Term Definition Extraction

Катя Гриневская, Таня Казакова

GitHub

Лингви́стика — наука, изучающая языки.

Актуальность

Результаты TDE можно использовать в:

- NLP tools:
  - word sense disambiguation
  - paraphrasing
- NLP resources:
  - WordNet
  - WikiData
- Reading tools:
  - scholarly papers
  - dictionaries, glossaries...

Языки:

- пробовалось для:
  - английского
  - немецкого
  - голландского
  - чешского
  - немного китайского
- НЕ пробовалось для:
  - **русского!**

Команда

Катя

идейный руководитель

- Ручная разметка статей
- Разметка wikiданных
- Baseline2 (берт+crf)
- CRFвариации
- Эксперименты

Таня

организатор деятельности (Trello, тг)

Отчётописатель

- Wiki данные
- Baseline1 (правила)
- Эксперименты

Гитхаб: https://github.com/KateGrinevskaya/NN_methods_project

Данные

Wikipedia

- **300** статей
- первое предложение точно содержит определение TERM - DEF
- 100500 способов дать определение

Архив статей по корпусной лингвистике

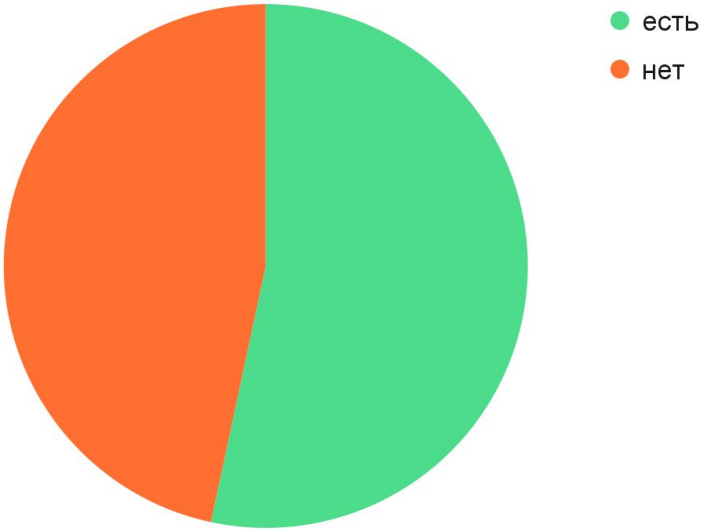
- ручная разметка (10 статей)

Храним в таблицах

id_sent	text_sent	has_def (0/1)

id_sent	text_token	tag

Данные:
926 предложений



Данные: идеологические проблемы

- Что такое определение? Здесь есть определение?

Пушкин - великий поэт. (?)

Пушкин - великий русский поэт и писатель. (Да!)

Данные: идеологические проблемы

Единицами семантики принято считать, с одной стороны, более простые (или даже элементарные) единицы — значения с их компонентами и различительными признаками (семами), а с другой стороны — правила, по которым из этих более простых единиц строятся более сложные содержательные образования — смыслы.

Два определения на один термин? Или термин с одним длинным определением?

Вложенные определения - тех.проблема.

Лингви́стика (от лат. lingua «язык»), языкозна́ние, языкове́дение — наука, изучающая языки.

Три термина на одно определение...

Что делать с разорванные определения?

Данные: идеологические проблемы

“Попробуй прочитать”

Мельчайшая единица синтаксиса — словоформа с её синтактикой (то есть свойствами сочетаемости) есть инвентарная номинативная единица и в то же время максимальная единица морфологии.

Сколько пар термин-определение? Каких?

Если человек не может понять, модель тоже не должна.

Решение идеологических проблем

(по мотивам статьи)

определение должно состоять из:

- гиперонима
- уточнения (не эпитет!)

Семантические кальки — это слова, которые получили новые, переносные значения под влиянием иностранного слова.

Может быть несколько терминов и определений.

Безумные случаи не берём. Храним отдельно, чтоб посмотреть, что с ними будет делать хорошая модель.

Кстати, вот что получилось!

Мельчайшая единица синтаксиса — словоформа с её синтактикой (то есть свойствами сочетаемости) есть инвентарная номинативная единица и в то же время максимальная единица морфологии.

Данные: тех. проблемы

- ручная разметка данных (готовых датасетов не найдено)

- сбалансированность данных?

- вложенные определения

Решения:

Разметка сначала правилами, потом ручная выверка

Взяли примерно одинаковое количество с определениями и без...

