An informativeness approach to Open IE evaluation

William Léchelle, Philippe Langlais

RALI — University of Montreal
{lechellw, felipe} @ iro.umontreal.ca

Abstract. Open Information Extraction (OIE) systems extract relational tuples from text without requiring to specify in advance the relations of interest. Systems perform well on widely used metrics such as precision and yield, but a close look at systems output shows a general lack of informativeness in facts deemed correct.
We propose a new evaluation protocol, based on question answering, that is closer to text understanding and end user needs. Extracted information is judged upon its capacity to automatically answer questions about the source text. As a showcase for our protocol, we devise a small corpus of question/answer pairs, and evaluate available state-of-the-art OIE systems on it. Performance-wise, our results are in line with previous findings. Furthermore, we are able to estimate recall for the task, which is novel. We distribute our annotated data and automatic evaluation program.

Keywords : Open Information Extraction · Evaluation · Question Answering

1 Introduction

OIE – information extraction without pre-specification of relations or entities to target – seeks to extract relational tuples from large corpora, in a scalable way and without domain-specific training [5] [3]. Recently, there has been a trend of successful use of OIE output as a text understanding tool, for instance in [4], [7] and [8].

However, a close look to systems output reveals that a large fraction of extracted facts, albeit correctly extracted from the text, are devoid of useful information. The main reason for this is lack of context : many extracted noun phrases, and facts, only have meaning in the context of their sentences. Once the source is lost, the remaining relation is empty, for factual purposes[1]. Figure 1 shows examples of uninformative facts. We discuss in section 2 how state-of-the-art metrics of extraction performance fail to account for meaningless extractions.

We propose an evaluation procedure for open information extractors that more tightly fits downstream user needs. The most direct usage of information is

[1] For automatic language modelling purposes, on the other hand, extracted facts are a great source of learning material, as demonstrated in [8].

Sentence : In response, a group of Amherst College students held a patriotism rally in October, reciting the Pledge of Allegiance.
Fact : (a group of Amherst College students ; held ; a patriotism rally)

Sentence : That's a lot of maybes in a sport where the right thing seldom happens, but [given X, Y should Z if he wants T].
Fact : (That ; is ; a lot of maybes)

Sentence : The scandal has now forced resignations at Japan's fourth-largest bank and three of Japan's Big Four brokerages.
Fact : (Japan ; has ; fourth-largest bank)

Sentence : This leads to one of two inescapable conclusions : Either the president reads BioScope or I got lucky.
Fact : (the president ; reads ; BioScope or I got lucky)

Fig. 1. Extracted facts deemed correct by previous manual evaluations, respectively from ReVerb and ClausIE. In the first extraction, even though it is true in itself, crucial information from another sentence is missing to give the fact its meaning. In the second, the first argument it at best a vague idea. In the third, the fact holds true for most countries and holds little information by itself (it would be more adequate at another level of abstraction, e.g. *countries have banks*). The last extracted fact does not reflect the actual sentence meaning.

answering questions about it, so we evaluate extractors' output on their capacity to answer questions asked about the text at hand.

This procedure serves two purposes not previously addressed :

1. incorporate informativeness in the judging criterion for correct extractions ;
2. estimate recall for the task.

Section 3 details the evaluation methodology we follow, and the guidelines that drive our annotations. Section 4 presents the dataset we built and our basic automatic evaluation procedure, and section 5 exposes our results.

2 Related work

Little work has directly addressed the issue of evaluating OIE performance. Up to now, performance of extractors mostly relied on 2 metrics : number of extracted facts, and precision of extraction, area under the precision-yield curve being a shorthand for both [5]. As the bulk of extractions are usually obtained with a precision in the 70-80% range, consecutive generations of extractors have mostly improved on the yield part :

> "Ollie finds 4.4 times more correct extractions than ReVerb and 4.8 times more than WOE$^{\text{parse}}$ at a precision of about 0.75". [5]

> "ClausIE produces 1.8–2.4 times more correct extractions than Ollie". [3]

Sentence : For the 2006-07 season, Pace played with the Nelson Giants in the New Zealand National Basketball League.
4-ary fact : (Pace ; played ; with the Nelson Giants ; for the 2006-07 season ; in the New Zealand National Basketball League)

Questions :
Who did Pace play with ?
When did Pace play with the Nelson Giants ?
In what league did Pace play with the Nelson Giants ?

