日本語転移学習モデルにおける事前学習コーパスのフィルタリング

早稲田大学
渡邊亞椰　河原大輔

研究背景

- 自然言語処理モデルの事前学習には大規模コーパスが必要
- Webクローリングで集めたテキストは汚いことが多い
  - ➡フィルタリングが必要
    - 例)顔文字、複数個繋がった長音等
- フィルタリングの有無がモデルに及ぼす影響とは？

関連研究

- C4 [Raffel 20]、mC4 [Xue 21]
  - Webクロールで集めた大規模コーパス。mC4は日本語を含む多言語版。フィルタリングしている
- Japanese Chit-chat Systems [Sugiyama 21]
  - 日本語モデル構築のために学習データ（Twitter）にフィルタリングしている
- mecab-ipadic-NEologd [Sato 15]
  - 形態素解析エンジン用の正規化処理方法を示している

事前学習コーパスのフィルタ設計

- かな漢字が少なすぎる行を削除
- 括弧を含む行を削除
- URLを含む行を削除
- 指定終端文字（。」．など）で終わらない行を削除
- 【】を含む行を削除
- 長音の規格化

フィルタリング前後のテキスト例

２０２０／０３／０６　２３：５４　－　ＮＹの小さな灯り　～ヘアメイク日記～
黒糖きな粉
冬限定の生八つ橋「ふゆおたべ」
日本から戻る時に大抵空港で生八つ橋か信玄餅のどちらかを必ず買って帰ってくるのですが、今回は珍しいバージョンを見かけたので生八つ橋を買う事にしました。それがこちら、冬限定の生八つ橋「ふゆおたべ」です。切り絵のデザインも素敵ですよね＾＾お味は黒豆と栗きんとんの二種類。おせち料理みたいです～♪栗きんとんは元々味がそんなに個性的でない事もあり、あまりよく分からなかったのですが（ふんわり甘くて美味しい．．．
２０１９／０３／０６　００：３０　－　ＮＹの小さな灯り　～ヘアメイク日記～

↓

冬限定の生八つ橋「ふゆおたべ」
日本から戻る時に大抵空港で生八つ橋か信玄餅のどちらかを必ず買って帰ってくるのですが、今回は珍しいバージョンを見かけたので生八つ橋を買う事にしました。それがこちら、冬限定の生八つ橋「ふゆおたべ」です。切り絵のデザインも素敵ですよね＾＾お味は黒豆と栗きんとんの二種類。おせち料理みたいですー♪栗きんとんは元々味がそんなに個性的でない事もあり、あまりよく分からなかったのですが（ふんわり甘くて美味しい．．．

フィルタリング前後のテキスト例

２０２０／０３／０６　２３：５４　－　ＮＹの小さな灯り　～ヘアメイク日記～
黒糖きな粉
冬限定の生八つ橋「ふゆおたべ」
日本から戻る時に大抵空港で生八つ橋か信玄餅のどちらかを必ず買って帰ってくるのですが、今回は珍しいバージョンを見かけたので生八つ橋を買う事にしました。それがこちら、冬限定の生八つ橋「ふゆおたべ」です。切り絵のデザインも素敵ですよね＾＾お味は黒豆と栗きんとんの二種類。おせち料理みたいです～♪栗きんとんは元々味がそんなに個性的でない事もあり、あまりよく分からなかったのですが（ふんわり甘くて美味しい．．．
２０１９／０３／０６　００：３０　－　ＮＹの小さな灯り　～ヘアメイク日記～

↓

冬限定の生八つ橋「ふゆおたべ」
日本から戻る時に大抵空港で生八つ橋か信玄餅のどちらかを必ず買って帰ってくるのですが、今回は珍しいバージョンを見かけたので生八つ橋を買う事にしました。それがこちら、冬限定の生八つ橋「ふゆおたべ」です。切り絵のデザインも素敵ですよね＾＾お味は黒豆と栗きんとんの二種類。おせち料理みたいです―♪栗きんとんは元々味がそんなに個性的でない事もあり、あまりよく分からなかったのですが（ふんわり甘くて美味しい．．．

