# Comparison of Methods for Automatic Relation Extraction in German Novels

**Krug, Markus**
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

**Wick, Christoph**
christoph.wick@uni-wuerzburg.de
Universität Würzburg, Deutschland

**Jannidis, Fotis**
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

**Reger, Isabella**
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

**Weimer, Lukas**
lukas.weimer@uni-wuerzburg.de
Universität Würzburg, Deutschland

**Madarasz, Nathalie**
nathalie.madarasz@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

**Puppe, Frank**
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Deutschland

## Einleitung

Die automatische Erkennung von spezifischen Relationen ermöglicht Einsichten über die Beziehungen zwischen Entitäten. Solche Informationen können nicht nur als Kantenbezeichner in sozialen Netzwerken fungieren, sondern auch als globale Constraints für das schwierige Problem der Coreference Resolution eingesetzt werden. Darüber hinaus kann eine Relationserkennung zur Beantwortung diverser literarischer Fragestellungen eingesetzt werden, z.B. ob eine Romangattung sich mit bestimmten Relationstypen befasst, oder ob die Arten der Relationen sich über die Jahrhunderte verändern. In dieser Arbeit stellen wir ein Label-Set für die Extraktion von binären Relationen zwischen Personen-Entitäten vor und vergleichen Feature-basierte Ansätze des maschinellen Lernens mit regelbasierten Ansätzen zur automatischen Erkennung dieser Relationen. Da Trainingsmaterial zur Verfügung steht, liegt der Fokus in dieser Arbeit auf dem Einsatz überwachter Methoden, d.h. unsere regelbasierten Verfahren sind ebenfalls auf einer zuvor abgetrennten Menge entwickelt worden. Wir verwenden ein neues Korpus, das manuell mit mehr als 50 verschiedenen, hierarchisch gegliederten Relationstypen annotiert wurde.

## Related Work

Eine Übersicht über Arbeiten zur Relationserkennung findet sich in [Jung et al. 2012] sowie [Bach und Badaskar 2007]. Sowohl für den überwachten, als auch den halb-überwachten Fall wurden erfolgreiche Methoden entwickelt. Da dieses Paper sich hauptsächlich auf überwachte Algorithmen bezieht, geben wir nur einen knappen Überblick über halb-überwachte Verfahren.

Algorithmen zur Relationsextraktion erhalten typischerweise zwei (oder mehr) Referenzen zu Entitäten (sogenannte Instanzen) als Input und sollen die Klasse, und das dazugehörige Label, vorhersagen, welche die Relation zwischen den Entitäten beschreibt. Die meisten Experimente wurden anhand englischer Texte und den Datensätzen der Automatic Content Extraction (ACE) Workshops 2004 und 2006 durchgeführt. Auf dem Datensatz von 2004 wurden Experimente zur Unterscheidung von 5 und 27 verschiedenen Klassen wie Arbeitsplatz-, körperliche, soziale, Mitgliedschafts- und Diskursrelationen (wobei manche Unterklassen von anderen sein können) betrachtet. Hierfür gibt es zahlreiche Ansätze, die jedoch alle versuchen, eine diskriminative Beschreibung der Instanzen zu erhalten und diese davon ausgehend zu klassifizieren:

- In der Feature-basierten Klassifikation wird eine Instanz (normalerweise zwei Referenzen zu Entitäten) durch einen Feature-Vektor mit manchmal mehr als einer Million Dimensionen repräsentiert und mit Methoden wie Maximum Entropy Modellen [Kambhatla 2004] oder Support Vector Machines [Jiang und Zhai 2007] klassifiziert. Der letztere Ansatz konnte auf den ACE2004-Daten einen F1-Score von 72,9% für die Erkennung von 7 verschiedenen Relationen erzielen. In unseren Experimenten verwenden wir für die Feature-basierten Methoden ähnliche Features wie Kambhatla [Kambhatla 2004].
- Kernel-basierte Klassifikation wurde häufig zur Relationsextraktion genutzt und liefert konkurrenzfähige Ergebnisse [Zhou et al. 2007, Zhang et al. 2006, Zhao und Grishman 2005]. Während Feature-basierte Verfahren die Instanz direkt repräsentieren, funktionieren Kernel-basierte Methoden etwas anders. Aus einer technischen Perspektive kann ein Kernel als eine Funktion betrachtet werden, die zwei Instanzen als Input erhält (also ein Paar von