Fig. 2. Some facts are intrinsically n-ary, and naturally answer many questions. An extractor capturing only binary relations would likely miss much context.

It is commonly lamented that absolute recall cannot be calculated for this task, because of the absence of a reference. We aim to address this issue.

Typically, precision is measured by sampling extractions and manually labelling them as correct or incorrect. As a rule of thumb, an extraction is deemed correct if it is implied by the sentence :

> "Two annotators tagged the extractions as correct if the sentence asserted or implied that the relation was true." [5]

> "We also asked labelers to be liberal with respect to coreference or entity resolution; e.g., a proposition such as ('he', 'has', 'office'), or any unlemmatized version thereof, is treated as correct." [3]

By contrast with previously employed criteria, we propose to incorporate the informativeness of extracted facts, as measured by their ability to answer relevant questions, in the judgement of their validity.

Figure 1 shows examples of previous facts manually labelled as correct, taken respectively from ReVerb[2] and ClausIE[3]. Except for the last, they pass the standard criteria for correctness. We seek to devise an evaluation protocol that would reject such extractions on the grounds that they are not informative.

As highlighted by [1], one major element of OIE performance is the handling of n-ary facts. Most extractors focus on binary relations, but many support n-ary relations to some extent. KrakeN [1] and Exemplar [6] are designed towards n-ary extractions. ClausIE [3] supports generation of n-ary propositions when optional adverbials are present and Ollie [5] can capture n-ary extractions by collapsing extractions where the relation phrase only differs by the preposition[4]. Though it is not our main focus, our evaluation protocol addresses this question in that n-ary facts that capture more information will answer more questions than binary facts that would leave out some of the arguments. Figure 2 shows an example of 4-ary fact, and the corresponding questions it answers. Information extractors will be evaluated on their ability to answer such questions.

[2] `http://reverb.cs.washington.edu/reverb_emnlp2011_data.tar.gz`
[3] `http://resources.mpi-inf.mpg.de/d5/clausie/`
[4] This was added to the distributed software since publication – see `https://github.com/knowitall/ollie`

In [6], the authors present an experimental comparison of several systems, over multiple datasets, on the very similar task of open relation extraction (that differs from OIE by only considering named entities as possible arguments). Their study focuses on the tradeoff between processing speed — depth of linguistical analysis — and accuracy. Much of their discussion stresses the difficulties of building a common evaluation methodology that is fair to various methods. In particular, their section 3.3 is a good illustration of several evaluation difficulties.

3 Proposed Methodology

3.1 Evaluation Protocol

As hinted in figure 2, the evaluation protocol we propose for OIE is as follows :

> Given some input text, annotate all factoid questions that can be answered by information contained in that text, and the answers. Run the extractors on the input text. The evaluation metric is the number of questions that can be answered using only the output of each system.

The step of using extracted tuples to answer questions raises issues. Of course manual matching of extractions and questions would be most precise, but unrealistically expensive in human labor. An automatic scoring system also makes for more objective and easily replicable results, albeit less precise. Still, automatic question answering is a notoriously difficult problem that we would rather not tackle. In order to avoid this difficulty, we design our questions to be in the simplest possible form. Figure 3 shows a sample of our annotations. The questions are worded in very transparent ways.

We describe the automatic scoring system we use for this paper at greater length in section 4.2, and release it along with our data.

3.2 Annotation Process

We wish to annotate all questions that can be answered by information contained in the input text, in a way that is easy to answer automatically.

Our primary goal being to evaluate OIE systems, we found useful to consider their output on the sentences at hand as a base for annotation. As stated before (figure 1), many extractions are not informative by themselves. Examples of correctly extracted facts that cannot answer real-world questions are showed in the first sentence of figure 4.

Given OIE output, all extracted relations that are factually informative are asked about, i.e. a question is added that is expected to be answered by it. Then, we also ask all other questions that can be answered with information contained in the sentence, without being overly specific.

One could argue that the annotator seeing the output of the systems introduces a bias in the evaluation procedure. We do not believe this to be the case. As all OIE output is considered, the annotator is blind to specific extractors and

S: Esaka and six other top executives will quit to take responsibility for 67.28 million yen in payoffs to corporate racketeer Ryuichi Koike, 54.

