Von komplexen Sätzen zu einer formalen semantischen Repräsentation mittels syntaktischer Textvereinfachung und Open Information Extraction[1]

Christina Niklaus[2]

Abstract: Moderne Systeme, die sich mit Inferenzen in Texten beschäftigen, benötigen automatisierte Methoden zur Extraktion von Bedeutungsrepräsentationen aus großen Textkorpora. Open Information Extraction (IE) ist eine führende Methode, um sämtliche in einem Text vorhandenen Relationen zu extrahieren. Open-IE-Ansätze haben sich im Laufe der Jahre weiterentwickelt, um Beziehungen zu erfassen, die über einfache Subjekt-Prädikat-Objekt-Tripel (SPO) hinausgehen. Dabei ist jedoch ein genauerer Blick auf die Extraktion von Verknüpfungen zwischen Klausen und Phrasen innerhalb eines komplexen Satzes vernachlässigt worden. Um diese Lücke zu schließen, wird ein neuartiges Open-IE-Framework vorgestellt, das komplexe Textdaten in eine leichtgewichtige semantische Repräsentation in Form von normalisierten und kontextwahrenden relationalen Tupeln transformiert. Das Framework nutzt einen diskursorientierten Ansatz, um komplexe Sätze in eine semantische Hierarchie von Minimalaussagen zu überführen. Diese weisen eine kanonische SPO-Struktur auf, wodurch die Extraktion von relationalen Tupeln erleichtert wird, was zu einer verbesserten Genauigkeit (engl. *"precision"*) (bis zu 32%) und einer höheren Erkennungsrate (engl. *"recall"*) (bis zu 30%) der extrahierten Relationen in einem großen Benchmark-Korpus führt. Darüber hinaus wird der semantische Kontext der extrahierten Tupel in Form von rhetorischen Strukturen und hierarchischen Beziehungen erfasst. Auf diese Weise wird die oberflächliche semantische Darstellung aktueller Open-IE-Systeme mit zusätzlichen Metainformationen angereichert und so wichtige Kontextinformationen bewahrt, die zur Extraktion von korrekten, aussagekräftigen und kohärenten relationalen Tupeln erforderlich sind.

1 Einführung

Nach Angaben des Weltwirtschaftsforums wird die täglich weltweit generierte Datenmenge bis zum Jahr 2025 voraussichtlich 463 Exabyte erreichen, wobei ein Großteil dieser Daten in Form von unstrukturiertem Text vorliegt.[3] Um diese Daten effektiv nutzen zu können, müssen sie in ein strukturiertes Format überführt werden, das von Maschinen gelesen und analysiert werden kann. Open IE ist eine führende Methode zur Transformation unstrukturierter Informationen in eine strukturierte Repräsentation in Form von Prädikaten und zugehörigen Argumenten. Diese bilden die grundlegenden Komponenten einer semantischen Darstellung eines Satzes und ermöglichen die automatische Verarbeitung der Daten. Die vorliegende Dissertation befasst sich daher mit der Frage, wie unstrukturierte Textdaten in Form von komplexen Sätzen in eine formale semantische Repräsentation überführt werden können, die von Maschinen interpretiert werden kann.

[1] Englischer Titel der Dissertation: "From Complex Sentences to a Formal Semantic Representation using Syntactic Text Simplification and Open Information Extraction" [Ni22]

[2] Universität Passau & Universität St.Gallen, christina.niklaus@unisg.ch

[3] https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Forschungslücken Diese Arbeit wird durch folgende Beobachtungen motiviert:

- *Fehlende Beziehungen und inkorrekte Extraktionen*: Lange, verschachtelte Satzstrukturen stellen ein großes Problem für Open-IE-Anwendungen dar. Studien haben gezeigt, dass die Erfassung von Relationen, die sich über solch komplexe Strukturen erstrecken oder in nicht-kanonischer Form vorliegen, eine schwierige Aufgabe ist, weshalb diese Beziehungen von den derzeitigen Ansätzen oft übersehen oder inkorrekt extrahiert werden.
- *Fehlender Kontext*: Bisherige Arbeiten im Bereich Open IE haben sich hauptsächlich auf die Extraktion isolierter relationaler Tupel konzentriert. Unter Vernachlässigung des kohäsiven Charakters von Texten, in denen wichtige Kontextinformationen über Klausen oder Sätze hinweg verteilt sind, generieren aktuelle Open-IE-Ansätze in der Regel nur eine unzusammenhängende Menge von Tupeln und damit unvollständige, uninformative oder inkohärente Extraktionen.

