Analysing Errors of Open Information Extraction Systems

Rudolf Schneider, Tom Oberhauser, Tobias Klatt,
Felix A. Gers and **Alexander Löser**
{rudolf.schneider, tom.oberhauser, tobias.klatt, gers, aloeser}@beuth-hochschule.de
Beuth University of Applied Sciences Berlin
Luxemburger Str. 10, 13353 Berlin, Germany

Abstract

We report results on benchmarking Open Information Extraction (OIE) systems using RelVis, a toolkit for benchmarking Open Information Extraction systems. Our comprehensive benchmark contains three data sets from the news domain and one data set from Wikipedia with overall 4522 labeled sentences and 11243 binary or n-ary OIE relations. In our analysis on these data sets we compared the performance of four popular OIE systems, ClausIE, OpenIE 4.2, Stanford OpenIE and PredPatt. In addition, we evaluated the impact of five common error classes on a subset of 749 n-ary tuples. From our deep analysis we unreveal important research directions for a next generation of OIE systems.

1 Introduction

Open Information Extraction (OIE) is an important intermediate step of the nlp stack for many text mining tasks, such as summarization, relation extraction, knowledge base construction and question answering (Mausam, 2016; Stanovsky et al., 2015; Khot et al., 2017). OIE systems are designed for extracting n-ary tuples from diverse and large amounts of text, without being restricted to a fixed schema or domain. These tuples consist of one predicate and n arguments e.g: *flew*(Obama; from Berlin; to New York).

Users often desire to select a suitable OIE system for their specific application domain. Making the right choice is not an easy task. Unfortunately, there is surprisingly little work on evaluating and comparing results among different OIE systems. Worse, most OIE methods utilize proprietary and unpublished data sets. In most cases users can only rely on publications and need to download, compile and apply existing systems to their own data sets.

Contribution Ideally, one could compare different OIE systems with a unified benchmarking suite. As a result, the user could identify "sweet spots" of each system but also weaknesses for common error classes. The benchmarking suite should feature a diverse set of gold annotations with several thousands of annotated sentences. By exploring results and errors, the user can learn how to design the next generation of OIE systems or how to combine several systems into an ensemble.

Our contributions are: (1) We report results of a quantitative analysis on four commonly used OIE systems: STANFORD OPENIE (SIE) (Angeli et al., 2015), OPENIE 4.2 (OIE)[1], CLAUSIE (CIE) (Del Corro and Gemulla, 2013) or PREDPAT (PP) (White et al., 2016). Which employ rule based as well as machine learning based methods on linguistic structures like dependency parses. These were applied on *4522* sentences and *11243* n-ary gold standard tuples. (2) We share in-depth insights on a qualitative error analysis of *749* n-ary tuples in *68* sentences from four gold standard data sets annotated by all four OIE systems. (3) We provide an integrated benchmark for OIE systems consisting of three news data sets NYT-222, WEB-500 (Mesquita et al., 2013), PENN-100 (Xu et al., 2013) and a large OIE benchmark from Newswire and Wikipedia (Stanovsky and Dagan, 2016) combined in our evaluation tool *RelVis*. Our benchmark tool will be provided to the community under an open source license.

The remainder of this paper is structured as follows: First, Section 2 gives detailed insights on methods used for qualitative and quantitative evaluation. Section 3 introduces our evaluation system

[1] https://github.com/allenai/openie-standalone

Name	Type	Domain	Sent.	# Tuple
NYT-222	n-ary	News	222	222
WEB-500	binary	Web/News	500	461
PENN-100	binary	Mixed	100	51
OIE2016	n-ary	Wiki	3200	10359

Table 1: Data sets in RelVis

in a demo walkthrough. We report in Section 4 in-depth insights on our experiment results. Finally, Section 5 concludes with design recommendations for next generation OIE systems.

2 Analysing Open IE Systems

We set up two experiments with four OIE systems STANFORD OPENIE (Angeli et al., 2015), OPENIE 4.2[2], CLAUSIE (Del Corro and Gemulla, 2013) and PREDPAT (White et al., 2016) and four gold standard data sets. The qualitative analysis was done by two human judges, who classified errors in the output of the systems into six categories. Our qualitative analysis includes gold labeled data sets from previous evaluations, shown in Table 1.