Q: Why will Esaka quit ?
A: to take responsibility for 67.28 million yen in payoffs to corporate racketeer Ryuichi Koike
X: 1

YQ: Will Esaka quit ?
A: Yes
X: 1

Q: What age is Ryuichi Koike ?
A: 54
X: 1

S: And he has eased up on team rules.

S: His teammates indeed loved the show.

S: Mrs. Yogeswaran was shot five times with a pistol near her Jaffna home on May 17, 1998.

Q: What was Mrs. Yogeswaran shot with ?
A: a pistol
X: 1

Q: How many times was Mrs. Yogeswaran shot ?
A: five
X: 1

S: Robert Barnard (born 23 November 1936) is an English crime writer, critic and lecturer.
Q: Who is Robert Barnard ?
A: an English crime writer
(there is no X: annotation on this Q&A pair)

Q: When is Robert Barnard born ?
A: 23 November 1936
X: 0

Fig. 3. Examples of annotated sentences, with questions and answers. Questions are worded following the original text so that the question answering step is simple to perform. Many sentences are embedded in so specific a context that they do not carry any extractable information, like the second and third sentences. Others usually yield a handful of facts. Lines are prefixed as **S**entences, **Q**uestions, and **A**nswers (**YQ** stands for yes/no questions). In our dataset, most questions (but not all) are tagged with an e**X**pected result of the Q&A system, given the extractions seen by the annotator (these are the **X**: lines – 1 if the answer will be found, 0 if it won't), for intrinsic evaluation purposes.

Sentence: While the Arab world is a rich prize in itself, Europe has been and remains the primary objective.
Extraction: (Europe ; remains ; the primary objective)
Extraction: (the Arab world ; is ; a rich prize)
Important context is missing for these extractions to have meaning.

Sentence: Wu worked as a reporter for United Press International from 1973 until 1978 when she joined WGBH-TV, Boston's public television station, as the Massachusetts State House reporter until 1983.
Extraction: (Wu ; worked as a reporter for ; United Press International)
Sentence: Daughter of the actor Ismael Sanchez Abellan and actress and writer Ana Maria Bueno, Gabriel was born in San Fernando, Cadiz ...
Extraction: (Gabriel ; was born in ; San Fernando)
"Wu" is at the threshold of being sufficiently determined to ask questions about. "Gabriel" being a common first name, it is just on the other side of our threshold.

Sentence: He was born in New York and died at Livonia, Michigan.
Extraction: (He ; was born in ; New York)
Sentence: He consults his family doctor for solution.
Sentence: Had I known then what I know now, I might have argued for a different arrangement.
Coreference resolution is key in many sentences.

Fig. 4. Ambiguity due to the loss of context is the main issue for annotation.

the resulting dataset is fair to all sytems. If proper attention is paid to asking questions about facts that are not correctly extracted, then the measure of recall isn't biased towards technology performance either. In favor of using the information, it is easier to ask about useful extracted facts (e.g. in the very terms of the fact), which means all due credit is given to system successes. Additionally, annotation so helps reflecting on OIE capabilities and limitations.

3.3 Dealing with Ambiguity

The major issue in our search for informativeness is ambiguity due to lack of context. It is a problem on various levels, as exemplified in figure 4.

In some cases, each element of the relation is understandeable, but the whole meaning cannot be understood out of context — for instance (*"the Arab world is a rich prize")* and (*"Europe remains the primary objective")* in figure 4. When we cannot understand what the text at hand is about, we naturally do not annotate anything. In the last sentence of figure 4, (*I ; might have argued for ; a different arrangement*) similarly relates to a lost specific context.

Often, and this is the key difficulty, the arguments themselves only make sense in context. In *"the scandal forced resignations ..."* (figure 1), what *"the scandal"* refers to maybe was obvious in the news context of the time, but is lost to us.

Therefore, there is a somewhat arbitrary line to draw regarding the amount of context we can assume a potential user has in mind when asking questions.

Amongst the many shades of ambiguity, *Albert Einstein* and *New York City* are utmostly self-explanatory[5]. *"I"*, to the contrary, is completely dependant on its particular utterance, and a clear definition of its reference will most often pertain to the metadata of the document at hand.