Beiträge Um diese Einschränkungen zu überwinden, haben wir ein Open-IE-Framework entwickelt, das komplexe Textdaten in eine neuartige leichtgewichtige semantische Repräsentation in Form von normalisierten und kontextwahrenden relationalen Tupeln transformiert. Mit unserer Arbeit leisten wir einen doppelten Beitrag: (1) Wir präsentieren einen diskursorientierten Ansatz zur Zerlegung von Sätzen, der eine semantische Hierarchie von syntaktisch vereinfachten Propositionen erstellt. Mit Hilfe klausaler und phrasaler Zerlegungsmechanismen werden komplexe Ausgangssätze in kleinere Einheiten mit einer kanonischen Syntax untergliedert. Die daraus resultierende normalisierte SPO-Struktur der vereinfachten Sätze kann von Open-IE-Anwendungen leichter verarbeitet werden. Dadurch wird die Komplexität bei der Bestimmung von Prädikat-Argument-Strukturen reduziert, was zu einer verbesserten Genauigkeit und höheren Erkennungsrate der extrahierten relationalen Tupel führt. (2) Um die Extraktion aussagekräftiger Tupel zu unterstützen, die eine sinngetreue Interpretation in nachgelagerten Anwendungen ermöglichen, führen wir eine Methode ein, die den vereinfachten Sätzen eine semantische Ebene hinzufügt, indem sie den semantischen Kontext und die Beziehungen zu anderen Einheiten in Form von rhetorischen Beziehungen erfasst. Diese Repräsentation wird anschließend verwendet, um relationale Tupel in ihrem semantischen Kontext zu extrahieren und so die Ausgabe aktueller Open-IE-Systeme mit zusätzlichen Metainformationen anzureichern, wodurch wichtige Kontextinformationen der extrahierten relationalen Tupel gewahrt werden. Auf diese Weise erhalten wir eine innovative leichtgewichtige semantische Repräsentation für Open IE, die die Extraktion von korrekten, aussagekräftigen und kohärenten Tupeln unterstützt.

2 Diskursorientierte Zerlegung komplexer Satzstrukturen

In einem ersten Schritt wird ein diskursorientierter Ansatz zur Zerlegung komplexer Sätze vorgestellt. Dieser wandelt Sätze, die eine komplexe Struktur aufweisen, in eine semantische Hierarchie von syntaktisch vereinfachten Minimalaussagen um. Das Framework unterscheidet sich von bisherigen Ansätzen durch einen linguistisch fundierten Transforma-

tionsprozess, der unter Verwendung von klausalen sowie phrasalen Zerlegungsmechanismen komplexe englische Sätze in atomare Aussagen mit einer kanonischen SPO-Struktur überführt. Dabei nimmt es einen Satz als Eingabe entgegen und führt einen rekursiven Transformationsprozess auf der Grundlage von 35 manuell definierten Grammatikregeln durch, die aus der Struktur von Sätzen abgeleitete syntaktische und lexikalische Merkmale kodieren. Jede Regel spezifiziert, (i) wie die Eingabe in syntaktisch vereinfachte Sätze zu zerlegen ist und (ii) wie eine kontextuelle Hierarchie zwischen den aufgespaltenen Komponenten aufzubauen und die semantische Beziehung zwischen ihnen zu ermitteln ist.

2.1 Auflösen komplexer Satzstrukturen: Zerlegung in Minimalaussagen

Komplexe verschachtelte Satzstrukturen stellen eine große Herausforderung für Open-IE-Anwendungen dar. Experimente haben gezeigt, dass es schwierig ist, Beziehungen zu erfassen, die sich über diese Strukturen hinweg erstrecken oder in nicht standardmäßiger Form vorliegen. Daher werden solche Beziehungen von aktuellen Ansätzen häufig übersehen oder inkorrekt extrahiert. Um dieser Einschränkung entgegenzuwirken, stellen wir einen Ansatz zur Textvereinfachung (engl. *"Text Simplification"*, TS) vor, der darauf abzielt, die Struktur komplexer Sätze zu vereinfachen, indem sie in atomare Einheiten zerlegt werden, die jeweils eine kanonische SPO-Struktur aufweisen. Dabei müssen die resultierenden vereinfachten Sätze folgende Eigenschaften besitzen: (i) syntaktische Korrektheit, (ii) semantische Korrektheit und (iii) Minimalität.

TRANSFORMATIONSREGEL	TREGEX-MUSTER	EXTRAHIERTER SATZ
Extraktor für **koordinierte Verbphrasen**	ROOT <<: (S < (NP $.. (VP < +(VP) (VP > VP ?$.. CC & $.. **VP**))))	NP + **VP**.
Extraktor für dem Hauptsatz **nachgestellte Adverbialsätze**	ROOT <<: (S < (NP $.. (VP < +(VP) (SBAR < (***S*** < ***(NP*** $.. ***VP)*****))))))	***S*** < ***(NP*** $.. ***VP)***.