Размечаем только верхний слой

Baseline1

“Правило” для есть/нет

Предложение с определением содержит:

[' — ', 'то есть', 'называют', 'называется', 'называются', 'является', 'являются’]

accuracy = 0,85

“Правила” для BIO:

term — (это(такой/такие)) def

term ((то есть / т.е.) definition)

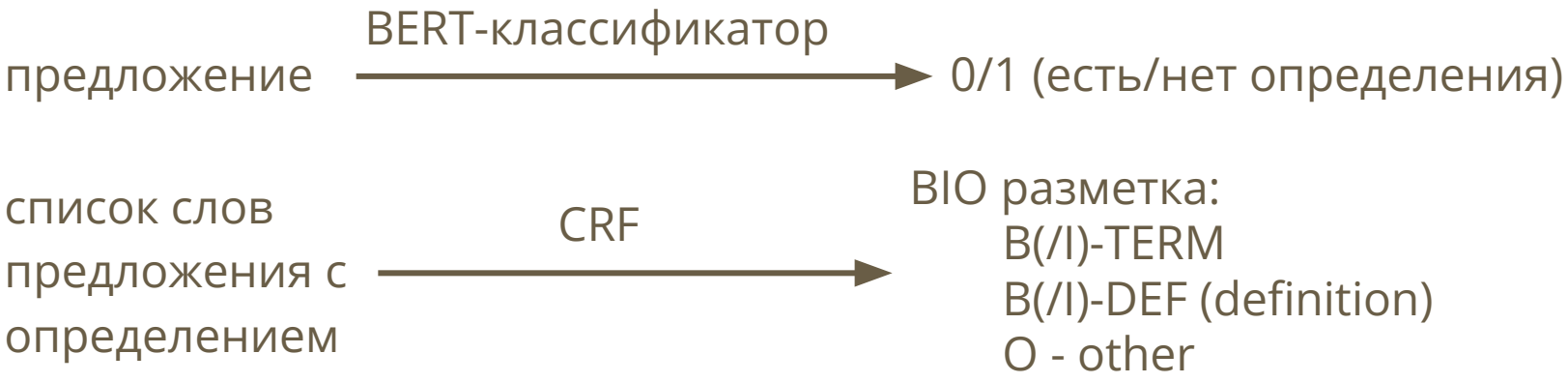
И ещё 7 др. шаблонов

Если разметилось определение, определение есть.

accuracy = 0,86

Улучшение0 или baseline2

Идея из статьи:



BERT-классификатор: rubert-base-cased, BertForSequenceClassification

Метрики

слот-теггер (TERM/DEF/O)	классификация предложений
macro-averaged precision	accuracy F-мера
macro-averaged recall	
macro-averaged F1	

Можно ещё было границы спанов отдельно оценивать.

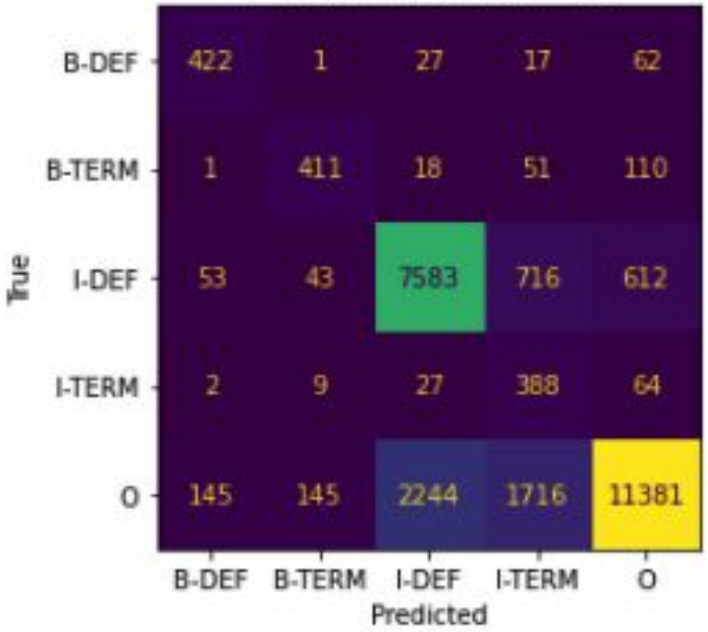
https://arxiv.org/pdf/2010.05129.pdf

baseline1 (правила): классификация

	precision	recall	f1-score	support
no term-def	0.93	0.76	0.84	432
term-def	0.82	0.95	0.88	494
accuracy			0.86	926
macro avg	0.87	0.86	0.86	926
weighted avg	0.87	0.86	0.86	926

