BERTScore とキーワード採用率を用いた 語義タグ付き用例文生成手法

長友日雅[1] 佐々木稔[2]
[1]茨城大学大学院理工学研究科情報工学専攻 [2]茨城大学工学部情報工学科
{23nm732f,minoru.sasaki.01}@vc.ibaraki.ac.jp

概要

近年の語義曖昧性解消タスクには、語義タグ付きコーパスを用いて機械学習を行ったモデルが問題解決に用いられている。しかし、語義タグ付きコーパスは基本的に人手でしか作成することができないため、データ数が限られている。そこで本研究では、語義タグ付きコーパスからキーワードを抽出し、それをもとにテキスト生成を行うことで、語義タグ付きコーパスのデータ数を機械的に増やす手法を提案し、生成したテキストの有用性を分析する。キーワードからテキスト生成を行うモデルには keytotext があるが、文章が流暢でない、文中にキーワードが含まれないなどの問題を抱えており、自然なテキストを生成することがほとんどない。そのため、本研究では、モデルの学習に使用するデータセットや損失関数を変更することで、モデルの性能がどのように変化するのかを定量的・定性的に評価する。検証の結果、学習に利用するデータセットを洗練し、検証時の損失関数に BertScore やキーワードの採用率といった、正解テキストを直接参照しない評価指標を用いて学習を行うことにより、既存モデルよりも高いスコアを持つモデルが得られることが分かった。

1 はじめに

文中の多義語がどの語義として使用されているのかを推定する語義曖昧性解消(WSD)では、語義タグ付きコーパスを用いて機械学習を行ったモデルが問題解決に用いられている。語義タグ付きコーパスとは、文章中の単語に語義を付与したもので、基本的に人手でしか作成することができないため、データ数が限られている。

そこで本研究では、語義タグ付きコーパスから語義タグのついた単語を含むキーワードを抽出し、それをもとにテキスト生成を行うことで、語義タグ付きコーパスのデータ数を機械的に増やす手法を提案し、生成したテキストの有用性を分析する。

キーワードからテキスト生成を行うモデルには keytotext(k2t)[4]があるが、文章が不自然になっている、キーワードの品詞を間違えている、文中にキーワードを含んでいないなど、多くの問題を抱えており、自然なテキストを生成することがほとんどない。そのため、本研究では、モデルの学習に使用するデータセットや損失関数を変更することで、モデルの性能がどのように変化するのかを定量的・定性的に評価し、この手法の有効性を検証する。

具体的には、k2t の学習に用いるデータセットや損失関数を変更して新たに k2t モデルを学習し、WordNet の例文を用いて、テキストの生成を行う。その後、生成したテキストを定量的・定性的に評価し、データセットや損失関数の変更により、モデルの性能がどのように変化したのか検証する。

2 関連研究・関連手法

2.1 keytotext

keytotext(k2t)は、キーワードをスペース区切りで入力すると、そのキーワードを含むテキストを出力するモデルである。

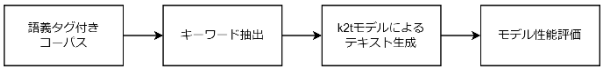
例えば、k2t モデルに”Japan,capital,Tokyo”のようにキーワードを入力すると、”The capital of Japan is Tokyo”のように、キーワードをすべて含むテキストが出力される。しかし、完璧に例文を生成できるのはごく一部のキーワードのみで、ほとんどの場合に、テキストが不自然になっている、キーワードの品詞を間違えている、文中にキーワードを含んでいないなどの問題が発生し、自然なテキストを生成することがほとんどない。そのため、本研究では学習に使用するデータセットや損失関数を変更することで、k2t モデルを改良し、より自然なテキストを生成できるモデルの作成を目指す。

k2t モデルの学習では、事前学習済みの T5 モデルに対する転移学習を行っている。学習データには、Hugging Face が提供するデータセット(common_gen, ag_news, cc_news)、損失関数にはクロスエントロピー損失を使用している。