Tab. 1: Eine Auswahl von Transformationsregeln. Ein fettgedrucktes Muster repräsentiert denjenigen Teil des Eingabesatzes, der extrahiert und in einen eigenständigen Satz umformuliert wird. Das umrahmte Muster verweist auf seinen Referenten. Ein unterstrichenes Muster (durchgezogene Linie) wird aus dem verbleibenden Teil der Eingabe gelöscht. Das mit einer gepunkteten Linie unterstrichene Muster dient als Schlagwortphrase für die Identifizierung der rhetorischen Beziehung. Ein kursives Muster wird als Kontextsatz gekennzeichnet.

Während die ersten beiden Eigenschaften vergleichsweise einfach zu charakterisieren sind, erfordert letztere eine genauere Definition. Wir haben daher eine umfassende linguistische Analyse durchgeführt, um die Eigenschaft der Minimalität sowohl auf syntaktischer[4] als auch semantischer[5] Ebene detailliert zu spezifizieren. Auf der Grundlage dieser Analyse haben wir dann insgesamt 35 linguistisch fundierte Grammatikregeln für die Zerlegung klausaler und phrasaler Komponenten und deren Umwandlung in eigenständige Sätze definiert. Zwei Beispiele für solche Transformationsregeln sind in Tabelle 1 zu finden.

[4] Aus syntaktischer Sicht ist ein Minimalsatz ein einfacher Satz, der auf seine wesentlichen Bestandteile reduziert wurde, d.h. auf diejenigen Elemente, die Teil seines Satztyps sind.

[5] Semantisch betrachtet kann ein Minimalsatz als eine Aussage verstanden werden, die ein einzelnes Ereignis beschreibt, das aus einem Prädikat und seinen zugehörigen Kernargumenten besteht.

Die Transformationsregeln definieren sowohl die Punkte, an denen ein Satz aufzuspalten ist, als auch die zur Rekonstruktion grammatikalisch korrekter Sätze erforderlichen Umformulierungen. Jede Regel nimmt den Parsebaum eines Satzes als Eingabe entgegen und spezifiziert ein Muster, das im Falle einer Übereinstimmung bestimmte Konstituenten aus dem Baum extrahiert. Die segregierten Komponenten sowie die verbleibenden Elemente werden anschließend in syntaktisch wohlgeformte, eigenständige Sätze transformiert. Zum besseren Verständnis des Zerlegungsverfahrens ist in Abbildung 1 das Ergebnis des ersten Transformationsschrittes für den Beispielsatz aus Abbildung 2 dargestellt.

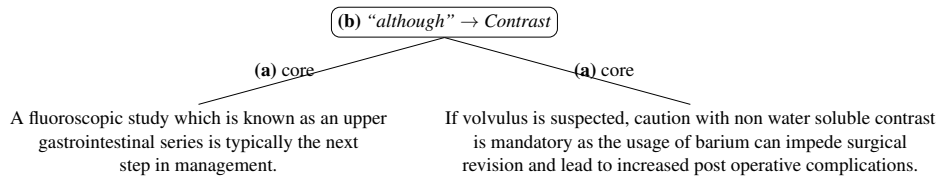

Abb. 1: Semantische Hierarchie nach dem ersten Transformationsdurchgang.

Beiträge Mit der Satzteilung wird eine Operation adressiert, die im Bereich der Textvereinfachung noch weitgehend unerforscht ist. Darüber hinaus wird die Aufgabe der Zerlegung von Sätzen erstmalig um das Konzept der Minimalität erweitert, einschließlich einer detaillierten Spezifikation dieser Eigenschaft sowohl auf syntaktischer als auch semantischer Ebene. Außerdem wird eine systematische Analyse der syntaktischen Phänomene komplexer Satzkonstruktionen durchgeführt, deren Ergebnis in einem Satz von 35 linguistisch begründeten Vereinfachungsregeln materialisiert wird. Schließlich hat eine umfassende Evaluierung unter Einbeziehen aktueller TS-Systeme gezeigt, dass der hier präsentierte Ansatz den Stand der Technik im Bereich der strukturellen Textvereinfachung übertrifft und zu einer vereinfachten Ausgabe führt, die weitgehend grammatikalisch korrekt ist, aus atomaren Sätzen besteht und die Bedeutung der Eingabe bewahrt.

2.2 Einbeziehen von Kontextinformationen: Aufbau einer semantischen Hierarchie

Wenn komplexe Eingabesätze in eine Abfolge selbstständiger Propositionen ohne Beibehaltung ihres semantischen Kontexts zerlegt werden, kann dies dazu führen, dass inkohärente Aussagen entstehen, denen Kontextinformationen fehlen. Diese sind für eine korrekte Interpretation jedoch erforderlich. Daher können solche isolierten Aussagen leicht zu falschen Schlussfolgerungen in nachgelagerten Open-IE-Anwendungen führen. Um dieser Herausforderung zu begegnen, entwickeln wir eine neuartige kontextwahrende Repräsentation, die eine semantische Ebene über die vereinfachten Sätze legt, indem sie eine semantische Hierarchie zwischen ihnen aufbaut. Zu diesem Zweck wird die Diskursstruktur eines Satzes bestimmt durch (i) das Errichten einer kontextuellen Hierarchie zwischen den aufgespaltenen Komponenten und (ii) die Identifizierung rhetorischer Beziehungen, die zwischen ihnen bestehen.

