P4-6

# 擬似データを用いた教師あり学習による語彙平易化

野口 夏希、梶原 智之（愛媛大）、荒瀬 由紀（阪大）、内田 諭（九大）、二宮 崇（愛媛大）

## 概要

- 語彙平易化とは、文中の難解語の意味を保持しつつ、文脈に適した平易語に言い換える技術である
- 学習用のデータが少ないため、既存研究では教師あり学習の手法が提案されていない
- 文レベルの大規模なテキスト平易化コーパスを活用した擬似的な教師あり学習によってマスク言語モデルを再訓練し、語彙平易化の性能を改善する

Much of the water carried by these streams is **diverted**.

⇩ 言い換え候補生成

used, redirected, diverted, channel, poll, less, …

⇩ リランキング

1. redirected, 2. diverted, 3. poll, 4. used, …

⇩

Much of the water carried by these streams is **redirected**.

## 関連研究

- Light-LS[1]
  文中の難解語に対し、単語分散表現の余弦類似度を用いて類義語を収集する
- BERT-LS[2]
  BERTを再訓練せずに用いて、文脈を考慮した言い換え候補を生成する
- Simple BERT[3]
  文中の平易語をランダムにマスクし、BERTのマスク言語モデリングを追加で事前訓練する

[1] Goran Glavaš and Sanja Štajner. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In Proc. of ACL, pp. 63–68, 2015.
[2] Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. Lexical Simplification with Pretrained Encoders. In Proc. of AAAI, pp. 8649–8656, 2020.
[3] Renliang Sun and Xiaojun Wan. SimpleBERT: A Pretrained Model That Learns to Generate Simple Words. arXiv:2204.07779, 2022.

## 提案手法

文平易化パラレルコーパスのうち、少数の単語の置換のみが行われている文対を抽出する

1. 難解文と平易文のペアからなるパラレルコーパスを用意する
   - Wikipedia：50万文対、Newsela：40万文対
2. 文長が同じ文対を抽出する
3. 異なる単語数kが1≦k≦5の文対を抽出する
4. このデータを用いてBERTを再訓練する

| | Wikipedia | Newsela |
|---|---|---|
| k=1 | 10,260 | 7,681 |
| k=3 | 16,446 | 11,614 |
| k=5 | 18,995 | 11,379 |

| | 難解文 | 平易文 |
|---|---|---|
| 使用する | Relax the current standards? | Lower the current limits? |
| 2. で削除 | A member of the Democratic Party, she is the first woman to serve as Governor of Rhode Island. | She is a member of the Democratic Party. |
| 3. で削除 | The underside of the wings is also black. | The underside of the wings is also black. |

## 実験結果

| | 訓練データ | LexMTurk P | LexMTurk R | LexMTurk F1 | BenchLS P | BenchLS R | BenchLS F1 | NNSeval P | NNSeval R | NNSeval F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Light-LS | - | 15.1 | 12.2 | 13.5 | 14.2 | 19.1 | 16.3 | 10.5 | 14.1 | 12.1 |
| BERT-LS | - | 29.6 | 23.0 | 25.9 | 23.6 | 32.0 | 27.2 | 19.0 | 25.4 | 21.8 |
| Simple BERT | - | 35.3 | 27.5 | 30.9 | 26.9 | 36.4 | 30.9 | - | - | - |
| 提案手法（k=1） | Wikipedia | 35.8 | 35.7 | **35.8** | 25.4 | 40.9 | **31.3** | 19.0 | 34.1 | **24.4** |
| 提案手法（k=3） | Wikipedia | 35.2 | 35.3 | 35.3 | 25.1 | 40.3 | 30.9 | 18.9 | 32.6 | 23.9 |
| 提案手法（k=5） | Wikipedia | 34.8 | 35.0 | 34.9 | 24.8 | 40.3 | 30.7 | 18.6 | 32.7 | 23.7 |
| 提案手法（k=1） | Newsela | 27.0 | 27.7 | 27.4 | 20.5 | 34.9 | 25.9 | 16.8 | 33.0 | 22.3 |
| 提案手法（k=3） | Newsela | 25.8 | 26.4 | 26.1 | 19.3 | 32.9 | 24.3 | 16.4 | 31.3 | 21.6 |
| 提案手法（k=5） | Newsela | 24.5 | 25.3 | 24.9 | 18.4 | 31.7 | 23.3 | 15.0 | 29.9 | 20.0 |

| | |
|---|---|
| 入力文 | The intent is to display the likeness, personality, and even the mood of the person. |
| 正解単語 | aim, meaning, idea, plan, goal, attempt, reason, point, purpose, desire, trotting |
| BERT-LS | intent, intention, **purpose**, objective, intended, **aim**, **desire**, object, emphasis, understanding |
| 提案手法（k=1） | **aim**, **goal**, **purpose**, **idea**, intention, object, objective, job, task, **plan** |

- 擬似的な教師あり学習による再訓練で語彙平易化の最高性能を達成した
- 評価データと同じドメインの訓練データを用い、ノイズが少ないk=1の設定が最高性能を達成した