Creating a Large Benchmark for Open Information Extraction

Gabriel Stanovsky and **Ido Dagan**
Computer Science Department,
Bar-Ilan University, Ramat Gan, Israel
gabriel.satanovsky@gmail.com
dagan@cs.biu.ac.il

Abstract

Open information extraction (Open IE) was presented as an unrestricted variant of traditional information extraction. It has been gaining substantial attention, manifested by a large number of automatic Open IE extractors and downstream applications. In spite of this broad attention, the Open IE task definition has been lacking – there are no formal guidelines and no large scale gold standard annotation. Subsequently, the various implementations of Open IE resorted to small scale post-hoc evaluations, inhibiting an objective and reproducible cross-system comparison. In this work, we develop a methodology that leverages the recent QA-SRL annotation to create a first independent and large scale Open IE annotation,[1] and use it to automatically compare the most prominent Open IE systems.

1 Introduction

Open Information Extraction (Open IE) was originally formulated as 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). Wu and Weld (2008) further defined that an Open IE extractor should "produce one triple for every relation stated explicitly in the text, but is not required to infer implicit facts". For example, given the sentence "*John managed to open the door*" an Open IE extractor should produce the tuple *(John;* **managed to open***; the door)* but is *not* required to produce the extraction *(John;* **opened***; the door)*.

[1]Publicly available at `http://www.cs.biu.ac.il/nlp/resources/downloads`

Following this initial presentation of the task, Open IE has gained substantial and consistent attention. Many automatic extractors were created (e.g., (Fader et al., 2011; Mausam et al., 2012; Del Corro and Gemulla, 2013)) and were put to use in various downstream applications.

In spite of this wide attention, Open IE's formal definition is lacking. There are no clear guidelines as to what constitutes a valid proposition to be extracted, and subsequently there is no large scale benchmark annotation. Open IE evaluations therefore usually consist of a post-hoc manual evaluation of a small output sample.

This evaluation practice lacks in several respects: (1) Most works provide a precision oriented metric, whereas recall is often not measured, (2) the numbers are not comparable across systems, as they use different guidelines and datasets, and (3) the experiments are hard to replicate.

In this work, we aim to contribute to the standardization of Open IE evaluation by providing a large gold benchmark corpus. For that end, we first identify consensual guiding principles across prominent Open IE systems, resulting in a clearer formulation of the Open IE task. Following, we find that the recent formulation of QA-SRL (He et al., 2015) in fact subsumes these requirements for Open IE. This enables us to automatically convert the annotations of QA-SRL to a high-quality Open IE corpus of more than 10K extractions, 13 times larger than the previous largest Open IE annotation.

Finally, we automatically evaluate the performance of various Open IE systems against our corpus, using a soft matching criterion. This is the first

time such a comparative evaluation is performed on a large scale gold corpus.

Future Open IE systems (and its applicative users) can use this large benchmark, along with the automatic evaluation measure, to easily compare their performance against previous baselines, alleviating the current need for ad-hoc evaluation.

2 Background

2.1 Open IE

Open Information Extraction (Open IE) was introduced as an open variant of traditional Information Extraction (Etzioni et al., 2008). As mentioned in the Introduction, its primary goal is to extract coherent propositions from a sentence, each comprising of a relation phrase and two or more argument phrases (e.g., (Barack Obama, **born in**, Hawaii)). Since its inception, Open IE has gained consistent attention, mostly used as a component within larger frameworks (Christensen et al., 2013; Balasubramanian et al., 2013).

In parallel, many Open IE extractors were developed. TextRunner (Banko et al., 2007) and WOE (Wu and Weld, 2010) take a self-supervised approach over automatically produced dependency parses. Perhaps more dominant is the rule based approach taken by ReVerb (Fader et al., 2011), OLLIE (Mausam et al., 2012), KrakeN (Akbik and Löser, 2012) and ClausIE (Del Corro and Gemulla, 2013).

Two recent systems take a semantically-oriented approach. Open IE-4[2] uses semantic role labeling to extract tuples, while Stanford Open Information Extraction (Angeli et al., 2015) uses natural logic inference to arrive at shorter, more salient, arguments.

Recently, Stanovsky et al. (2016b) presented PropS, a proposition oriented representation, obtained via conversion rules from dependency trees. Performing Open IE extraction over PropS structures is straightforward – follow the clearly marked predicated nodes to their direct arguments.