2.1 Data sets

Our evaluation process for Open Information Extraction systems should be convenient and comparable. To meet this goal, we deliver supplementary scripts to import commonly used data sets with our evaluation system RelVis. The unified data model enables the user to perform quantitative comparisons and extensive analyses on widely used data sets. We used in our experiments four data sets, see Table 1, of which two feature only binary relations with two arguments. Data sets *NYT-222* and *OIE2016* also contain n-ary relations. These labeled data sets origin from Mesquita et al. (2013) and Stanovsky and Dagan (2016).

2.2 Measuring OIE Systems

A naive way to match a tuple to a gold standard is an *equal match*. Enforcing equal matching of boundaries in text to a gold standard delivers exact results for computing precision. However, this strategy penalizes other, potentially correct, boundary definitions beyond the gold standard. Dealing with multiple OIE systems and their different annotation styles requires a less restrictive matching strategy. A second strategy is a *containment match* where an argument or predicate is considered correct, if it at least contains a gold standard annotation. Hence, spans from the gold standard may be contained (fully) inside the spans of the annotation from the OIE system. However, this strategy may label over-specific tuples as correct and may lead to a lower precision and penalizes binary systems on n-ary data sets.

Therefore we introduce a *relaxed containment strategy* which removes a penalty for wrong boundaries especially for over-specific extractions. This strategy counts an extraction correct, even when the number of arguments doesn't match the gold standard. For example, Stanford OIE, a system that only returns binary OIE tuples, performs well on *NYT-nary (b)*, an n-ary data set and yields large parts of relatively short sentences as one argument. With the relaxed matching strategy Stanford OIEs binary extractions are counted correct as long as they contain all gold standard arguments.

The approach of Mesquita et al. (2013) has simplified the task by replacing all entities in the test set with the words "Europe" and "Asia". In our opinion this decision is contrary to the definition of OpenIE given by Banko et al. (2007) which describes OIE as *"domain-independent discovery of relations extracted from text and readily scales to the diversity and size of the Web corpus."* and may hide or even cause problems in the analysed systems.

Measurements. In our quantitative evaluation we calculate precision, recall and F_2 measure at sentence level. Following Pink et al. (2014), we choose F_2 instead of F_1 measure because it gives the recall a larger impact. The basic intuition is that a high recall of an OIE system is critical to the performance of any downstream application that can apply additional filters.

2.3 Common Error Classes.

Authors of OIE systems distinguish among six major error classes, Table 3 reports errors for our four OIE systems. In the following paragraphs we describe each error class in detail.

Wrong Boundaries. Banko et al. (2007) describe this error as *too large or too small boundaries for an argument or predicate of an OIE extraction*. In each of the four OIE systems we observe wrong boundaries for at least one third of the results. This indicates that OIE systems often

[2] https://github.com/allenai/openie-standalone

fail in generalizing to unseen word distributions. This might be caused by errors in used intermediate structures, such as dependency parses, or overestimation of boundaries. Incorrect boundaries for relation arguments can prohibit fusing, linking or aggregating tuples for the same predicate. As a consequence, an additional system needs to filter out incorrect boundaries which may cause a drastic recall loss.

A solution proposed in the literature is to 'wait' until intermediate systems, such as dependency parser, POS tagger etc., provide an improved generalization. However, this may not always be the case for niche domains, such as medical text or text in enterprise scenarios, where often no labeled corpora for those intermediate systems exist.

Following Arnold et al. (2016a,b) we suggest end-to-end architectures, such as TASTY, an end-to-end named entity recognition and named entity linking system. TASTY leverage stacked deep learning architectures and requires only a few hundred labeled annotations to reach high F-measures in various domains and languages.

Redundant Extraction. In absence of a schema, OIE systems output redundant extractions for the same sentence, such as for the same subject-predicate structure. For example, in the sentence *"Additionally, we included some other relevant results from the 2005 survey in Antwerp."* SIE yields two times the tuple (we, included, other relevant results). These OIE systems are tuned towards high recall and leave the decision to filter out redundant tuples to a downstream application (Del Corro and Gemulla, 2013). The OIE system SIE which returns in extreme cases up to 140 tuples for the same sentence. Our results indicate that this error class has been resolved to a large extent in most systems by filtering and aggregating results from multiple similar extraction rules.