Referenzen) und direkt einen Wert berechnet, der auf der "Ähnlichkeit" dieser Instanzen basiert, wobei einer höherer Wert eine größere Ähnlichkeit anzeigt. Es wurden zahlreiche Kernel für die Relationsextraktion vorgeschlagen; eine tiefgehende Analyse und Erklärung findet sich in Jung et al. [Jung et al. 2012].

- Die regelbasierte Klassifikation verwendet eine für den Menschen lesbare Repräsentation durch Regeln, die entweder manuell erstellt oder gelernt wurden. Als Vorteile können die inhärente Erklärungsfähigkeit und die einfache Integration in Feature-basierte Machine Learning-Verfahren gesehen werden.

Im Folgenden vergleichen wir die genannten Methoden anhand eines Label-Sets zur Erkennung binärer Relationen zwischen Figuren in manuell annotierten Abschnitten von deutschsprachigen Romanen.

# Annotation, Datensatz und Vorverarbeitung

Da Textstellen, an denen Relationen zwischen Entitäten explizit benannt werden, in Romanen typischerweise rar sind, ist es nicht sinnvoll, komplette Romane zu annotieren, da der Ertrag an Daten zu gering wäre. Aus diesem Grund wurde zunächst eine kleine Teilmenge per Hand annotiert und dann genutzt, um mit einem MaxEnt Classifier in einer Active Learning-Umgebung neue Sätze zum Labeln vorschlagen zu können. (Ein Überblick hierzu findet sich in Finn und Kushmerick [Finn und Kushmerick 2003]). Diese Umgebung erhielt Sätze aus 312 verschiedenen Romanen von Projekt Gutenberg und 215 Zusammenfassungen aus dem Kindler Literatur Lexikon Online. Daraus entstand ein Korpus mit 2412 Sätzen, die insgesamt 1265 Relationen enthalten (was wiederum die Knappheit an Daten illustriert). 33 Texte wurden zufällig für die Testmenge ausgewählt, sodass es feste Test- und Trainingsdaten gibt (1988 respektive 424 Sätze mit 1070 respektive 195 Relationen). Die verwendeten Label sind ähnlich zu Massey et al. [Massey et al. 2015]. Die Relationen werden durch eine Ontologie mit momentan 57 verschiedenen Relationstypen repräsentiert, die hierarchisch geordnet sind (beispielsweise ist die Relation "Tochter" der Relation "Familie" untergeordnet). Abbildung 1 zeigt die oberste Ebene des Label-Sets, mit den gleichen Kategorien wie in Massey et al. [Massey et al. 2015] und einer zusätzlichen Relation "Liebe".

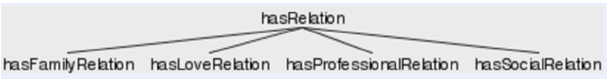


Abbildung 1: Die ersten beiden Ebenen unseres verwendeten Label-Sets mit den vier Haupttypen, die sich weiter in insgesamt 57 Relationstypen untergliedern lassen.

Eine Relation wurde von einem Annotator als ein benannter, gerichteter Bogen zwischen zwei Entitäten in einem Satz gelabelt, sofern sie explizit im Text beschrieben ist. Es wurde immer das spezifischste Label verwendet, da die übergeordneten Relationstypen (vgl. Abbildung 1) daraus abgeleitet werden können. Abbildung 2 zeigt ein Beispiel einer Relation, wie sie in unserem Korpus annotiert ist.

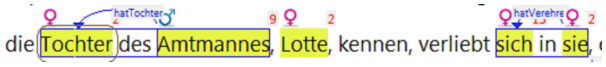


Abbildung 2: Zwei gelabelte Instanzen von Relationen in unserem Datensatz. Die erste zeigt die Relation “hatTochter” und die zweite die Relation “hatVerehrer”.

