More Informative Open Information Extraction via Simple Inference

Hannah Bast and Elmar Haussmann

Department of Computer Science
University of Freiburg
79110 Freiburg, Germany
{bast, haussmann}@informatik.uni-freiburg.de

Abstract. Recent Open Information Extraction (OpenIE) systems utilize grammatical structure to extract facts with very high recall and good precision. In this paper, we point out that a significant fraction of the extracted facts is, however, not *informative*. For example, for the sentence *The ICRW is a non-profit organization headquartered in Washington*, the extracted fact *(a non-profit organization) (is headquartered in) (Washington)* is not informative. This is a problem for semantic search applications utilizing these triples, which is hard to fix once the triple extraction is completed. We therefore propose to integrate a set of simple inference rules into the extraction process. Our evaluation shows that, even with these simple rules, the percentage of informative triples can be improved considerably and the already high recall can be improved even further. Both improvements directly increase the quality of search on these triples.[1]

1 Introduction

Information extraction (IE) is the task of automatically extracting relational tuples, typically triples, from natural language text. In recent years, the trend has been towards Open Information Extraction (OpenIE), where identifying the predicate and hence the relation is part of the problem. For example, from the sentence

The ICRW is a non-profit organization headquartered in Washington.

the following triples might be extracted:

#1: *(The ICRW) (is) (a non-profit organization)*
#2: *(a non-profit organization) (is headquartered in) (Washington)*

Extracted triples are an important source of information for many information retrieval (IR) systems, in particular in the area of semantic search. For example, systems for the semantic search challenges in the SemSearch 2010/2011 [1] and TREC Entity Track 2009/2010 [2] perform search on triples. A public demo of a search on triple extractions of the ReVerb OpenIE system [3] is available at: http://openie.cs.washington.edu/. Semantic search systems like Broccoli [4] search in triples or triple-like excerpts extracted from the full text as well. All these approaches rely on the usefulness of extracted triples, usually indicated by how much facts were extracted (*recall*) and whether they are correct (*precision*).

[1] A demo of our system is available via http://ad.informatik.uni-freiburg.de/publications

Early approaches to OpenIE focused on extracting triples with high precision but comparably low recall [5]. Later systems focused on improving recall at best possible precision. Newer systems also addressed other important quality aspects, for example in [3] *incoherent* extractions were addressed and [6] considers the *context* of triples. The by far highest recall (at reasonably good precision) was recently achieved with rule-based approaches utilizing grammatical structure, namely ClausIE [7] and CSD-IE [8]. They rely on grammatical rules based on deep parses of a sentence to extract *direct* facts, as e.g., triples #1 and #2 above. The facts are direct in the sense that subject, predicate (possibly implicit) and object are in some form directly connected via a grammatical relation. With respect to these direct facts, the systems achieve almost perfect recall which makes them suitable for a wide range of applications in IR. However, these systems ignore various quality aspects from earlier work.

In this paper, we show that a significant amount of the extracted facts is not *informative*. In the example above, triples #1 and #2 can both be considered correct, but only triple #1 is, by itself, informative. For all practical purposes, the fact that some non-profit organization is headquartered in Washington is useless. This is a serious problem for systems utilizing these triples for search. For example, a search for where the ICRW is headquartered would not be possible to answer from the extracted triples above. An informative extraction that instead can be inferred is:

#3: *(The ICRW) (is headquartered in) (Washington)*

With this, the extracted triple #2 becomes superfluous - all information of the sentence is covered in a precise form in triples #1 and #3. Note that inferring this is only possible while processing the sentence, when individual subjects, objects and their relations are uniquely identified. Afterwards, multiple facts extracted from different sentences mentioning (say) *a non-profit organization* cannot be guaranteed to refer to the same organization.

Based on the observation that this phenomenon is frequent we propose to integrate simple inference into the extraction process of an OpenIE system. Our approach utilizes some of the generic rules used in large scale inference systems, see Section 3. The process is simple and fast, only uses few inference rules and already shows good results. We provide a brief overview of related work in the next section, describe our approach in Section 3 and provide an evaluation in Section 4.

2 Related Work

For an elaborate overview of recent OpenIE systems we refer to [7]. To the best of our knowledge no existing OpenIE system addresses the issue of inference during its extractions process. Some earlier systems, e.g. [9], extract "indirect" facts, similar to inferred facts, using learned patterns. This only works if the text pattern learned for extraction is part of the used training set. In contrast, our inference rules are generic and independent of the exact text surface of a relation.

A lot of work on inferring new information from triples or knowledge bases exists, e.g. [10, 11]. The goal is usually to infer facts from triples extracted from different sources in order to, for example, extend knowledge bases or perform question answering. Our goal is to improve the informativeness of extracted triples in the first place, not

to perform elaborate inference. Furthermore, as we argued in Section 1, some information can only be inferred while extracting triples, and is irrevocably lost afterwards.