3 提案手法

本研究では、語義タグ付きコーパスからキーワードの抽出を行い、k2t モデルを用いて抽出したキーワードをすべて含むテキストの生成を行う。テキスト中に含まれるターゲット単語(語義タグの付いた単語)の語義は抽出したキーワード同士の共起関係から同じ語義であると考えられる。以下に提案する手法の概要を示す。

図 1:提案手法の概要



3.1 キーワードの抽出

語義タグ付きコーパスからのキーワード抽出には、BERT を用いたキーフレーズ抽出パッケージである KeyBERT を使用する。KeyBERT はテキストを入力すると、指定の個数までキーワードを抽出することができる。そこで語義タグ付きテキストからすべての語義についてターゲット単語を含むキーワードを 3 つ抽出し、テキスト生成に使用する。キーワードを 3 つ抽出できない短いテキストや、ターゲット単語がキーワードに含まれないテキストは生成不可として除外する。

3.2 k2t モデルの学習

k2t モデルの学習は、T5-base に対する転移学習で行う。既存の k2t モデルの学習には Huggingface が提供する、common_gen, ag_news,cc_news の 3 つのデータセットを利用している[4]。common_gen はテキストとテキストのキーワード 3 つが一組になったデータセットで、キーワードを入力データ、テキストを正解テキストとして学習を行う。ag_news と cc_news は、ニュース記事の本文をデータセットとしたものであるため、KeyBERT を用いてキーワードを抽出し、入力データと正解テキストを取得している。しかし、ag_news と cc_news から得られるテキストは文章が長く、記号が多く含まれるため、学習に適さないものであった。そこで本研究では、学習に使用するデータセットを common_gen のみに限定している。

また、既存の k2t では学習時・検証時の損失関数にクロスエントロピー損失用いているが、k2t では正解テキストを再現することではなく、キーワードを含む新たなテキストを生成することを目的としているため、正解テキストを直接参照するクロスエントロピー損失は k2t モデルの学習・検証には適していないと考えられる。そこで本研究では、学習時や検証時に使用する損失関数を変更することでモデルの性能がどのように変化するかを調査する。

モデルの学習には、既存の k2t-trainer の損失の計算やデータセットの読み込みを行う部分を改変して使用した。batch_size は 2、early_stopping は設定せずに 20 エポック学習し、最も検証時の損失が少ないモデルを選択し評価に用いる。

3.3 k2t モデルの損失関数

本研究では以下の 2 つの損失関数を用いて学習を行う。

1) クロスエントロピー損失

k2t で使用されている損失関数であるため、ベースラインとして使用する。

2) キーワードのクロスエントロピー損失

キーワードが正解テキストの何番目に使用されているのかを確認し、キーワードが使用されている部分のクロスエントロピー損失を学習時の損失関数に加算する。キーワードが使用されている部分のクロスエントロピー損失を二重に計算することで、キーワードが使用されているかどうかで損失が大きく変わるようになり、キーワードをすべて含むテキストを生成する事例を増やすことを目的としている。

検証には以下の 3 つの損失関数を使用し、それぞれ損失が最も少ないモデルを評価に使用する。

1) クロスエントロピー損失

学習時と同様に、k2t で使用されている損失関数であるため、ベースラインとして使用する。

2) BERTScore による類似度

BERTScore を用いて生成したテキストと正解テキスト間の類似度を計算し、検証時の損失に用いる。正解テキストと生成したテキストが完全に一致していなくても、文意が近ければ高いスコアが得られるため、”car”と”vehicle”のような類義語の言い換えに対応することを目的としている。

3)　キーワードの採用率

　生成に使用したキーワードがどれだけテキストに採用されているかの割合を計算し、検証時の損失に用いる。キーワードの採用率を直接検証時の損失に使用することで、キーワードをすべて含むテキストを生成する事例を増やすことを目的としている。