Aufbau einer kontextuellen Hierarchie Um die Kohärenz der zerlegten Einheiten zu wahren, wird eine kontextuelle Hierarchie zwischen den aufgespaltenen Sätzen aufgebaut. Hierfür werden sie mit Informationen über ihren Konstituententyp verbunden, die die relative Wichtigkeit des in der jeweiligen Proposition ausgedrückten Inhalts angeben: Zwei Texteinheiten gleicher Relevanz werden als *Kernsätze* (engl. *"core"*) bezeichnet; dasselbe gilt für Einheiten, die die Hauptaussage der Eingabe widerspiegeln. Im Gegensatz dazu werden Aussagen, die eine Hintergrundinformation liefern, als *Kontextsätze* (engl. *"context"*) bezeichnet. Um zwischen diesen beiden Arten von Konstituenten zu unterscheiden, kodieren die Transformationsregeln eine syntaxbasierte Methode, bei der untergeordnete Klausen und Phrasen als Kontextsätze klassifiziert werden, während ihre übergeordneten Gegenstücke sowie koordinierte Klausen und Phrasen als Kernsätze bezeichnet werden.[6] Dieser Ansatz ermöglicht die Unterscheidung zwischen Schlüsselinformationen, die die Kernaussage der Eingabe enthalten, und Kontextinformationen, die Zusatzinformationen liefern, was zu einer zweistufigen hierarchischen Darstellung in Form von Kernfakten und begleitenden Kontextinformationen führt.

Bewahren der semantischen Beziehungen Um die semantischen Beziehungen zwischen den zerlegten Komponenten aufrechtzuerhalten, werden die rhetorischen Relationen, die zwischen den vereinfachten Sätzen bestehen, identifiziert und klassifiziert. Dabei werden sowohl syntaktische als auch lexikalische Merkmale, die in den Transformationsregeln kodiert sind, verwendet. Lexikalische Merkmale werden basierend auf der Arbeit von [KD94] aus dem Parsebaum in Form von Schlagwörtern bzw. -phrasen extrahiert. Anhand einer vordefinierten Liste von rhetorischen Stichwörtern [TD13] werden dann die Arten von rhetorischen Beziehungen ermittelt. Die zweite Regel in Tabelle 1 gibt beispielsweise an, dass im Beispielsatz aus Abbildung 1 *"although"* (dt. *"obwohl"*) als Stichwort dient, welches auf eine *Gegensatz*-Beziehung abgebildet wird (siehe (b)). Darüber hinaus sind einige der Regeln explizit darauf ausgelegt, ausgewählte rhetorische Beziehungen zu identifizieren, die stark auf syntaktischen Merkmalen beruhen, darunter die Beziehungen *"Zweck"* und *"Zuschreibung"*. In einem solchen Fall wird kein Schlagwort extrahiert. Stattdessen erfolgt die Bestimmung der zugehörigen rhetorischen Relation anhand der Struktur des Parsebaums.

Verknüpfter Propositionsbaum als Ergebnis Abbildung 1 zeigt das Ergebnis des ersten Transformationsdurchgangs für den Beispielsatz aus Abbildung 2, bei dem eine semantische Hierarchie zwischen dem aufgegliederten Satzpaar aufgebaut wird. Die resultierenden Blattknoten werden anschließend rekursiv in einem Top-down-Ansatz weiter

[6] Dabei sind jedoch zwei Ausnahmen zu beachten. Im Falle einer Zuschreibung (z. B. *"Obama noted that he would resign his Senate seat."*) wird die untergeordnete Klause, die die Schlüsselinformation des Ausgangssatzes enthält (*"He would resign his Senate seat."*), mit einem Kernlabel versehen, während die übergeordnete Klause, die die sich äußernde Entität benennt (*"Obama announced"*), als Kontextinformation gekennzeichnet wird, da sie als weniger relevant angesehen wird. Darüber hinaus werden bei der Zerlegung von Adverbialsätzen des Gegensatzes (engl. *"Contrast"*) sowohl die über- als auch die untergeordnete Klause als Kernsatz gekennzeichnet (siehe (a) in Abbildung 1). Dies erfolgt in Übereinstimmung mit der Rhetorischen Strukturtheorie (RST) [MT88], bei der Textabschnitte, die in einer Gegensatz-Beziehung stehen, beide als Kern angesehen werden.

