

Open Information Extraction

обзор: ключевые статьи, инструменты, наборы данных

лаб. искусственного интеллекта
ПОМИ РАН им. В.А. Стеклова
Антон Алексеев

Санкт-Петербург
4 августа 2021 г.

DISCLAIMERs

- 1) Если вы всерьёз занимаетесь извлечением отношений, то едва ли в этом обзорном докладе для вас будет много нового;
- 2) подробно мы остановимся лишь на паре статей, прочее будет изложено очень поверхностно -- только чтобы понять эволюцию предметной области;
- 3) докладчик — совсем не лингвист, поэтому возможна терминологическая небрежность;
- 4) план доклада будет через четыре слайда.

Artificial Intelligence Lab, PDMI RAS



Laboratory Head: **Sergey Nikolenko, Ph.D.**

Author of [170+ research papers](#) in machine learning (ICML, CVPR, ACL, SIGIR, WSDM...) and algorithms (SIGCOMM, INFOCOM, ICNP...), several [books](#), including a bestselling “Deep Learning” book, lecture courses in ML, DL and other branches of computer science in St. Petersburg State Univ., NRU Higher School of Economics, Harbour Space University, and much more. Extensive experience in leading research and industrial projects in AI/ML.



[PDMI RAS website](#)

Core researchers of the *NLP gang* in the AI Lab



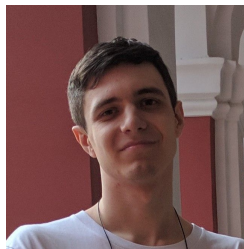
Elena Tutubalina

Ph.D., 40+ research papers, including ACL, ECIR, COLING, and top journals, head of several research projects and grants. Interests: NLP, multilingual models, drug discovery, etc.



Anton Alekseev

10+ research papers, including ECIR and COLING; exp. in research, software development, and teaching AI/ML. Interests: NLP, deep learning, digital humanities.



Ilya Shenbin

Publications at WSDM and ECIR. Interests: Bayesian learning, reinforcement learning, collaborative filtering.



Valentin Malykh

Ph.D., 30+ research papers, including NeurIPS, ACL, COLING, ECIR. Organizer of workshops and shared tasks. Interests: NLP, recsys, deep learning, reinforcement learning



Michael Vasilkovsky

The invited rock star from R&D, the actual OpenIE wizard. Ex-Neuromation, currently Snap Inc.

Our Computer Vision powerhouse is led by prof. Andrey Savchenko (D.Sc.).

We also work with the invited DL practitioners and academic researchers: Yaroslav Chizh, Zulfat Miftahutdinov, etc.

Постановка задачи

a function from a document to a **set of tuples indicating a semantic relation between a predicate phrase and its arguments** (Banko et al.,2007).

Mausam et al. 2012: Open IE looks for a phrase that expresses a **relation between a pair of arguments**

Wu and Weld (2008): an Open IE extractor should “produce one triple for every relation stated explicitly in the text, but is **not required to infer implicit facts**”

“John managed to open the door”

надо бы извлечь:

(John; managed to open; the door)

извлекать необязательно:

(John; opened; the door)



Взято из статьи Stanovsky-Dagan 2016

Пригодится впоследствии: что такое SRL?

Semantic Role Labeling aka *shallow semantic parsing*

То есть разметка слов согласно их семантической роли в предложении, пример:

Кошка съела рыбу

кошка = агентс, agent

рыба = пациентс (“жертва”), patient

Таких ролей много (детали нам сейчас не важны):

https://en.wikipedia.org/wiki/Thematic_relation



**кроси
вое**

Зачем это нужно? Примеры

• Question Answering

- Yan Z. et al. **Assertion-based QA with question-aware open information extraction** //Proceedings of the AAIL Conference on Artificial Intelligence. – 2018. – Т. 32. – №. 1.
- Khot T., Sabharwal A., Clark P. **Answering complex questions using open information extraction** //arXiv preprint arXiv:1704.05572. – 2017.

• Event Schema Induction

- Balasubramanian N. et al. **Generating coherent event schemas at scale** //Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. – 2013. – С. 1721-1731.

• Fact Salience

- Ponza M., Del Corro L., Weikum G. **Facts that matter** //Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. – 2018. – С. 1043-1048.

Question	who killed jfk
Method	Answer
PBQA	A ten-month investigation from November 1963 to September 1964 by the Warren Commission concluded that Kennedy was assassinated by Lee Harvey Oswald, acting alone, and that Jack Ruby also acted alone when he killed Oswald before he could stand trial.
MRC	Lee Harvey Oswald
ABQA	<Kennedy; was assassinated; by Lee Harvey Oswald>

полуструктурированные ответы с подробностями часто использовать удобнее, чем краткий ответ или целый абзац

ESI = выделение событий и их участников “без учителя”; в работе выделяют с помощью OpenIE и группируют по соупоминаниям

This work introduces fact salience: The task of generating a machine-readable representation of the most prominent information in a text document as a set of facts. We also present SALIE, the first fact salience system. SALIE is

План

1. OpenIE до 2016 года
 - a. TextRunner
 - b. ReVerb
 - c. OLLIE
 - d. OpenIE-4
2. Датасеты и бенчмарки
 - a. OIE2016
 - b. WiRe57
 - c. CaRB
 - d. Иное
3. Победный марш глубокого обучения
 - a. SpanOIE
 - b. IMoJIE
 - c. Mult^2OIE
 - d. OpenIE6
4. А что с русским языком?
5. Важные работы, о которых не говорили



Сверим терминологию

- Extractions *aka* извлечения *aka* тройки (кортежи)

(subject, relation, object) ≅ (arg0, predicate, arg1 [, arg2, arg3, ...])

Говорят, это не одно и то же, но мы будем использовать взаимозаменяемо

- **Part-of-speech (POS) tagging** = частеречная разметка
- **Syntax parsing** = синтаксический разбор
(чаще всего у нас будет в рамках грамматики зависимостей, см. далее)
- **Noun phrase (NP)** = именная группа
(словосочетание, в котором имя существительное является вершиной, то есть *главным словом*, определяющим характеристику всей составляющей)

2007 [TextRunner] Banko, Cafarella, Soderland, Broadhead, Etzioni

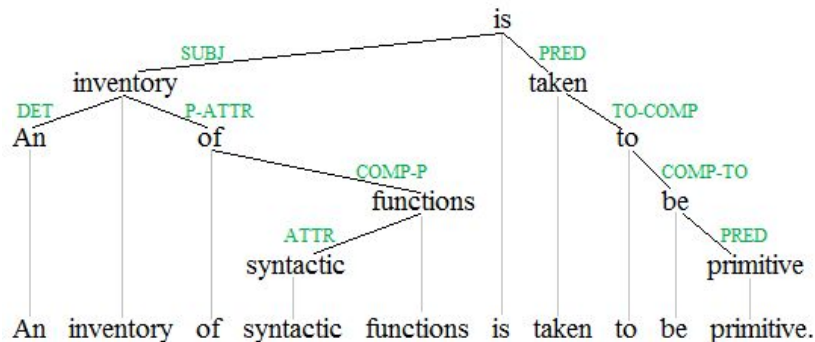
Open information extraction from the web. [IJCAI 2007](#), 2607 citations

- Вводят OpenIE как новую парадигму Information Extraction:
примерно до этого момента набор **видов отношений был конечным**
- Для работы с текстами интернетных качества и масштабов старые подходы с синтаксическими парсерами и NER-ами не подходили
- Предлагают сообществу инструмент TextRunner, “масштабируемый и не зависящий от предметной области”

2007 [TextRunner] Кратко об устройстве

1. Self-Supervised Learner:

- a. Строим деревья в рамках грамматики зависимостей по набору из нескольких тысяч предложений
- b. Находим базовые **именные группы**, т.е. без вложенных именных групп
- c. Для каждой пары именных групп в предложении ищем в дереве зависимостей кандидата в **связывающее их отношение**
- d. С помощью заранее заданных правил размечаем таких кандидатов-троек на подходящие и неподходящие
- e. Каждой тройке сопоставляем вектор **несинтаксических и нелексических признаков**, и обучаем наивный байесовский классификатор

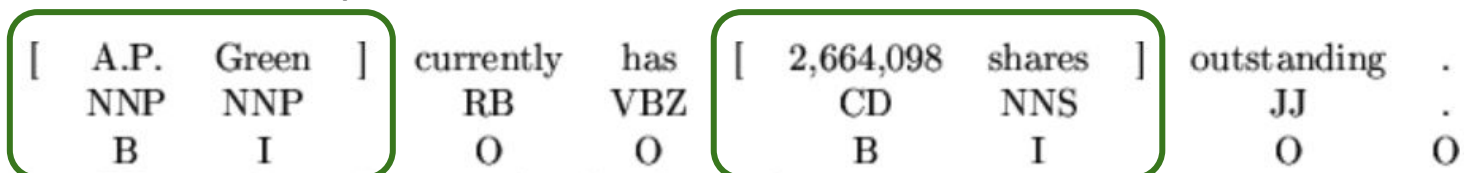
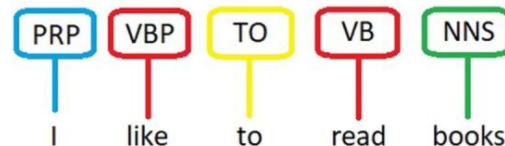


2007 [TextRunner] Кратко об устройстве

2. Single-Pass Extractor:

- Производим частеречную разметку
- На её основе выделяем именные группы (noun chunking)
- Эвристиками отбрасываем лишнее
- Строим тройки-кандидаты на основе разметки, соединяя именные группы
- Вычисляем оценки обученного на прошлом шаге классификатора для этих троек
- Тройки с низкими оценками отбрасываем, остальное сохраняем

POS Tagging



Base NP transformation to a classification task

2007 [TextRunner] Кратко об устройстве

3. Redundancy-based Assessor:

“схлопывает” неточные дубликаты извлечённых из всего корпуса троек и запоминает для каждой, из какого числа предложений каждая извлечена; назначает вероятности

Сравнивали с другой “не вполне unsupervised” системой, которая извлекала факты из веба, KnowItAll: зафиксировали 10 отношений (count > 1000 в 9-милл. интернет-корпусе)

Остальной анализ -- в статье

was originally developed by =>
was developed by

	Average Error rate	Correct Extractions
TEXTRUNNER	12%	11,476
KNOWITALL	18%	11,631

Table 1: Over a set of ten relations, TEXTRUNNER achieved a 33% lower error rate than KNOWITALL, while finding approximately as many correct extractions.

2011 [ReVerb] Fader, Soderland, Etzioni

Identifying relations for open information extraction. [1441 цитата](#) EMNLP2011

- На тот момент -- новый мощный SoTA
- TextRunner и WOE
 - эвристиками размечали подкорпус
 - обучали извлекатель отношений (~ relation, predicate)
 - для пары кандидатов аргументов (именных групп, NP) извлекателем выделяли часть предложения как relation

...как следствие -- много шума, и распределение троек фактически задано правилами-эвристиками; есть и проблема “поломанных” текстов

ReVerb

1. Найти отношения, удовлетворяющие **синтаксическим и лексическим ограничениям**
2. Подобрать для них именные группы (NP) как аргументы
3. Оценить “уверенность” в построенной тройке с помощью логистической регрессии

2011 [ReVerb] Синтаксические ограничения

$V \mid VP \mid VW^*P$

V = verb particle? adv?

