが，単純に逆翻訳を行うだけでは文中の複単語表現が別の表現に言い換えられることは保証されない．そこで制約付き文生成を用いて，中間言語から英語に翻訳する際に複単語表現の出現を制限してやることで，複単語表現が別の単語やフレーズに言い換えられることを期待する．図 2 は逆翻訳と制約付き文生成に基づく手法の概観を表す．

逆翻訳に用いるモデルは [Ng 19] らの学習済みモデルを使用し，制約付き文生成は [Post 18, Hu 19a] らの手法を用いる [*1]．また逆翻訳における中間言語は，[Zhang 19] らに倣い入力文が英語である場合における翻訳の質の高さからドイツ語を選択する．さらに，文中の複単語表現をそのまま制限するだけでは，“take on” という複単語表現に対し “took on” の様な活用形の変換しか行わない様な文生成を行う可能性があるため，複単語表現中の単語の活用形も含めて出現しない様に制約付き文生成を行う [*2]．

3.2 BERT によるマスクの予測に基づく文生成

BERT の事前学習である Masked LM に基づいた文生成手法を提案する．まず元となる文中の複単語表現を 1 つの “[MASK]” トークンに置き換える．置き換えた文を事前学習済みの BERT(bert-base-cased[*3]) に入力し，マスクした箇所に何の単語が入るかを予測させ，その単語で複単語表現を置き換えることで文を生成する．これにより，元の文中の複単語表現は必ずある 1 つの単語に置き換えられる．BERT による単語の予測は周辺の文脈を考慮していることから，生成した文はある程度文法的・意味的な流暢さが保証されると考えられるが，元の複単語表現を考慮した置き換えではないため，元の文とは類似しない文が多く生成されると考えられる．実験では，BERT が予測した上位 5 件の単語を採用し，1 つの複単語表現を含む文から 5 つの文を生成する．

3.3 文法的・意味的に問題のある文の排除

上記の 2 つの手法で生成した文はどちらも翻訳モデルや言語モデルによって自動で生成したものであるため，文法的または意味的に破綻した文になっている可能性がある．そこでクラウドソーシングを用いて文のフィルタリングを行う．クラウドソーシングは Amazon Mechanical Turk 上で行い，各文に対して 5 人のワーカーを割り当て，「文法的な誤りがないか」，「ネイティブスピーカーにとって流暢で自然な文か」の 2 点を満たすかどうかについてアノテーションを依頼する．アノテーションの質を保証するため，MACE[Hovy 13] を用いて信頼度の低い下位 30%ワーカーの回答を除外し，その上で，残りのワーカー全員が上記 2 つの要件を満たすとアノテーションした文のみを採用する．

3.4 文類似度のアノテーション

3.3 節のアノテーションで意味的・文法的に問題がないと判定された文をその文の生成元となった文とペアにし，それらの文ペアの類似度のアノテーションを依頼する．文類似度の基準は [Agirre 12] らの用いた定義を参考にし，2 つの文がどの程度類似しているかを表す 0 から 5 の整数値とする．(値が大きいほど 2 つの文は類似している．) アノテーションは 3.3 節と同様に Amazon Mechanical Turk 上で行い，各文ペアに対して 5 人のワーカーを割り当て，各ワーカーの判定した文類似度の平均値を最終的な文類似度として採用する．

4. 実験結果

4.1 使用したデータ

複単語表現の箇所の情報が付与された文のセットとして，[Tu 12] のデータセットを用いた．このデータセット中の文は動詞とそれに連続する前置詞句を必ず含み，その組み合わせが複単語表現の一種である句動詞として用いられているかどうかについてアノテーションされている．データセットは “take on”,“get over”,“make out” といった頻出する 23 種類の句動詞となる候補を含む 1,348 文から成り，その候補が句動詞であるとアノテーションされている文は 878 文存在した．また本

*1 実装はすべて fairseq[Ott 19] を用いて行う．

*2 活用形の取得には以下を利用した：https://github.com/bjascob/pyInflect

*3 https://huggingface.co/bert-base-cased

	逆翻訳と制約付き文生成	BERT によるマスクの予測
生成した文ペア	593	2,965
文法的・意味的に問題のある文の排除		
通過した文	239	854
排除された文	354	2,111
Krippendorff's alpha	0.390	0.388
文類似度		
0 以上 1 未満	3	14
1 以上 2 未満	5	201
2 以上 3 未満	44	267
3 以上 4 未満	93	244
4 以上 5 以下	94	128

表 1: 文法的・意味的に問題のある文の排除と，2 つの文類似度のアノテーション結果

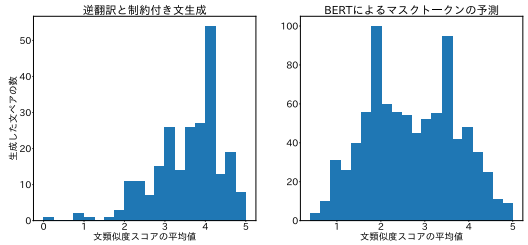

図 3: 文類似度の平均値の分布．逆翻訳と制約付き文生成に基づく手法 (左) と BERT によるマスクトークンの予測に基づく手法 (右).