3.4 定量的評価方法

　テキスト生成の定性的評価には、BLEU や ROUGE のような生成されたテキストと正解テキスト間でどの程度相違があるかを評価する指標が広く用いられている。しかし、本研究では抽出したキーワードを含む正解テキストとは別のテキストを生成することが目標であるため、これらの評価指標は、本実験の定量的評価には適さない。そこで、本実験では以下の 4 つをテキスト生成に失敗する要因として定め、生成したテキストのうち何文が失敗要因に該当するかを計測する。

1) テキストにキーワードが含まれていない。

　本研究では、キーワードをすべて含むテキスト生成を行うことを目的としているため、キーワードが含まれていないテキストは失敗とする。

2) 同じ単語がテキストに 2 回以上登場する。

　既存の k2t モデルでは、同じ単語を繰り返し並べただけのテキストが見られたため、単語を繰り返し使用していることも失敗要因の一つとする。ただし、自然なテキストの生成に成功したテキストも単語を繰り返し使用する場合があるため、正解テキストと比較してどの程度同じ単語を繰り返したテキストが多いのかを評価対象とする。

3) be 動詞が含まれる。

　既存の k2t モデルでは、キーワードの品詞を問わず be 動詞で繋げただけのテキストが見られたため、be 動詞を含むことも失敗要因の一つとする。ただし、2)と同様に正解テキストと比較してどの程度 be 動詞を含むテキストが多いのかを評価対象とする。

4) 固有名詞が含まれる

　既存の k2t モデルでは、固有名詞でないキーワードを””で囲み固有名詞として使用するテキストが見られたため、””を含むことも失敗要因の一つとする。

以上に該当するテキストが少ないほどモデルの性能が高いと言える。

3.5 定性的評価方法

3.4 節で紹介した定量的な評価のみでは、WSD の学習データに利用できるテキストが生成できているか評価できない。

そこで本実験では、各モデルが生成したテキストからランダムに 15 文を抽出し、以下の二つの項目について、定性的な評価を行う。

1) テキストの流暢さ

　生成されたテキストの流暢さを 5 段階で評価する。

2) ターゲット単語の語義

　語義タグが付与されたキーワードの語義が元の文と同じ語義であるかを 5 段階で評価する。

以上の平均スコアが高いほど、モデルの性能が高いと言える。

4 実験

本研究では、学習時の損失関数を以下のように変更し、学習を行う。

1) クロスエントロピー損失

2) キーワードのクロスエントロピー損失

また、検証時に以下の損失関数を用いてモデル性能を検証し、検証時の損失が最も小さいモデルを選択する。

1) クロスエントロピー損失

2) BERTScore による類似度

3) キーワードの採用率

5 実験結果

定量的評価結果を表 1、定性的評価結果を表 2 に示す。

k2t-base はキーワードの採用率が高いが、内容は単なるキーワードの羅列であることが多く、例文には適さないため、定性的な評価ではスコアが低くなっている。本研究で作成したモデルは、そのような傾向が無くなっていたため、common_gen が学習に適したデータセットであると考えられる。

次に、損失関数を比較する。キーワードのクロスエントロピー損失は、キーワードの採用率を向上させることを目的としていたが、キーワードの採用率の向上にはつながらなかった。これは、正解テキストでキーワードが使用されている部分にキーワードが生成されるとは限らないためだと考えられる。また、定性的評価においてもスコアが低下しているため、モデルの学習には適していないと考えられる。

設定		設定			
学習時の損失関数	検証時の損失関数	1)	2)	3)	4)
クロスエントロピー損失	クロスエントロピー損失	7929	7690	5540	0
クロスエントロピー損失	BERTScore による類似度	9014	8912	6263	0
クロスエントロピー損失	キーワードの採用率	8738	7443	6381	9
キーワードの クロスエントロピー損失	クロスエントロピー損失	7340	4170	3402	0
キーワードの クロスエントロピー損失	BERTScore による類似度	9619	6416	4892	0
キーワードの クロスエントロピー損失	キーワードの採用率	9127	5830	4307	0
k2t-base(ベースライン)		4171	14101	14536	3301
WordNet(元のテキスト)		-	3172	3705	-