W = (noun | adj | adv | pron | det)

P = (prep | particle | inf. marker)

The syntactic constraint requires the relation phrase to match the POS tag pattern shown in Figure 1. The pattern limits relation phrases to be either a verb (e.g., *invented*), a verb followed immediately by a preposition (e.g., *located in*), or a verb followed by nouns, adjectives, or adverbs ending in a preposition (e.g., *has atomic weight of*). If there are multiple possible matches in a sentence for a single verb, the longest possible match is chosen. Finally, if the pattern matches multiple adjacent sequences, we merge them into a single relation phrase (e.g., *wants to extend*). This refinement enables the model to readily handle relation phrases containing multiple verbs. A



После частеречной разметки отношение должно удовлетворять регулярному выражению

Ограничение -- очень строгое, опыт показал, что может теряться **15% извлечений**

Зато тогда не будет таких случаев:

The syntactic constraint eliminates the incoherent relation phrases returned by existing systems. For example, given the sentence

Extendicare agreed to buy Arbor Health Care for about US \$432 million in cash and assumed debt.

TEXTRUNNER returns the extraction

(Arbor Health Care, for assumed, debt).

2011 [ReVerb] Лексические ограничения

...но заданное регуляркой правило будет извлекать подчас слишком редкие и длинные отношения!

Хотелось бы извлекать отношения, которые присущи **самым разным парам аргументов**

- 1) Наберём заранее кандидатов в предикаты,
- 2) подберём для них соседние именные группы как аргументы,
- 3) запомним только те отношения, с которыми попадаются не менее **k = 20 различных пар аргументов**,
- 4) их далее и будем извлекать.

The Obama administration is offering only modest greenhouse gas reduction targets at the conference.

The POS pattern will match the phrase:

is offering only modest greenhouse gas reduction targets at

2011 [ReVerb] Итог

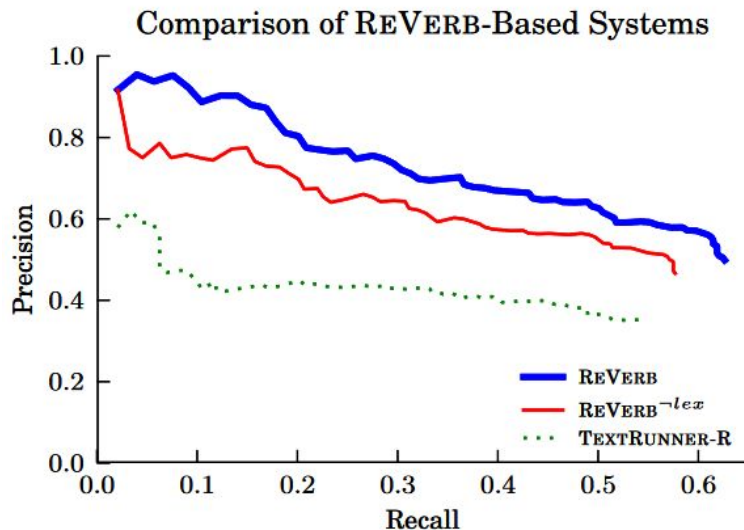


Figure 3: The lexical constraint gives REVERB a boost in precision and recall over REVERB^{-lex}. TEXTRUNNER-R is unable to learn the model used by REVERB, which results in lower precision and recall.

- Для оценки уверенности логистической регрессией используются вручную подготовленные признаки и 1000 размеченных предложений
- Извлечения из 500 предложений оценили ассессоры

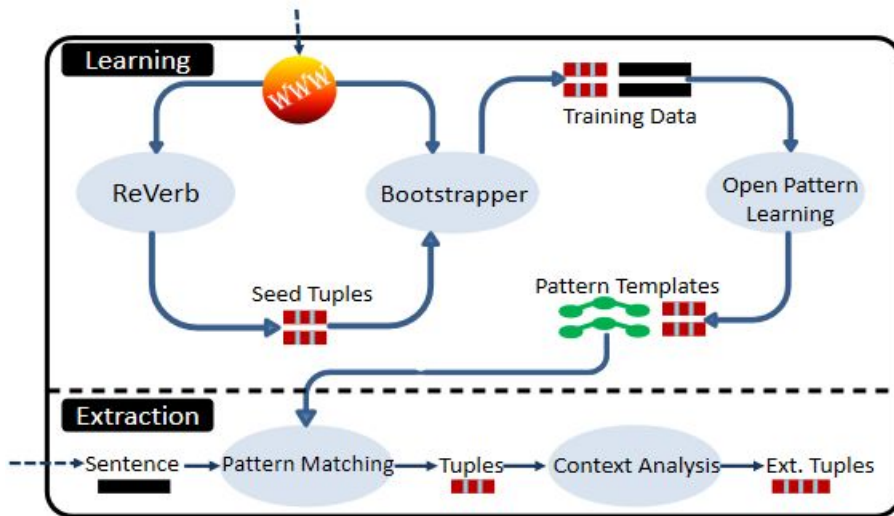
In sum, our model is by no means complete. However, we have empirically shown that the majority of binary verbal relation phrases in a sample of Web sentences are captured by our model. By

2012 [OLLIE] Mausam, Schmitz, Soderland, Bart, Etzioni

Open Language Learning for Information Extraction [783 цитаты](#) EMNLP2021

Предшественники ориентируются только на глаголы как на предикаты и не отличают, например, косвенную речь от фактографической информации

OLLIE целенаправленно
это исправляет



2012 [OLLIE]

W = WOE

R = ReVerb

O = Ollie

Многие отношения не под силу ReVerb и WOE

WOE работает похожим на OLLIE образом: ищет в тексте википедии объекты из инфобоксов и пытается вывести правила извлечения на основе синтаксических деревьев

1. “After winning the Superbowl, the Saints are now the top dogs of the NFL.”
O: (the Saints; win; the Superbowl)
2. “There are plenty of taxis available at Bali airport.”
O: (taxis; be available at; Bali airport)
3. “Microsoft co-founder Bill Gates spoke at ...”
O: (Bill Gates; be co-founder of; Microsoft)
4. “Early astronomers believed that the earth is the center of the universe.”
R: (the earth; be the center of; the universe)
W: (the earth; be; the center of the universe)
O: ((the earth; be the center of; the universe)
AttributedTo believe; Early astronomers)
5. “If he wins five key states, Romney will be elected President.”
R,W: (Romney; will be elected; President)
O: ((Romney; will be elected; President)
ClausalModifier if; he wins five key states)

2012 [OLLIE] Bootstrapping

1. ReVerb извлекаются тройки из веб-корпуса [ClueWeb](#)
2. Оставляют только те,
 - a. которые встретились хотя бы дважды и
 - b. у которых аргументы -- имена собственные
3. Делаем запрос в корпус со словами из каждой тройки (18 млн. предложений)
4. Т. о. имеем различные варианты формулировок одних и тех же фактов (не всегда так, некоторое отбрасываем эвристиками)

Paul Annacone is the coach of Federer

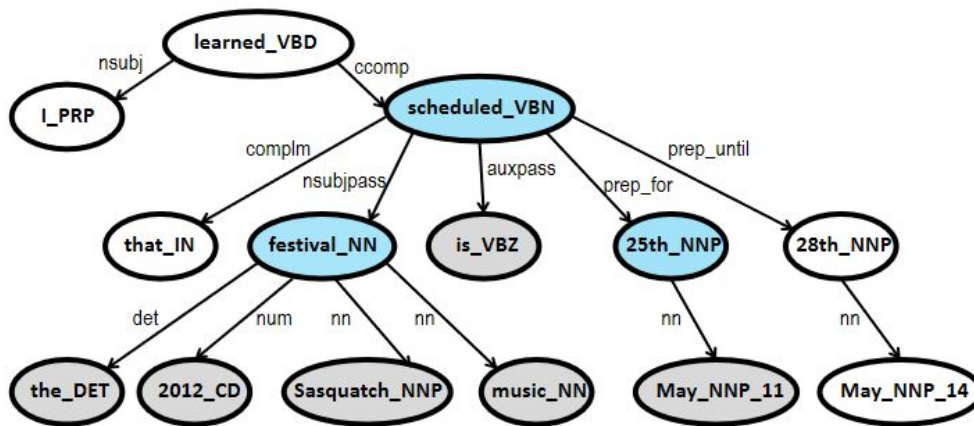
(Paul Annacone; is the coach of; Federer)

Query: Paul, Federer, Annacone, coach

*e.g.: Now coached by Annacone,
Federer is winning more titles than ever*

2012 [OLLIE] Open Pattern Learning

Как и в WOE, строим правила извлечения интересующих нас вещей с помощью синтаксиса и лекс. ограничений (много тонкостей, так что без подробностей)



(the 2012 Sasquatch Music Festival; is scheduled for; May 25th)

Extraction Template	Open Pattern
1. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓{prep_*}↓ {arg2}
2. (arg1; {rel}; arg2)	{arg1} ↑nsubj↑ {rel:postag=VBD} ↓doobj↓ {arg2}
3. (arg1; be {rel} by; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓agent↓ {arg2}
4. (arg1; be {rel} of; arg2)	{rel:postag=NN;type=Person} ↑nn↑ {arg1} ↓nn↓ {arg2}
5. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {slot:postag=VBN;lex ∈ announce name choose...}

2012 [OLLIE] Results

...Context analysis -- частично правила, частично логрег

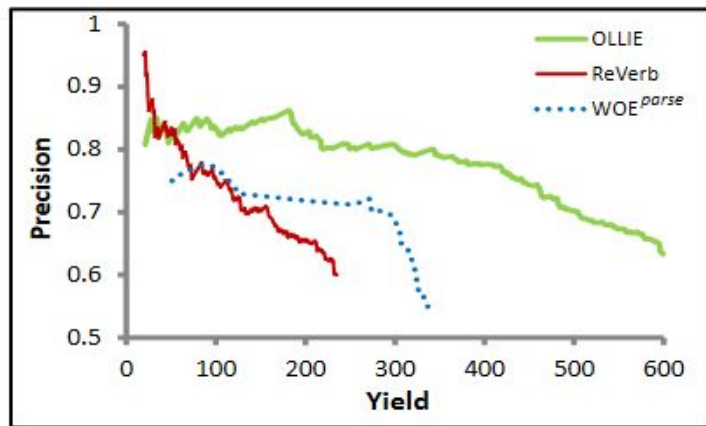


Figure 5: Comparison of different Open IE systems. OLLIE achieves substantially larger area under the curve than other Open IE systems.

Relation	OLLIE	REVERB	incr.
<i>is capital of</i>	8,566	146	59x
<i>is president of</i>	21,306	1,970	11x
<i>is professor at</i>	8,334	400	21x
<i>is scientist of</i>	730	5	146x

Figure 6: OLLIE finds many more correct extractions for relations that are typically expressed by noun phrases – up to 146 times that of REVERB. WOE^{parse} outputs no instances of these, because it does not allow nouns in the relation. These results are at point of maximum yield (with comparable precisions around 0.66).

2012 [OLLIE] Results

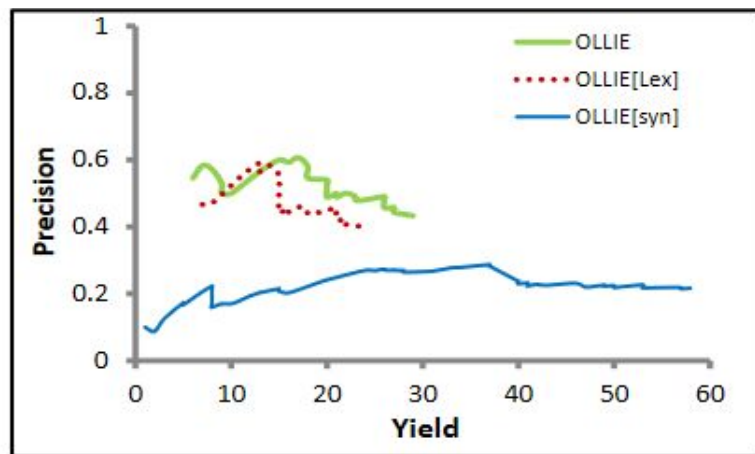


Figure 7: Results on the subset of extractions from patterns with semantic/lexical restrictions. Ablation study on patterns with semantic/lexical restrictions. These patterns without restrictions (OLLIE[syn]) result in low precision. Type generalization improves yield compared to patterns with only lexical constraints (OLLIE[lex]).