Основные ошибки: находит лишние с тире, ‘то есть’ и глаголом ‘называть’

baseline1 (правила): классификация+BIO

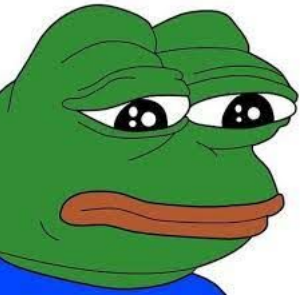


	precision	recall	f1-score	support
B-DEF	0.677	0.798	0.733	529
B-TERM	0.675	0.695	0.685	591
I-DEF	0.766	0.842	0.802	9007
I-TERM	0.134	0.792	0.230	490
O	0.931	0.728	0.817	15631
accuracy			0.769	26248
macro avg	0.637	0.771	0.653	26248
weighted avg	0.848	0.769	0.796	26248

Основные ошибки: выделяет больше I-TERM и I-DEF, чем нужно

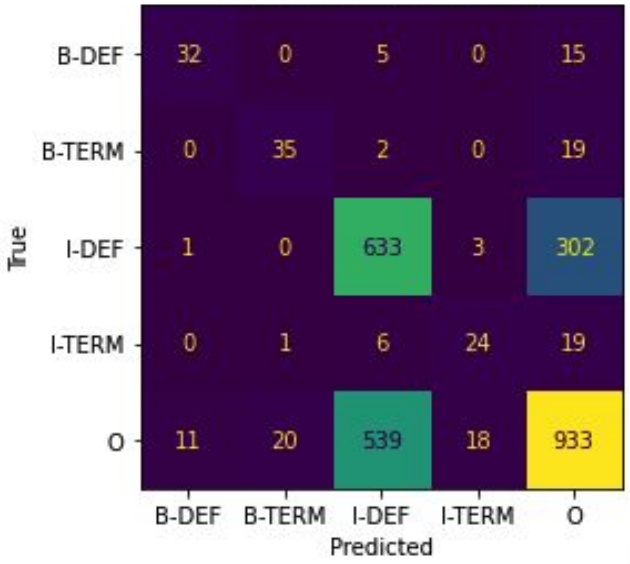
Улучшение0 или baseline2: классификация

DeepPavlov/rubert-base-cased, недообученный



	precision	recall	f1-score	support
0	0.29	0.11	0.16	44
1	0.49	0.76	0.59	49
accuracy			0.45	93
macro avg	0.39	0.43	0.38	93
weighted avg	0.40	0.45	0.39	93

Улучшение0 или baseline2: классификация + BIO



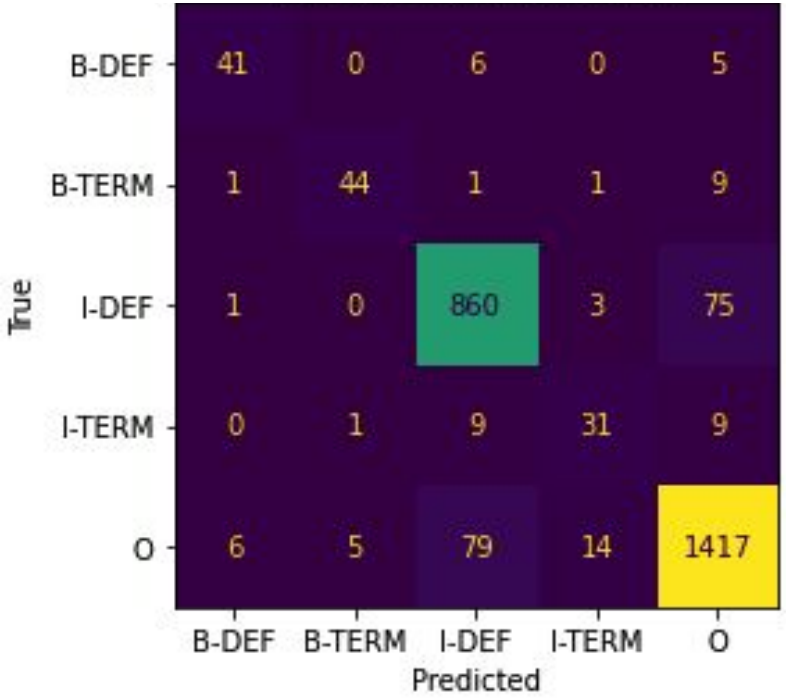
	precision	recall	f1-score	support
O	0.724	0.613	0.664	1521
B-DEF	0.727	0.615	0.667	52
I-DEF	0.534	0.674	0.596	939
B-TERM	0.625	0.625	0.625	56
I-TERM	0.533	0.480	0.505	50
accuracy			0.633	2618
macro avg	0.629	0.602	0.611	2618
weighted avg	0.650	0.633	0.636	2618