Informativeness of extracted triples has previously been addressed in [3]. Triples were considered uninformative if they omit critical information, for example, when relation phrases are underspecified. The informativeness of extracted triples was evaluated as part of correctness, i.e. uninformative triples were labeled incorrect. We take this one step further and consider a triple uninformative if there is a more precise triple that should be extracted instead, e.g., if the subject should be different (as in the example in Section 1). In our evaluation we explicitly label informativeness as well as correctness of extracted triples.

3 Simple Inference for OpenIE

Our approach consists of three straightforward steps, which are comparably simple yet effective (as our quality evaluation in section 4 shows). The steps are performed after subjects, predicates and objects of all triples in a sentence have been identified, but when all other information (underlying parse tree, supporting data structures etc.) is still available. Given the predicate of each triple in a sentence we first classify the predicate into one of several *semantic relation classes*. Based on the semantic relation we apply a set of inference rules to derive new triples. In the final step we remove existing triples that we consider uninformative depending on whether and how they were used to derive new triples. The next subsections each describe one of the three steps.

3.1 Identifying Semantic Predicate Class

We first classify the predicate of each triple into one of five semantic relation classes shown in Table 1. The relations have previously been successfully used for inference [10] and allow deriving generic, domain-independent inference rules.

To identify the relations we match simple indicator words and patterns. The patterns are implemented as regular expressions over text or parse tree fragments using Tregex [12].

Table 1. Semantic relation classes and patterns for identification

	Semantics	Pattern
SYN	synonymy	*is, was, has|have been, are, nicknamed, known as*
IS-A	hyponymy	*(has|have been, are, is) a|an*
PART-OF	meronymy	*part of, consist* of*
IN	containment or placement	** in*
OTHER	all other relations	*

3.2 Inferring New Triples

Given the triples with identified semantic predicates we infer new triples using a set of generic inference rules. Table 2 shows the rules used. For our example from the

introduction, the last rule matches because the semantic relation IS-A holds between *The ICRW* (A) and *a non-profit organization* (B) and C can be bound to *Washington*. As a result, it is inferred that *The ICRW is headquartered in Washington*.

These rules are similar to the up-ward monotone rules from [10], but have been extended with an additional rule to reason over IS-A relations. The implementation differentiates between lexically identical subjects and objects that occur in different places of a sentence. This is a fundamental difference to approaches inferring information after triple extraction, where this information is no longer available.

Table 2. Inference rules for new triples

OTHER(A', B)	$\leftarrow$	OTHER$(A, B) \wedge$ SYN(A, A')
OTHER(A', B)	$\leftarrow$	OTHER$(A, B) \wedge$ SYN(A', A)
OTHER(A, B')	$\leftarrow$	OTHER$(A, B) \wedge$ SYN(B, B')
OTHER(A, B')	$\leftarrow$	OTHER$(A, B) \wedge$ SYN(B', B)
IN(A, C)	$\leftarrow$	IN$(A, B) \wedge$ PART-OF(B, C)
IN(A, C)	$\leftarrow$	IN$(A, B) \wedge$ IS-A(B, C)
OTHER(A, C)	$\leftarrow$	IS-A$(A, B) \wedge$ OTHER(B, C)

Table 3. Rule for deleting triples

		IS-A(A, B)
$remove($OTHER$(B, C))$	$\leftarrow$	$\wedge$OTHER(B, C)
		$\wedge$OTHER(A, C)

3.3 Removing Uninformative Triples

As described in Section 1 some triples become redundant after they were used to infer additional information. These triples should not be part of the output of the system. This is often the case for IS-A relations and we use a single rule shown in Table 3 to remove triples from our result list.

4 Evaluation

We evaluate the quality of extracted triples with respect to correctness and informativeness. A system similar to the OpenIE system in [8] was used to integrate inference as described above. We compared it against the OpenIE system without inference.

As dataset we used 200 random sentences from Wikipedia. The sentences contain only few incorrect grammatical constructions and cover a wide range of complexity and length. This is the exact same dataset that has already been used in [7].

For each extracted triple we manually assigned two labels: one for correctness (yes or no) and one for informativeness (yes or no). We follow the definition of [5] and consider a triple correct if it is consistent with the truth value of the corresponding sentence. A correct triple is considered informative if there is no extraction that is more precise, according to the sentence it was extracted from. For example, in the sentence from the introduction, triples #1 and #2 would be considered correct, but only triple #1 would be considered informative and triple #3 would be considered both, correct and informative. From the labeled triples we calculated precision of correct triples and estimate recall using the number of extracted correct triples. We also calculated corresponding breakdown statistics for triples that are informative (inf.) as well as correct (corr.). Table 4 shows overall results and Table 5 provides detailed information about the inferred triples.

We first discuss the results in Table 4. Without inference, a large fraction of 10% of correct triples is not informative (prec-corr. inf.). This means that, on average, every 10th extracted correct triple is more or less useless. Using inference the overall number of extracted facts increases from 649 to 762, a relative increase of 17%. The number of correct facts (#facts corr.) also increases: from 429 to 484, corresponding to a relative increase of 13%. The relative increase in correct triples is smaller, because a small number of incorrect triples are inferred (see next paragraph). This is also the reason for the small decrease in the percentage of correct triples (prec corr.) from 66% to 64% (at a 13% higher recall, however). Overall, the number of triples that are both correct and informative (#facts corr. + inf.) increases from 385 to 444: a 15% increase. This is a major improvement, caused by the large number of correct informative triples that were inferred and the uninformative triples removed. Correspondingly, the percentage of correct triples that are also informative (prec-corr. + inf.) increases from 90% to 92%.

