Porting an Open Information Extraction System from English to German

Tobias Falke† Gabriel Stanovsky‡ Iryna Gurevych† Ido Dagan‡

†Research Training Group AIPHES and UKP Lab
Computer Science Department, Technische Universität Darmstadt

‡Natural Language Processing Lab
Department of Computer Science, Bar-Ilan University

Abstract

Many downstream NLP tasks can benefit from Open Information Extraction (Open IE) as a semantic representation. While Open IE systems are available for English, many other languages lack such tools. In this paper, we present a straightforward approach for adapting PropS, a rule-based predicate-argument analysis for English, to a new language, German. With this approach, we quickly obtain an Open IE system for German covering 89% of the English rule set. It yields 1.6 n-ary extractions per sentence at 60% precision, making it comparable to systems for English and readily usable in downstream applications.¹

1 Introduction

The goal of Open Information Extraction (Open IE) is to extract coherent propositions from a sentence, each represented as a tuple of a relation phrase and one or more argument phrases (e.g., born in (Barack Obama; Hawaii)). Open IE has been shown to be useful for a wide range of semantic tasks, including question answering (Fader et al., 2014), summarization (Christensen et al., 2013) and text comprehension (Stanovsky et al., 2015), and has consequently drawn consistent attention over the last years (Banko et al., 2007; Wu and Weld, 2010; Fader et al., 2011; Akbik and Löser, 2012; Mausam et al., 2012; Del Corro and Gemulla, 2013; Angeli et al., 2015).

Although similar applications of Open IE in other languages are obvious, most previous work focused on English, with only a few recent exceptions (Zhila and Gelbukh, 2013; Gamallo and Garcia, 2015). For most languages, Open IE systems are still missing. While one could create them from scratch, as it was done for Spanish, this can be a very laborious process, as state-of-the-art systems make use of hand-crafted, linguistically motivated rules. Instead, an alternative approach is to transfer the rule sets of available systems for English to the new language.

In this paper, we study whether an existing set of rules to extract Open IE tuples from English dependency parses can be ported to another language. We use German, a relatively close language, and the PropS system (Stanovsky et al., 2016) as examples in our analysis. Instead of creating rule sets from scratch, such a transfer approach would simplify the rule creation, making it possible to build Open IE systems for other languages with relatively low effort in a short amount of time. However, challenges we need to address are differences in syntax, dissimilarities in the corresponding dependency representations as well as language-specific phenomena. Therefore, the existing rules cannot be directly mapped to the German part-of-speech and dependency tags in a fully automatic way, but require a careful analysis as carried out in this work. Similar manual approaches to transfer rule-based systems to new languages were shown to be successful, e.g. for temporal tagging (Moriceau and Tannier, 2014), whereas fully automatic approaches led to less competitive systems (Strötgen and Gertz, 2015).

Our analysis reveals that a large fraction of the PropS rule set can be easily ported to German, requiring only small adaptations. With roughly 10%...
of the effort that went into the English system, we could build a system for German covering 89% of the rule set. As a result, we present PropsDE, the first Open IE system for German. In an intrinsic evaluation, we show that its performance is comparable with systems for English, yielding 1.6 extractions per sentence with an overall precision of 60%.

2 Background

Open Information Extraction Open IE was introduced as an open variant of traditional Information Extraction (Banko et al., 2007). Since its inception, several extractors were developed. The majority of them, namely ReVerb (Fader et al., 2011), KrakeN (Akbik and Löser, 2012), Exemplar (Mesquita et al., 2013) and ClausIE (Del Corro and Gemulla, 2013), successfully used rule-based strategies to extract tuples. Alternative approaches are variants of self-supervision, as in TextRunner (Banko et al., 2007), WOE (Wu and Weld, 2010) and OLLIE (Mausam et al., 2012), and semantically-oriented approaches utilizing semantic role labeling (Open IE-4\(^2\)) or natural logic (Angeli et al., 2015). While TextRunner and ReVerb require only POS tagging as preprocessing to allow a high extraction speed, the other systems rely on dependency parsing to improve the extraction precision.

