

Bachelorarbeit zur Erlangung des akademischen Grades  
**Bachelor of Science**  
**des Fachbereiches Computerlinguistik**  
der Humanwissenschaftlichen Fakultät der Universität Potsdam

# **Eigennamen- und Zitaterkennung in Rechtstexten**

Elena Leitner  
Matrikel-Nr.: 781063

**Erstgutachter: Dr. Thomas Hanneforth**  
Fachbereich der Computerlinguistik der Universität Potsdam  
**Zweitgutachter: Dr. Georg Rehm**  
Deutsches Forschungszentrum für Künstliche Intelligenz

# Inhaltsverzeichnis

<b>Zusammenfassung</b>	<b>vi</b>
<b>Abstract</b>	<b>vii</b>
<b>1 Einleitung</b>	<b>1</b>
<b>2 Forschungsstand der Eigennamenerkennung</b>	<b>3</b>
2.1 Begriffsdefinition und Entstehungsgeschichte . . . . .	3
2.2 Aktuelle Ansätze der Eigennamenerkennung . . . . .	3
2.3 Ansätze der Eigennamenerkennung in Rechtstexten . . . . .	6
<b>3 Formale Definition der Sequenzmodelle</b>	<b>8</b>
3.1 Begriffsdefinition und Aufgabenbeschreibung . . . . .	8
3.2 Formale Definition von CRFs . . . . .	8
3.3 Formale Definition von LSTMs . . . . .	9
3.4 Formale Definition von LSTM-CRFs . . . . .	10
<b>4 Juristisches Korpus</b>	<b>12</b>
4.1 Rechtstexte als Fachtexte und ihre Besonderheiten . . . . .	12
4.2 Semantische Klassen juristischer Entitäten . . . . .	12
4.3 Beschreibung und Eigenschaften des Korpus . . . . .	16
<b>5 Evaluation und Ergebnisse</b>	<b>19</b>
5.1 Evaluationsmetriken der Eigennamenerkennung . . . . .	19
5.2 Experimentbeschreibung . . . . .	19
5.3 Ergebnisse der CRF-Modelle . . . . .	20
5.4 Ergebnisse der BLSTM-Modelle . . . . .	22
5.5 Analyse der richtigen und falschen Vorhersagen . . . . .	23
5.6 Diskussion . . . . .	27
<b>6 Zusammenfassung und Ausblick</b>	<b>29</b>
6.1 Zusammenfassung . . . . .	29
6.2 Ausblick . . . . .	30
<b>Literaturverzeichnis</b>	<b>31</b>
<b>A Anhang</b>	<b>35</b>

# Abbildungsverzeichnis

3.1 Schematische Darstellung der Eigennamenerkennung.	8
3.2 Aufbau der LSTM-Speicherzelle (Quelle: Überarbeitete Darstellung in Anlehnung an Greff et al., 2017, S. 2)	9
3.3 BLSTM-Modell (links) und BLSTM-CRF-Modell (rechts).	10
4.1 Datenformat des juristischen Korpus.	18
A.1 Quelldaten im XML-Format am Beispiel der Datei WBRE201800396.xml.	35
A.2 Quelldaten im PDF-Format am Beispiel der Datei WBRE201800396.xml.	36
A.3 Konfusionsmatrix des CRF-Fs für die feinkörnigen Klassen.	37
A.4 Konfusionsmatrix des CRF-Fs für die grobkörnigen Klassen.	37
A.5 Konfusionsmatrix des CRF-FGs für die feinkörnigen Klassen.	38
A.6 Konfusionsmatrix des CRF-FGs für die grobkörnigen Klassen.	38
A.7 Konfusionsmatrix des CRF-FGLs für die feinkörnigen Klassen.	39
A.8 Konfusionsmatrix des CRF-FGLs für die grobkörnigen Klassen.	39
A.9 Konfusionsmatrix des BLSTM-CRFs für die feinkörnigen Klassen.	40
A.10 Konfusionsmatrix des BLSTM-CRFs für die grobkörnigen Klassen.	40
A.11 Konfusionsmatrix des BLSTM-CRFs+ für die feinkörnigen Klassen.	41
A.12 Konfusionsmatrix des BLSTM-CRFs+ für die grobkörnigen Klassen.	41
A.13 Konfusionsmatrix des BLSTM-CNN-CRFs für die feinkörnigen Klassen.	42
A.14 Konfusionsmatrix des BLSTM-CNN-CRFs für die grobkörnigen Klassen.	42