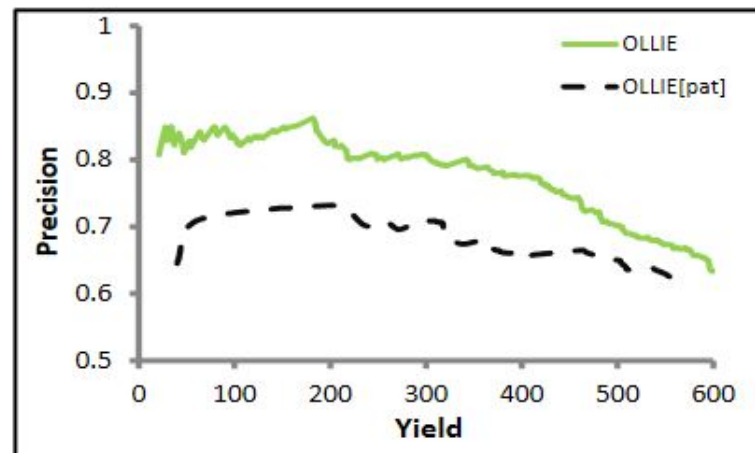


Figure 8: Context analysis increases precision, raising the area under the curve by 19%.

2016 [OpenIE-4] Christensen, Mausam, Soderland, Etzioni, Pal

An Analysis of Open Information Extraction based on Semantic Role Labeling". KCAP2011
"Demonyms and Compound Relational Nouns in Nominal Open IE" AKBC workshop at NAACL2016

github: [knowitall/openie](https://github.com/knowitall/openie) ИЛИ [allenai/openie-standalone](https://github.com/allenai/openie-standalone) (use with care)

- Наследник Ollie, важный бейзлайн, который не так-то и просто побить
- Является комбинацией **SRLIE (2011)**...

Выводы обученных на PropBank SRL-моделей преобразовываются под задачу OpenIE и при неограниченном времени работы работают прекрасно; ещё лучше, если комбинировать с **TextRunner**

arg0 Eli Whitney
rel created
arg1 the cotton gin

binary tuple

arg0 Eli Whitney
rel created (arg1) in
arg1 the cotton gin
arg2 1793

n-ary tuple

vs

A0 Eli Whitney
verb created
A1 the cotton gin
temporal in 1793

2016 [OpenIE-4]

...и Relnoun

Иногда отношения сформулированы не в формате подлежащее-сказуемое-дополнение

Phrase	RELNOUN 1.1	RELNOUN 2.2
“ <i>United States President Obama</i> ”		(Obama, [is] President [of], United States)
“ <i>Seattle historian Feliks</i> ”	(Feliks, [is] historian [of], Seattle)	(Feliks, [is] historian [from], Seattle)
“ <i>Japanese foreign minister Kishida</i> ”		(Kishida, [is] foreign minister [of], Japan)
“ <i>GM Deputy Chairman Lutz</i> ”		(Lutz, [is] Deputy Chairman [of], GM)

Table 2: Comparison of RELNOUN 1.1 and RELNOUN 2.2 on some phrases

Правила и бутстреппинг

План

- ~~1. OpenIE до 2016 года~~
 - ~~a. TextRunner~~
 - ~~b. ReVerb~~
 - ~~c. OLLIE~~
 - ~~d. OpenIE-4~~
2. Датасеты и бенчмарки
 - a. OIE2016
 - b. WiRe57
 - c. CaRB
 - d. Иное
3. Победный марш глубокого обучения
 - a. SpanOIE
 - b. IMoJIE
 - c. Mult^2OIE
 - d. OpenIE6
4. А что с русским языком?
5. Важные работы, о которых не говорили

Отступление: поговорим о данных

- Несмотря на то, что некоторые системы на правилах в OpenIE -- сильные бейзлайны, **сейчас** без непосредственного обучения в извлечении отношений никуда
- Почти все рассмотренные работы (и часть других, см. конец презентации) считали оценки качества по-разному и на разных тестовых выборках

Примерно с 2016 года всё

- обучается на синтетике и
- оценивается на одних и тех же тестовых выборках одними и теми же скриптами (или их апгредами)

2016 [OIE2016] Stanovsky & Dagan

Creating a Large Benchmark for Open Information Extraction [85 цитат](#) EMNLP2016

- Мотивация:
 - давно был нужен единый датасет для оценки -- и больше, чем прочие,
 - ...а также единый способ оценки,
 - ...который бы ориентировался в том числе и на полноту.
- Взяли сравнительно большой набор данных QA-SRL и преобразовали его в датасет для OpenIE, **OIE16**
- Вместе с ним выпустили скрипт оценки качества: [gabrielStanovsky/oie-benchmark](https://github.com/gabrielStanovsky/oie-benchmark)
- Несколько лет был золотым стандартом, используют до сих пор

2016 [OIE2016] Уточнение постановки задачи (наканецта)

Присущие всем более ранним системам свойства, важные для OpenIE:

1. **Assertedness.** Извлечённое должно подтверждаться текстом, никаких подразумеваемых смыслов. В предикаты входят `not` и модальные глаголы

(Sam; **succeeded in convincing**; John)
(**Sam; convinced; John**)
(John; **could not join**; the band)

2. **Minimal propositions.** Чем меньше извлечено, тем лучше -- но без потери информации

Bell distributes electronic and building products

(Bell, distributes, electronic products)
(Bell, distributes, building products).

3. **Completeness and open lexicon.** Максимальное число пропозиций из текста, желательно без привязки к наперёд заданным словарям

tence. In practice, most current Open IE systems limit their scope to extracting verbal predicates, but consider all possible verbs without being bound to a pre-specified lexicon.

3. Кроссворд "По мотивам поэзии Пушкина".

По горизонтали:

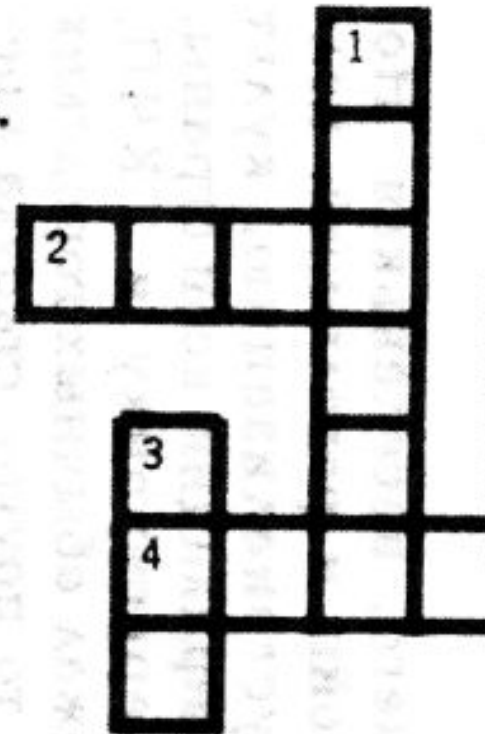
2. А мать грозит ему во что?

4. А кто грозит ему в окно?

По вертикали:

1. А мать чего ему в окно?

3. А мать грозит кому в окно?



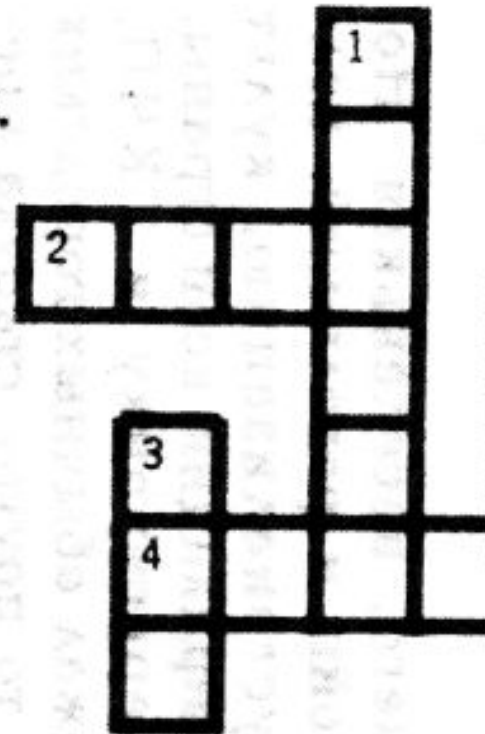
3. Кроссворд "По мотивам поэзии Пушкина".

По горизонтали:

2. А мать грозит ему во что?
4. А кто грозит ему в окно?

По вертикали:

- ~~1. А мать чего ему в окно?~~
3. А мать грозит кому в окно?



2016 [OIE2016] Как это сделано

SRL часто рассматривается как задача
ответа на **role questions**

В QA-SRL (He et al. 2015) 3200 предложений,
к каждому предикату каждого предложения
есть список вопросов, к каждому вопросу есть список ответов

В QA-SRL -- предикаты и аргументы, но *не вполне* подходящие под 3 требования

Но несколькими простыми правилами строятся тройки; например,

- в качестве аргументов не рассматриваются местоимения,
- кореференция не разрешается и т. п.

Consider the sentence “*Giles Pearman, Microsoft’s director of marketing, left his job*” and the target predicate **left**. The QA-SRL annotation consists of the following pairs: (1) *Who left something?* {**Giles Pearman; Microsoft’s director of marketing**} and (2) *what did someone leave?* **his job**.⁵

2016 [OIE2016] Как это оценивать

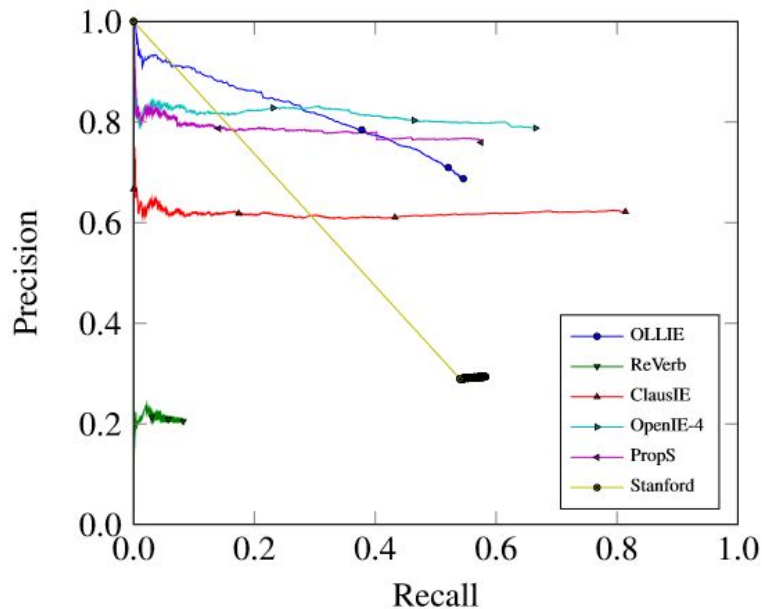
Итог: 10,359 Open IE extractions
over 3200 sentences from 2 domains
(Wall Street Journal & Wikipedia)

Судя по скрипту*, извлечённое сопоставляется
“золотой” тройке только если

1. хотя бы четверть* токенов в тройке совпадают,
2. есть пересечение хотя бы по одному “осмысленному” токену в предикате

Доля угаданных

- относительно золотого стандарта -- полнота
- относительно числа извлечённых -- точность



3. *Stanford Open IE* assigns confidence of 1 to 94% of its extractions, explaining its low precision.