Um solche Relationen automatisch erkennen zu können, müssen die Texte eine große Zahl an Vorverarbeitungsschritten durchlaufen. Wir verwenden die Figurenerkennung von Jannidis et al. [Jannidis et al. 2015] und die gleiche Vorverarbeitung wie in [Krug et al. 2016].

# Experimente

Wir verwenden einen regelbasierten Ansatz mit manuell erstellten Regeln und zwei Feature-basierte Lernverfahren (Maximum Entropy, MaxEnt und Support Vector Machines, SVM). Der regelbasierte Ansatz nutzt sowohl die textuelle Repräsentation, als auch den kürzesten Pfad im Dependency-Baum und formuliert die Regel auf Basis dieser Repräsentationen und der Repräsentationen aus dem reinen Text. Das folgende Beispiel zeigt Regeln, die zu den Relationen aus Abbildung 2 passen:

- Tochter des <Entität> => hatTochter(2,1)
- Pfad: <Entität>->verliebt<->in<-<Entität> => hatVerehrer(2,1)

Die erste Regel basiert auf der angepassten Text-Repräsentation, während die zweite Regel sich auf den kürzesten Dependency-Pfad zwischen "sich" und "sie" bezieht. Die Zahlen in runden Klammern geben die Richtung an (in beiden Fällen von Entität 2 auf Entität 1). Die Regeln wurden manuell auf den zuvor gewählten Trainingsdaten erzeugt. Insgesamt wurden fast 500 solcher Regeln ermittelt. Der Großteil der Relationen konnte jedoch mit 3 Regeln (ab hier sogenannte Core-Regeln) abgedeckt werden, die Possessiv- und Genitivkonstruktionen abbilden.

Die Feature-basierten Ansätze wurden in zwei Szenarien evaluiert: a) nur mit bereits bekannten Features aus Related Work und b) mit zusätzlichen Booleschen Features (eines pro Regel), falls eine der 500 Regeln passt.

Tabelle 1 zeigt die Evaluationsergebnisse der verschiedenen Methoden für drei hierarchische Ebenen (alle Relationen, Relationen der obersten Ebene, alle 57 Relationstypen) und Tabelle 2 die Ergebnisse für die

vier Relationstypen der obersten Ebene. Während die Verwendung aller Regeln zu einem F1-Score von 71% für alle Relationen und 59% für die vier übergeordneten Relationstypen führt, erreicht der Feature-basierte Ansatz mit MaxEnt mit einem Booleschen Feature für jede Regel etwas bessere Ergebnisse (F1 von 73,6% und 61,2%). Ohne die Regel-Features liegt der Score der Lernverfahren deutlich niedriger. Die SVM erreicht teilweise eine höhere Precision als MaxEnt, aber im Allgemeinen einen signifikant geringeren F1-Wert.

Tabelle 1: Ergebnisse der verschiedenen Ansätze für drei verschiedene Evaluationsszenarien: binär (das reine Vorliegen einer Relation), für die 4 Haupttypen und für alle 57 Relationstypen insgesamt.

Tabelle 2: Ergebnisse für die verschiedenen Ansätze, aufgeschlüsselt nach den 4 Haupttypen. Familienrelationen erreichen sehr gute Ergebnisse mit einem F1-Wert von fast 80% und einer Precision von bis zu 95%. Liebesrelationen sind schwerer zu erkennen, liegen aber dennoch bei 56,3% F1. Die anderen Relationstypen fallen in der Qualität ab, sind aber gleichzeitig weniger relevant.

Sehr auffällig ist das gute Ergebnis für die drei Core-Regeln und dabei besonders die hervorragende Precision von 96,2% für Familien-Relationen. Eine genauere Betrachtung der False Positives (FP) in Tabelle 3 zeigt, dass diese Relationen fast immer syntaktisch korrekt erkannt wurden, aber semantisch irrelevant und daher nicht im Goldstandard annotiert sind (z.B. "mein Gott"). Hier zeigt sich eine Schwachstelle dieser Arbeit: teilweise unpräzise Richtlinien für die Annotation von Relationen. Das ist jedoch ein sehr schwieriges Problem, das eventuell umgangen werden kann, wenn die Relationserkennung kein Ziel in sich, sondern eine untergeordnete Aufgabe im Zuge der Erkennung der Hauptfiguren und deren Beziehungen in Romanen ist.