vereinfacht und damit schrittweise in ihre grundlegenden semantischen Bausteine mit einer kanonischen SPO-Struktur zerlegt. Wenn keine der Transformationsregeln mehr mit den vereinfachten Sätzen übereinstimmt, hält der Algorithmus an und gibt den generierten verknüpften Propositionsbaum (engl. *"Linked Proposition Tree", LPT*) aus. Der finale Baum unseres Beispiels ist im oberen Teil von Abbildung 2 zu sehen. Seine Blattknoten repräsentieren die im Zuge des Transformationsprozesses generierten minimalen Propositionen $prop \in PROP$. Interne Knoten werden durch Kästchen dargestellt, die den Konstituententyp sowie die identifizierte rhetorische Relation $rel \in REL$ enthalten, die die zugehörigen Sätze miteinander verbindet. Jeder Knoten des Baumes ist durch eine Kante mit seinem Elternknoten verbunden, die ein Kern- oder Kontextlabel $c \in CL$ darstellt.

Beiträge Es wird eine feingranulare Darstellung komplexer Sätze erzeugt in Form von (i) hierarchisch geordneten und (ii) semantisch miteinander verknüpften Sätzen. Dies ist das erste Mal, dass syntaktisch komplexe Sätze innerhalb ihres semantischen Kontexts zerlegt werden, so dass eine neuartige kontextwahrende Repräsentation entsteht, die eine semantische Schicht über die vereinfachten Sätze legt.

3 Extraktion semantisch typisierter Relationen

In einem zweiten Schritt nutzen wir die semantische Hierarchie von Minimalaussagen für die *Extraktion semantisch typisierter relationaler Tupel*. Dieser Ansatz basiert auf der Idee, dass die feingranulare Repräsentation komplexer Aussagen in Form von hierarchisch geordneten und semantisch verknüpften Sätzen mit einer vereinfachten Syntax die Aufgabe von Open IE in zweierlei Hinsicht unterstützt: (1) Verbesserung der Genauigkeit und der Erkennungsrate der extrahierten Relationen und (2) Generieren einer kontextwahrenden Repräsentation durch das Anreichern der Ausgabe mit semantischen Informationen.

3.1 Anreichern der Ausgabe mit semantischen Informationen

Um semantisch typisierte relationale Tupel aus einem komplexen Satz zu extrahieren, wird nacheinander jede im vorherigen Schritt generierte minimale Proposition $prop \in PROP$ als Eingabe an das Open-IE-System gesendet, dessen Ausgabe mit semantischen Informationen angereichert werden soll. Nach der Identifikation der Prädikat-Argument-Strukturen in den vereinfachten Sätzen müssen die extrahierten relationalen Tupel nun auf ihren semantischen Kontext abgebildet werden. Dieser kann dem Propositionsbaum *LPT* in Form von semantischen Beziehungen $rel \in REL$ zwischen den aufgespaltenen Propositionen $prop \in PROP$ und ihrer hierarchischen Ordnung $c \in CL$ entnommen werden. Da Open-IE-Systeme oft mehr als ein relationales Tupel unterschiedlicher Granularität aus einem gegebenen vereinfachten Satz extrahieren, muss zunächst bestimmt werden, welches von ihnen die Hauptaussage des zugrunde liegenden Minimalsatzes darstellt. Diese Entscheidung wird auf Basis folgender Heuristik getroffen: Es wird angenommen, dass eine Extraktion die Hauptaussage eines Satzes verkörpert, wenn (1) das Hauptverb des Eingabesatzes in der relationalen Phrase rel_phrase des extrahierten Tupels enthalten ist (z. B.