Uninformative Extraction. Following Fader et al. (2011), *uninformative extractions are extractions that omit critical information.* This type of error is caused by improper handling of relation phrases that are expressed by a combination of a verb with a noun, such as light verb constructions (LVCs). Adding syntactic and lexical constraints may solve this problem to certain extent.

Missing Extraction - False Negatives. This class describes relations which were not found by a particular system. According to Fader et al. (2011), missing extractions are often caused by argument-finding heuristics choosing the wrong arguments, or failing to extract all possible arguments. One example is the case of coordinating conjunctions. CIE and OIE can spot certain cases of coordinating conjunctions and do miss fewer tuples. Other sources of this error are lexical constraints filtering out a valid relation phrase, another source are errors in dependency parsing. Overall, we observe a trade-off among OIE systems between utilizing lexical constraints for filtering out uninformative tuples and thereby creating false negatives. Our results indicate that system OIE handles this trade-off slightly better than other systems.

Wrong Extraction. Stanovsky and Dagan (2016) consider a tuple as correct as long as it shares a specified threshold of characters with a gold annotation. However, this policy may lead to emitting large parts of a sentence as one argument and poses additional computation effort to a downstream application. We focus on sentence-level correctness (Mesquita et al., 2013; Angeli et al., 2015) and define a tuple as correct if the following conditions are met:

1. The selected matching strategy yields a match for the predicate.
2. The number of arguments aligns with the gold standard.
3. The selected matching strategy yields a match for all arguments.

This error class is critical since it is not possible to recover from a error of this class and it emits a wrong signal which might trigger additional errors in downstream tasks.

Out of Scope. We observe in Table 3 that our OIE systems yield more correct extractions as recognized by authors of gold data sets. For these additional annotations, we introduce an *out of scope* category. This label does not indicate an error, but it helps us from distinguishing errors of gold labels and additional annotations of a particular OIE system that are not present in the gold standard. Our two judges marked an annotation, in the qualitative evaluation, as out of scope if it is valid and provides an information gain. If marked as out of scope, no other error category is applied to the extraction.

3 The RelVis System

In the following Section, we guide through our OIE benchmark system which was used to perform the quantitative and qualitative analysis. We show how the system can support a user in such sophisticated evaluation processes.

Startup. At system initialisation, RelVis reads gold-annotations. Next, the system stores extraction and gold annotations in a RDBMS from which a web based front end visualizes text data and annotations.

Dashboards for exploring annotations. Now, the user can start exploring results and understanding the behaviour of each system. Figure 1 visualizes in a dashboard example sentences, precision, recall and F_2 scores for each OIE system and for each error class.

Please note, RelVis plots error distributions as a Kiviat diagram and draws bar charts for comparing error class impacts for each OIE system. In addition, the user can export results as tables and CSV files from the database, as shown in Table 3 and Table 2.

Understanding and adding a single annotation. RelVis visualizes OIE extractions on sentence level. Figure 3 shows how the dashboard visualizes example sentences. For each hit by a system, the user can drill down into a single sentence and can understand extraction predicates or arguments. Next, she can dive down into correct or incorrect annotations, can add labels for error classes for incorrect annotations or may leave a comment, see also Figure 2. We permit the user to apply multiple error classes to each part of an annotation, such as a predicate or argument. Next, she can focus on a sentence of interest and can compare extractions between different OIE systems.

We permit the user to update or add new annotations with a BRAT style functionality (Stenetorp et al., 2012), optimized for n-ary OIE relations.

Figure 3 shows a screenshot to illustrate the process. The user selects a sentence to annotate and starts with the first annotation by clicking on the "Add new OIE Relation" button (6). Next, she marks the predicate and arguments in the sentence for her first annotation by selecting them with the cursor and interacting with Button (1) and (2). The system indicates predicates (4) in green and arguments (5) in blue colour.