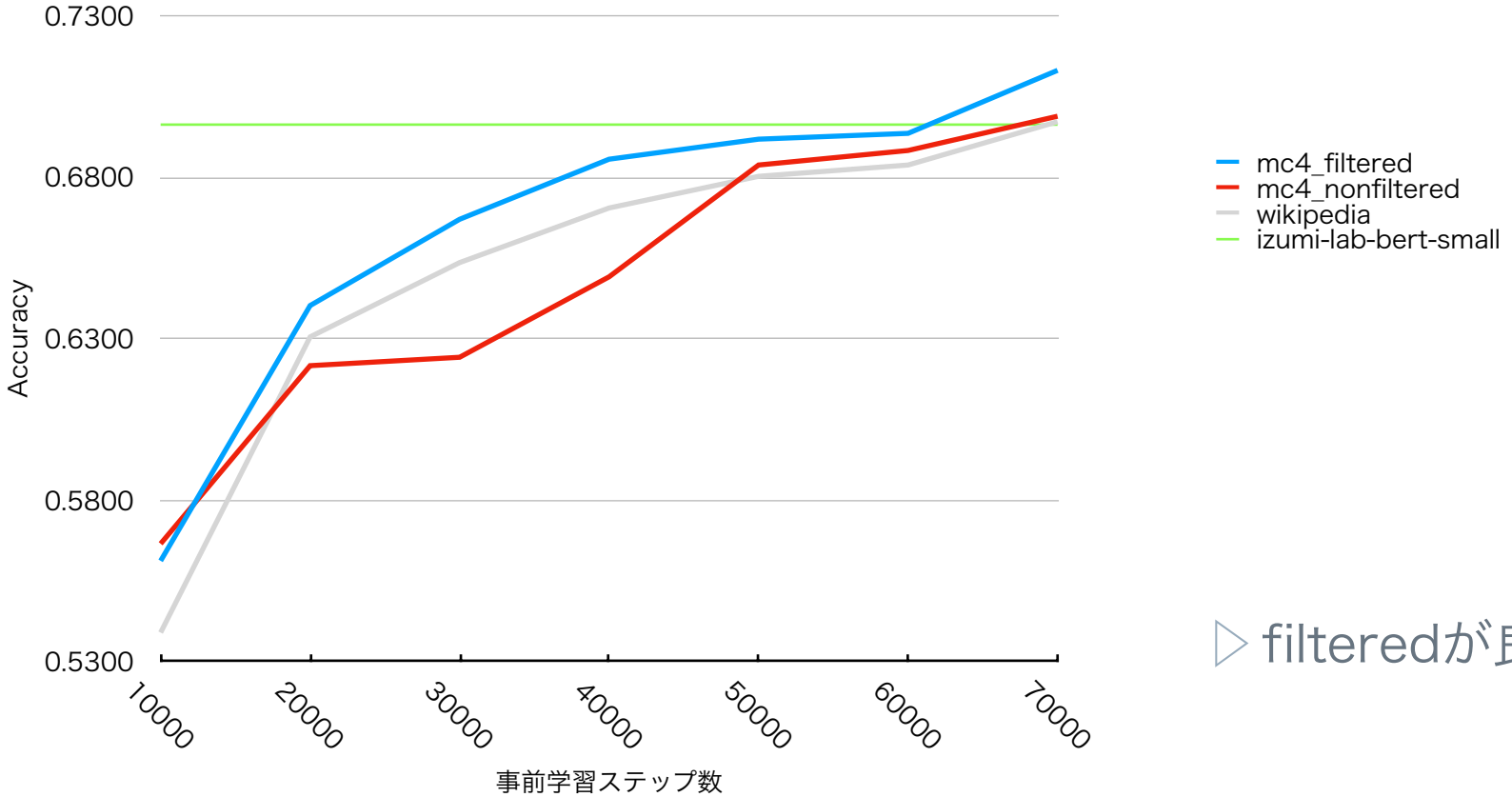
実験

- 以下のコーパスで同じ構成のモデルを事前学習
  - フィルタリングしたmC4日本語部分（一部）
  - フィルタリングしていないmC4日本語部分（一部）
  - Wikipediaコーパス（全文）
- 全て大きさは4GB
- モデルはRoBERTa [Liu 19] (small)