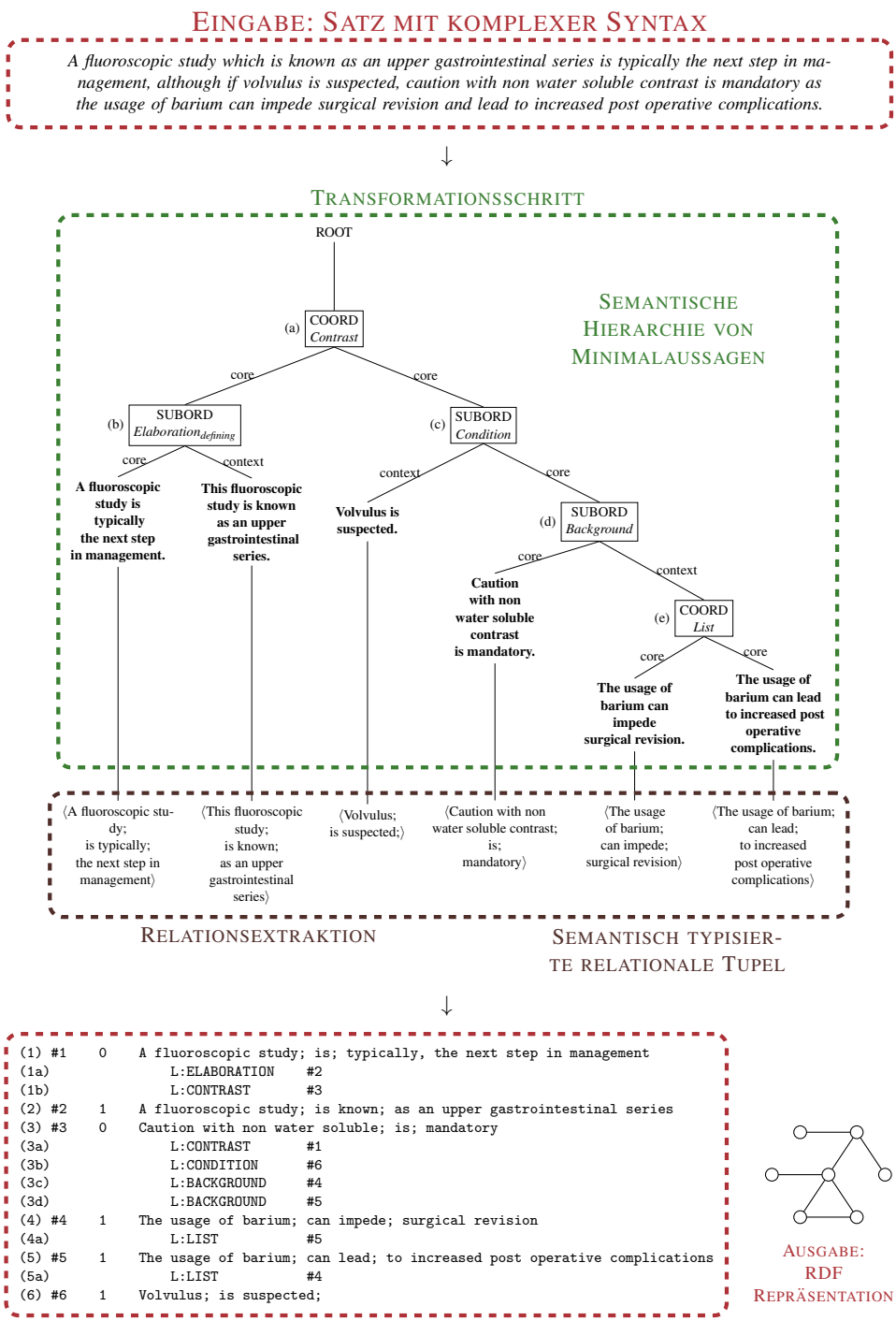

Abb. 2: TS-Open-IE-Pipeline. Ein komplexer Eingabesatz wird zunächst in eine semantische Hierarchie von Minimalpropositionen zerlegt. Diese dient Open-IE-Systemen anschließend als Grundlage zur Extraktion semantisch typisierter relationaler Tupel. Auf diese Weise werden komplexe Textdaten in eine neuartige leichtgewichtige Bedeutungsrepräsentation in Form von normalisierten und kontextwahrenden Prädikat-Argument-Strukturen überführt.

⟨*Volvulus*; *is* *suspected*; ⟩); oder (2) das Hauptverb des Eingabesatzes dem Objektargument arg_{obj} des extrahierten Tupels entspricht (z. B. ⟨*Volvulus*; *is*; *suspected*⟩).

Jede Extraktion, die eines der oben genannten Kriterien erfüllt, fungiert als Haupttupel *htuple(s)* des entsprechenden vereinfachten Satzes *s* und wird somit auf diesen abgebildet. Sobald die Zuordnung von Extraktionen zu den Minimalaussagen $prop \in PROP$ aus dem verknüpften Propositionsbaum *LPT* erfolgt ist, kann die kontextuelle Hierarchie $c \in CL$ sowie die semantische Beziehung $rel \in REL$, die zwischen einem Paar von zerlegten Satzeinheiten besteht, direkt auf die extrahierten relationalen Tupel übertragen werden. Auf diese Weise werden sie in den semantischen Kontext, der aus dem Ausgangssatz abgeleitet wurde, eingebettet. Extraktionen, die kein Haupttupel darstellen, werden verworfen. Das Ergebnis des Relationsextraktionsschritts ist im mittleren Teil von Abbildung 2 dargestellt.