For non-English Open IE, ExtrHech has been presented for Spanish (Zhila and Gelbukh, 2013). Similar as the English systems, it uses a set of extraction rules, specifically designed for Spanish in this case. More recently, ArgOE (Gamallo and Garcia, 2015) was introduced. It manages to extract tuples in several languages with the same rule set, relying on a dependency parser that uses a common tagset for five European languages. However, an evaluation for English and Spanish revealed that this approach cannot compete with the systems specifically built for those languages. To the best of our knowledge, no work on Open IE for German exists.

Open IE with PropsDE Stanovsky et al. (2016) recently introduced Props, a rule-based converter turning dependency graphs for English into typed graphs of predicates and arguments. An example is shown in Figure 1 (in German). Compared to a dependency graph, the representation masks non-core syntactic details, such as tense or determiners, unifies semantically equivalent constructions, such as active/passive, and explicates implicit propositions, such as indicated by possessives or appositions.

The resulting graph can be used to extract Open IE tuples in a straightforward way. Every non-nested predicate node pred in the graph, together with its n argument-subgraphs argi, yields a tuple \(\text{pred}(\text{arg}_1; \ldots; \text{arg}_n)\). With this approach, Props is most similar to KrakeN and ClausIE, applying rules to a dependency parse. However, due to additional nodes for implicit predicates, it can also make extractions that go beyond the scope of other systems, such as has (Michael; bicycle) from Michael’s bicycle is red. In line with more recent Open IE systems, this strategy extracts tuples that are not necessarily binary, but can be unary or of higher arity.

3 Analysis of Portability

Approach For each rule of the converter that transforms a dependency graph to the Props graph, we assess its applicability for German. A rule is applied to a part of the graph if certain conditions are
fulfilled, expressed using dependency types, POS
tags and lemmas. As we already pointed out in
the introduction, several differences between the de-
pendency and part-of-speech representations for En-
lish and German make a fully automatic translation
of these rules impossible. We therefore manually
alyzed the portability of each rule and report the
findings in the next section.

While using Universal Dependencies (Nivre et al.,
2016) could potentially simplify porting the rules,
we chose not to investigate this option due to the on-
go ing nature of the project and focused on the estab-
lished representations for now. In line with the En-
lish system, that works on collapsed Stanford de-
pendencies (de Marneffe and Manning, 2008), we
assume a similar input representation for German
that can be obtained with a set of collapsing and
propagation rules provided by Ruppert et al. (2015)
for TIGER dependencies (Seeker and Kuhn, 2012).

Findings Overall, we find that most rules can be
used for German, mainly because syntactic differ-
ces, such as freer word order (Kübler, 2008), are
already masked by the dependency representation
(Seeker and Kuhn, 2012). About 38% of the rule set
can be directly ported to German, solely replacing
dependency types, POS tags and lemmas with their
German equivalents. As an example, the rule remov-
ing negation tokens looks for neg dependencies in
the graph, for which a corresponding type NG ex-
ists in German. We found similar correspondences
to remove punctuation and merge proper noun and
number compounds. In addition, we can also handle
appositions and existentials with direct mappings.

For 35% of the English rules, small changes are
necessary, mainly because no direct mapping to the
German tag set is possible or the annotation style
differs. For instance, while English has a specific
type det to link determiners to their governor, a more
generic type (NK) is used in German. Instead, deter-
miners can be detected by part-of-speech:

Stanford dependencies, main verbs govern all aux-
iliaries, whereas in TIGER dependencies, an auxiliary
heads the main verb. The above example shows this
for gone and am. Therefore, all rules identifying and
removing auxiliaries and modals have to be adapted
to account for this difference.