実験

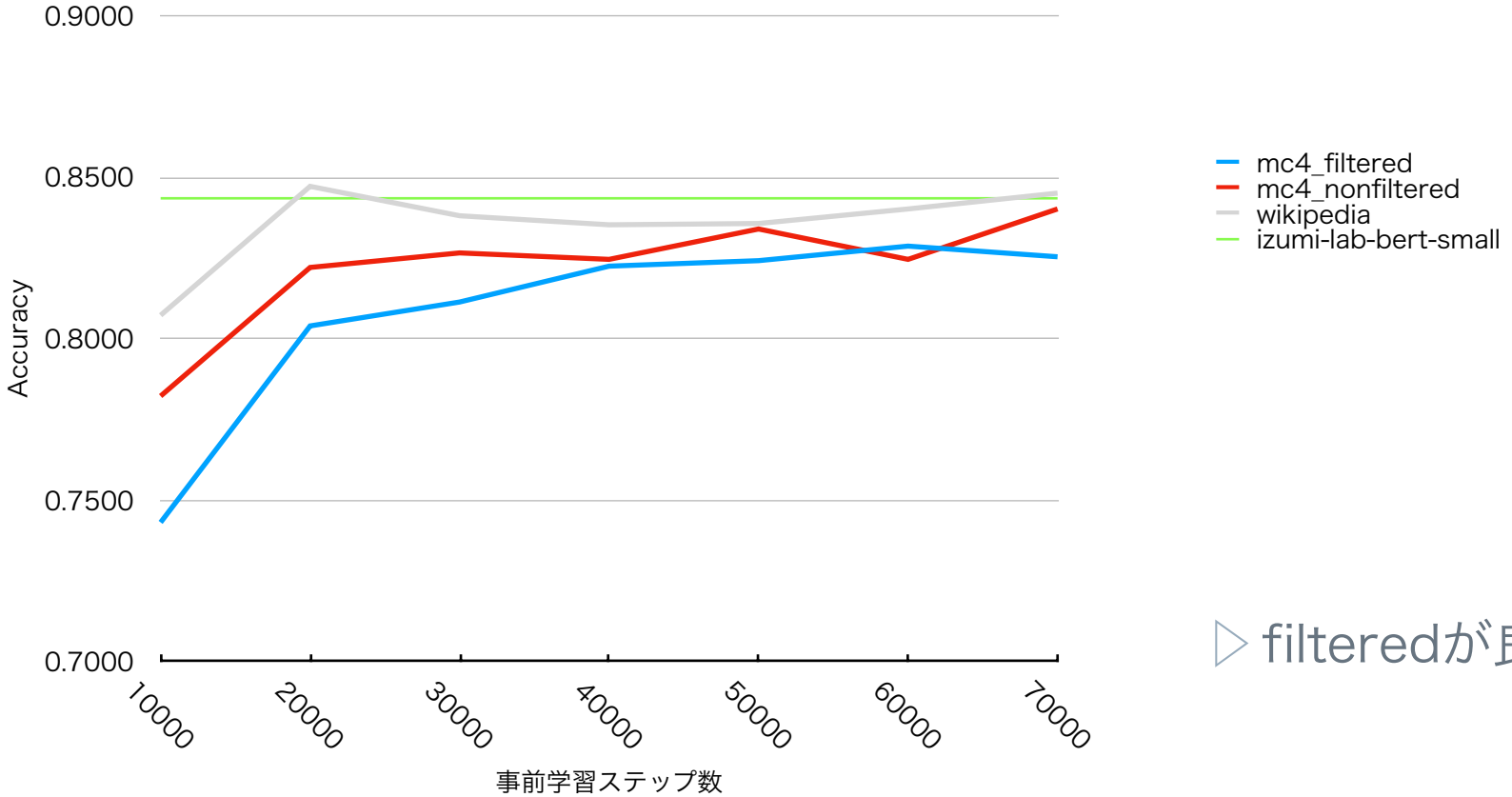
- 日本語言語理解ベンチマークであるJGLUE [栗原 22] タスクでファインチューニングし、評価
  - JCommonsenceQA: 常識推論タスク
  - JNLI: 文関係推論タスク
  - JSTS: 文類似性推定タスク
- 評価は事前学習10,000ステップごとに行う
- ベースライン: izumi-lab-bert [Suzuki 21] (small)

結果(JCommonsenceQA)



