Lexical Complexity Prediction

Курсова работа по Извличане на информация

Задача 1, SemEval 2021

Симона Михайлова, ФН 26432 (ИИ)

Ива Борисова, ФН 26494 (ИИ)

Джовани Чемишанов, ФН 26415 (ИИОЗ)

Съдържание

- Задача на курсовата работа
- Основен dataset
- Възможни подходи за решаването на задачата
- Избран подход и инструменти за реализация
- План за реализация/експерименти

Задача за курсовата работа - 1

Да се подобри точността на класификатора, описан в оригиналната статия.

- Подзадача 1: Предсказване на сложността на отделни думи
- Подзадача 2: Предсказване на сложността на изрази, съставени от няколко думи

Задача на курсовата работа - 2

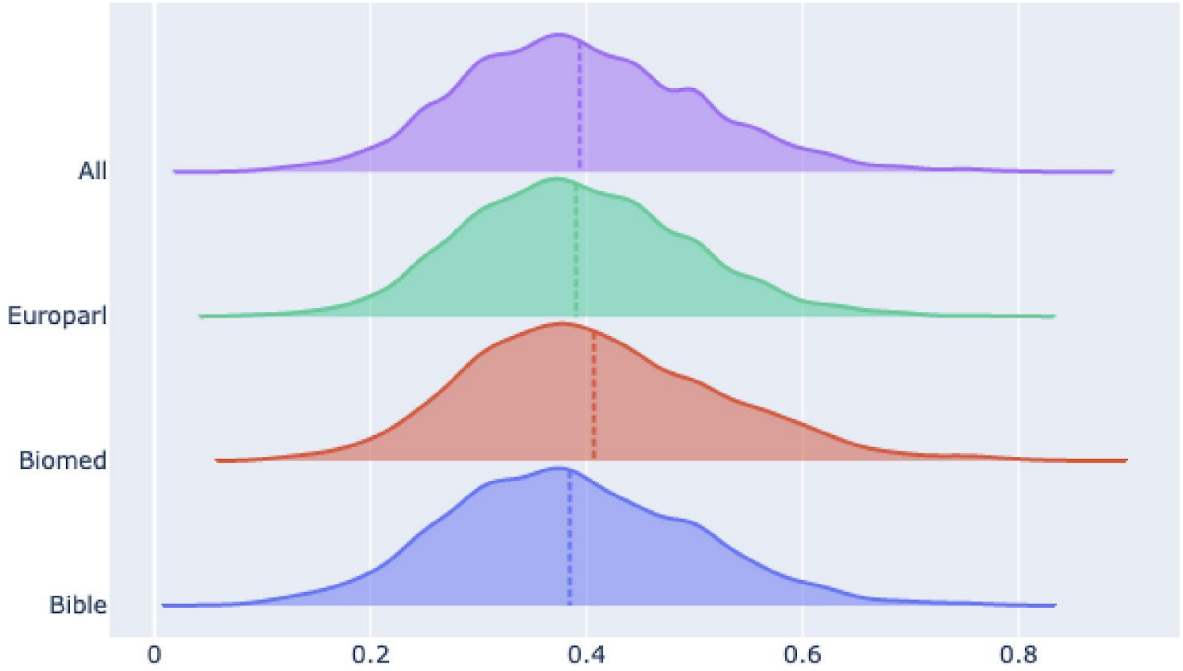
Оценяване на сложността на дума или израз

Пример

Corpus	Context	Complexity
Bible	This was the **length** of Sarah's life.	0.125
Biomed	[...] cell **growth** rates were reported to be 50% lower [...]	0.125
Europarl	Could you tell me under which rule they were enabled to extend this item to have four rather than three **debates**?	0.208
Europarl	These agencies have gradually become very important in the **financial world**, for a variety of reasons.	0.438
Biomed	[...] leads to the **hallmark loss** of striatal neurons [...]	0.531
Bible	The **idols** of Egypt will tremble at his presence [...]	0.575
Bible	This is the law of the **trespass offering**.	0.639
Europarl	They do hold elections, but candidates have to be endorsed by the conservative clergy, so **dissenters** are by definition excluded.	0.688
Biomed	[..] due to a reduction in **adipose** tissue.	0.813

Основен dataset

- Библия (Christodouloupoulos and Steedman, 2015).
- Europarl (Koehn, 2005)
- CRAFT corpus (Bada et al., 2012)



	Contexts	Unique Words	Median Annotators
All	9476 / 7974 / 1500	5166 / 3903 / 1263	7 / 7 / 7
Europarl	3496 / 2896 / 600	2194 / 1693 / 501	7 / 7 / 7.5
Biomed	2960 / 2480 / 480	1670 / 1250 / 420	7 / 7 / 7
Bible	3020 / 2600 / 420	1705 / 1362 / 343	7 / 7 / 8

План за реализация

1. Имплементация на описаната в статията базова система
2. Намиране на допълнителни корпуси и обработка на новите данни;
3. Оценяване на точността на базовата система при по-голям обем от данни
4. Подобряване на базовата система чрез Bert

Предложени подходи*

- Линейна регресия на базата на GloVe и InferSent Embeddings
- Евристични предиктори като:
  - дължина на думата, брой срички
  - честота на срещане в универсалния домейн, представляван от универсалния индекс за извършвани търсения на Гугъл

** в оригиналната статия, Shardlow et al. (2020), "CompLex: A New Corpus for Lexical Complexity Prediction from Likert Scale Data"*

Избрани подходи - 1

Bert:

1. Трениран е върху голямо множество от данни;
2. По-малка чувствителност към грешки в данните;
3. Bidirectional подход - взема предвид контекста на думата в зависимост от съседните ѝ;

Избрани подходи - 2

1. Пример:

   amazingly - дълга дума с четири срички -> по-висока оценка

   amaze - по-кратка дума -> по-ниска оценка

   Други форми на думата: amazed, amazing, amazingly...

2. Проблем: значителна разлика в оценките, на думите, които имат един и същ корен и смисъл;

3. Идея за решение: отчитане на различните глаголни и граматически форми и формиране на обща оценка за думите с един и същ корен и смисъл

Източници

[1] Matthew Shardlow, *"CompLex: A New Corpus for Lexical Complexity Prediction from Likert Scale Data" (2020)*

[2] Jeffrey Pennington, *"GloVe: Global Vectors for Word Representation"* (2014)

[3] Alexis Conneau, *"Supervised Learning of Universal Sentence Representations from Natural Language Inference Data"*, (2017)

[4] Jacob Devlin, *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"*, (2018)