2018 [WiRe57] L chelle, Gotti, Langlais

WiRe57 : A **Fine-Grained** Benchmark for Open Information Extraction
(Proceedings of the 13th Linguistic Annotation Workshop)

arxiv: [1809.08962](https://arxiv.org/abs/1809.08962) github: [rali-udem/WiRe57](https://github.com/rali-udem/WiRe57)

- Мотивация -- такая же
- Показаны недостатки OIE2016
- Разобраны 57 предложений и 347 кортежей*
- Предложен новый способ оценки
- Опубликована [бесценная инструкция по разметке](#)
- Пересмотрены прежние результаты

* Wi Re = 3 текста из **Wikipedia** + 2 из **Reuters**

2018 [WiRe57] Критика OIE2016

- QA-SRL привязан к конкретным **предикатам**, поэтому теряется часть отношений
- Из QA-SRL проникли слова, которых нет в исходных данных

В примере в статье глаголы **said**, **overcome** рассмотрены (так как к ним есть вопросы в данных QA-SRL), а **is** -- нет, при этом *is*-а -- важное отношение

pressed facts in the end. For instance, the uninformative triple (*a manufacturer ; might get ; something*) is generated from the sentence “...and if a manufacturer is clearly trying to get something out of it ...”, with the same added “*might*”.

2018 [WiRe57] Критика OIE2016 (бадибэг)

OIE2016 не штрафует за слишком избыточные извлечения и за попадание токена из отношения в аргумент

Легко хакнуть кодом в 25 строк: если предложение -- последовательность токенов $w_0 w_1 \dots w_n$, то скрипт, который будет генерировать все тройки вида

$(w_0; w_1; w_2 \dots w_n)$
 $(w_0; w_1 w_2; w_3 \dots w_n)$
...

всех победит

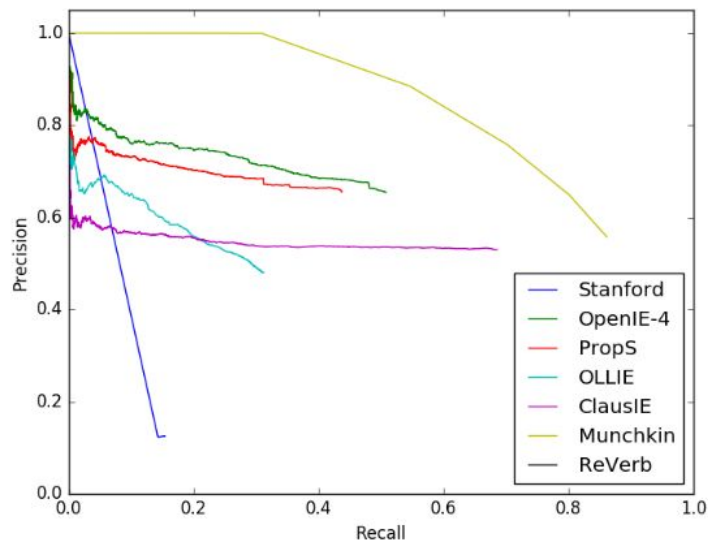


Figure 1: Performance metrics must take span precision into account. The 25-line long Munchkin script returns variations of the full sentence (with decreasing confidence) and is not penalized by the evaluation script of the latest benchmark (Stanovsky and Dagan, 2016). Its superior performance is artificially inflated.

2018 [WiRe57] Scorer

Оценивают качество, используя оценки в том числе на уровне токенов

Точность -- доля правильно выделенных токенов в предсказании

Полнота -- доля правильно выделенных токенов в золотом стандарте

Считаем ВОЗМОЖНО совпадающими, если в предикате и первых двух аргументах есть соответствующие пересечения хотя бы по одному токenu

Вычисляем для таких F1 и **жадно** **исключаем от лучших к худшим**

A predicted tuple t_i may match a reference tuple g_j from the same sentence if they share at least one word from each of the relation, first and second arguments, that is iff (w_{a_1}, w_r, w_{a_2}) exist such that $w_1 \in g_j^{a_1} \cap t_i^{a_1}$, $w_r \in g_j^r \cap t_i^r$ and $w_2 \in g_j^{a_2} \cap t_i^{a_2}$.

$$\text{precision}(t_i, g_j) = \frac{\sum_k |t_i^{p_k} \cap g_j^{p_k}|}{|t_i|}$$

$$\text{recall}(t_i, g_j) = \frac{\sum_k |t_i^{p_k} \cap g_j^{p_k}|}{|g_j|}$$

$$F_1 = \frac{2 p r}{p + r}.$$

2018 [WiRe57] Итоговая оценка

$m(i)$ -- функция, сопоставляющая наш жадный выбор

Т. е. имеем что-то вроде micro-averaging, используя количество токенов вместо количества извлечённых кортежей

$$\text{precision}_{\text{sys}} = \frac{\sum_i^n \left(\sum_k |t_i^{p_k} \cap g_{m(i)}^{p_k}| \right)}{\sum_i^n |t_i|}$$

$$\text{recall}_{\text{sys}} = \frac{\sum_j^N \left(\sum_k |t_{m(j)}^{p_k} \cap g_j^{p_k}| \right)}{\sum_j^N |g_j|}$$

$$F_{1\text{sys}} = \frac{2 p_{\text{sys}} r_{\text{sys}}}{p_{\text{sys}} + r_{\text{sys}}}$$

2018 [WiRe57] Собственно датасет

5 документов, 57 предложений, 347 кортежей-извлечений

- Вставлены (в квадратных скобках) слова, нужные для осмысленности извлечённого кортежа, которых нет в исходном предложении **[scorer не использует]**
- Помечен факт “непрямого” высказывания
- Разрешена кореференция **[scorer не использует]**
- Явно выполнена токенизация
- Проставлены соответствия токенам в исходном предложении, если это возможно
- Для воспроизводимости выложены предсказания нескольких систем

Two annotators (authors of this paper) first independently extracted tuples from the documents, based on a first version of the annotation guidelines which quickly proved insufficient to reach any significant agreement. The two sets of annotations were then merged, and the guidelines rectified along the way in order to resolve the issues that arose. After merging, a quick test on a few additional sentences from a different document showed a much improved agreement, more than half of extractions matching exactly and the remaining missing a few details. The guidelines are detailed in the next sections.

2019 [CaRB] Bhardwaj, Aggarwal, Mausam

A crowdsourced benchmark for open IE [9 цитат](#) EMNLP2019
github: [dair-iitd/CaRB](https://github.com/dair-iitd/CaRB)

- OIE2016 и ReVis имеют схожие проблемы в данных и “скоринге”, WiRe57 жадно сопоставляет и вообще слишком маленький
- Краудсорсинг! Amazon MTurk, разметка в три шага:
 - identifying the relation,
 - identifying the arguments for that relation, and
 - optionally identifying the location and time attributes for the tuple
- Потом сверка со своей разметкой 50 предложений
- Иная оценка качества (см. дальше)

2019 [CaRB] Оценка качества (scoring)

Все аргументы, начиная со второго, объединяются в один (как и везде)

Оценка:

1. Для каждой пары извлечений вычисляем точность и полноту по токенам; матрицы: **колонки** -- предсказания, **строки** -- золотой стандарт
2. Полнота = средняя максимальная полнота по **строкам**
3. Точность = средняя точность при жадном сопоставлении
4. При этом -- **tuple match**, а не объединение всего в один мешок, как в OIE2016

Зачем **multi-match** для полноты?

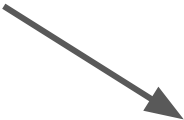
не наказываем за слияние информации из неск. gold-кортежей в один предсказанный

Почему **single-match** для точности?

В CaRB кортежи атомарны, разделять нельзя + помогает от избыточных кортежей

2019 [CaRB] Оценка качества (scoring)


Системой №1 информация не утеряна, поэтому здесь полнота = 1, а у системы №2 -- меньше



Sentence	<i>I ate an apple and an orange.</i>	(prec,rec)	
Gold	(I; ate; an apple) (I; ate; an orange)	OIE2016	CaRB
System 1	(I; ate; an apple and an orange)	(1,0.5)	(0.57,1)
System 2	(I; ate; an apple)	(1,0.5)	(1,0.87)

Table 2: One-to-One Match vs. Multi Match

Справедливое наказание за перепутанные аргументы



Sentence	<i>I ate an apple.</i>	(prec,rec)	
Gold	(I; ate; an apple)	OIE2016	CaRB
System 1	(I; ate; an apple)	(1,1)	(1,1)
System 2	(ate; an apple; I)	(1,1)	(0,0)

Table 3: Tuple Match vs. Lexical Match

2019 [CaRB] Перестановка!

Неизменны только успехи OpenIE4

Позвали 4-х аннотаторов с MTurk, прошедших обучение, и попросили их сравнить выводы ClausIE и PropS на 100 предложениях

Со счётом 69:15 (16 -- ничьи) победила ClausIE, что лишний раз **валидирует предложенный бенчмарк**

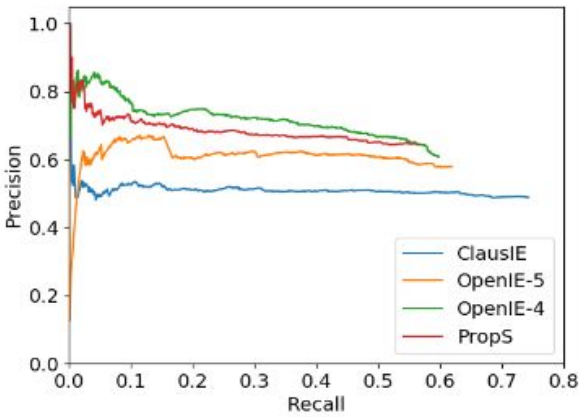


Figure 1: Comparison of Open IE systems using DIE2016

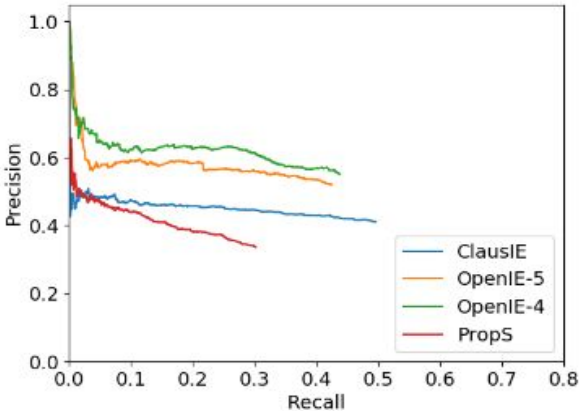


Figure 2: Evaluation of Open IE systems using CaRB

Бенчмарки, которые также представляют интерес

2017	ReIVis	Schneider R. et al. Analysing Errors of Open Information Extraction Systems //Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems. – 2017. – С. 11-18.	1707.07499	SchmaR/ReIVis (пусто, а видео недоступно)	4522 sentences and 11243 n-ary tuples, большая часть из OIE16 , но другой скоринг	
2020	Re-OIE2016	Zhan J., Zhao H. Span model for open information extraction on accurate corpus //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 05. – С. 9523-9530.	1901.10879	zhanjunlang/ Span_OIE	Переразмеченный OIE2016, его критикуют за не адекватную задачу обработки сочинения (coordination)	
2021	LSOIE	Solawetz J., Larson S. LSOIE: A Large-Scale Dataset for Supervised Open Information Extraction //Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. – 2021. – С. 2595-2600.	2101.11177	Jacobsolawetz/ large-scale-oie	QA-SRL 2.0, переделанный в OpenIE; отличная от более ранних датасетов политика; зато в 20 раз больше	
			LSOIE-wiki	Wiki, Wikinews	24,296	56,662
			LSOIE-sci	Science	47,998	97,550

План

1. ~~OpenIE до 2016 года~~
 - a. ~~TextRunner~~
 - b. ~~ReVerb~~
 - c. ~~OLLIE~~
 - d. ~~OpenIE-4~~

2. ~~Датасеты и бенчмарки~~
 - a. ~~OIE2016~~
 - b. ~~WiRe57~~
 - c. ~~CaRB~~
 - d. ~~Иное~~

3. Победный марш глубокого обучения
 - a. SpanOIE
 - b. IMoJIE
 - c. Mult^2OIE
 - d. OpenIE6

4. А что с русским языком?