With similar changes as discussed for deter-
miners, we can also handle possessive and copular con-
structions. The graph for Michael’s bicycle is red,
for example, features an additional predicate have to
explicate the implicit possessive relation, while red
becomes an adjectival predicate, omitting is:

Moreover, conditional constructions can be pro-
cessed with slight changes as well. Missing a coun-
terpart for the type mark, we instead look for sub-
ordinating conjunctions by part-of-speech. In fact,
we found conditionals to be represented more con-
sistently across different conjunctions, making their
handling in German easier than in English.

More substantial changes are necessary for the
remaining 27% of the rules. To represent active and
passive in a uniform way, in passive clauses, PropS
turns the subject into an object and a potential by-
clause into the subject. For English, these cases
are indicated by the presence of passive dependen-
cies such as nsubjpass. For German, however, no
counterparts exist. As an alternative strategy, we
instead look for past participle verbs (by POS tag)
that are governed by a form of the auxiliary werden
(Schäfer, 2015). Instances of the German static pas-
sive (Zustandspassiv) are, in contrast, handled like
copulas. Another deviation from the English system
is necessary for relative clauses. PropS heavily relies
on the Stanford dependency converter, which propa-
gates dependencies of the relative pronoun to its ref-
rent. The German collapser does not have this fea-
ture, and we therefore implement it as an additional
transformation (see subjliegen;Orte) in Figure 1).

To abstract away from different tenses, PropS rep-
resents predicates with their lemma, indicating the
original tense as a feature, as detected with a set of
rules operating on POS tags. For German, no tense
information is contained in POS tags, but instead, a
morphological analysis can provide it. Determining
the overall tense of a sentence based on that requires
a new set of rules, as the grammatical construction of
tenses differs between German and English. PropS
also tries to heuristically identify raising construc-
tions, in which syntactic and semantic roles of argu-
ments differ. In German, this phenomenon occurs
in similar situations, such as in Michael scheint zu
lächeln (Michael seems to smile), in which Michael
is not the semantic subject of scheinen, though syn-
tactically it is. To determine these cases heuristi-
cally, an empirically derived list of common raising
verbs, such as done by Chrupała and van Genabith
(2007) for English, needs to be created.

An additional step that is necessary during the
lemmatization of verbs for German is to recover sep-
arated particles. For example, a verb like ankomy-
men (arrive) can be split in a sentence such as Er
ekam an (He arrived), moving the particle to the end
of the sentence, with a potentially large number of
other tokens in between. We can reliably reattach
these particles based on the dependency parse. An-
other addition to the rules that we consider important
is to detect subjunctive forms of verbs and indicate
the mood with a specific feature for the predicate.
A morphological analysis provides the necessary in-
put. Compared to English, the usage of the subjunc-
tive is much more common, usually to indicate ei-
ther unreality or indirect speech (Thieroff, 2004).

4 German Open IE System

Following our analysis, we implemented a German
version of PropS, named PropsDE. It uses mate-
tools for POS tagging, lemmatizing and parsing
(Bohnet et al., 2013). Dependencies are collapsed
and propagated with JoBimText (Ruppert et al.,
2015). The rule set covers 89% of the English rules,
lacking only the handling of raising-to-subject verbs
and more advanced strategies for coordination con-
structions and tense detection. To assign confidence
scores, PropsDE uses a logistic regression model
trained to predict the correctness of extractions. Fig-
ure 1 illustrates some extracted tuples. Based on
 correspondence with the authors of the English sys-
tem, we conclude that we were able to implement
the German version with roughly 10% of the effort
they reported. This shows that our approach of man-
ually porting a rule-based system can overcome the
lack of a tool for another language with reasonable
effort in a short amount of time.

5 Experiments

Experimental Setup Following the common eval-
uation protocol for Open IE systems, we manu-
ally label extractions made by our system. For
this purpose, we created a new dataset consisting
of 300 German sentences, randomly sampled from
three sources of different genres: news articles from
TIGER (Brants et al., 2004), German web pages
from CommonCrawl (Habernal et al., 2016) and fea-
tured Wikipedia articles. For the treebank part, we
ran our system using both gold and parsed depen-
dencies to analyze the impact of parsing errors.