表 1: 定量的評価結果 1)テキストにキーワードが含まれていない 2) 同じ単語がテキストに 2 回以上登場する 3) be 動詞が含まれる 4) 固有名詞が含まれる

設定		定性的評価結果	
学習時の損失関数	検証時の損失関数	1)	2)
クロスエントロピー損失	クロスエントロピー損失	3.50	2.92
クロスエントロピー損失	BERTScore による類似度	3.42	3.06
クロスエントロピー損失	キーワードの採用率	3.54	2.82
キーワードの クロスエントロピー損失	クロスエントロピー損失	3.32	2.98
キーワードの クロスエントロピー損失	BERTScore による類似度	3.24	2.88
キーワードの クロスエントロピー損失	キーワードの採用率	3.22	2.88
k2t-base(ベースライン)		1.06	1.20

表 2: 定性的評価結果 1) テキストの流暢さ 2) キーワード単語の語義

最後に、検証時の損失関数を比較する。
キーワードをすべて含むテキストの生成ではクロスエントロピー損失を用いたモデルが高いスコアを獲得し、キーワードの採用率を直接検証に用いたものよりも高いスコアを獲得している。
テキストの流暢さでは BERTScore を用いたモデルが高いスコアを獲得していた。BRETScore を用いたモデルはキーワード自体を採用しなくても、キーワードに類似した単語を使用すれば高いスコアを獲得できるため、キーワード採用率が低くなる代わりに定性的評価のスコアが高くなったと考えられる。
キーワードが採用されなかったテキストを機械的に削除するのであれば、学習にクロスエントロピー損失、検証に BERTScore を用いたモデルが例文生成に最も適していると考えられる。

6 おわりに

本研究では、キーワードからテキスト生成を行うモデルである keytotext(k2t)を改良し、WSD モデルの学習に適したテキストを生成することを目的として実験を行った。その結果、学習に利用するデータセットを common_gen のみに絞り、検証時の損失関数に BertScore のように、正解テキストを直接参照しない評価指標を用いて学習を行うことにより、既存モデルよりも高いスコアを持つモデルが得られることが分かった。今後の展望としては、キーワード抽出時に単語間の共起関係を考慮した指標を追加することや、モデルの性能評価を実際に WSD モデルの学習に使用することで評価を行うことを考えている。

参考文献

[1]Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. *Just "OneSeC" for producing multilingual sense-annotated data.* In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 699–709.

[2]Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi and Xiang Ren. 2019. *CommonGen: A constrained text generation challenge for generative commonsense reasoning*. arXiv preprint arXiv:1911.03705.

[3]Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. *Exemplification Modeling: Can You Give Me an Example, Please?* In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), pages 3779-3785.

[4]gagan3012. 2021. *keytotext.* https://github.com/gagan3012/keytotext

Navigli, R. (2009). *Word sense disambiguation: A survey, ACM Computing Surveys*, vol. 41, no. 2, pp. 10:1-10:69.

[5]Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. 2020. *Exploring the limits of transfer learning with a unified text-to-text transformer*. The Journal of Machine Learning Research, 21(1), 5485-5551.

[6]Yarowsky, D. (1995). *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods.* Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics pp. 189--196.

[7]Yufei Wang, Ian Wood, Stephen Wan, Mark Dras, and Mark Johnson. 2021. *Mention Flags (MF): Constraining Transformer-based Text Generators.* In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 103–113, Online. Association for Computational Linguistics.

[8]Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. 2019. *Bertscore: Evaluating text generation with bert.* arXiv preprint arXiv:1904.09675.