Contrary to the vast interest in Open IE, its task formulation has been largely overlooked. There are currently no common guidelines defining a valid extraction, which consequently hinders the creation of an evaluation benchmark for the task. Most Open IE extractors[3] evaluate performance by manually examining a small sample of their output. Table 1 summarizes the evaluations taken by the most prominent Open IE systems.

2.2 QA-SRL

Semantic Role Labeling (SRL) (Carreras and Màrquez, 2005) is typically perceived as answering **argument role questions**, such as *who, what, to whom, when*, or *where*, regarding a target predicate. For instance, PropBank's ARG0 for the predicate **say** answers the question "*who* **said** *something?*".

QA-SRL (He et al., 2015) suggests that answering explicit role questions is an intuitive means to solicit predicate-argument structures from non-expert annotators. Annotators are presented with a sentence in which a target predicate[4] was marked, and are requested to annotate argument role questions and corresponding answers.

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]

He et al. assessed the validity of QA-SRL by annotating 3200 sentences from PropBank and Wikipedia, showing high agreement with the PropBank annotations. In the following section we automatically derive an Open IE benchmark from this QA-SRL annotation.

3 Creating an Open IE Benchmark

3.1 Open IE Guidelines

Before creating a generic benchmark for evaluating Open IE systems, it is first needed to obtain a clearer specification of the common task that they address. Despite some nuances, we identified the following core aspects of the Open IE task as consensual across all systems mentioned in Section 2:

[2] https://github.com/knowitall/openie

[3] Except for (Wu and Weld, 2010) who evaluated recall.

[4] Currently consisting of automatically annotated verbs.

[5] Three cases give rise to multiple answers for the same question: appositives (as illustrated in this example), co-reference ("***Jimmy Hendrix*** *played the guitar,* ***he*** *was really good at it*"), and distributive coordinations ("***Bob*** *and* ***Mary*** *were born in America*").

System	#Sentences	Genre	Metric	#Annot.	Agreement
TextRunner	400	Web	% Correct	3	-
WOE	300	Web, Wiki, News	Precision / Recall	5	-
ReVerb	500	Web	Precision / AUC	2	86%, .68 k
KrakeN	500	Web	% Correct	2	87%
Ollie	300	News, Wiki, Biology	Precision/Yield AUC	2	96%
ClauseIE	300	Web, Wiki, News	Precision/Yield	2	57% / 68% / 63%

Table 1: The post-hoc evaluation metrics taken by the different systems described in Section 2. In contrast, Stanford Open IE and PropS took an extrinsic evaluation approach.

Assertedness Extracted propositions should be asserted by the original sentence. For example, given the sentence "*Sam succeeded in convincing John*", ReVerb and ClausIE produce the extraction: *(Sam;* **succeeded in convincing***; John)*. Most Open IE systems do not attempt to recover implied embedded propositions (e.g., *(Sam;* **convinced***; John)*), but rather include matrix verbs (e.g., **succeeded**) in the predicate slot. Other elements that affect assertedness, like negations and modals, are typically included in the predicate slot as well (e.g. *(John;* **could not join***; the band)*).

Minimal propositions Open IE systems aim to "break down" a sentence into a set of small isolated propositions. Accordingly, the span of each individual proposition, and hence the span of each of its predicate and argument slots, should be as minimal as possible, as long as the original information (truth conditions) is preserved. For example, this leads to splitting distributive coordination in the sentence "*Bell distributes electronic and building products*", for which ClausIE produces: *(Bell,* **distributes***, electronic products)* and *(Bell,* **distributes***, building products)*. Having shorter entities as Open IE arguments was further found to be useful in several semantic tasks (Angeli et al., 2015; Stanovsky et al., 2015).

Completeness and open lexicon Open IE systems aim to extract all asserted propositions from a sentence. 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.

3.2 From QA-SRL to Open IE

SRL and Open IE have been defined with different objectives. Particularly, SRL identifies argument role labels, which is not addressed in Open IE. Yet, the two tasks overlap as they both need to recover predicate-argument structures in sentences. We now examine the above Open IE requirements and suggest that while they are only partly embedded within SRL structures, they can be fully recovered from QA-SRL.

Asserted (matrix) propositions appear in SRL as non-embedded predicates (e.g., **succeeded** in the "*Sam succeeded to convince John*"). However, SRL's predicates are grounded to a lexicon such as PropBank (Palmer et al., 2005) or FrameNet (Baker et al., 1998), which violates the *completeness and open lexicon* principle. Further, in contrast to the *minimal propositions* principle, arguments in SRL annotations are inclusive, each marked as full subtrees in a syntactic parse.