4 Experiment Results

In a **quantitative evaluation** we report precision, recall and F_2 scores on all four data sets. Table 2 reports overall results for four OIE systems on all four data sets, with the limitation that only a subset of OIE2016, containing 1768 sentences, was available to us. We conduct our experiments with an exact (a) and relaxed (b) containment match strategy.

For the **qualitative evaluation** we execute four OIE systems on 17 sentences of each data set. This resulted in 749 predicted extractions which we evaluate and classify into error categories by two human judges, as shown in Table 3. Additionally, Figure 4 gives an overview of the general performance of all tools over all data sets. We apply a strict containment match strategy in this evaluation. Observing that multiple errors can happen to a single extraction, we assign in these cases more than one error category.

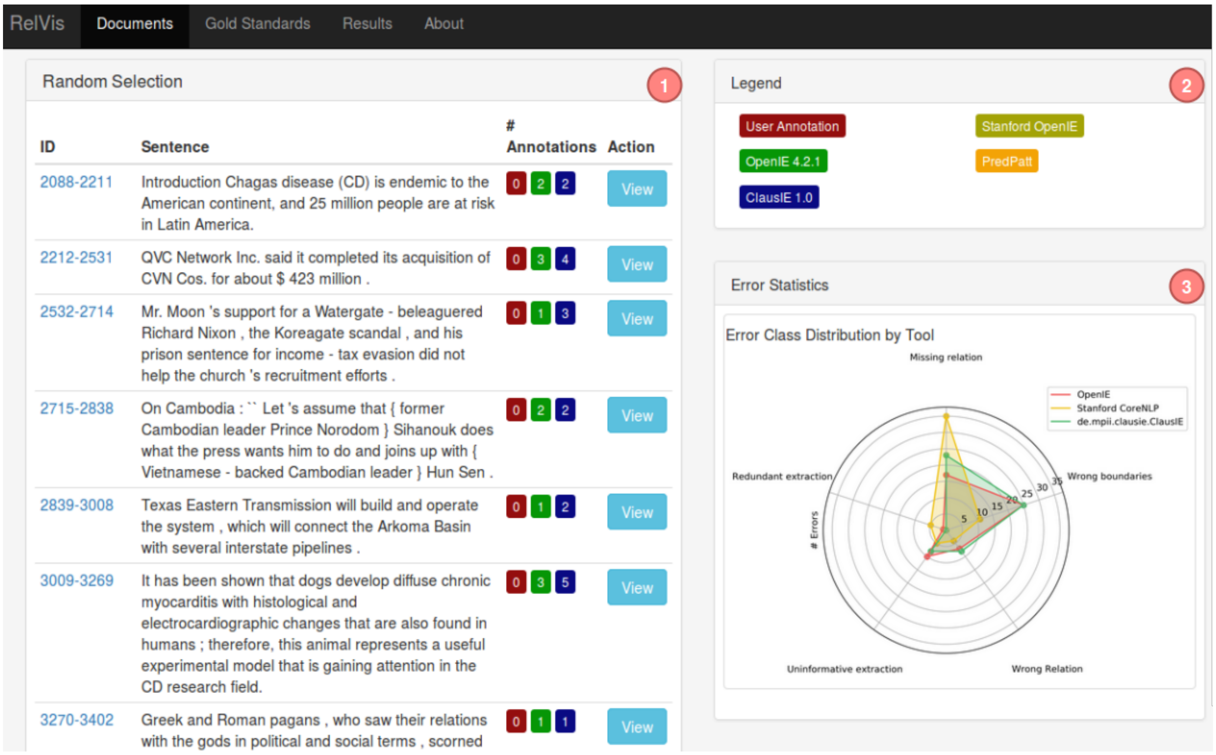
Note, for both experiments, we configure system CIE to binary extraction mode for binary data sets and otherwise in n-ary mode.

4.1 General Findings

We observe no clear overall winner: Each OIE system works best on a particular data set, and no OIE system significantly outperforms on two or more data sets.