Standardisiertes Open-IE-Ausgabeschema Die von einem Open-IE-System aus einer Menge vereinfachter Sätze extrahierten relationalen Tupel werden anschließend in eine leichtgewichtige Bedeutungsrepräsentation in Form von semantisch typisierten Prädikat-Argument-Strukturen umgewandelt. Ein binäres relationales Tupel $t \leftarrow (arg_{subj}, rel\_phrase, arg_{obj})$ wird dabei erweitert um (i) einen eindeutigen Bezeichner *id*, (ii) Informationen über die kontextuelle Hierarchie, die sog. *Kontextschicht cl*, sowie (iii) zwei Arten von semantisch klassifizierten Kontextargumenten, C_S (*einfache Kontextargumente*) und C_L (*verlinkte Kontextargumente*). Die aus einem komplexen Satz extrahierten Tupel werden damit als eine Menge von (id, cl, t, C_S, C_L)-Tupeln repräsentiert, wobei die *Kontextschicht cl* die hierarchische Ordnung von Kern- und Kontextfakten kodiert. Tupel der Kontextebene 0 enthalten die Kerninformationen eines Satzes, während Tupel der Kontextschicht $cl > 0$ Kontextinformationen über Extraktionen der Kontextebene $cl - 1$ bereitstellen. Beide Arten von Kontextargumenten, C_S und C_L, liefern semantisch klassifizierte Kontextinformationen über die in *t* ausgedrückte Aussage. Während ein einfaches Kontextargument $c_S \in C_S, c_S \leftarrow (s, rel)$ eine Texteinheit *s* enthält, die durch die semantische Beziehung *rel* klassifiziert ist, bezieht sich ein verlinktes Kontextargument $c_L \in C_L, c_L \leftarrow (id(z), rel)$ auf den Inhalt, der in einem anderen relationalen Tupel *z* wiedergegeben wird.

Die Bedeutungsrepräsentation des Beispielsatzes ist im unteren Teil von Abbildung 2 zu sehen. Relationale Tupel werden durch tabulatorgetrennte Zeichenfolgen repräsentiert, die den Identifikator *id*, die Kontextebene *cl* und die durch ein binäres relationales Tupel $t \leftarrow (arg_{subj}, rel\_phrase, arg_{obj})$ repräsentierte Kernextraktion (z. B. #1 und #2) spezifizieren. Kontextargumente (C_S und C_L) werden durch eine zusätzliche Einrückungsebene zu ihren übergeordneten Tupeln gekennzeichnet (z. B. #1a und #1b). Ihre Darstellung besteht aus einer Typzeichenkette und einem durch Tabulatoren getrennten Inhalt. Der Typ-String kodiert sowohl den Kontexttyp (`S` für ein einfaches Kontextargument $c_S \in C_S$ und `L` für ein verknüpftes Kontextargument $c_L \in C_L$) als auch die klassifizierte semantische Beziehung (z. B. *Ursache*, *Zweck*). Der Inhalt eines einfachen Kontextarguments ist eine einzelne Phrase (z. B. *"[This was] on August 6, 2009."*), während der Inhalt eines verknüpften Kontextarguments der Bezeichner der jeweiligen Zielextraktion ist, der wiederum ein vollständiges relationales Tupel repräsentiert (z. B. #1a).

Beispiel Auf diese Weise kann jedes beliebige Open-IE-System das diskursorientierte TS-Framework nutzen, um seine Extraktionen mit semantischen Informationen anzureichern. Die semantisch typisierte Ausgabe, die von dem Open-IE-System RnnOIE [St18] bei Verwendung des Frameworks erzeugt wird, ist im unteren Teil von Abbildung 2 dargestellt. Im Vergleich dazu zeigt Abbildung 3 die Menge der von RnnOIE ausgegebenen Tupel, wenn es direkt mit dem komplexen Ausgangssatz arbeitet. In diesem Fall wird lediglich eine unzusammenhängende Anordnung von relationalen Tupeln generiert, die schwer zu interpretieren und damit anfällig für falsche Schlussfolgerungen sind, da sie den Kontext ignorieren, in dem eine Extraktion vollständig und korrekt ist. So ist es beispielsweise für ein korrektes Verständnis des Beispielsatzes wichtig, zwischen Informationen, die im Satz direkt festgestellt werden, und Informationen, die nur bedingt wahr sind, zu unterscheiden, wie z. B. im Falle der von der vorangestellten *if*-Klause abhängigen Aussage *"caution with non water soluble contrast is mandatory."* Wie die in Abbildung 3 dargestellte Ausgabe zeigt, ist RnnOIE jedoch nicht in der Lage, diese Beziehung zu erfassen. Mit Hilfe der vom TS-Ansatz generierten semantischen Hierarchie kann die Ausgabe jedoch leicht mit Kontextinformationen angereichert werden, die es ermöglichen, die semantische Beziehung zwischen den extrahierten Relationen zu erfassen und somit ihre Interpretierbarkeit in nachgelagerten Anwendungen zu erhalten.

```
RnnOIE (eigenständig):
(1) (A fluoroscopic study; known; as an upper gastrointestinal series)
(2) (caution with non water soluble contrast; is; mandatory as the usage of barium)
(3) (as the usage; of barium can impede; surgical revision and lead)
(4) ( ; to increased; post operative complications)
```

Abb. 3: Von RnnOIE extrahierte relationale Tupel *ohne* Verwendung des TS-Ansatzes.