In between, consider *"Wu"* in the second sentence of figure 4. We consider this name to be on the threshold of acceptable ambiguity for our purpose. Using context, we can trace her to be Janet Wu, an American television reporter from the Boston area[6]. Also consider *"Gabriel"* [7] in the following sentence, which we consider to be on the other side of the threshold.

We envision two possible policies regarding context requirements.

What we did is the following : assume there is no word sense disambiguation. Every argument phrase used in our dataset should refer to a single entity. When an input sentence is about the most common use of its words (like *Europe* for the continent), we consider it well defined and annotate it, using its words in those senses. When sentences are about less common senses of their words (like *Gabriel* for Ruth Gabriel[7] or *New York City* for the video game), we do not annotate them, on the grounds that further processing would be required to make use of such annotations, that is outside the scope of this work. It is normal for OIE to lose the context of extractions, and informativeness judgement calls shall take that into account.

Hence, by the fact that there is only one *"Wu"* in our corpus, it is a valid designation for an entity. *"Gabriel"*, on the other hand, is more often a common first name than refers to Ruth Gabriel, so we would not ask about her.

A looser policy would be to assume that the user that asks questions has in mind the same context as the writter of the original text. Within that view, some user may ask *"did the scandal force resignations ?"*, having in mind the japanese banking public embarrassment of figure 1, or *"what is the primary objective ?"*, thinking of them who sell to the Arab world in figure 4.

3.4 Coreference

In a way, coreferential mentions are the extreme case of the ambiguity-without-context problem just mentioned.

Currently available OIE systems don't resolve coreferential mentions, and a significant portion of extracted facts have "it", "he" or "we" as arguments. As such, we consider that these extractions cannot answer questions, as the reference of such mentions is lost. On the Wu sentence, systems extract (**she** ; *joined ; WGBH-TV as the Massachusetts State House reporter*), but we lack the means of answering *"did Wu join WGBH-TV ?"* with that fact.

[5] although they could refer to an American actor and a 1984 Atari video game.

[6] The sentence is from `https://en.wikipedia.org/wiki/Janet_Wu_(WCVB)`. Incidentally, `https://en.wikipedia.org/wiki/Janet_Wu_(WHDH)` also is an American television reporter who worked in the Boston area. We consider this to be an ironic coincidence, but stand by our arbitrary line.

[7] Ruth Gabriel is a Spanish actress.

In our dataset, we ask such questions that would require coreference to be resolved at extraction time when it happens inside a sentence. As all sentences in our dataset were randomly picked from documents, all the cross-sentential references are lost, and we do not ask questions about them. In figure 4, we can't ask about the main theme of the last three sentences because of that.

It would be natural to extend the evaluation protocol to such facts, and it would enrich the measure of recall to examine how many facts are spread over several sentences, by annotating a whole document. Considering that the issue is not currently addressed by available systems (they treat all sentences independently), it would be a moot point for now.

4 Evaluation Data and Program

4.1 Question-Answer Dataset

As a proof of concept, we annotated slightly more than 100 Q&A pairs on a corpus previously employed for OIE evaluation. Rather than aiming for these to become a standard dataset, we encourage other researchers to write their own questions datasets, tailored to the needs of their particular OIE systems, and enrich the pool of available resources for evaluation.

The data we annotate is that distributed[3] by [3]. Sentences were randomly picked from 3 sources :

- 500 sentences are the so-called "ReVerb dataset", obtained from the web via a Yahoo random-link service ;
- 200 sentences come from Wikipedia ;
- 200 sentences come from the New York Times.

A sample of annotations is shown in figure 3. To give an idea and as discussed in section 3.3, about half of sentences are not suitable to ask meaningful questions about. On the other sentences, we typically find 2-4 questions that can be answered with their information (2.3 on average on the reverb dataset, 3 on the wikipedia sentences).

The data is available at `http://www.CICLing.org/2016/data/92`.

4.2 Question Matcher

With annotated Q&A pairs and extracted information in hand, the remaining step is to match and evaluate answers to the questions. We develop a very basic Q&A system, based on string matching. In short, it matches the text of extracted facts to questions, and assumes that a good match indicates the presence of an answer. It is not a very good Q&A system, but it is a decent evaluation script. Its most important features are to be fair to all systems, and easy to use, understand and replicate. Figure 5 illustrates how the evaluation system works.