Every tuple extracted from this set of 300 sen-
tences was labeled independently by two annota-
tors as correct or incorrect. In line with previous
work, they were instructed to label an extraction as
incorrect if it has a wrong predicate or argument,
including overspecified and incomplete arguments,
or if it is well-formed but not entailed by the sen-
tence. Unresolved co-references were not marked
as incorrect. We observed an inter-annotator agree-
ment of 85% (κ = 0.63). For the evaluation, we
merged the labels, considering an extraction as cor-
rect only if both annotators labeled it as such. Re-
sults are measured in terms of precision, the fraction
of correct extractions, and yield, the total number of
extractions. A precision-yield curve is obtained by
decreasing a confidence threshold. The confidence
 predictor was trained on a separate development set.

Results From the whole corpus of 300 sentences,
PropsDE extracted 487 tuples, yielding on average
1.6 per sentence with 2.9 arguments. 60% of them
were labeled as correct. Table 1 shows that most ex-
tractions are made from Wikipedia articles, whereas
the highest precision can be observed for newswire
text. According to our expectations, web pages are
most challenging, presumably due to noisier lan-
guage. These differences between the genres can
also be seen in the precision-yield curve (Figure 2).

For English, state-of-the-art systems show a sim-
ilar performance. In a direct comparison of sev-
eral systems carried out by Del Corro and Gemulla
(2013), they observed overall precisions of 58%
(Reverb), 57% (ClausIE), 43% (WOE) and 43%

(OLLIE) on datasets of similar genre. The reported yield per sentence is higher for ClausIE (4.2), OL-
LIE (2.6) and WOE (2.1), but smaller for Reverb (1.4). However, we note that in their evaluation, they
configured all systems to output only two-argument-
extractions. For example, from a sentence such as

\textit{The principal opposition parties boycotted
the polls after accusations of vote-rigging.}

OLLIE can either make two binary extractions

\textit{boycotted ( the principal opposition parties ;
the polls )
boycotted the polls after ( the principal oppo-
sition parties ; accusations of vote-rigging )}

or just a single extraction with three arguments. PropS always extracts the combined tuple

\textit{boycotted ( the principal opposition parties ,
the polls , after accusations of vote-rigging )},

which is in line with the default configuration of
more recent Open IE systems.

For the sake of comparability, we conjecture that
the yield of our system would increase if we broke
down higher-arity tuples in a similar fashion: Assu-
ming that every extraction with \( n \) arguments, \( n > 2 \),
can be split into \( n - 1 \) separate extractions, our
system’s yield would increase from 1.6 to 3.0. That
is in line with the numbers reported above for the
binary configuration for English. Overall, this indi-
icates a reasonable performance of our straightfor-
ward porting of PropS to German.

Extractions were most frequently labeled as in-
correct due to false relation labels (32%), overspeci-
fied arguments (21%) and wrong word order in argu-
ments (19%). Analyzing our system’s performance
on the treebank, we can see that the usage of gold de-
pendencies increases the precision by 8 percentage
points, making parsing errors responsible for about
28% of the incorrect extractions. Since the mate-
tools parser is trained on the full TIGER treebank,
including our experimental data, its error contribu-
tion on unseen data might be even higher.

6 Conclusion

Using PropS and German as examples, we showed
that a rule-based Open IE system for English can be
ported to another language in a reasonable amount
of time. As a result, we presented the first Open
IE system for German. In the future, studies tar-
geting less similar languages could further evaluate
the portability of PropS. Directions for future work
on PropsDE are extensions of the rule set to better
cover complex coordination constructions, nested
sentences and nominal predicates.

Acknowledgments

This work has been supported by the DFG-funded
research training group “Adaptive Preparation of In-
formation from Heterogeneous Sources” (AIPHES,
GRK 1994/1), by the German Research Founda-
tion through the German-Israeli Project Cooperation
(DIP, grant DA 1600/1-1 and grant GU 798/17-1)
and by the Volkswagen Foundation as part of the
Lichtenberg-Professorship Program under grant No.
I/82806.