Yet, QA-SRL seems to bridge this gap between traditional SRL structures and Open IE requirements. Its predicate vocabulary is open, and its question-answer format solicits *minimal propositions*, as was found in a recent study by (Stanovsky et al., 2016a). This correlation suggests that the QA-SRL methodology is in fact also an attractive means for soliciting Open IE extractions from non-experts annotators. Evidently, it enables automatically deriving high quality Open IE annotations from (current or future) QA-SRL gold annotations, as described in the following section

3.3 Generating Open-IE Extractions

Formally, we extract an Open-IE dataset from the QA-SRL dataset by the following algorithm, which is illustrated in more detail further below:

1. Given:
   - s - a sentence from the QA-SRL dataset.
   - p - a predicate in s.
   - $\{q_1, ..., q_n\}$ - a list of questions over p.
   - $\{\{a_{1,1}, ..., a_{1,l_1}\}, ...\{a_{n,1}, ..., a_{n,l_n}\}\}$ - a list of sets of corresponding answers, where question q_i has l_i answers.
2. If p is a non-embedded (matrix) verb:
   (a) Remove answers which are composed only of pronouns, as these are not expected to be extracted by Open-IE (and accordingly adjust the l_i's).
   (b) Return extractions composed of p and every combination of answers in $\{\{a_{1,1}, ..., a_{1,l_1}\} \times ... \times \{a_{n,1}, ..., a_{n,l_n}\}\}$ (the Cartesian product of the answers). This process results in a list of $l_1 \cdot l_2 \cdot ... \cdot l_n$ Open IE extractions.

For example, consider the sentence: "*Barack Obama, the U.S. president, was* **determined** *to win the majority vote in Washington and Arizona*". The questions corresponding to the predicate **determine** are: {*who was determined?*, *what was someone determined to do?*}, and the corresponding answer sets are: {{"Barack Obama", "the U.S president"}, {"win the majority vote in Washington", "win the majority vote in Arizona"}}.

Following, our algorithm will produce these Open IE extractions: *(Barack Obama;* **was determined***; to win the majority vote in Washington)*, *(the U.S. president;* **was determined***; to win the majority vote in Washington)*, *(Barack Obama;* **was determined***; to win the majority vote in Arizona)*, and *(the U.S. president;* **was determined***; to win the majority vote in Arizona)*.

Note that we do not produce extractions for embedded predicates (e.g., **win**) to conform with the *assertedness* principle, as discussed earlier.

With respect to pronoun removal (step 2(a)), we would remove the pronoun "he" as the answer to the question *who was tired?* in "*John went home,* ***he*** *was tired*". Notice that in this sentence "John" would be a second answer for the above question, yielding the extraction *(John;* ***was tired****)*. When the only answer to a question is a pronoun this question will be ignored in the extraction process, since the QA-SRL corpus does not address cross-sentence co-references. This issue may be addressed in future work.

Corpus	WSJ	WIKI	ALL
#Sentences	1241	1959	3200
#Predicates	2020	5690	7710
#Questions	8112	10798	18910
#Extractions	**4481**	**5878**	**10359**

Table 2: Corpus statistics.

System	#Extractions		
	WSJ	WIKI	ALL
Stanford	6423	14104	20527
ClausIE	5295	8265	13560
Open IE4	3634	5113	8747
OLLIE	2976	5250	8226
PropS	2852	4990	7842
ReVerb	1624	2552	4716

Table 3: The yield of the different Open IE systems.

Applying this process to the QA-SRL corpus yielded a total of 10,359 Open IE extractions over 3200 sentences from 2 domains (see Table 2). This corpus is about 13 times larger than the previous largest annotated Open IE corpus (Fader et al., 2011). The corpus is available at: `http://www.cs.biu.ac.il/nlp/resources/downloads`.

Corpus validation We assess the validity of our dataset by performing expert annotation[6] of Open IE extractions, following the principles discussed in Section 3.1, for 100 random sentences. We find that our benchmark extractions, derived automatically from QA-SRL, highly agree with the expert annotation, reaching 95.8 F1 by the head-agreement criterion defined in the next section.