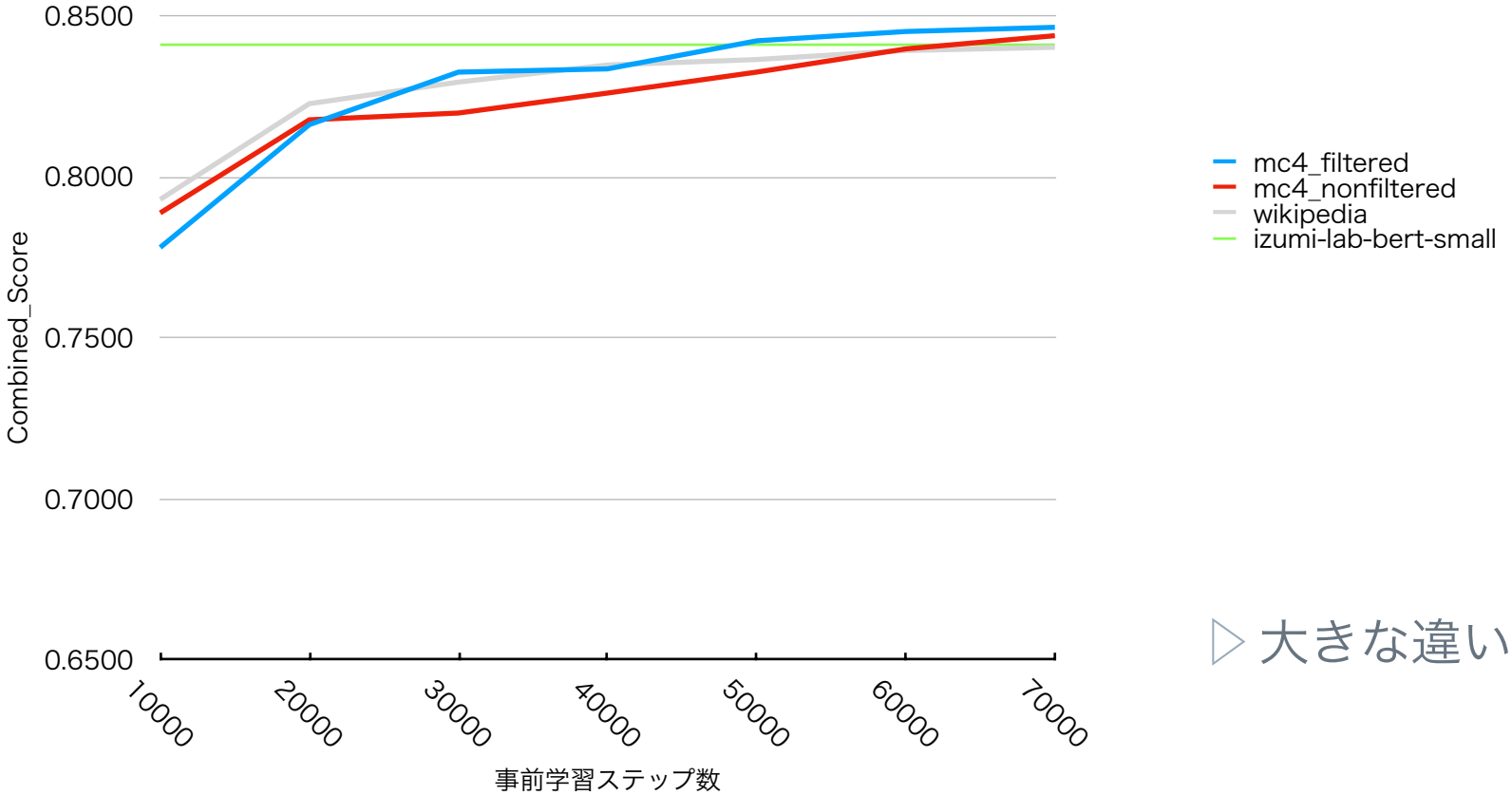
▷ filteredが良い

結果(JNLI)



▷ filteredが良くない

結果(JSTS)



▷ 大きな違いはない

まとめ

- フィルタリングは必ず良い方向に働くわけではないが、モデルに影響を与える
  - JCommonsenceQAにおいてはフィルタリングをした方が良い
  - JNLIにおいてはフィルタリングをしない方が良い
  - JSTSは他2つほどの差がない
- 事前学習コーパスのフィルタリングによる違いは低ステップでより出やすいため、小規模な事前学習においてより重要

参考文献

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Trans- former”, Journal of Machine Learning Research , Vol. 21, No. 140, pp. 1–67 (2020)
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C., “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”, in Proceed- ings of the 2021 Conference of the North American Chapter of the Association for Computa- tional Linguistics: Human Language Technologies , pp. 483–498, Online (2021), Association for Computational Linguistics
- Sugiyama,H.,Mizukami,M.,Arimoto,T.,Narimatsu,H.,Chiba,Y.,Nakajima,H., and Meguro, T., “Empirical Analysis of Training Strategies of Transformer-based Japanese Chit- chat Systems”, arXiv preprint arXiv:2109.05217 (2021)
- Sato, T., “Neologism dictionary based on the language resources on the Web for Mecab” (2015), https://github.com/neologd/mecab-ipadic-neologd
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettle- moyer, L., and Stoyanov, V., “Roberta: A robustly optimized bert pretraining approach”, arXiv preprint arXiv:1907.11692 (2019)
- 栗原健太郎, 河原大輔, 柴田知秀, “JGLUE: 日本語言語理解ベンチマーク”, 言語処理学会第 28 回年次大会 (2022)