研究の目的は複単語表現の文類似度に対する影響のみを調査できるようなデータセットの構築であるため，複単語表現以外の多様な現象が混在し得る文長の長い文と短い文を排除した．採用する文長の基準は 5 単語以上かつ 30 単語以下とし，最終的に使用したデータは 593 文となった．また，今回の実験では複単語表現の一種である句動詞を含む文を用いたが提案手法は句動詞に限定されるものではなく，一般的な複単語表現に適用可能である．

4.2 アノテーション結果

表 1 は，2 つの文生成手法に対する 2 種類のアノテーション結果を表す．逆翻訳と制約付き文生成に基づく手法では，生成した 593 文の内 239 文が文法的・意味的に問題が無いと判定され，BERT によるマスクトークンの予測に基づく手法では，生成した 2,965 文の内 854 文が文法的・意味的に問題が無いと判定された．ワーカーの回答の一致度の指標として，Krippendorff's α を用いたが，それぞれ手法に対して 0.390, 0.388 と一致度は低くなっていた．文類似度の平均値の分布は図 3 のようになっており，逆翻訳と制約付き文生成に基づく手法における文類似度の平均値は 3.55 であり，BERT によるマスクトークンの予測に基づく手法では 2.70 となった．このことから逆翻訳と制約付き文生成に基づく手法は，BERT によるマスクトークンの予測に基づく手法に比べ，元の文と類似した文を生成しやすい傾向があることが確認できた．

4.3 生成した文ペアの観察

表 2 は 2 つの手法により生成された文ペアの例を表す．表 2 中の逆翻訳と制約付き文生成に基づく手法では，“got out” という複単語表現が “left” に，“gave in” という複単語表現が “succumbed” に言い換えられており，文ペアの意味はおおよそ保存されていることが確認できた．表 2 中の BERT によるマスクトークンの予測に基づく手法では，“get out” という複単語表現が “stay” という逆の意味の単語に置き換えられている例のように，文ペアの意味が保存されていないような例が多く確認できた．また，“take on” から “hire” の言い換えの例のように類似した文ペアも多く生成されていたが，これは周りの文脈から複単語表現の箇所の意味を予測しやすい場合には，マスクトークンを複単語表現に類似した意味の単語で置き換えるためであると考えられる．

5. 結論

本研究では，複単語表現の言い換えを含む類似した文ペアと，無関係の句で置き換えた類似しない文ペアを生成するための 2 つの文生成手法を提案した．前者の手法では類似した文を生成する傾向があり，後者の手法では前者の手法に比べ類似していない文を生成する傾向にあることがわかった．従って，2 つの手法で生成した文ペアを組み合わせることで，正例と負例のバランスの取れたデータセットを構築できるということが確認できた．今回の実験では，頻出する句動詞をターゲットとして 2 つの手法の検証を行ったが，これらの手法は句動詞だけではなく一般的な複単語表現に対しても適用可能である．今後は様々な種類の複単語表現をターゲットとし，あらゆる複単語表現に対する頑健性を評価できるようなデータセット構築を目標とする．

6. 今後の展望

論文中で提案した文類似度評価データ構築手法に関して，改良したい点として以下の 2 点を考えている．

1 点目は逆翻訳と制約付き文生成に基づく手法に関する問題点である．逆翻訳と制約付き文生成に基づく手法は，複単語表現の言い換えを含む類似した文ペアを生成するように設計しているが，これは逆翻訳に用いる翻訳モデルが複単語表現の意味を翻訳の際に保存できることを前提としている．もし翻訳モデルがある複単語表現の意味を保存できなければ，その複単語表現の言い換えを含む類似した文を生成できないという問題点がある．従って，現状の翻訳モデルが意味を捉えられないような複単語表現の言い換えを含む類似した文ペアの生成手法，例えばクラウドソーシングを用いてワーカーに直接言い換え文を生成してもらうといったような手法を検討する必要がある．