## Tabellenverzeichnis

2.1	$F_1$ -Werte von Modellen für die Eigennamenerkennung, die auf dem englischen und deutschen Testkorpus CoNNL 2003 evaluiert wurden. Das mit $\star$ gekennzeichnete Ergebnis ist aus Riedl und Padó (2018) entnommen.	5
4.1	Grob- und feinkörnige Klassen juristischer Entitäten.	13
4.2	Korpusgröße: die Anzahl Dokumente, Token mit und ohne Satzzeichen, Sätze und die Verteilung der juristischen Entitäten.	16
4.3	Verteilung der grob- und feinkörnigen Klassen in den sieben Bundesgerichten.	17
5.1	Precision, Recall und $F_1$ -Werte der CRF-Modelle für die feinkörnigen Klassen.	21
5.2	Precision, Recall und $F_1$ -Werte der CRF-Modelle für die grobkörnigen Klassen.	21
5.3	Precision, Recall und $F_1$ -Werte der BLSTM-Modelle für die feinkörnigen Klassen.	22
5.4	Precision, Recall und $F_1$ -Werte der BLSTM-Modelle für die grobkörnigen Klassen.	23

# Abkürzungsverzeichnis

<b>LER</b>	Legal-Entity-Recognition
<b>NER</b>	Eigennamenerkennung
<b>POS-Tag</b>	Part-Of-Speech-Tag
<b>POS-Tagging</b>	Part-Of-Speech-Tagging

## Bundesgerichte

<b>BAG</b>	Bundesarbeitsgericht
<b>BFH</b>	Bundesfinanzhof
<b>BGH</b>	Bundesgerichtshof
<b>BPatG</b>	Bundespatentgericht
<b>BSG</b>	Bundessozialgericht
<b>BVerfG</b>	Bundesverfassungsgericht
<b>BVerwG</b>	Bundesverwaltungsgericht

## Semantische Klassen

<b>AN</b>	Anwälte
<b>EUN</b>	Europäische Normen
<b>GRT</b>	Gerichte
<b>GS</b>	Gesetze
<b>INN</b>	Institutionen
<b>IS</b>	Institute
<b>KONT</b>	Kontinente
<b>LD</b>	Länder
<b>LDS</b>	Landschaften
<b>LIT</b>	Rechtsliteratur
<b>LOC</b>	Orte
<b>MRK</b>	Marken
<b>MUS</b>	Museen
<b>NRM</b>	Rechtsnormen
<b>ORG</b>	Organisationen
<b>OTH</b>	Verschiedenes
<b>PER</b>	Personen
<b>REG</b>	Einzelfallregelungen
<b>RR</b>	Richter
<b>RS</b>	Rechtsprechungen
<b>ST</b>	Städte
<b>STR</b>	Straßen
<b>UN</b>	Unternehmen
<b>UNI</b>	Universitäten
<b>VO</b>	Rechtsverordnungen
<b>VS</b>	Vorschriften
<b>VT</b>	Verträge

## Sequenzmodelle

<b>BLSTM</b>	Bidirektionales Long-Short-Term-Memory-Netz
--------------	---

---

<b>BLSTM-CNN</b>	Kombination von BLSTM und CNN
<b>BLSTM-CNN-CRF</b>	Kombination von BLSTM, CNN und CRF
<b>BLSTM-CRF</b>	Kombination von BLSTM und CRF
<b>BLSTM-CRF+</b>	Kombination von BLSTM und CRF, inkl. Buchstabeneinbettungen
<b>CNN</b>	Convolutional-Neural-Network
<b>CRF</b>	Conditional-Random-Field
<b>CRF-F</b>	Conditional-Random-Field mit Features
<b>CRF-FG</b>	Conditional-Random-Field mit Features und Namenslisten
<b>CRF-FGL</b>	Conditional-Random-Field mit Features, Namenslisten und Lookup-Tabelle
<b>LSTM</b>	Long-Short-Term-Memory-Netz
<b>LSTM-CRF</b>	Kombination von LSTM und CRF
<b>RNN</b>	Rekurrentes Neuronales Netz