5. Важные работы, о которых не говорили

Отступление: два подхода к OpenIE

Generation: порождение по слову за раз

- + позволяет бороться с проблемой избыточности вывода
- извлечение, как правило, очень медленное

Пример: IMoJIE, CopyAttention

Labeling: каждое слово помечается как одно из:

S (subject), **R** (relation), **O** (object), **N** (none)

- + извлечение происходит **быстро**
- много дублирования и избыточности, качество, как правило, ниже

Пример: RnnOIE

2019 [SpanOIE] Junlang Zhan, Hai Zhao

Span Model for Open Information Extraction on Accurate Corpus

arxiv: [1901.10879](https://arxiv.org/abs/1901.10879) github: [zhanjunlang/Span_OIE](https://github.com/zhanjunlang/Span_OIE)

Замысел: выделяем

- 1) потенциальные отношения (т.е. relation, predicate, etc.)
как **участки предложения**,
- 2) затем пытаемся классифицировать остальные участки
как аргументы (subject & object)

Иногда выделяют в отдельный span-based-подход

(Также в этой работе представляют Re-OIE2016)

2019 [SpanOIE] Перебор подстрок

Перебираются подстроки
для поиска предиката, затем
для прочих аргументов

В ходе обучения есть
ограничения на перебор
“участков”: длина, пересечение
с “участком” предиката,
синтаксические ограничения

$$\arg \max_{(i',j') \in S} SCORE_l(i', j'), l \in L \quad (1)$$

where

$$\begin{aligned} SCORE_l(i, j) &= P_\theta(i, j|l) \\ &= \frac{\exp(\phi_\theta(i, j, l))}{\sum_{(i',j') \in S} \exp(\phi_\theta(i', j', l))} \end{aligned} \quad (2)$$

and ϕ_θ is a trainable scoring function with parameters θ . To train the parameters θ , in the training set, for each sample X and the gold structure Y^* , we minimize the cross-entropy loss:

$$l_\theta(X, Y^*) = \sum_{(i,j,l) \in Y^*} -\log P_\theta(i, j|l) \quad (3)$$

2019 [SpanOIE] Модель: Bi-LSTM over GloVe

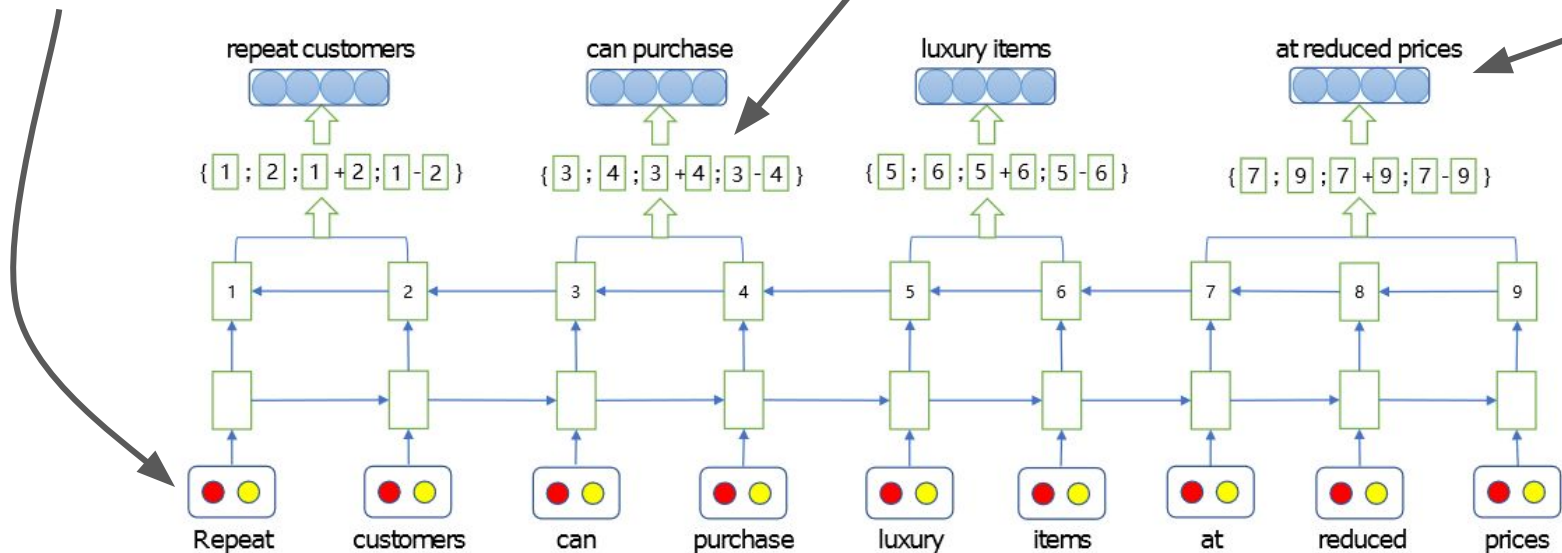
GloVe и эмбединги части речи, факта присутствия в предикате и типа зависимости

Представления по двум направлениям biLSTM конкатенируются в h , а потом эти h для краёв участка i и j конкатенируются вот так:

$$x_i = emb(w_i) \oplus emb(pos(w_i)) \oplus emb(p(w_i)) \oplus emb(dp(w_i))$$

$$f_{span}(s_{i:j}) = h_i \oplus h_j \oplus h_i + h_j \oplus h_i - h_j$$

Линейный слой и софтмакс для каждого рассматриваемого участка



2019 [SpanOIE] Результаты

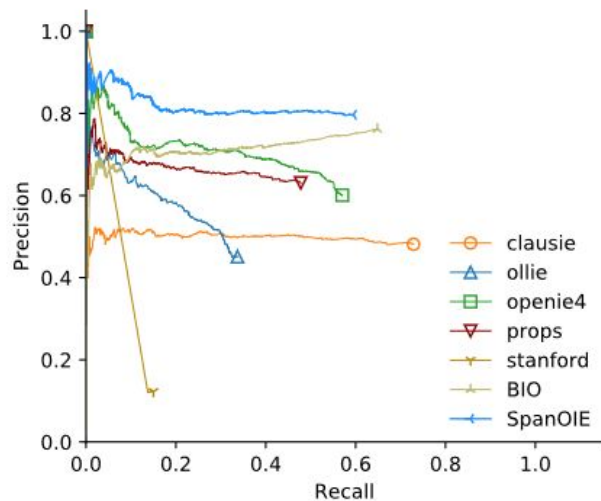


Figure 2: The P-R curve of different Open IE systems on OIE2016

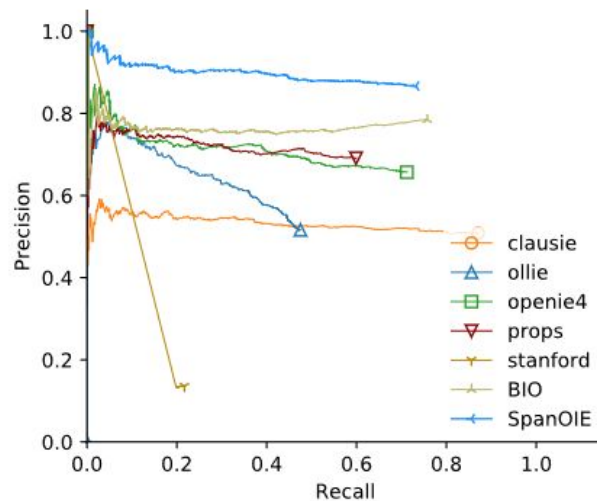


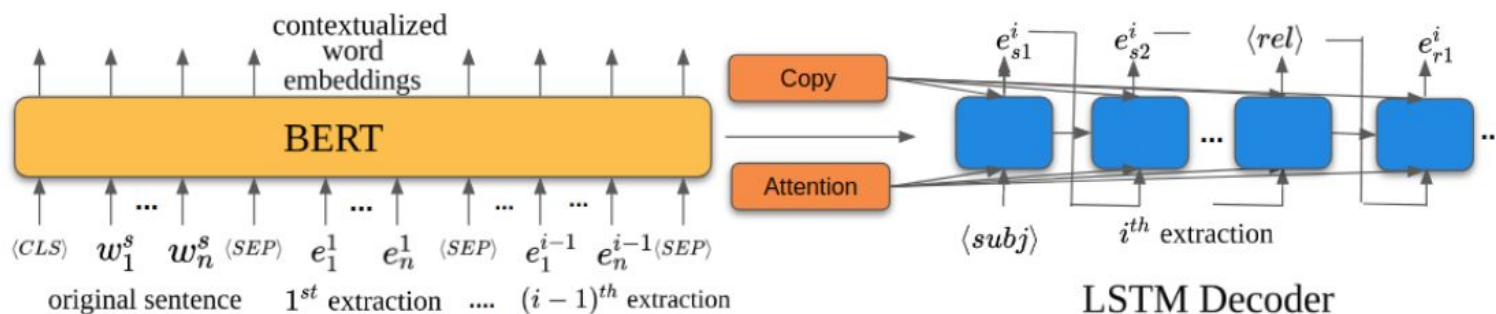
Figure 3: The P-R curve of different Open IE systems on Re-OIE2016

2020 [IMoJIE] Kolluru, Aggarwal, Rathore, Mausam, Chakrabarti

Iterative Memory-Based Joint Open Information Extraction [7 цитат](#) ACL2020
arxiv: [2005.08178](#) github: [dair-iitd/imojie](#)

- Позиционируют себя как улучшение CopyAttention (по сути, seq2seq)
- Проблемы CopyAttention:
 - не учитывает, что из длинных предложений обычно больше извлечений,
 - “заикается” -- порождает избыточные результаты из-за использования Beam Search
- ...Т.е. надо сделать так, чтобы декодировщик помнил, что уже извлечено из текущего предложения
- Улучшают датасет для обучения: вместо отбора вывода OpenIE4 -- механизм **Score-and-Filter** для объединения RnnOIE, OpenIE4 (высокая точность) и ClausIE (высокая полнота))

2020 [IMoJIE] Модель



- Наконец используется BERT
- Просто добавляем эмбединги очередного извлечения в кодировщик, пока не декодируется **EndOfExtractions**

