Term Definition Extraction

Катя Гриневская, Таня Казакова

Актуальность

Результаты TDE можно использовать в:

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

Команда

Катя

идейный руководитель

- Ручная разметка статей
- Baseline2 (берт)
- Улучшения

Таня

организатор деятельности (Trello, тг)

Отчётописатель

- Wiki данные
- Baseline1 (правила)
- Улучшения

Данные

- WikiData
  - N статей
  - первое предложение точно содержит определение, остальные в абзаце - нет
  - обычно: TERM - DEF (+ глазами просмотреть)
- Архив статей по корпусной лингвистике
  - ручная разметка

id_sent	text_sent	has_def (0/1)

слово	id_sent	тег(B/I-TERM / B/I-DEF / O)

Data Statistics посчитаем, когда дособираем данные

Baseline1

“Правила”

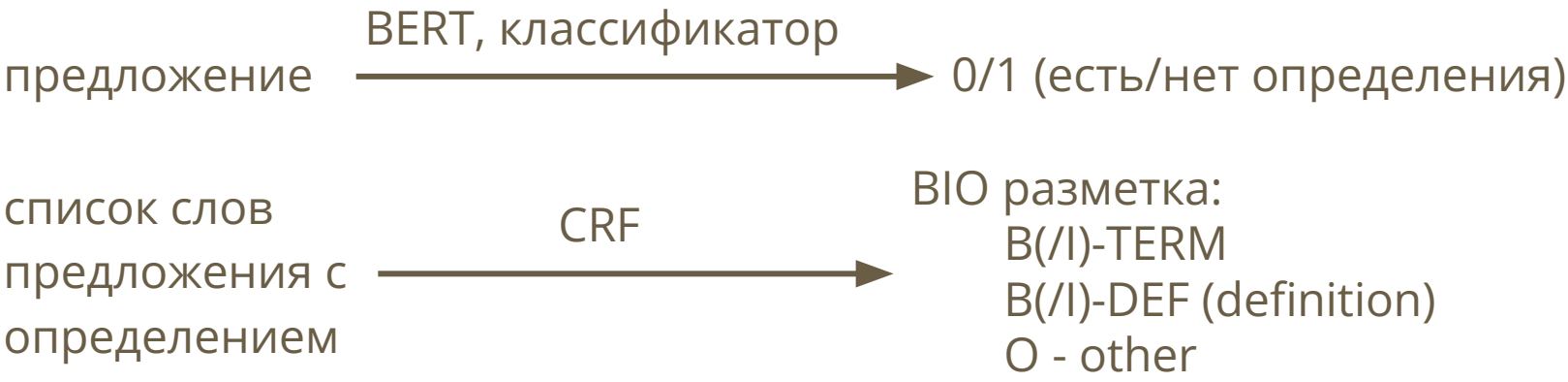
… - (это) …

АКМР (аббревиатура, которую можно расшифровать)

другие?

Улучшение0 или baseline2

Идея из [статьи]:



Если нет T и D тегов - предложение убираем. has_def = 0
Если I-TERM или I-DEF разорваны другими словами, включаем их в T или D.
Верим, что может быть несколько терминов и определений.

Придумать, что делать с пунктуацией:

1. попробовать с (но в Википедии все наши примеры с тире)
2. попробовать без (но тогда мы лишаем сеть важной информации)

3. Предобучить с, дообучить без?
4. Дать модели много примеров с тире, но без определений (из НКРЯ)?

Улучшения

- попробовать разные BERT’ы
- добавить синтаксическую разметку в качестве признаков
- добавить куда-нибудь CNN
- давать программе только предложения с терминами (NER)
- ...

Метрики

слот-теггер (TERM/DEF/O)	слот-теггер (границы спанов)	классификация предложений
macro-averaged precision	partial F (каждый выделенный спан к соответствующему ему в gold)	accuracy
macro-averaged recall		
macro-averaged F1		

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

Данных немного, поэтому, возможно, сделаем кросс-валидацию и усредним оценки (а возможно и нет)

Литература

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

Планируем посмотреть ещё 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

Планы

Февраль

31	1	2	3 данные	4	5	6
7	8 baseline	9	10	11	12	13 baseline2
14	15	16	17 другие векторы	18	19	20
21	22	23	24	25 +признаки	26	27
28	1 пунктуация	2	3	4	5	6

- до 20.03 анализировать, улучшать ещё как-то

См. Trello. Еженедельный созвон: четверг 21:30