Исправление опечаток и грамматических ошибок в русскоязычных текстах при помощи BERT

Бунин Дмитрий, группа 792

Научный руководитель: Сорокин А. А.

Анализ предметной области

Виды ошибок

1. Грамматические ошибки (grammatical)
   Нарушение правил грамматики. Например, неправильное образование и употребление форм слова.
2. Орфографические ошибки (spelling)
   Неверное написание слов.
3. Опечатки (typo)
   Ошибки в печатном тексте в результате случайности.

Анализ предметной области

Задачи

1. GEC – grammatical error correction
2. Spelling correction

Анализ предметной области

Метрики

1. F-мера

   Пусть имеется текст из n предложений, тогда обозначим g_i – множество корректных исправлений предложения i, а e_i – множество наших исправлений.

$$R = \frac{\sum_{i=1}^{n} |g_i \cap e_i|}{\sum_{i=1}^{n} |g_i|},$$

$$P = \frac{\sum_{i=1}^{n} |g_i \cap e_i|}{\sum_{i=1}^{n} |e_i|}.$$

2. GLEU

   Аналог BLEU для машинного перевода.

Анализ предметной области

Датасеты, Английский язык

1. CoNLL-2014
   - Статья: Ng и др. (2014)
   - Метрика: $F_{1/2}$
   - Train: 1M токенов
   - Test: 30k токенов
2. JFLEG
   - Статья: Napoles, Sakaguchi и Tetreault (2017)
   - Метрика: GLEU
   - All: 1.5k предложений
3. BEA-2019
   - Страница соревнования
   - Метрика: $F_{1/2}$
   - Train: 628k токенов
   - Validation: 87k токенов
   - Test: 86k токенов

Анализ предметной области

Датасеты, Русский язык

1. SpellRuEval
   - Статья: Sorokin, Baytin и др. (2016)
   - Страница соревнования
   - Validation: 2k предложений
   - Test: 2k предложений

Анализ предметной области I

Существующие подходы

1. Стандартные решения
   - GNU Aspell
   - Hunspell
   - JamSpell
   - Яндекс.Спеллер
2. Модель шумного канала
   - Модель на основе взвешенного расстояния Дамерау – Левенштейна: Kernighan, Church и Gale (1990).
   - Улучшенная модификация с более сложной моделью ошибок: Brill и Moore (2000).
3. Поиск кандидатов, ранжирование
   Схема была предложена в Flor и Fugati (2012). Решение задачи состоит из этапов:
   - Поиск кандидатов для исправления ошибки
   - Ранжирование кандидатов

Анализ предметной области II

Существующие подходы

4. Трансформеры
   Задачу можно рассматривать, как машинный перевод.
   Улучшения подобных моделей:
   - Копирование исходного текста: Zhao и др. (2019).
   - Генерация синтетических данных для обучения: Kiyono и др. (2020).
5. Sequence labeling
   При введении определенных классов трансформаций можно рассматривать GEC, как задачу sequence labeling.
   Модели:
   - Parallel Iterative Edit Model: Awasthi и др. (2020).
   - GECToR: Omelianchuk и др. (2020).

Анализ предметной области

Результаты, Английский язык

GEC system	Ens.	CoNLL-2014 (test)	BEA-2019 (test)
Copy Aug. Transformer		59.8	–
PIE		59.7	–
Transformer (synt. data)		61.3	64.2
GECToR		**65.3**	**72.4**
Copy Transformer	✓	61.2	–
PIE	✓	61.2	–
Transformer (synt. data)	✓	65.0	70.2
GECToR	✓	**66.5**	**73.6**

Таблица: Сравнение результатов различных моделей для английского языка

Анализ предметной области

Результаты, Русский язык

GEC system	**Precision**	**Recall**	F_1
Yandex.Speller	**83.09**	59.86	69.59
JamSpell	44.57	35.69	39.64
SpellRuEval Baseline	55.91	46.41	50.72
SpellRuEval Winner	81.98	**69.25**	**75.07**

Таблица: Сравнение результатов различных моделей для русского языка

Исследование

Цель работы

Исследование применимости модели BERT к задаче исправления опечаток и грамматических ошибок в русскоязычных текстах.