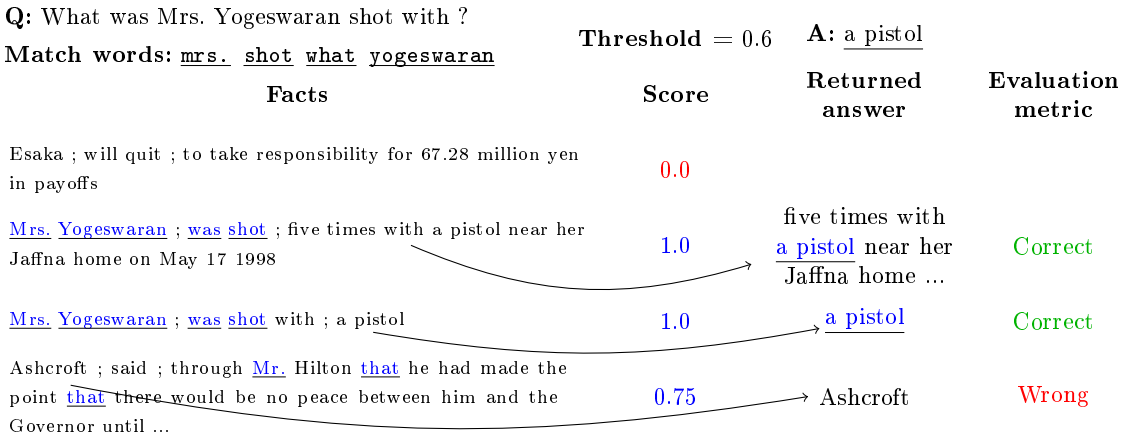

Fig. 5. Q&A script based on string matching. A fact that matches the words of a question past a given threshold is assumed to contain an answer to it. The part of that fact (arg1, rel, or arg2) that least matched the question is picked as the answer. A candidate answer is correct if it contains all the words of the reference answer.

As mentioned in section 3.1, and by contrast with the task of open-domain question answering, e.g. studied in [4], we deliberately do not address the difficult problem of question understanding. Instead, questions were written in transparent ways so as to facilitate their automatic answering (see figure 3).

In order to answer a given question, the system attempts to match each extracted fact to it, at the word level. A fact that matches more than 60%[8] of a question's words is assumed to contain an answer to the question. Stopwords are excluded, using NLTK [2]. Edit distance is used to relax the words matching criteria[9], to make up for slight morphological variations between the words of interrogative questions and that of affirmative facts.

When a fact matches a question, we look for the part (first argument, relation, second argument) that least matches the question, and pick it as an answer (typically the second argument, arguments are favored in case of equality, as in the last line of figure 5). We consider an answer to be correct when all the gold answer words are contained in the returned answer.

Were we to build a true question-answering system, we would need to pick or order the various candidate answers gathered for each question. In practice, we seek to devise an evaluation script, and there are only a handful[10] of answers per question, so we consider that a question is correctly answered if any of its candidate answers is correct.

[8] We examine the impact of this factor in section 5.2.

[9] Words differing by 1 or less of their characters are considered to match, as *Mrs.* and *Mr.* in figure 5.

[10] See table 2 as the exact figure is directly dependant on the matching threshold.

5 Results

5.1 Performance

The results of the evaluation procedure are showed in table 1. On factoid questions, the automatic answering system finds candidate answers to 35-70% of questions, depending on the information extractor. When candidate answers are found by the answering system, at least one is correct in 20-50% of cases.

Most importantly, recall measures how much of the sentences' relevant information was captured by the extractions, and we can see that combining all the systems output, nearly 40% of the information is captured.

Our results fall in line with previous authors' findings. Pattern learning makes Ollie a significant improvement over the simplistic mechanism of ReVerb (at the computational price of dependency parsing), making it both more precise and yielding more facts, leading to a large increase in recall. ClausIE extracts more facts than Ollie at a similar level of precision, further boosting recall.

Practically though, both being open-source software, Ollie runs significantly faster than ClausIE, due to the difference in the embedded parsers they use.