Tabelle 3: Auswertung der drei Core-Regeln auf unserem Datensatz

| Regel | Beispiel | TP | FP |
|---|---|---|---|
| <Possessive> … <Entity> | seine liebe Mutter [his loved mother] | 83 | 22 |
| <Entity_Noun>.+<GENITIV_Noun> | Frau des Kanzlers [wife of the chancellor] | 21 | 7 |
| <GENITIV_Noun> <Entity_NN> | Peters Frau [Peter’s wife] | 19 | 5 |

# Fazit und zukünftige Arbeiten

Dieses Paper hat gezeigt, dass automatische Relationserkennung eine Herausforderung darstellt. Einfache Regeln können jedoch bereits einen wesentlichen Teil der Relationen mit hoher Precision erkennen. Dennoch ist der Bedarf an weiteren Verbesserungen durch fortschrittliche Methoden hier deutlich. Zudem ist die Evaluation der Relationserkennung an sich schwierig und kann besser im Kontext eines übergeordneten Ziels wie der automatischen Erstellung eines Netzwerks der Hauptfiguren eines Romans [Krug 2016] oder der Gattungsklassifikation [Hettinger et al. 2015] eingebracht werden.

# Bibliography

**Bach, Nguyen / Badaskar, Sameer** (2007): „A review of relation extraction“, in: *Literature review for Language and Statistics* II.

**Finn, Aidan / Kushmerick, Nicolas** (2003): „Active learning selection strategies for information extraction“, in: *Proceedings of the International Workshop on Adaptive Text Extraction and Mining (ATEM-03)*.

**Hettinger, Lena / Becker, Martin / Reger, Isabella / Jannidis, Fotis / Hotho, Andreas** (2015): „Genre classification on German novels“, in: *Proceedings of the 12th International Workshop on Text-based Information Retrieval*.

**Jannidis, Fotis / Krug, Markus / Reger, Isabella / Toepfer, Martin / Weimer, Lukas / Puppe, Frank** (2015): „Automatische Erkennung von Figuren in deutschsprachigen Romanen“, in: *DHd 2015: Von Daten zu Erkenntnissen*.

**Jiang, Jing / Zhai, ChengXiang** (2007): „A Systematic Exploration of the Feature Space for Relation Extraction“, in: *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*.

**Jung, Hanmin / Choi, Sung-Pil / Lee, Seungwoo / Song, Sa-Kwang** (2012): „Survey on Kernel-Based Relation Extraction“, in: Sakurai, Shigeaki (ed.): *Theory and Applications for Advanced Text Mining*. InTech Open Science 10.5772/51005.

**Kambhatla, Nanda** (2004): „Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations“, in: *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*.

**Krug, Markus / Fotis, Jannidis / Reger, Isabella / Weimer, Lukas / Macharowsky, Luisa / Puppe, Frank** (2016): „Attribuierung direkter Reden in deutschen Romanen des 18.-20. Jahrhunderts. Methoden zur Bestimmung des Sprechers und des Angesprochenen“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung*.

**Krug, Markus / Fotis, Jannidis / Reger, Isabella / Weimer, Lukas / Macharowsky, Luisa / Puppe, Frank** (2016): „Comparison of Methods for the Identification of Main Characters in German Novels“, in: *DH2016: Converence Abstracts*.

**Massey, Philip / Xia, Patrick / Bamman, David / Smith, Noah A.** (2015): „Annotating Character Relationships in Literary Texts“, in: *arXiv*, arXiv:1512.00728.

**Zhao, Shubin / Grishman, Ralph** (2005): „Extracting relations with integrated information using kernel methods“, in: *Proceedings of ACL-2005*.