## Zusammenfassung

Die vorliegende Arbeit widmet sich der automatischen Erkennung von Eigennamen und Zitaten in deutschen gerichtlichen Entscheidungen. Dafür wird ein juristisches Korpus konstruiert, das aus ca. 67 Tsd. Sätzen besteht. Die Entscheidungen sind mit 19 Klassen, zu denen Namen sowie Zitate von Personen, Orten, Organisationen, Rechtsnormen, Einzelfallregelungen, Rechtsprechungen und Rechtsliteratur gehören, manuell annotiert. Insgesamt enthält das Korpus rund 54 Tsd. annotierte Entitäten. Als automatische Verfahren werden Conditional-Random-Fields (CRFs) und bidirektionale Long-Short-Term-Memory-Netze (BLSTMs) getestet. Für diese Modelfamilien werden jeweils drei verschiedene Modelle erarbeitet. Die BLSTM-Modelle erreichen die beste Performanz mit einem  $F_1$ -Wert von 95,46 %. Die CRF-Modelle erreichen dagegen maximal 93,23 %.

## Abstract

The present thesis is dedicated to automatic legal entity recognition in German court decisions. For this purpose, a legal dataset, which consists of approximately 67 thousand sentences was compiled. Court decisions were manually annotated with 19 classes, including names and citations of people, places, organizations, legal norms, case-by-case regulations, case laws, and legal literature. Overall, the corpus contains around 54 thousand annotated entities. Conditional Random Fields (CRFs) and bidirectional Long-Short-Term Memory Networks (BLSTMs) are being tested as automatic methods. Three different models are developed for each of these model families. The BLSTM models achieve the best performance with an  $F_1$  value of 95.46 %. By contrast, the CRF models reach a maximum of 93.23 %.

# 1 Einleitung

In der Ära der Digitalisierung wachsen Datenmengen enorm an. Menschen produzieren täglich tausende Texte wie Bücher, Artikel, E-Mails, Posts, Reviews u. Ä., die in verschiedenen Sprachen erstellt und unterschiedlichen Themen gewidmet sind. Trotz des freien Zugangs zu vielen Textsammlungen ist die Analyse relevanter Inhalte durch Menschen kaum denkbar, da dies zu zeit- und kostenaufwendig ist. Dank moderner Verfahren zur automatischen Textanalyse können aber große Datenmengen schnell maschinell verarbeitet werden. Insbesondere ermöglicht es die automatische Textanalyse, wertvolle Informationen wie Eigennamen sowie numerische Ausdrücke zu extrahieren. Diesen Prozess, nämlich das Erkennen von semantischen Konzepten in Texten, nennt man Informationsextraktion (engl. *information extraction*) (Jurafsky und Martin, 2014).

Die Identifikation und Klassifikation von Eigennamen, bekannt als Eigennamenerkennung, ist ein zentraler Baustein und eine wichtige Voraussetzung für viele Verfahren zur automatischen Textanalyse wie z. B. die Textzusammenfassung, die maschinelle Übersetzung, die automatische Beantwortung von Fragen sowie für Suchmaschinen. Prototypisch zählt man zu den semantischen Konzepten (insbesondere in Zeitungstexten) Personen, Orte und Organisationen. Die Eigennamenerkennung umfasst u. a. auch die Identifikation von Medikamenten, Proteinen, chemischen Verbindungen, die in der biomedizinischen Domäne von großer Bedeutung sind.

Die Untersuchung der Eigennamenerkennung hat eine 20-jährige Geschichte. Die ausgearbeiteten Verfahren für die Identifikation und Klassifikation von Eigennamen beruhen auf linearen statistischen Modellen wie z. B. Maximum-Entropie-Modellen (Bender et al., 2003; Clark, 2003), Hidden-Markov-Modellen (Klein et al., 2003; Mayfield et al., 2003) und Conditional-Random-Fields (McCallum und Li, 2003). Heutzutage verwendet man jedoch Neuronale Netze und insbesondere Long-Short-Term-Memory-Netze (Huang et al., 2015; Chiu und Nichols, 2016; Lample et al., 2016; Ma und Hovy, 2016). Für das Englische erzielen moderne, dem Forschungsstand entsprechende Modelle einen  $F_1$ -Wert von ca. 90 % (das  $F_1$ -Maß ist eine standardisierte Evaluationsmetrik in der Eigennamenerkennung), sodass dieses Problem zumindest für Zeitungstexte als gelöst gilt (Didakowski et al., 2006). In der gleichen Domäne weichen die Ergebnisse für das Deutsche ab. Die  $F_1$ -Werte liegen bestenfalls bei 80 %.