Table 1. OIE systems performance results. Answered is the proportion of factoid questions for which at least one answer was proposed (correct or not) ; Precision is the proportion of answered questions to which one or more answers were correct ; and Recall is the proportion of questions for which at least one candidate answer is correct. As a matter of fact, given the way metrics are computed, answered × precision = recall.

	Answered	Precision	Recall
ReVerb	35%	19%	6%
Ollie	39%	43%	17%
ClausIE	68%	42%	29%
All	71%	53%	38%

5.2 Analysis

In order to assess the quality of the evaluation script in terms of desired behaviour, manual assessment of whether a correct answer would be found or not was annotated on a sample of questions (80 out of 106). These are the **X:** lines in figure 3. This would be similar to a human-judged step of answering the questions given the facts, and comparison of the result of the automatic procedure with respect to the manual evaluation (but the annotator tagged it's expectation of the automatic procedure, rather than the desired behavior as in the manual judge case). On this sample, the evaluation system performed as predicted in upwards of 95% of cases, which is satisfying.

An important parameter of our approach, mentionned in section 4.2, is the matching threshold past which a fact is assumed to contain an answer to the

question it matched. Table 2 shows the impact of this parameter on our results, using extractions from all systems.

As expected, the lower the threshold, the looser the answers, and the higher the recall. We retained 0.6 as threshold for performance measures, for it has the highest precision, and above all maximises recall while keeping the average number of candidate answers reasonable (less than 5 rather than more than 20).

Table 2. Impact of matching threshold on evaluation metrics.

Matching threshold	0.5	**0.6**	0.7	0.8
Questions answered	95%	**71%**	48%	31%
Answers per question	23.	**3.7**	2.5	2.0
Precision	51%	**53%**	46%	38%
Recall	48%	**38%**	22%	12%

6 Conclusion and Future Work

We presented a new protocol for evaluation of OIE, that consists in annotating questions about the relevant information contained in input text, and automatically answering these questions using systems' output. Our performance metric more closely matches the usefulness of OIE output to end users than the previously employed methodology, by incorporating the informativeness of extracted facts in the annotation process. In addition, our protocol permits to estimate the recall of extraction in absolute terms, which to the best of our knowledge had never been performed. According to our results, about 40% of pieces of knowledge present in sentences are currently extracted by OIE systems.

We annotate a small dataset with Q&A pairs, and present our annotation guidelines, as well as the evaluation script we developed, in the form of a rudimentary Q&A system. We distribute our annotations and evaluation system to the community[11].

As directions for future work, we would like to annotate whole documents rather than isolated sentences, and measure the proportion of cross-sentential information. Our framework also naturally allows for evaluation of other text understanding systems, such as semantic parsers, or full-fledged question answering systems in the place of our own, which would be interesting to perform.

[11] `http://www.CICLing.org/2016/data/92`

Bibliography

[1] Akbik, A., Löser, A.: Kraken: N-ary facts in open information extraction. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. pp. 52–56. AKBC-WEKEX '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), http://dl.acm.org/citation.cfm?id=2391200.2391210

[2] Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly, Beijing (2009), http://www.nltk.org/book

[3] Del Corro, L., Gemulla, R.: Clausie: Clause-based open information extraction. In: Proceedings of the 22Nd International Conference on World Wide Web. pp. 355–366. WWW '13, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2013), http://dl.acm.org/citation.cfm?id=2488388.2488420

[4] Fader, A., Zettlemoyer, L., Etzioni, O.: Open question answering over curated and extracted knowledge bases. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1156–1165. KDD '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2623330.2623677

[5] Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open language learning for information extraction. In: Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL) (2012)

[6] Mesquita, F., Schmidek, J., Barbosa, D.: Effectiveness and efficiency of open relation extraction. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 447–457. Association for Computational Linguistics (October 2013)

[7] Soderland, S., Gilmer, J., Bart, R., Etzioni, O., Weld, D.S.: Open information extraction to KBP relations in 3 hours. In: Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19, 2013. NIST (2013), http://www.nist.gov/tac/publications/2013/participant.papers/UWashington.TAC2013.proceedings.pdf

[8] Stanovsky, G., Dagan, I., Mausam: Open ie as an intermediate structure for semantic tasks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 303–308. Association for Computational Linguistics, Beijing, China (July 2015), http://www.aclweb.org/anthology/P15-2050