Исследование

Архитектура модели

Было решено начать с модели, основанной на ранжировании в связи с небольшим количеством данных и хорошими показателями согласно SpellRuEval (см. Sorokin и Shavrina (2016)).
Поиск кандидатов будет осуществляться на основании расстояния Дамерау – Левенштейна при помощи префиксного бора. За основу будет взят spelling-correction модуль из библиотеки DeepPavlov.
Полученные кандидаты будут отранжированны на основе признаков:

1. Взвешенное расстояние Дамерау – Левенштейна (см. Brill и Moore (2000)).
2. Вероятность BERT MLM (см. Devlin и др. (2019)).

Веса для BERT будут взяты из RuBERT.

Исследование

План работ

1. Создание модели, ранжирующей на основе BERT MLM, ее тестирование на русскоязычных и англоязычных датасетах.
2. Введение дополнительных признаков, как в решении-победителе SpellRuEval.
3. Дообучение BERT на MLM для датасета.
4. Возможно, испытание других языковых моделей (например, GPT).

Литература I

Awasthi, Abhijeet и др. (2020). «Parallel iterative edit models for local sequence transduction». в: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, с. 4260—4270. DOI: 10.18653/v1/d19-1435. arXiv: 1910.02893.

Brill, Eric и Robert C. Moore (2000). «An improved error model for noisy channel spelling correction». в: Kukich 1992, с. 286—293. DOI: 10.3115/1075218.1075255.

Devlin, Jacob и др. (2019). «BERT: Pre-training of deep bidirectional transformers for language understanding». в: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1.Mlm, с. 4171—4186. arXiv: 1810.04805.

Flor, Michael и Yoko Fugati (2012). «On using context for automatic correction of non-word misspellings in student essays.». в: *Proceedings of the 7th workshop on innovative use of {NLP} for {Building} {Educational} {Applications}*, с. 105—115. URL: http://aclweb.org/anthology/W/W12/W12-2012.pdf.

Kernighan, Mark D., Kenneth W. Church и William A. Gale (1990). «A spelling correction program based on a noisy channel model». в: с. 205—210. DOI: 10.3115/997939.997975.

Kiyono, Shun и др. (2020). «An empirical study of incorporating pseudo data into grammatical error correction». в: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, с. 1236—1242. DOI: 10.18653/v1/d19-1119. arXiv: 1909.00502.

Литература II

Napoles, Courtney, Keisuke Sakaguchi и Joel Tetreault (2017). «JFLEG: A fluency corpus and benchmark for grammatical error correction». в: *arXiv* 2, с. 229—234.

Ng, Hwee Tou и др. (2014). «The CoNLL-2014 Shared Task on Grammatical Error Correction». в: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*. July. Stroudsburg, PA, USA: Association for Computational Linguistics, с. 1—14. DOI: 10.3115/v1/W14-1701. URL: http://aclweb.org/anthology/W14-1701.

Omelianchuk, Kostiantyn и др. (2020). «GECToR – Grammatical Error Correction: Tag, Not Rewrite». в: April, с. 163—170. DOI: 10.18653/v1/2020.bea-1.16. arXiv: 2005.12592.

Sorokin, A. A., A. V. Baytin и др. (2016). «SPELLRUEVAL: The first competition on automatic spelling correction for Russian». в: *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, с. 660—673. ISSN: 20757182.

Sorokin, A. A. и T. O. Shavrina (2016). «Automatic spelling correction for Russian social media texts». в: *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, с. 688—701. ISSN: 20757182.

Zhao, Wei и др. (2019). «Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data». в: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1, с. 156—165. DOI: 10.18653/v1/n19-1014. arXiv: 1903.00138.