3.2 Verbessern der Genauigkeit und Erkennungsrate der extrahierten Relationen

In einer umfassenden vergleichenden Analyse aktueller Open-IE-Systeme auf zwei großen Open-IE-Benchmark-Korpora [SD16] [BAM19] wird gezeigt, dass die kanonische Struktur der vereinfachten Sätze die Extraktion von relationalen Tupeln erleichtert. Es wird eine um bis zu 32% höhere Genauigkeit erzielt, woraus sich schließen lässt, dass Open-IE-Systeme bei Verwendung des TS-Ansatzes in der Lage sind, fehlerhafte Extraktionen zu korrigieren. Die um bis zu 30% höhere Erkennungsrate zeigen darüber hinaus, dass nun Relationen, die zuvor aufgrund komplexer Strukturen von den Open-IE-Systemen übersehen wurden, erkannt und korrekt extrahiert werden.

4 Fazit

Diese Arbeit befasst sich mit der Herausforderung, unstrukturierte Textdaten in Form von komplexen Sätzen in eine formale semantische Repräsentation zu überführen, die von Maschinen interpretiert werden kann. Sie ist motiviert durch die Unzulänglichkeiten bestehender Open-IE-Systeme, die häufig (i) Beziehungen übersehen oder inkorrekte Extraktionen produzieren, wenn sie auf langen, verschachtelten Satzstrukturen arbeiten; und (ii) keine Kontextinformationen berücksichtigen, was zu unvollständigen, uninformativen oder inkohärenten Extraktionen führt. Um diese Einschränkungen zu überwinden,

präsentieren wir ein Open-IE-Framework, das eine neuartige leichtgewichtige semantische Repräsentation in Form von normalisierten und kontextwahrenden relationalen Tupeln erstellt. Unser Ansatz leistet zwei wesentliche Beiträge. Erstens stellen wir eine diskursorientierte Methode zur Zerlegung komplexer Sätze vor, durch die die Komplexität der Extraktion von Relationen reduziert wird. Zweitens führen wir einen Ansatz ein, der den vereinfachten Sätzen eine semantische Ebene hinzufügt, indem er ihren semantischen Kontext erfasst, was zu einer informativeren Ausgabe führt, die wichtige Kontextinformationen der extrahierten relationalen Tupel bewahrt. Unser Framework stellt somit eine innovative Lösung für Open-IE dar, die die Extraktion von korrekten, aussagekräftigen und kohärenten relationalen Tupeln unterstützt.

Literaturverzeichnis

[BAM19] Bhardwaj, Sangnie; Aggarwal, Samarth; Mausam, Mausam: CaRB: A Crowdsourced Benchmark for Open IE. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. ACL, Hong Kong, China, S. 6262–6267, November 2019.

[KD94] Knott, Alistair; Dale, Robert: Using linguistic phenomena to motivate a set of coherence relations. Discourse processes, 18(1):35–62, 1994.

[MT88] Mann, William C; Thompson, Sandra A: Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3):243–281, 1988.

[Ni22] Niklaus, Christina: From Complex Sentences to a Formal Semantic Representation using Syntactic Text Simplification and Open Information Extraction. Dissertation, University of Passau, 2022.

[SD16] Stanovsky, Gabriel; Dagan, Ido: Creating a Large Benchmark for Open Information Extraction. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. ACL, Austin, Texas, S. 2300–2305, November 2016.

[St18] Stanovsky, Gabriel; Michael, Julian; Zettlemoyer, Luke; Dagan, Ido: Supervised Open Information Extraction. In: Proceedings of the 2018 Conference of the North American Chapter of the ACL. ACL, New Orleans, Louisiana, S. 885–895, Juni 2018.

[TD13] Taboada, Maite; Das, Debopam: Annotation upon Annotation: Adding Signalling Information to a Corpus of Discourse Relations. D&D, 4(2):249–281, 2013.

Christina Niklaus wurde 1987 in Schweinfurt geboren. Sie studierte Informatik an der Otto-Friedrich-Universität Bamberg, der Universität Passau und der ENSEEIHT Toulouse. Anschließend begann sie 2016 an der Universität Passau ihre Promotion in der von Prof. Siegfried Handschuh geleiteten Natural Language Processing and Semantic Computing Gruppe. Zwei ihrer insgesamt 20 Publikationen wurden ausgezeichnet (Best Survey Paper Award COLING 2018, Honourable Mention Award CHI 2020). 2022 schloss sie ihre Promotion an der Universität Passau mit Auszeichnung ab. Seit 2022 ist sie Assistenzprofessorin für Informatik mit Fokus auf Datenbanken und Data Engineering an der Universität St.Gallen.