Boundary Errors. We observe that an OIE system causes boundary errors often by over- or under-specific argument spans. In more rare cases the source for this error are predicate spans. Both, argument and predicate related errors can be caused by wrong intermediate structures in a particular OIE system. Another source of the problem could be the argument candidate generation, which overestimates the size of an argument span, so that it envelops multiple distinct arguments. Further causes for a boundary error are different annotation styles, which appear among systems as well as among gold standard data sets.

As one possible source for the overall bad results on the NYT-222 dataset, we pinpoint the differing styles of conjunction extraction. Consider a gold standard which expects a single extraction with multiple arguments for the sentence: "*DENVER BRONCOS signed LB Kenny Jackson, DT Garrett Johnson and CB Sam Young.*" like e.g. *signed(DENVER BRONCOS; Kenny Jackson; Garrett Johnson; Sam Young)*. Systems CIE and



Random Selection

ID	Sentence	# Annotations	Action
2088-2211	Introduction Chagas disease (CD) is endemic to the American continent, and 25 million people are at risk in Latin America.	0 2 2	View
2212-2531	QVC Network Inc. said it completed its acquisition of CVN Cos. for about $ 423 million .	0 3 4	View
2532-2714	Mr. Moon 's support for a Watergate - beleaguered Richard Nixon , the Koreagate scandal , and his prison sentence for income - tax evasion did not help the church 's recruitment efforts .	0 1 3	View
2715-2838	On Cambodia : `` Let 's assume that { former Cambodian leader Prince Norodom } Sihanouk does what the press wants him to do and joins up with { Vietnamese - backed Cambodian leader } Hun Sen .	0 2 2	View
2839-3008	Texas Eastern Transmission will build and operate the system , which will connect the Arkoma Basin with several interstate pipelines .	0 1 2	View
3009-3269	It has been shown that dogs develop diffuse chronic myocarditis with histological and electrocardiographic changes that are also found in humans ; therefore, this animal represents a useful experimental model that is gaining attention in the CD research field.	0 3 5	View
3270-3402	Greek and Roman pagans , who saw their relations with the gods in political and social terms , scorned	0 1 1	View

Figure 1: Screenshot of the sentence selection view of RelVis. (1) For each sentence in the document we show text and number of extractions by system. (2) The "Legend panel" denotes various OIE systems with different colours. (3) The lower right hand side shows visualizations of error evaluation statistics.

Dataset	**ClausIE (%)**			**OpenIE 4.2 (%)**			**Stanford OIE (%)**			**PredPatt (%)**		
	P	R	F_2	P	R	F_2	P	R	F_2	P	R	F_2
PENN-100 (a)	4.00	21.15	11.39	12.41	36.54	26.31	**14.85**	**57.69**	**36.58**	6.83	42.30	20.75
PENN-100 (b)	4.00	21.15	11.39	13.07	38.46	27.70	**14.85**	**57.69**	**36.59**	7.76	48.08	23.58
WEB-500 (a)	**16.33**	**46.70**	**34.03**	12.83	19.62	17.74	13.65	40.72	29.16	5.18	13.43	10.19
WEB-500 (b)	**16.33**	**46.70**	**34.03**	13.39	20.47	18.51	13.65	40.72	29.16	6.09	15.78	11.97
NYT-222 (a)	1.64	5.85	3.87	**2.86**	7.66	5.73	0	0	0	2.22	**13.51**	**6.71**
NYT-222 (b)	4.69	16.67	11.03	11.28	30.18	22.60	**13.37**	**73.87**	**38.77**	8.47	51.35	25.51
OIE2016 (a)	14.81	13.67	13.89	**24.85**	**18.69**	**19.67**	0.80	1.49	1.27	7.26	12.39	10.86
OIE2016 (b)	20.38	18.81	19.10	**39.58**	**29.76**	**31.31**	3.83	7.10	6.07	13.52	23.09	20.23

Table 2: Quantitative Evaluation. The (b) variant are results with relaxed containment match strategy and (a) are those with the strict containment strategy.

OIE yield persons and their positions as one large argument in a binary relation: *signed(DENVER BRONCOS; LB Kenny Jackson, DT Garrett Johnson and CB Sam Young.)*. On the contrary, System PP implements another style extracting every person of the sample sentence in an own binary relation.