Сразу SoTa, но работает долго

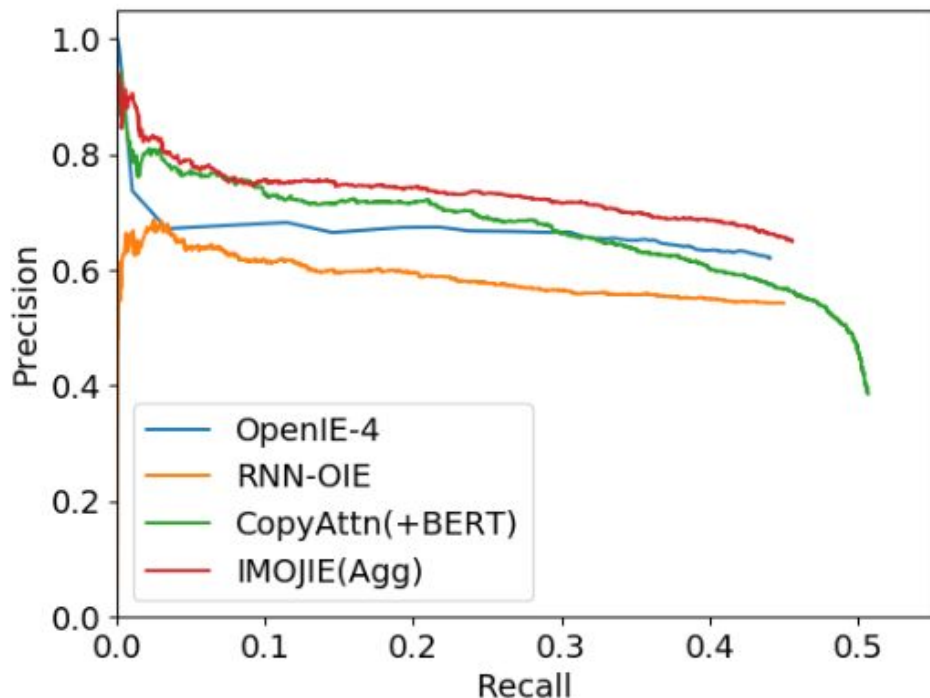
LSTM Decoder

The overall process can be summarized as:

1. Pass the sentence through the Seq2Seq architecture to generate the first extraction.
2. Concatenate the generated extraction with the existing input and pass it again through the Seq2Seq architecture to generate the next extraction.
3. Repeat Step 2 until the *EndOfExtractions* token is generated.

IMoJIE is trained using a cross-entropy loss between the generated output and the gold output.

2020 [IMoJIE] Результаты на CaRB



System	Metric		
	Opt. F1	AUC	Last F1
Stanford-IE	23	13.4	22.9
OllIE	41.1	22.5	40.9
PropS	31.9	12.6	31.8
MinIE	41.9	-*	41.9
OpenIE-4	51.6	29.5	51.5
OpenIE-5	48.5	25.7	48.5
ClausIE	45.1	22.4	45.1
CopyAttention	35.4	20.4	32.8
RNN-OIE	49.2	26.5	49.2
Sense-OIE	17.2	-*	17.2
Span-OIE	47.9	-*	47.9
CopyAttention + BERT	51.6	32.8	49.6
IMOJIE	53.5	33.3	53.3

Также значительная часть статьи посвящена методу Score-and-Filter, использующему целочисленное программирование для выбора извлечений, и оценке качества для разных комбинаций моделей-источников обучающих данных. Интересно.

2020 [Multi²OIE] Ro, Lee, Kang

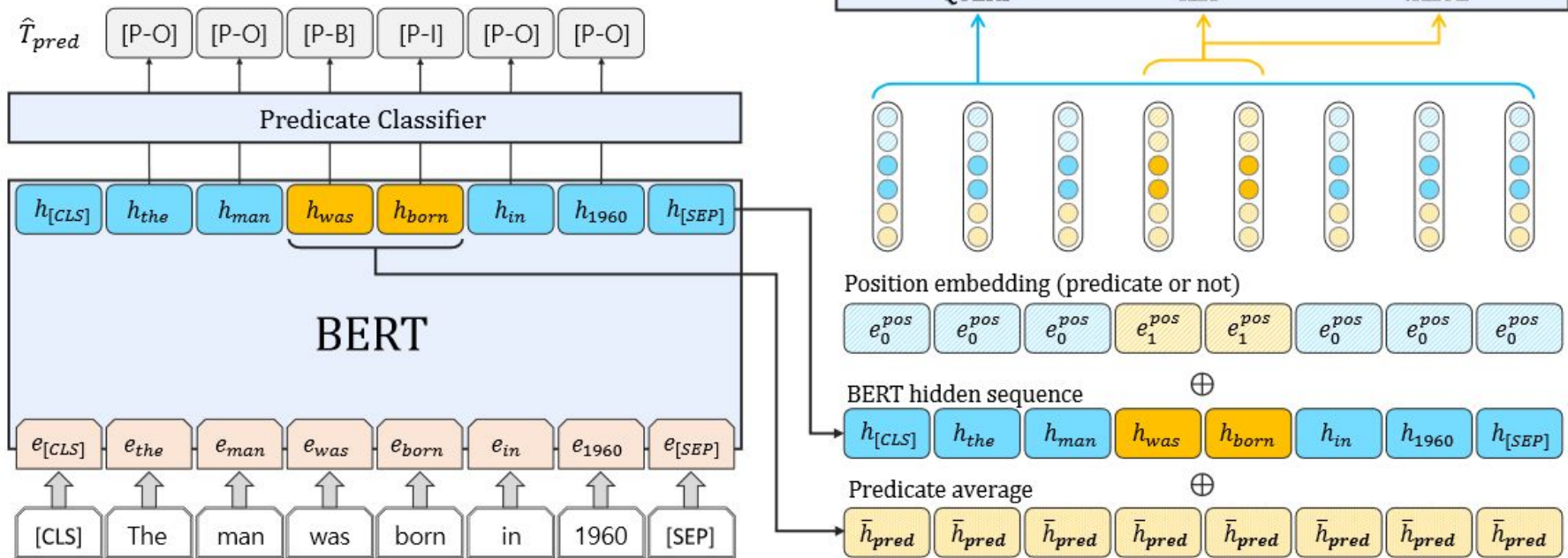
Multilingual Open Information Extraction based on Multi-Head Attention with BERT
[5 цитат](#) EMNLP2020, arxiv:[2009.08128](#), github: [youngbin-ro/Multi2OIE](#)

С каким прицелом разрабатывалось:

- Multi-Head Attention “смотрит” сразу на всю последовательность, в отличие от рекуррентных сетей
- многоязычный BERT в основе позволяет работать с другими языками без обучения на данных на этих языках
- “несколько менее” авторегрессивное порождение извлечений ускоряет предсказания

Перевели Re-OIE2016 на испанский и португальский

- **Sentence** : < The man was born in 1960 >
- **Predicate** : < was born >
- **Argument0** : < The man >
- **Argument1** : < in 1960 >



Процесс повторяется для каждого предиката снова

Напоминание

...как self-attention устроен
в “обычном” трансформере

К каждому эмбедингу x применяется
свой линейный слой, получаем q, k, v

Скалярное q на k^T , масштабируем
корнем из размерности, применяем
софтмакс, домножаем на v
(взвешиваем)

То есть для каждого из параллельных
блоков (“голов”) h имеем

$$Z_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_h}}\right) V_h$$

Input

Embedding

Queries

Keys

Values

Score

Divide by 8 ($\sqrt{d_k}$)

Softmax

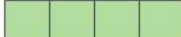
Softmax

X

Value

Sum

Thinking

x_1 

q_1 

k_1 

v_1 

$q_1 \cdot k_1 = 112$

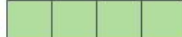
14

0.88

v_1 

z_1 

Machines

x_2 

q_2 

k_2 

v_2 

$q_2 \cdot k_2 = 96$

12

0.12

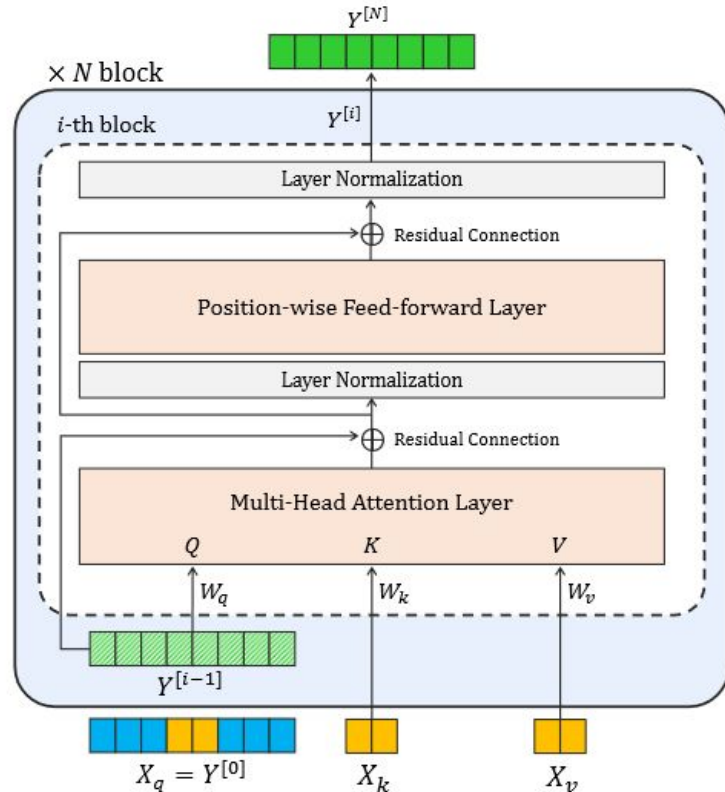
v_2 

z_2 

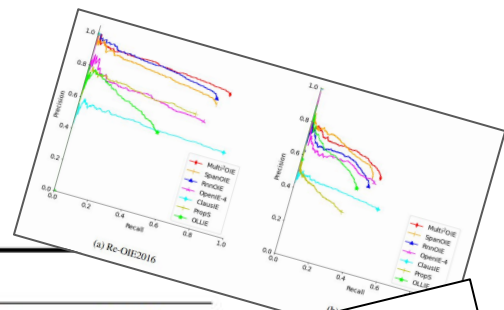
2020 [Mult2OIE] Немного странное self-attention

Query здесь -- вся подготовленная последовательность

Key и **Value** -- берутся только из “участка” предиката



2020 [Multi²OIE] Результаты



	Re-OIE2016				CaRB			
	AUC	F1	PREC.	REC.	AUC	F1	PREC.	REC.
Stanford	11.5	16.7	-	-	13.4	23.0	-	-
OLLIE	31.3	49.5	-	-	22.4	41.1	-	-
PropS	43.3	64.2	-	-	12.6	31.9	-	-
ClausIE	46.4	64.2	-	-	22.4	44.9	-	-
OpenIE4	50.9	68.3	-	-	27.2	48.8	-	-
RnnOIE	68.3	78.7	84.2	73.9	26.8	46.7	55.6	40.2
BIO	71.9	80.3	84.1	76.8	27.7	46.6	55.1	40.4
BIO+MH	71.3	81.5	87.0	76.6	27.3	47.5	57.2	40.7
SpanOIE	65.8	77.0	79.7	74.5	30.0	49.4	60.9	41.6
SpanOIE+MH	68.0	78.8	83.1	74.9	30.2	50.0	62.2	41.8
BERT+BiLSTM	72.1	81.3	86.0	77.0	30.6	50.6	61.3	43.1
Multi²OIE (ours)	74.6	83.9	86.9	81.0	32.6	52.3	60.9	45.8

с PR-кривыми тоже всё хорошо и убедительно

the use of multi-head attention is superior to simple concatenation in terms of utilizing predicate information

⁵IMoJIE achieved (AUC, F1) of (33.3, 53.5) on the CaRB dataset.