Table 4. Quality evaluation results with inference (top row) and without inference (bottom row) over the labels correct (corr.) and informative (inf.). *prec corr.* refers to the percentage of all triples labeled correct, *prec-corr. inf.* to the percentage of correct triples labeled informative.

	#facts	#facts corr.	#facts corr. + inf.	prec corr.	prec-corr. inf.
No Inference	649	429	385	66%	90%
Inference	762	484	444	64%	92%

Table 5. Detailed statistics for inferred triples. *prec inf.* refers to the percentage of inferred triples labeled correct and *prec-corr. inf.* to the percentage of inferred correct triples also labeled informative.

#inferred	#inferred corr.	prec inf.	#inferred corr. + inf.	prec-corr. inf.
127	69	54%	59	85%

Table 5 shows the statistics for inferred triples. Note that, as described in Section 3.3, during inference previously extracted triples may be removed. Therefore, the number of extracted facts with inference does not equal the sum of facts extracted without inference and inferred facts (see Table 4). Overall, about 54% of inferred triples are correct (prec. inf.). A preliminary investigation shows that this is mainly caused by mistakes in preceding phases, in particular wrong parses, wrong identification of objects or predicates and wrong mapping of predicates to their semantic class (see Section 3.1). Eliminating these errors should be part of our next steps (see Section 5). About 85% of correct inferred triples are also informative (prec-corr. inf.). Closer analysis shows that, due to inference, 32% of the triples that were correct but uninformative were removed and replaced with informative triples. Together with the inferred triples, this causes the increase in the percentage of correct triples that are also informative (prec-corr. + inf. in Table 4).

In some cases uninformative triples were inferred. For example, from the sentence *She joined WGBH-TV, Boston's public television station* it is inferred that *(She) (joined) (Boston's public television station)*. Given that our current approach does not differen-

tiate between concrete and abstract subjects and objects (and therefore the "direction" of inference) the high percentage of informative triples derived is remarkable. Further work should, however, try to prevent these extractions.

5 Conclusions

We have presented a simple yet effective way to increase the informativeness of extracted triples for a recent OpenIE system. Using only a few simple inference rules integrated into triple extraction can increase the number of extracted informative triples by 15%. There are a lot of promising directions to improve our work.

A preliminary error analysis shows that most mistakes happen in preceding extraction stages, in particular the precise identification of predicates and objects. Improvements in these areas will likely obviate the small negative effect on precision. To improve the recognition of semantic relations, utilizing existing collections of semantic patterns, such as provided in [13], seems promising. Our manually designed inference rules could be exchanged for automatically derived rules, e.g., as suggested in [14]. Finally, to derive additional facts and distinguish between abstract and concrete facts utilizing information from named entity recognition seems promising.

References

1. Tran, T., Mika, P., Wang, H., Grobelnik, M.: Semsearch'11: the 4th Semantic Search Workshop. In: WWW. (2011)
2. Balog, K., Serdyukov, P., de Vries, A.P.: Overview of the TREC 2010 Entity Track. In: TREC. (2010)
3. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: EMNLP. (2011) 1535–1545
4. Bast, H., Bäurle, F., Buchhold, B., Haussmann, E.: Broccoli: Semantic full-text search at your fingertips. CoRR (2012)
5. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI. (2007) 2670–2676
6. Mausam, Schmitz, M., Soderland, S., Bart, R., Etzioni, O.: Open language learning for information extraction. In: EMNLP-CoNLL. (2012) 523–534
7. Corro, L.D., Gemulla, R.: ClausIE: clause-based open information extraction. In: WWW. (2013) 355–366
8. Bast, H., Haussmann, E.: Open information extraction via contextual sentence decomposition. In: ICSC. (2013)
9. Banko, M., Etzioni, O.: The tradeoffs between open and traditional relation extraction. In: ACL. (2008) 28–36
10. Schoenmackers, S., Etzioni, O., Weld, D.S.: Scaling textual inference to the web. In: EMNLP. (2008) 79–88
11. Lao, N., Mitchell, T.M., Cohen, W.W.: Random walk inference and learning in a large scale knowledge base. In: EMNLP. (2011) 529–539
12. Levy, R., Andrew, G.: Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In: LREC. (2006) 2231–2234
13. Nakashole, N., Weikum, G., Suchanek, F.M.: PATTY: A taxonomy of relational patterns with semantic types. In: EMNLP-CoNLL. (2012) 1135–1145
14. Schoenmackers, S., Davis, J., Etzioni, O., Weld, D.S.: Learning first-order horn clauses from web text. In: EMNLP. (2010) 1088–1098