SIE, a binary extraction system, performs surprisingly well on this data set with the relaxed containment match strategy and on *NYT-222 (b)*. With a strict containment match strategy, NYT-222 (a), the system was not able to find a correct extraction, because the data set does not contain binary relations. Using a relaxed containment match strategy, system SIE outperforms all other systems, by extracting large, over-specific arguments. This shifts additional effort for further processing towards downstream applications. This shows the importance of taking boundaries into account in an evaluation. However, system SIE fails on the extraction of OIE2016, which contains more complex sentences, including numerical values and multiple gold annotations in comparison to NYT-222.

Dataset	**NYT-222** (n-ary)				**OIE2016** (n-ary)				**PENN-100** (binary)				**WEB-500** (binary)			
# Relations	17				29				17				17			
	CIE	OIE	PP	SIE	CIE	OIE	PP	SIE	CIE	OIE	PP	SIE	CIE	OIE	PP	SIE
# Predicted	42	35	68	74	28	30	57	91	63	34	61	49	33	22	24	38
# Correct	2	1	**6**	0	8	**12**	6	5	4	8	10	**11**	5	4	3	**10**
# Redundant	0	0	0	**5**	0	0	0	**18**	1	0	0	**4**	**2**	0	0	0
# Uninformative	4	2	**8**	0	2	0	**6**	1	**9**	3	**9**	4	0	0	0	**3**
# Boundaries	11	17	18	**39**	11	11	21	**69**	**14**	5	9	**14**	8	**9**	**9**	**9**
# Wrong	2	1	3	**5**	1	1	**6**	3	3	1	**10**	4	1	**2**	**2**	**2**
# Out of Scope	24	17	**34**	30	7	6	**21**	13	**33**	17	31	18	**19**	8	12	14
# Missed	4	1	**5**	**5**	8	4	7	**12**	**14**	6	6	7	8	3	**11**	6

Table 3: Occurrences of extraction errors found in the qualitative analysis of four OIE systems on 17 sentences drawn from four gold standard datasets. 749 predicted extractions were evaluated in total. Note: multiple errors per predicted extraction are possible and that number of missed extractions is naturally not contained in # Predicted.

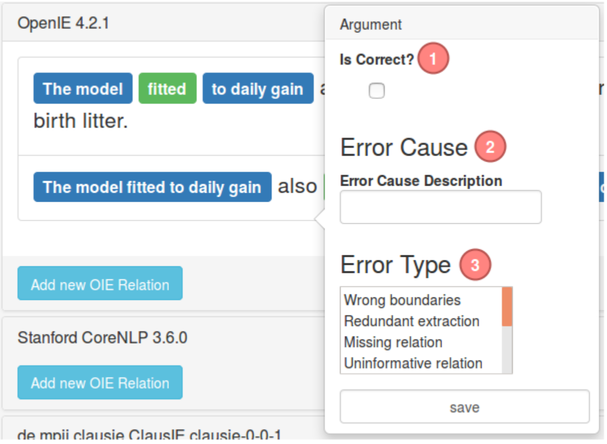


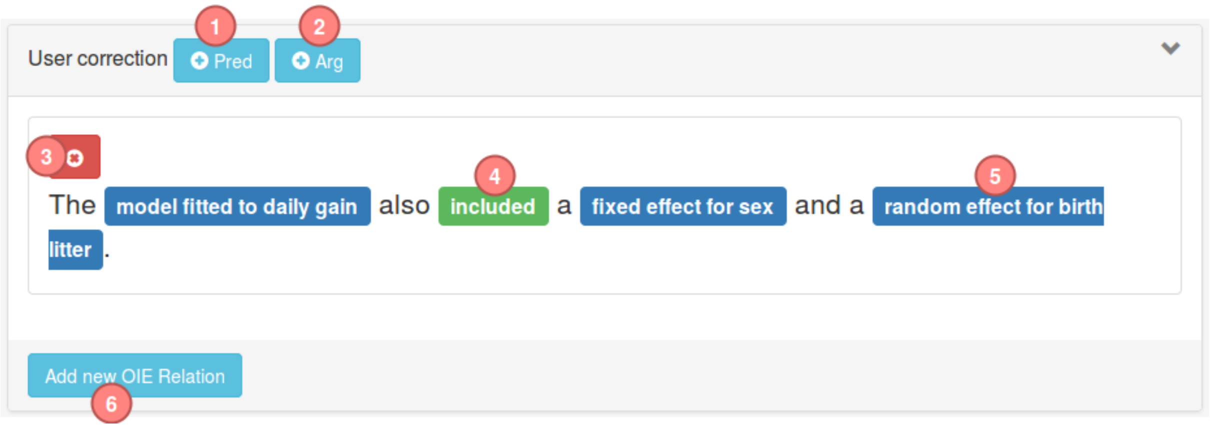
Figure 2: Interface for specifying the correctness (1), comment on error cause (2) and error class (3) of an OIE extraction.