2020 [Multi²OIE] Multilingual

Re-OIE2016 переводили гугл-транслейтом
и чинили

Сравнили с языконезависимыми
системами -- 2015 и 2016 соответственно

Sentence	<i>When the explosion tore through the hut, Stauffenberg was convinced that no one in the room could have survived.</i>
English	<i>(tore; the explosion; through the hut) (was convinced; Stauffenberg; that no one in the room could have survived) (could have survived; no one in the room)</i>
Spanish	<i>(desgarró; la explosión; a través de la cabaña) (estaba convencido; Stauffenberg; de que nadie en la habitación podría haber sobrevivido) (podría haber sobrevivido; nadie en la habitación)</i>
Portuguese	<i>(rasgou; a explosão; através da cabana) (estava convencido; Stauffenberg; de que ninguém na sala poderia ter sobrevivido) (poderia ter sobrevivido; ninguém na sala)</i>

Lang.	System	F1	PREC.	REC.
EN	ArgOE	43.4	56.6	35.2
	PredPatt	53.1	53.9	52.3
	Multi²OIE	69.3	66.9	71.7
ES	ArgOE	39.4	48.0	33.4
	PredPatt	44.3	44.8	43.8
	Multi²OIE	60.2	59.1	61.2
PT	ArgOE	38.3	46.3	32.7
	PredPatt	42.9	43.6	42.3
	Multi²OIE	59.1	56.1	62.5

Table 8: Binary extraction performance without confidence scores on the multilingual Re-OIE2016 dataset.

2020 [OpenIE6] Kolluru, Adlakha, Aggarwal, Mausam, Chakrabarti

OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction; arxiv: [2010.03147](https://arxiv.org/abs/2010.03147) github: [dair-iitd/openie6](https://github.com/dair-iitd/openie6)

SoTA по извлечению троек для английского языка

Замысел — **взять лучшее от generation и labeling:**

1. Предлагают решение как “разметку решётки”
2. Вводят языкозависимые штрафы как доп. слагаемые в невязку (loss)
3. Обучают SoTA для Coordination Analysis и применяют его к решению

Что это всё значит?

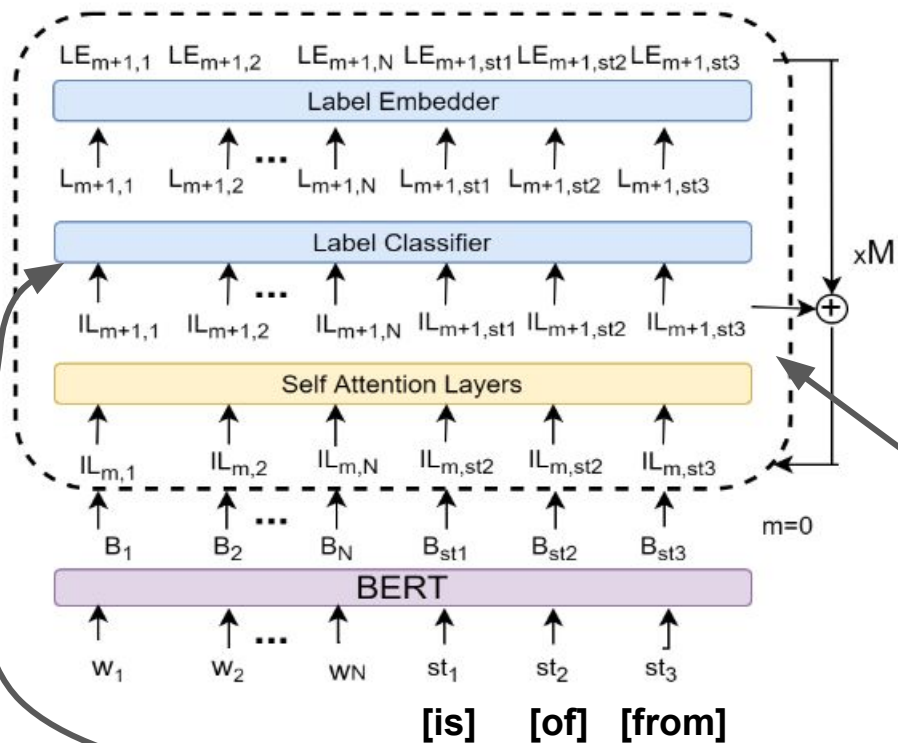
2020 [OpenIE6] “Разметка решётки”: Iterative Grid Labeling

Subject	Relation	Object									
Rome	the capital	of	Italy	is known	for	it's	rich	history	[is]		
Rome	the capital	of	Italy	is known	for	it's	rich	history	[is]		
Rome	the capital	of	Italy	is known	for	it's	rich	history	[is]		

M возможностей для извлечения, **N** слов в предложении

Замысел: размечать последовательно за **M** шагов с оглядкой на прошлый шаг

2020 [OpenIE6] Что значит “с оглядкой”?

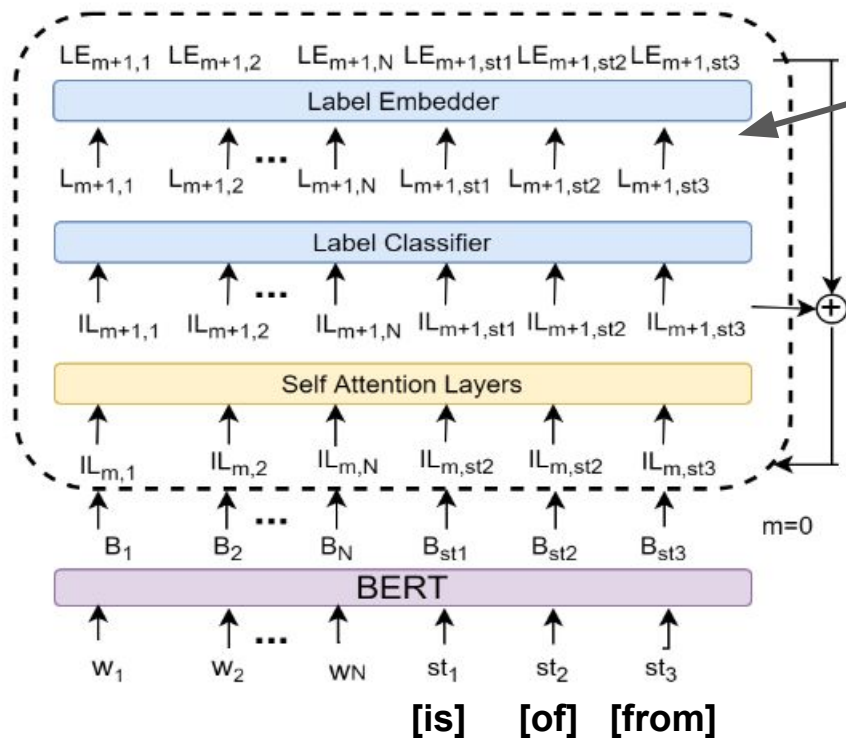


1. На вход берём представления “слов” от BERT. В конец добавляем токены, которые чаще всего приходится добавлять: “[is]”, “[of]”, “[from]”

```
# Remove the brackets from the inserted [is], [of], [from], etc
for bracket in re.findall(r'\([.*?]\)', extraction):
    if bracket[0] in orig_sent:
        continue
    extraction = extraction.replace(bracket[0], bracket[1])
```

2. Через два трансформера получаем представления IL
3. Предсказываем метки: S, R, O, N, применив к IL полносвязный слой

2020 [OpenIE6] Что значит “с оглядкой”?



3. Строим эмбединги меток LE

4. Складываем IL с LE и передаём на вход для следующей такой итерации

Так мы учитываем на каждой итерации, что мы уже извлекли ранее, чтобы не повторяться

Функция потерь -- сумма всех кросс-энтропий с каждого уровня

Confidence scores -- сумма логарифмов вероятностей меток (кроме N), нормализованная по длине извлечения

2020 [OpenIE6] Ограничения

Эксперименты показали, что IGL в чистом виде теряет много информации, нельзя столько упускать!

Замысел: выскажем пожелания и добавим их в функцию потерь

- **POSC**: все существительные, глаголы, прилагательные и наречия должны быть частью хотя бы одной извлечённой тройки

$$J_{posc} = \sum_{n=1}^N x_n^{imp} \cdot posc_n, \text{ where}$$
$$posc_n = 1 - \max_{m \in [1, M]} \left(\max_{k \in \{S, R, O\}} Y_{mn}(k) \right)$$

индикатор, что у слова “важная” часть речи

максимум по итерациям

максимальная вероятность получить “непустую” метку

2020 [OpenIE6] Ограничения

- **HVC**: каждый “осмысленный” глагол (*head verb*) должен быть представлен хотя бы в некоторых (но не во многих) предикатах

индикатор, что является “осмысленным” глаголом

tractions. This penalty is aggregated over head verbs, $J_{hvc} = \sum_{n=1}^N x_n^{hv} \cdot hvc_n$, where $hvc_n = \left| 1 - \sum_{m=1}^M Y_{mn}(R) \right|$.

сумма вероятностей быть Relation = предикатом

2020 [OpenIE6] Ограничения

- **HVC**: иметь метку R может лишь один “осмысленный” глагол

сумма по ИТЕРАЦИЯМ

$$\text{I.e., } J_{hve} = \sum_{m=1}^M hve_m, \text{ where}$$
$$hve_m = \max \left(0, \left(\sum_{n=1}^N x_n^{hv} \cdot Y_{mn}(R) \right) - 1 \right)$$

плохо, если на данной итерации по “осмысленным” глаголам наберётся сумма вероятностей больше единицы

2020 [OpenIE6] Ограничения

- **ЕС:** троек должно быть извлечено не меньше, чем в предложении “осмысленных” глаголов

$$ec_m = \max_{n \in [1, N]} (x_n^{hv} \cdot Y_{mn}(R))$$

$$J_{ec} = \max \left(0, \sum_{n=1}^N x_n^{hv} - \sum_{m=1}^M ec_m \right)$$

количество “осмысленных” глаголов

сумма их вероятностей оказаться в предикате
(по всем итерациям!)

2020 [OpenIE6] CIGL-OIE

loss function is $J = J_{CE} + \lambda_{posc} J_{posc} + \lambda_{hvc} J_{hvc} + \lambda_{hve} J_{hve} + \lambda_{ec} J_{ec}$, where λ_* are hyperparameters.