References

Alan Akbik and Alexander Löser. 2012. Kraken: N-
ary Facts in Open Information Extraction. In Proceed-
ings of the Joint Workshop on Automatic Knowledge
Base Construction & Web-scale Knowledge Extrac-
tion, pages 52–56, Montreal, Canada.
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 and the 7th International Joint Conference on Natural Language Processing, pages 344–354, Beijing, China.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In Proceedings of the 20th International Joint Conference on Artifical Intelligence, pages 2670–2676, Hyderabad, India.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richard Farkas, Filip Ginter, and Jan Hajic. 2013. Joint Morphological and Syntactic Analysis for Richly Inflected Languages. Transactions of the Association for Computational Linguistics, 10(1):415–428.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther Koenig, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. Research on Language and Computation, 2(4):597–620.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards Coherent Multi-Document Summarization. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1163–1173, Atlanta, GA, USA.

Grzegorz Chrupala and Josef van Genabith. 2007. Using Very Large Corpora to Detect Raising and Control Verbs. In Proceedings of the Lexical Functional Grammar 2007 Conference, pages 597–620, Stanford, CA, USA.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In Proceedings of the 22nd International Conference on Computational Linguistics, pages 1–8, Manchester, United Kingdom.

Luciano Del Corro and Rainer Gemulla. 2013. ClausIE: Clause-Based Open Information Extraction. In Proceedings of the 22nd International Conference on the World Wide Web, pages 355–366, Rio de Janeiro, Brazil.

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, Edinburgh, United Kingdom.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1156–1165, New York, NY, USA.

Pablo Gamallo and Marcos Garcia. 2015. Multilingual Open Information Extraction. In Proceedings of the 17th Portuguese Conference on Artificial Intelligence, volume 9273 of Lecture Notes in Computer Science, pages 711–722, Coimbra, Portugal.

Ivan Habernal, Omnia Zayed, and Iryna Gurevych. 2016. C4Corpus: Multilingual Web-size corpus with free license. In Proceedings of the 10th International Conference on Language Resources and Evaluation, pages 914–922, Portorož, Slovenia.

Sandra Kübler. 2008. The PaGe 2008 shared task on parsing German. In Proceedings of the ACL-08: HLT Workshop on Parsing German (PaGe-08), pages 55–63, Columbus, OH, USA.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, 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.

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, pages 447–457, Seattle, WA, USA.

Véronique Moriceau and Xavier Tannier. 2014. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 3239–3243, Reykjavik, Iceland.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of the 10th International Conference on Language Resources and Evaluation, pages 1659–1666, Portorož, Slovenia.

Eugen Ruppert, Jonas Klesy, Martin Riedl, and Chris Biemann. 2015. Rule-based Dependency Parse Collapsing and Propagation for German and English. In Proceedings of the GSCL 2015, pages 58–66, Duisburg, Germany.

Roland Schäfer. 2015. Einführung in die grammatische Beschreibung des Deutschen. Language Science Press, Berlin, Germany.

Wolfgang Seeker and Jonas Kuhn. 2012. Making Ellipses Explicit in Dependency Conversion for a Ger-
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 and the 7th International Joint Conference on Natural Language Processing, pages 303–308, Beijing, China.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016. Getting More Out Of Syntax with PropS. arXiv:1603.01648.

Jannik Strötgen and Michael Gertz. 2015. A Baseline Temporal Tagger for all Languages. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 541–547, Lisbon, Portugal.

Rolf Thieroff. 2004. The subjunctive mood in German and in the Germanic languages. In Focus on Germanic Topology, pages 315–358. Akademie Verlag, Berlin, Germany.

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.

Alisa Zhila and Alexander Gelbukh. 2013. Comparison of open information extraction for English and Spanish. In Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue 2013), pages 714–722, Bekasovo, Russia.