Missed Extractions. Noisy text, wrong intermediate structures and different annotation styles among gold data sets often trigger this error. We report a significant drop in recall for all systems on the WEB-500 dataset compared to PENN-100, except for CIE, see Table 2, even though both data sets show a similar annotation style. However, the WEB-500 data set is quite noisy and contains HTML-character encodings, unfinished sentences or headlines with special characters. Those artifacts cause errors in intermediate structures, like dependency parses or POS tags, which causes the systems to fail. In particular, the n-ary systems OIE or PP do not seem to be robust to such noisy data.

Another source for missed relations is a mismatch between annotation styles. For example, system CIE shows a different style as the gold annotation in PENN-100, NYT-222 and WEB-500 data sets. A closer inspection reveals that CIE's verb centric extraction behaviour handles nominal or adjectival triggered relations (Peng et al., 2014) in a different style as the gold standard data set. Its design triggers inserting an artificial predicate (Del Corro and Gemulla, 2013) which can cause many missed annotations in our evaluation. For example, consider the following sentence: "*At least one potential GEC partner, Matra, insists it isn't interested in Ferranti.*" System CIE extracts the tuple: *is(one potential GEC partner; Matra)*, but the style of the gold standard expects: *partner(GEC; Matra)*. We explain the increase of all scores of system CIE by the larger number of gold annotations, compared to PENN-100, which does not interfere with the annotation style of system CIE.

Wrong and uninformative unary extractions. Wrong extraction errors are in many cases complex and caused by other errors. For example, a boundary error often leads to missing a important information like a negation. Furthermore, we observe problems in the predicate candidate selection process for unary extractions which leads to wrong extractions.

Uninformative extractions are mostly yielded by systems CIE and PP. In many cases, these errors are triggered in possessive relations without resolved co-references or relations with adjectival triggers, e.g. *first(world war)*. To overcome these problems, we suggest to improve filtering for uninformative unary relations, supply additional



The model fitted to daily gain also included a fixed effect for sex and a random effect for birth litter.

Figure 3: Visualization of OIE annotations. The predicate is marked green (4) and arguments blue (5). Buttons for adding a predicate (1) or argument (2) are on top. The button for adding another OIE Annotation (6) is on the bottom. In the top left corner is the delete annotation button (3).

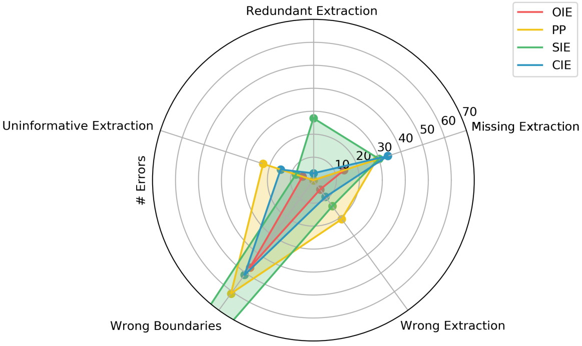


Figure 4: Error occurrence of all OIE systems on 68 sentences of four data sets. Error categories described in **??** are plotted along five axes. The system with the smallest covered area makes the least errors. We crop the diagram at 70 occurrences for easier interpretation. SIE hits 131 times in total the Wrong Boundaries category.

checks for missed negations or important arguments and integrate a co-reference resolution components into next generation OIE systems.

Redundant Extractions Redundant extractions exclusively occur in systems SIE and CIE[3].

