Supervised Open Information Extraction

Gabriel Stanovsky[2,3], Julian Michael[2], Luke Zettlemoyer[2], Ido Dagan[1]

github.com/gabrielStanovsky/supervised-oie

[1]Bar-Ilan University

[2]University of Washington

[3]Allen Institute for Artificial Intelligence

Open Information Extraction

Aims to extract asserted propositions from unstructured text:

*"Barack Obama, a former U.S president, **was born in** Hawaii."*

1. (Barack Obama; **was born in**; Hawaii)
2. (a former U.S. president; **was born in**; Hawaii)

BIO Encoding

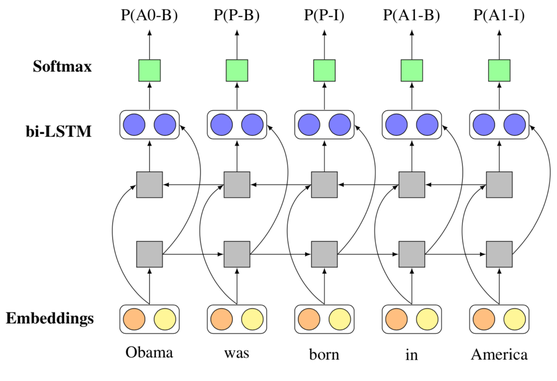
Each tuple is encoded with respect to a single predicate, where argument labels indicate their position in the tuple.

Labels repeat when a single predicate participates in multiple propositions

$$\text{Barack}_{A0-B}\ \text{Obama}_{A0-I}\ ,_{O}\ \text{a}_{A0-B}\ \text{former}_{A0-I}\ \text{U.S.}_{A0-I}\ \text{president}_{A0-I}\ ,_{O}\ \text{was}_{P-B}\ \text{born}_{P-I}\ \text{in}_{P-I}\ \text{Hawaii}_{A1-B}$$

Multi-word predicates are allowed

RNN-OIE: Bi-LSTM Sequence Tagger Inspired by recent state of the art in Semantic Role Labelling (Zhou and Xu, 2015; He et al., 2017).



Features: Concatenated pretrained embeddings of current word and target predicate (identified by a verb POS).

Decoding: Ignores malformed spans - if an A0-I label is not preceded by A0-I or A0-B, we treat it as O.

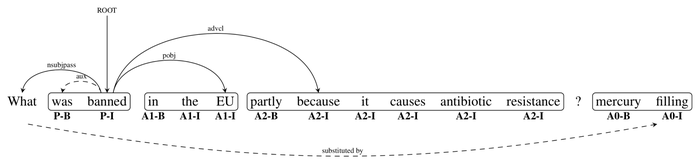
Confidence: Estimated for an extraction E by $\Pi_{l \in E} P(l)$

Training Data

We used the QA-SRL to Open IE conversion (**OIE2016,** Stanovsky and Dagan, 2016) to train our model. This consists of verbal propositions, automatically extracted from template QA-SRL annotations.

Augmenting with QAMR annotations

In addition, we converted the Question-Answer Meaning Representation bank (Michael et al, 2018 – **Come see our poster tomorrow!**), consisting of free-form question-answer format over a wide range of predicates. The conversion was achieved with heuristics over the QA parse tree.



(mercury filling; **was banned**; in the EU; partly because it causes antibiotic resistance)

Resulting Training Corpus

Dataset	Domain	#Sentences	#Tuples
OIE2016	News, Wiki	3200	5077
QAMR	Wikinews, wiki	3300	12952

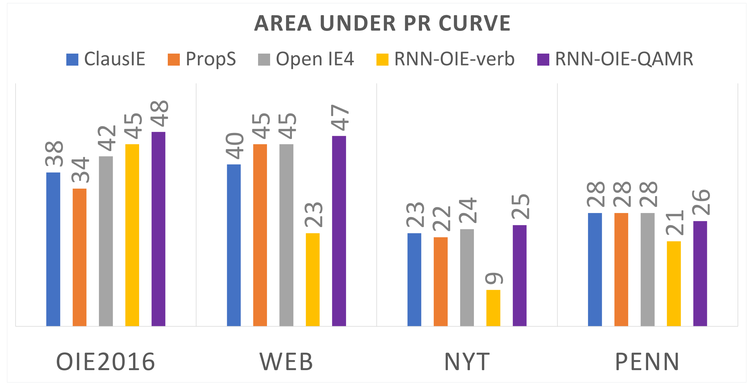
Test Data

We test our model on four publicly available Open IE corpora, following (Schneider et al., 2017).

Dataset	Domain	#Sentences	#Tuples
OIE2016	News, Wiki	3200	1729
WEB	News, Web	500	461
NYT	News, Wiki	222	222
PENN	Mixed	100	51

Evaluation

We compare RNN-OIE against top performing Open IE systems:



RNN-OIE performs competitively across all test sets, outperforming all other systems on the larger test sets. **QAMR improves performance**, especially on more diverse test sets.

Run-time Analysis

Rnn-OIE is able to leverage GPU architecture to achieve a 10 times improvement over the previous fastest system (measured in sentences per second).

	ClausIE	PropS	Open IE4	RNN-OIE
CPU	4.07	4.59	**15.38**	13.51
GPU	---	---	---	**149.25**

Error Analysis

An analysis of 100 gold propositions which were missed by **all** systems (i.e., recall errors) reveals that they all struggle with noun relations, sentence-level inference and long or informal sentences.