Die Digitalisierung betrifft heute auch den Bereich Wirtschaft und Verwaltung und viele Unternehmen müssen sich dem Wandel stellen, um wettbewerbsfähig zu bleiben. Die automatische Verarbeitung großer Datenmengen ist eine der Voraussetzungen dieses Wandels, welche wertvolle Zeit und Kosten einsparen kann. In diesem Zusammenhang werden fortlaufend weitere Verfahren zur automatischen Textanalyse entwickelt, die die charakteristischen Besonderheiten dieser Daten berücksichtigen. So ist die vorliegende Arbeit der Erkennung von semantischen Konzepten – Eigennamen und Zitaten – in deutschen Rechtstexten gewidmet. Die Rechtssprache ist einzigartig und unterscheidet sich stark von der Alltags- bzw. Zeitungssprache. Einerseits betrifft dies den Gebrauch von Eigennamen in Rechtstexten, die relativ selten vorkommen. Andererseits findet man solche semantischen Konzepte wie Bezeichnungen von Rechtsnormen und Verweise auf andere Rechtstexte (Gesetze, Rechtsverordnungen, Vorschriften, Rechtsprechungen usw.), die eine essentielle Rolle spielen. Trotz der Entwicklung der Eigennamenerkennung (auch in anderen Domänen) findet man wenig wissenschaftliche Literatur (Dozier et al., 2010; Cardellino et al., 2017; Glaser et al., 2018), die diesem Thema in der juristischen Domäne gewidmet ist. Die Forschung der Eigennamenerkennung in Rechtstexten ist auch dadurch erschwert, dass (1) keine einheitliche Klassifikation der semantischen Konzepte, die in Rechtstexten zu identifizieren und zu klassifizieren sind, vorhanden ist; (2) weder frei verfügbare juristische Korpora, in denen Eigennamen oder andere Kategorien annotiert wurden, noch einheitliche Annotationsrichtlinien existieren.

In diesem Zusammenhang besteht das Ziel der vorliegenden Arbeit darin, die Eigennamen- und Zitaterkennung in deutschen Rechtstexten zu untersuchen. Dies beinhaltet einerseits die Ausarbeitung der entsprechenden semantischen Klassen (inkl. ihrer Annotationsrichtlinien), und andererseits die Konstruktion eines juristischen Korpus. Zudem werden Modelle basierend auf aktuellen wissenschaftlichen Ansätzen zur Eigennamenerkennung getestet. Daraus ergeben sich folgende Fragen:

1. Welche theoretischen Ansätze sind in der Eigennamenerkennung erfolgreich? Welche wurden für die Eigennamenerkennung in Rechtstexten entwickelt? Inwieweit entsprechen diese dem Forschungsstand?
2. Welche Besonderheiten weisen Rechtstexte allgemein und insbesondere hinsichtlich der Eigenna-

menerkennung auf? Inwieweit unterscheiden sich Rechtstexte im Gebrauch von Eigennamen und Zitaten? Welche Rechtstexte eignen sich für das juristische Korpus am besten?

3. Welche semantischen Kategorien, die Eigennamen und Zitate umfassen, sind für Rechtstexte typisch? Welche Klassen sind im Rahmen der Eigennamen- und Zitaterkennung in Rechtstexten zu identifizieren und zu klassifizieren?
4. Welche Performanz haben aktuelle Modelle? Wie werden unterschiedliche semantische Kategorien erkannt? Welche Klassen werden besser erkannt als andere und bei welchen Klassen gibt es Verbesserungspotenzial?

Die folgende Arbeit ist in fünf Hauptteile gegliedert. Kapitel 2 leitet in das Problem der Eigennamenerkennung ein. So wird zuerst kurz der Begriff und die Entstehung der Eigennamenerkennung behandelt. Danach werden aktuelle Ansätze sowie Systeme mit besonders guter Performanz, die auf Conditional-Random-Fields (CRFs), uni- bzw. bidirektionalen Long-Short-Term-Memory-Netzen (LSTMs, BLSTMs) basieren und die für das Englische und Deutsche entwickelt wurden, in den Mittelpunkt der Betrachtung gerückt. Anschließend werden bereits bestehende Ansätze für die Eigennamenerkennung in Rechtstexten angeführt, wobei deren Vor- bzw. Nachteile und Grenzen diskutiert werden. Kapitel 3 beinhaltet eine formale Definition der aktuellen Modelle in der Eigennamenerkennung und beschreibt deren Funktionsweise.