Обученный BERT лучше

DeepPavlov/rubert-base-cased, 2 эпохи

	precision	recall	f1-score	support
0	0.91	0.89	0.90	44
1	0.90	0.92	0.91	49
accuracy			0.90	93
macro avg	0.90	0.90	0.90	93
weighted avg	0.90	0.90	0.90	93

Обученный BERT лучше



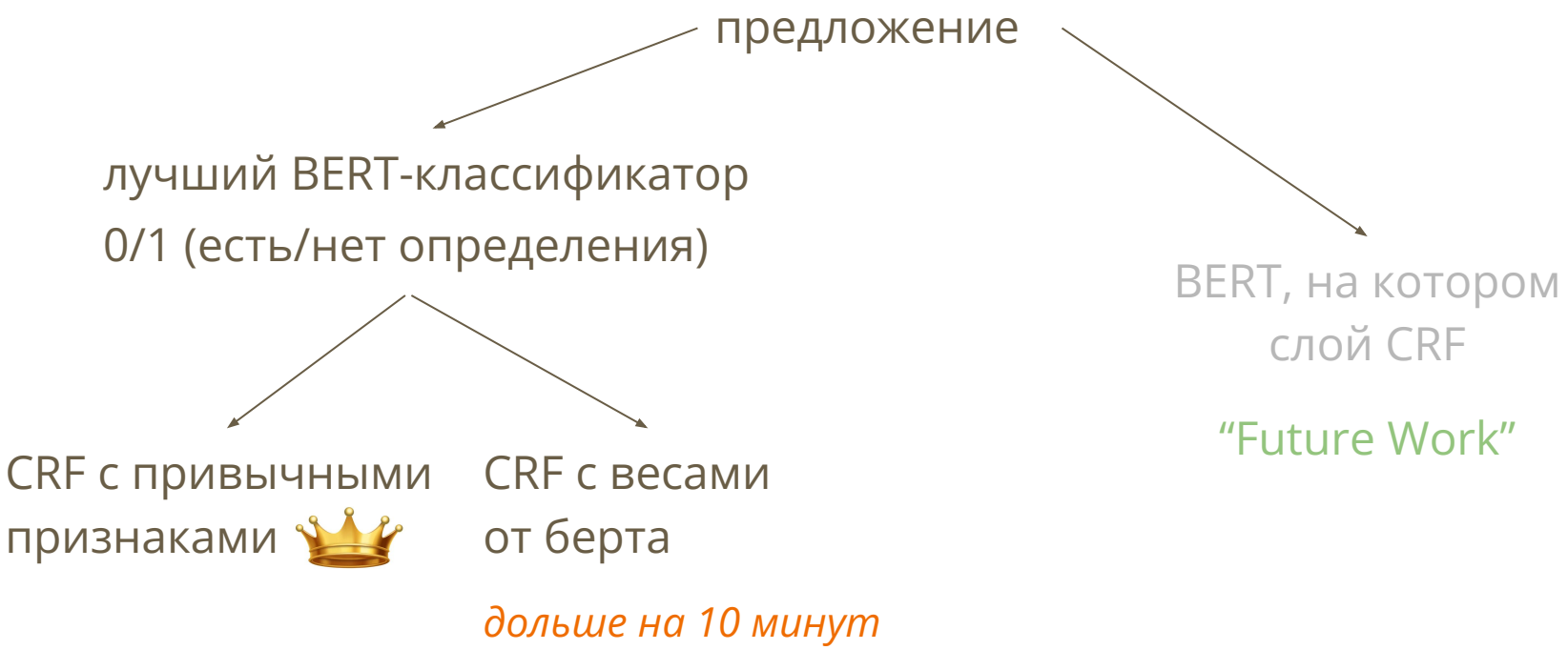
DeepPavlov/rubert-base-cased, 2 эпохи

	precision	recall	f1-score	support
O	0.935	0.932	0.933	1521
B-DEF	0.837	0.788	0.812	52
I-DEF	0.901	0.916	0.908	939
B-TERM	0.880	0.786	0.830	56
I-TERM	0.633	0.620	0.626	50
accuracy			0.914	2618
macro avg	0.837	0.808	0.822	2618
weighted avg	0.914	0.914	0.914	2618