2 点目は提案した 2 つの手法により生成される文ペアの語彙の重複率の違いである．逆翻訳と制約付き文生成に基づく手法は，複単語表現の言い換え以外にも他の箇所の単語やフレーズの言い換えを行う可能性が高い．対して BERT のマスクトークンの予測に基づく手法では，複単語表現の箇所以外は確実に変化しない．つまり逆翻訳と制約付き文生成によって生成された文ペアは，BERT による手法によって生成された文ペアに比べ語彙の重複率が小さいということになる．また，逆翻訳と制約付き文生成に基づく手法は類似した文ペアを生成しやすく，BERT による手法は類似しない文ペアを生成しやすいという傾向があるため，2 つの手法を用いて作成したデータセットを何らかの手法を用いて解かせる上で，語彙の重複率が文類似度判定の手がかりとなってしまう．従って，2 つの手法によ

文生成手法	元文と生成された文	文類似度
逆翻訳と 制約付き文生成	I think that's why Richard got out, because he's terrified of rats. I think that's why Richard left because he's afraid of rats.	4.2
逆翻訳と 制約付き文生成	How many Englishmen gave in to their emotions like that? How many Englishmen succumbed to emotions like that?	5.0
BERT による マスクの予測	I want to get out there. I want to stay there.	1.6
BERT による マスクの予測	The Stoke Row Garage is having to take on extra staff to deal with servicing demand. The Stoke Row Garage is having to hire extra staff to deal with servicing demand.	4.0

表 2: 2 つの手法により生成された文ペアの例

り生成される文ペアの語彙の重複率を統一するような工夫が必要である．例えば逆翻訳の際のサンプリング数を増やし，複数の文ペアの中からできるだけ語彙の重複率が大きい文ペアを選ぶことで，BERT による手法で生成される文ペアの語彙の重複率との差を小さくするといったような手法が考えられる．

参考文献

[Agirre 12] Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity, Technical report (2012)

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, Stroudsburg, PA, USA (2019), Association for Computational Linguistics

[Guu 20] Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W.: REALM: Retrieval-Augmented Language Model Pre-Training (2020)

[Hovy 13] Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E.: Learning Whom to Trust with MACE, Technical report (2013)

[Hu 19a] Hu, J. E., Khayrallah, H., Culkin, R., Xia, P., Chen, T., Post, M., and Van Durme, B.: Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting, in *Proceedings of the 2019 Conference of the North*, Vol. 1, pp. 839–850, Stroudsburg, PA, USA (2019), Association for Computational Linguistics

[Hu 19b] Hu, J. E., Singh, A., Holzenberger, N., Post, M., and Van Durme, B.: Large-Scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering, in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 44–54, Stroudsburg, PA, USA (2019), Association for Computational Linguistics

[Mathur 20] Mathur, N., Tian, J., Wei, Z., Freitag, M., Ma, Q., and Bojar, O.: Results of the WMT20 Metrics Shared Task, Technical report (2020)

[Ng 19] Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S.: Facebook FAIR ' s WMT19 News Translation Task Submission, in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 第 2 巻, pp. 314–319, Stroudsburg, PA, USA (2019), Association for Computational Linguistics

[Ott 19] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M.: fairseq: A Fast, Extensible Toolkit for Sequence Modeling, in *Proceedings of the 2019 Conference of the North*, pp. 48–53, Stroudsburg, PA, USA (2019), Association for Computational Linguistics

[Papineni 01] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, *ACL*, pp. 311–318 (2001)

[Post 18] Post, M. and Vilar, D.: Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation, in *Proceedings of the 2018 Conference of the North American Chapter of*
} the Association for Computational Linguistics: Human Language
} Technologies, Volume 1 (Long Papers), Vol. 1, pp. 1314–1324, Stroudsburg, PA, USA (2018), Association for Computational Linguistics

[Sag 02] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D.: Multiword expressions: A pain in the neck for NLP, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 2276, pp. 1–15, Springer Verlag (2002)*

[Sennrich 16] Sennrich, R., Haddow, B., and Birch, A.: Improving Neural Machine Translation Models with Monolingual Data, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96 (2016)*

[Shwartz 19] Shwartz, V. and Dagan, I.: Still a Pain in the neck: Evaluating text representations on lexical composition (2019)

[Tu 12] Tu, Y. and Roth, D.: Sorting out the Most Confusing English Phrasal Verbs, Technical report (2012)

[Zhang 19] Zhang, Y., Baldridge, J., and He, L.: PAWS: Paraphrase Adversaries from Word Scrambling, in Proceedings of the 2019 Conference of the North*, pp. 1298–1308, Stroudsburg, PA, USA (2019), Association for Computational Linguistics*