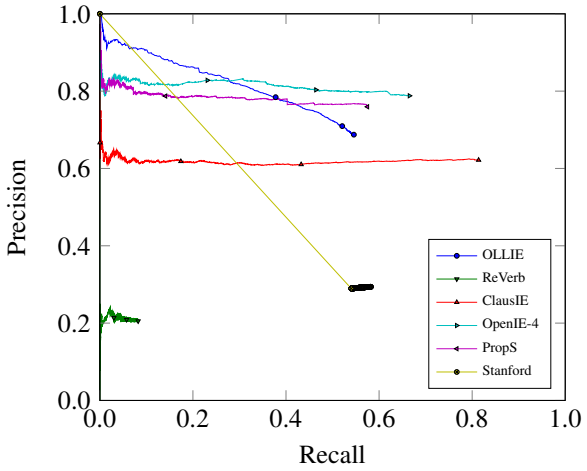

Figure 1: Precision-recall curve for the different Open IE systems on our corpus (see discussion in Section 4).

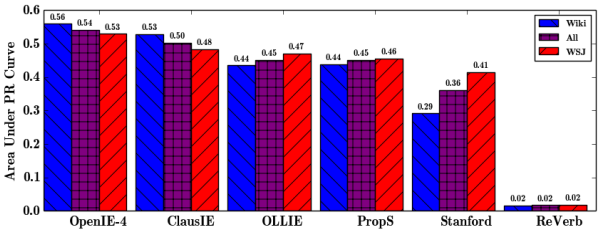

Figure 2: Area Under the PR Curve (AUC) measure for the evaluated systems.

4 Comparative Evaluation

In this section, we illustrate the utility of our new corpus by testing the performance of 6 prominent Open IE systems: OpenIE-4, ClausIE, OLLIE, PropS, Stanford, and ReVerb (see Section 2).[7]

In order to evaluate these systems in terms of precision and recall, we need to match between their automated extractions and the benchmark extractions. To allow some flexibility (e.g., omissions of prepositions or auxiliaries), we follow (He et al., 2015) and match an automated extraction with a gold proposition if both agree on the grammatical head of all of their elements (predicate and arguments). We then analyze the recall and precision of Open IE systems on different confidence thresholds (Figure 1). Furthermore, we calculate the area under the PR curve for each of the different corpora (Figure 2) and the explicit yield per system (Table 3).

To the best of our knowledge, this is the first objective comparative evaluation of prominent Open-IE systems, over a large and independently created dataset. This comparison gives rise to several observations; which can be useful for future research and for choosing a preferred system for a particular application setting, such as:

1. *Open IE-4* achieves best precision above 3% recall ($\geqslant$ 78.67) and best AUC score (54.02),
2. *ClausIE* is best at recall (81.38), and
3. *Stanford Open IE* assigns confidence of 1 to 94% of its extractions, explaining its low precision.

5 Conclusions

We presented the first independent and large scale Open IE benchmark annotation, and tested the most prominent systems against it. We hope that future Open IE systems can make use of this new resource to easily and objectively measure and compare their performance.

Acknowledgments

We would like to thank Mausam for fruitful discussions, and the anonymous reviewers for their helpful comments.

This work was supported in part by grants from the MAGNET program of the Israeli Office of the Chief Scientist (OCS), the Israel Science Foundation grant 880/12, and the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1).

References

Alan Akbik and Alexander Löser. 2012. Kraken: N-ary facts in open information extraction. In *NAACL-HLT 2012: Proceedings of the The Knowledge Extraction Workshop*.

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

[6]Carried by the first author.

[7]Currently, we test only the common case of verbal predicates.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of ACL*, pages 86–90. Association for Computational Linguistics.

Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *EMNLP*, pages 1721–1731.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2670–2676.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of CONLL*, pages 152–164.

Janara Christensen, Stephen Soderland Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *HLT-NAACL*, pages 1163–1173. Citeseer.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. International World Wide Web Conferences Steering Committee.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the Web. *Communications of the ACM*, 51(12):68–74.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea, July. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Gabriel Stanovsky, Ido Dagan, and Mausam. 2015. Open IE as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Gabriel Stanovsky, Ido Dagan, and Meni Adler. 2016a. Specifying and annotating reduced argument span via qa-srl. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016b. Getting more out of syntax with props. *arXiv preprint*.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden, July. Association for Computational Linguistics.

Fei Wu, Raphael Hoffmann, and Daniel S Weld. 2008. Information extraction from wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 731–739. ACM.