Разные BERTы: классификация

Модель	Эпох	f1-score weighted avg
DeepPavlov/rubert-base-cased	2	0,9
DeepPavlov/rubert-base-cased	3	0.94626
bert-base-multilingual-cased	2	0,88
bert-base-multilingual-cased	3	0,92
bert-base-multilingual-uncased	2	0,88
bert-base-multilingual-uncased	3	0.94626

Улучшения “CRFы”



DeepPavlov, 3 эпохи

модель **без** эмбеддингов

	precision	recall	f1-score
O	0.953	0.844	0.895
B-DEF	0.857	0.808	0.832
I-DEF	0.797	0.948	0.866
B-TERM	0.852	0.821	0.836
I-TERM	0.608	0.620	0.614
accuracy			0.876
macro avg	0.813	0.808	0.809
weighted avg	0.886	0.876	0.877

модель **с** эмбеддингами

	precision	recall	f1-score
	0.950	0.804	0.871
	0.857	0.808	0.832
	0.755	0.949	0.841
	0.852	0.821	0.836
	0.667	0.640	0.653
			0.853
	0.816	0.804	0.807
	0.871	0.853	0.855

Multilingual uncased, 3 эпохи

модель **без** эмбеддингов

	precision	recall	f1-score
O	0.954	0.937	0.946
B-DEF	0.857	0.808	0.832
I-DEF	0.912	0.945	0.928
B-TERM	0.868	0.821	0.844
I-TERM	0.620	0.620	0.620
accuracy			0.929
macro avg	0.842	0.826	0.834
weighted avg	0.929	0.929	0.929

разница с CRF - 5%

модель **с** эмбеддингами

	precision	recall	f1-score
O	0.947	0.938	0.943
B-DEF	0.878	0.827	0.851
I-DEF	0.914	0.934	0.924
B-TERM	0.868	0.821	0.844
I-TERM	0.673	0.660	0.667
accuracy			0.927
macro avg	0.856	0.836	0.846
weighted avg	0.927	0.927	0.927

разница с CRF - 7%

Что хорошо работает, а что не очень?

- определения длиннее, чем должны быть, там, где не надо, и наоборот
- и термины, и определения начинаются там, где надо
- буквы некириллических и нелатинских алфавитов вызывают трудности
- ловим термины и их синонимы

Стяже́ние (контракция) — слияние двух смежных гласных в один гласный или в дифтонг.

[['B-TERM', 'O', 'B-TERM', 'O', 'O', 'B-DEF', 'I-DEF', 'I-DEF', 'I-DEF', 'I-DEF', 'I-DEF', 'I-DEF', 'I-DEF', 'I-DEF', 'I-DEF', 'O']]

Сложности

- сильная зависимость от пунктуации
- сбалансированность данных (как в реальности?)
- **вложенные определения**
- BERT… не присоединился слой CRF

Как улучшить

Собрать слой CRF для BERTa

Использовать синтаксические зависимости

Больше данных

Подумать про CRF не с BERTом, чем-то другим

Дальше: восстанавливать опущенную вершину, приводить к начальной форме

Евклидова геометрия - это геометрическая теория, основанная на системе аксиом, изложенной в “Началах” Евклида, а Неевклидова - любая система, которая отличается от геометрии Евклида.

Приятности

	Term P/R/F	Definition P/R/F	Macro P/R/F	Classification
результаты ребят из статьи	71.1 / 72.1 / 70.9	75.4 / 74.6 / 74.2	72.9 / 74.3 / 73.4	85.1
наши результаты	**75.1** / **72.6** / **73.8**	**90.9** / **93.8** / **92.3**	**84.2** / **82.6** / **83.4**	**94.6**

Литература

Основная идея, baseline2, метрики: Document-Level Definition Detection in Scholarly Documents: Existing Models, Error Analyses, and Future Directions

Про rule-based подходы:

- https://aclanthology.org/P10-1134/
- https://aclanthology.org/E09-3011.pdf
- http://www.lrec-conf.org/proceedings/lrec2008/pdf/783_paper.pdf
- https://aclanthology.org/W12-3206/

И про более современные подходы:

- https://aclanthology.org/D13-1073/ (CRF)
- https://aclanthology.org/N18-2061/
- https://aclanthology.org/2020.lrec-1.256.pdf
- https://arxiv.org/abs/1911.01678

Спасибо за внимание!