Теперь это называется Constrained Iterative Grid Labeling OpenIE Extractor — но и это ещё не OpenIE6

Моделям, в которых явно не обрабатываются сочинительные связи/структуры (например, однородные члены предложения через запятую, союзы “и” или “или”), непросто их разрешать

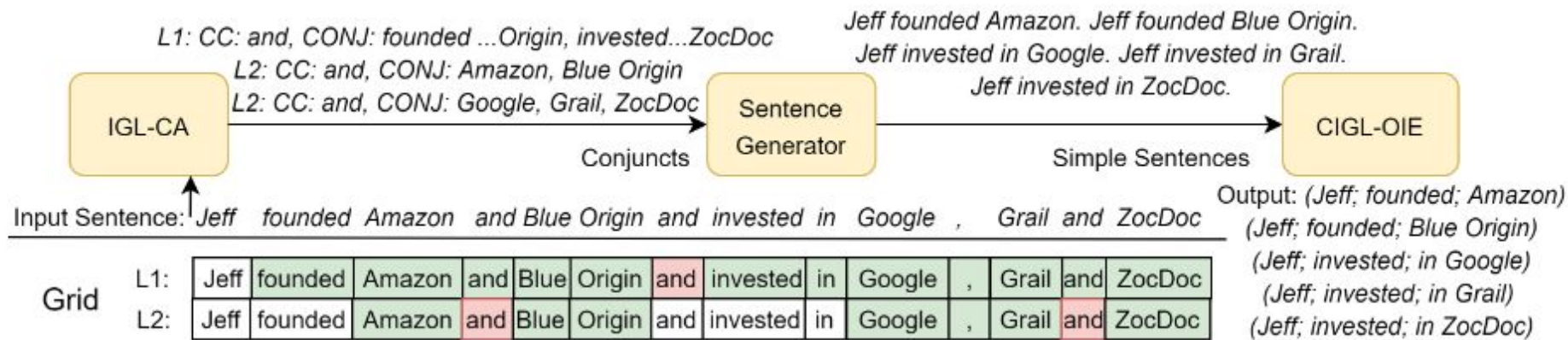
В работе об OpenIE6 делают это **явно**, используя **всё тот же IGL**

2020 [OpenIE6] Coordination Analysis

Это помогает, показано ещё в работе о CalmIE (Saha&Mausam, 2018)

В 2016 году (Ficler and Goldberg, 2016) разметили PTB (Penn Treebank) метками, указывающими границы сочинительных структур

Сочинительные структуры бывают вложенными, то есть задача по сути иерархическая; уровень иерархии будем укладывать как итерацию



2020 [OpenIE6] Coordination Analysis

Такой подход IGL-CA походя побил SoTA (на CYK parsing) по качеству

System	Precision	Recall	F1
(Teranishi et al., 2017)	71.5	70.7	71.0
(Teranishi et al., 2019)	75.3	75.6	75.5
BERT-Base:			
(Teranishi et al., 2019)	83.1	83.2	83.1
IGL-CA	86.3	83.6	84.9
BERT-Large:			
(Teranishi et al., 2019)	86.4	86.6	86.5
IGL-CA	88.1	87.4	87.8

Table 5: P, R, F1 of the system evaluated on Penn Tree Bank for different systems. We use both BERT-Base and BERT-Large as the encoder

2020 [OpenIE6] Finally

- Для OpenIE с помощью IGL-CA строятся простые (неконъюнктивные) предложения,
- к ним применяется CIGL-OIE,
- фильтруются тройки-дубликаты.

Но из-за такого подхода confidence scores приходится пересчитывать отдельной моделью

Sentence	Other signs of lens subluxation include mild conjunctival redness, vitreous humour degeneration, and an increase or decrease of anterior chamber depth .
IGL	(Other signs of lens subluxation; include; mild conjunctival redness, vitreous humour degeneration)
IGL +Constraints	(Other signs of lens subluxation; include; mild conjunctival redness, vitreous humour degeneration, and an increase or decrease of anterior chamber depth)
IGL +Constraints +Coordination Analyzer	(Other signs of lens subluxation; include; mild conjunctival redness) (Other signs of lens subluxation; include; vitreous humour degeneration) (Other signs of lens subluxation; include; an increase of anterior chamber depth) (Other signs of lens subluxation; include; an decrease of anterior chamber depth)

Table 1: For the given sentence, IGL based OpenIE extractor produces an incomplete extraction. Constraints improve the recall by covering the remaining words. Coordination Analyzer handles hierarchical conjunctions.

2020 [OpenIE6] Данные и оценка качества

Обучали на том же, на чём и IMoJIE -- смесь выходов разных моделей, 190'661 троек, 92774 статьи английской Википедии

Оценивали качество **на данных CaRB**, используя в качестве оценки качества скрипты: исходный и “жадный” CaRB, WiRe57, OIE16

Также

- измеряли время работы,
- перебирали наборы ограничений (смотрели число “нарушений” и влияние на качество),
- подкладывали IMoJIE и CIGL-OIE другие анализаторы сочинительных структур

2020 [OpenIE6] Данные и оценка качества

System	CaRB		CaRB(1-1)		OIE16-C		Wire57-C	Speed
	F1	AUC	F1	AUC	F1	AUC	F1	Sentences/sec.
MinIE	41.9	-	38.4	-	52.3	-	28.5	8.9
ClausIE	45.0	22.0	40.2	17.7	61.0	38.0	33.2	4.0
OpenIE4	51.6	29.5	40.5	20.1	54.3	37.1	34.4	20.1
OpenIE5	48.0	25.0	42.7	20.6	59.9	39.9	35.4	3.1
SenseOIE	28.2	-	23.9	-	31.1	-	10.7	-
SpanOIE	48.5	-	37.9	-	54.0	-	31.9	19.4
RnnOIE	49.0	26.0	39.5	18.3	56.0	32.0	26.4	149.2
(Cui et al., 2018)	51.6	32.8	38.7	19.8	53.5	37.0	33.3	11.5
IMoJIE	53.5	33.3	41.4	22.2	56.8	39.6	36.0	2.6
IGL-OIE	52.4	33.7	41.1	22.9	55.0	36.0	34.9	142.0
CIGL-OIE	54.0	35.7	42.8	24.6	59.2	40.0	36.8	142.0
CIGL-OIE + IGL-CA (OpenIE6)	52.7	33.7	46.4	26.8	65.6	48.4	40.0	31.7

2020 [OpenIE6] Итоги

- Модель классно работает, побить на этих датасетах трудно
- Быстрее предыдущей SoTA-модели, но всё-таки очень медленная
- Зависит от языка, так как использует списки “лёгких” глаголов, частеречную разметку, самые востребованные для вставки слова ([is] [of] [from]) и так далее

OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction

Keshav Kolluru^{1*}, Vaibhav Adlakha^{1*}, Samarth Aggarwal¹, Mausam¹, and Soumen Chakrabarti²

¹ Indian Institute of Technology Delhi

keshav.kolluru@gmail.com, vaibhavadlakha95@gmail.com

samarth.aggarwal.2510@gmail.com, mausam@cse.iitd.ac.in

² Indian Institute of Technology Bombay

План

1. ~~OpenIE до 2016 года~~

- a. ~~TextRunner~~
- b. ~~ReVerb~~
- c. ~~OLLIE~~
- d. ~~OpenIE-4~~

2. ~~Датасеты и бенчмарки~~

- a. ~~OIE2016~~
- b. ~~WiRe57~~
- c. ~~CaRB~~
- d. ~~Иное~~

3. ~~Победный марш глубокого обучения~~

- a. ~~SpanOIE~~
- b. ~~IMoJIE~~
- c. ~~Mult²OIE~~
- d. ~~OpenIE6~~

4. А что с русским языком?

5. Важные работы, о которых не говорили

А как дела в русском OpenIE?

- Работ по IE для русского много; NER, например, занимаются повсеместно; есть ряд работ по извлечению “типизированных” отношений
- Опубликованных работ, посвящённых исключительно русскому **OpenIE**, по моим сведениям, **нет**
- При этом наборы данных для дообучения и оценки качества существуют:

WMORC_{manual} — на французском, русском и хинди

WMORC_{auto} — автоматически для 61 языка

Faruqui M., Kumar S. Multilingual Open Relation Extraction Using Cross-lingual Projection

//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2015. – С. 1351-1356.

А как дела в русском OpenIE?

- Rule-based semantic parsers:
 - AOT.ru (Sokirko, A. 2001)
 - The parser of ISA FRC CSC RAS (Shelmanov and Smirnov, 2014)
 - etc.
- Known corpora annotated with semantic roles:
 - The corpus from ISA FRC CSC RAS
 - Shelmanov and Smirnov, 2014
 - FrameBank
 - Lyashevskaya, 2012
 - Lyashevskaya and Kashkin, 2015
- Data-driven semantic role labelers:
 - SVM-based parser + feature engineering (Kuznetsov I., 2015) trained on pre-release version of FrameBank
 - The parser of ISA FRC CSC RAS (Shelmanov and Smirnov, 2014) – bootstrapping based on automatic annotation of SynTagRus using rule-based semantic parser

Из слайдов к докладу

Shelmanov A., Devyatkin D.
Semantic role labeling with neural networks for texts in Russian //Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017). – 2017. – Т. 16. – С. 245-256.

Важные модели и работы, о которых речь не зашла

- **2010 [WOE^{pos}, WOE^{parse}]** Wu F., Weld D. S. Open information extraction using wikipedia //Proceedings of the 48th annual meeting of the association for computational linguistics. 2010. С. 118-127.

Значительный прирост в качестве относительно TextRunner за счёт использования синтаксических признаков

- **2013 [ClausIE]** Del Corro L., Gemulla R. Clausie: clause-based open information extraction //Proceedings of the 22nd international conference on World Wide Web. – 2013. – С. 355-366.

Во весь рост используется английская грамматика, результаты лучше TextRunner, WOE, OLLIE и ReVerb; кажется, первая работа, в которой явно обрабатывают сочинительные структуры

Важные модели и работы, о которых речь не зашла

- **2015 [Stanford OpenIE]** Angeli G., Premkumar M. J. J., Manning C. D. Leveraging linguistic structure for open domain information extraction //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). – 2015. – С. 344-354.

Слово авторам: We replace this large pattern set with a few patterns for canonically structured sentences, and shift the focus to a classifier which learns to extract self-contained clauses from longer sentences. We then run natural logic inference over these short clauses to determine the maximally specific arguments for each candidate triple

- **2017 [MinIE]** Gashteovski K., Gemulla R., del Corro L. MinIE: Minimizing Facts in Open Information Extraction //Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. – 2017. – С. 2630-2640.

Много трюков, в том числе направленных на отбрасывание малозначительных токенов

- **2018 [OpenIE5]** Комбинация многих работ, улучшающая OpenIE4; коллаборация University of Washington (UW) and Indian Institute of Technology, Delhi (IIT Delhi)
<https://github.com/dair-iitd/OpenIE-standalone>

Важные модели и работы, о которых речь не зашла

- **2018 [RnnOIE]** Stanovsky G. et al. Supervised open information extraction //Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). – 2018. – С. 885-895.

Используют датасет AW-OIE, полученный преобразованием QAMR (Question Answering Meaning Representation) для обучения, формулируют задачу как разметку последовательностей, бьют SoTA на момент 2018 года

Также, вероятно, стоит ознакомиться с работой половины этого коллектива авторов: Gabriel Stanovsky, Jessica Fidler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props. [arXiv preprint](#)

Важные модели и работы, о которых речь не зашла

- **2018 [Seq2seq OIE OR CopyAttention]** Cui L., Wei F., Zhou M. Neural Open Information Extraction //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2018. – С. 407-413.

Строят обучающий датасет с помощью OpenIE4 и Википедии, оставляя извлечения с “уверенностью” выше 0.9. Затем решают задачу как sequence-to-sequence, используя OpenNMT (и даже не подкладывают эмбединги, похоже)

IN: “deep learning is a subfield of machine learning”.

OUT: “<arg1>deep learning</arg1><rel>is a subfield of</rel><arg2>machine learning</arg2>”.

Также есть ряд работ, извлекающих информацию из **особых типов текстов** (например, [question-answer pairs](#)), но при этом всё-таки укладываемых в OpenIE; их можно найти, переходя по ссылкам из разделов Related work в свежих статьях

Спасибо за внимание!

Open Information Extraction

обзор: ключевые статьи, инструменты, наборы данных

лаб. искусственного интеллекта
ПОМИ РАН им. В.А. Стеклова
Антон Алексеев

<https://alexeyev.github.io/>
anton.m.alexeyev@gmail.com

The content is available
under CC BY-SA 3.0