4.2 Data set Specific Findings

OIE systems are still designed towards binary tuples. Very first OIE systems had been designed towards emitting binary OIE tuples. Therefore, we observe that all systems achieve a better recall score on the binary data sets when the strict containment strategy is used. This is caused by larger number of possible errors in an n-ary task. Additionally, inconsistent extraction styles for n-ary relations in both, systems and gold standards, cause errors.

Out of Scope. The PENN-100 data set supplies for every sentence just one gold extraction. In most cases it represents a non verbal triggered relation. This leads to many out of scope extractions, because most of the systems perform well in extracting verbal triggered relations. Each OIE system yields out of scope extractions in particular on the NYT-222 data set, which shows that the gold annotations in this data set do not cover capabilities of modern OIE systems.

Data set OIE2016 features the lowest number of out of scope extractions overall. It provides multiple gold annotations per sentence and covers a wide variety of extractions, starting with unary up to 7-ary tuples. System PP yields non verbal triggered unary extractions more often than other systems, which is the reason for its steady high number of out of scope extractions.

5 Conclusion

To our best knowledge this is the first attempt of a comprehensive in-depth error analysis, containing quantitative and qualitative evaluations, of four OIE systems on four data sets. In our future work we will publish our benchmark system RelVis, data sets and adapters under an open source licence for the general OIE community.[4]

Because of the nature of the OIE task, we conclude that there is a lack in stringent annota-

[3] in binary extraction mode

[4] https://github.com/SchmaR/RelVis

tion policies, which makes a comparative analysis but also the design of OIE system often difficult. Moreover, each tested OIE system depends on syntactic taggers that often propagate errors towards the logic for extracting OIE tuples. We also observe fewer errors among binary OIE tuples. This indicates that current OIE systems have not reached an effective design yet for extracting higher order n-ary tuples. Only system PP leverages well researched ideas from normal forms in data base theory in its design.

We suggest designers of next generation OIE systems to test their systems against various data sets, even data sets in idiosyncratic domains not included in this benchmark. Moreover, next generation OIE systems should offer some convenient 'knobs' for tuning it towards common downstream tasks, such as populating a knowledge base or extracting typed relations against a schema.

Acknowledgements

Our work is funded by the German Federal Ministry of Economic Affairs and Energy (BMWi) under grant agreement 01MD16011E (Medical Allround-Care Service Solutions), grant agreement 01MD15010B (Smart Data Web) and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 732328 (FashionBrain).

References

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. *Linguistics* (1/24).

Sebastian Arnold, Robert Dziuba, and Alexander Löser. 2016a. TASTY: Interactive Entity Linking As-You-Type. In *COLING Demos 2016*. pages 111–115.

Sebastian Arnold, Felix A. Gers, Torsten Kilias, and Alexander Löser. 2016b. Robust Named Entity Recognition in Idiosyncratic Domains. In *arXiv:1608.06757 [Cs.CL]*.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *IJCAI*. volume 7, pages 2670–2676.

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

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*. Association for Computational Linguistics, pages 1535–1545.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering Complex Questions Using Open Information Extraction. *arXiv:1704.05572 [cs]* .

Mausam Mausam. 2016. Open Information Extraction Systems and Downstream Applications. New York.

Filipe Mesquita, Jordan Schmidek, and Denilson Barbosa. 2013. Effectiveness and Efficiency of Open Relation Extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 447–457.

Yifan Peng, Manabu Torii, Cathy H. Wu, and K. Vijay-Shanker. 2014. A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems. *BMC Bioinformatics* 15:285.

Glen Pink, Joel Nothman, and James R. Curran. 2014. Analysing Recall Loss in Named Entity Slot Filling. In *EMNLP'14*. ACL, Doha, Qatar, pages 820–830.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a Large Benchmark for Open Information Extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Austin, Texas, page (to appear).

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)*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. Association for Computational Linguistics, Avignon, France.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Ying Xu, Mi-Young Kim, Kevin Quinn, Randy Goebel, and Denilson Barbosa. 2013. Open Information Extraction with Tree Kernels. In *HLT-NAACL*. pages 868–877.