Kapitel 4 beleuchtet die Entwicklung des juristischen Korpus. Dies schließt eine kurze Beschreibung der Besonderheiten von Rechtstexten und eine Überlegung zur Auswahl der Quelltexte ein, welche sich für die Eigennamen- und Zitaterkennung am besten eignen. Jedoch wird die Aufmerksamkeit auf die Ausarbeitung der juristischen Klassifikation bestehend aus semantischen Kategorien, die für Quelltexte typisch und relevant sind, gerichtet. Zusammenfassend wird auf die Eigenschaften und die Größe des juristischen Korpus eingegangen, das im Rahmen dieser wissenschaftlichen Arbeit entstanden ist.

Kapitel 5 beschreibt das Experiment und seinen Ablauf. Es wird kurz über standardisierte Evaluationsmetriken in der Eigennamenerkennung gesprochen. Danach werden Ergebnisse von CRF- und BLSTM-Modellen dargestellt, die mit dem juristischen Korpus trainiert und evaluiert wurden. Erstens werden diese Modellfamilien und zweitens zwei unterschiedliche Klassifikationen, nämlich eine grob- und eine feinkörnige, verglichen. Darüber hinaus wird anhand der richtig und falsch erkannten Beispiele aus dem Testkorpus diskutiert, wie gut die einzelnen Klassen erkannt wurden und welche möglichen Gründe dafür anzuführen sind.

In Kapitel 6 folgt eine Zusammenfassung der zentralen Ergebnisse, nämlich der Korpuskonstruktion und der Performanz der CRF- und BLSTM-Modelle. Darüber hinaus werden Möglichkeiten und Limitierungen der vorliegenden Arbeit diskutiert. Es wird auch besprochen, wie die Erkennung von Eigennamen und Zitaten in Rechtstexten verbessert werden könnte.

## 2 Forschungsstand der Eigennamenerkennung

In diesem Kapitel wird über die Entstehung der Eigennamenerkennung als automatisches Verfahren berichtet. Es wird auf den Begriff, die typischen semantischen Kategorien, die für Zeitungstexte entwickelt wurden, sowie auf die Besonderheiten der Eigennamenerkennung in anderen Domänen eingegangen. Es wird auch ein kurzer Überblick über Ansätze zur Eigennamenerkennung gegeben. Danach werden Ansätze basierend auf Conditional-Random-Fields und Long-Short-Term-Memory-Netzen näher behandelt, die dem aktuellen Forschungsstand entsprechen. Schließlich wird sich mit Ansätzen zur Eigennamenerkennung in Rechts- texten auseinandersetzt, wobei automatische Verfahren und Klassifikationen der semantischen Kategorien in den Mittelpunkt gerückt werden.

### 2.1 Begriffsdefinition und Entstehungsgeschichte

Die Eigennamenerkennung (engl. *named entity recognition*, NER)<sup>1</sup> beschäftigt sich mit der automatischen Erkennung von Eigennamen in Texten und mit ihrer Zuordnung zu semantischen Kategorien. Als typische Klassen in der Eigennamenerkennung gelten die semantischen Kategorien der Personen **PER**, Orte **LOC**, Organisationen **ORG** und des Verschiedenen **OTH** (Tjong Kim Sang, 2002; Sang und Meulder, 2003; Benikova et al., 2014b; Piskorski et al., 2017). Der Begriff ‚Eigennamenerkennung‘ wurde im Jahre 1996 in der sechsten Message-Understanding-Konferenz formuliert (Grishman und Sundheim, 1996). In den nächsten Jahren wurden wissenschaftliche Wettbewerbe, sogenannte Shared-Tasks<sup>2</sup>, für die Erforschung von Eigennamen in verschiedenen Sprachen wie Englisch, Deutsch, Indisch, Arabisch usw. und in verschiedenen Domänen, insbesondere im journalistischen und biomedizinischen Bereich, veranstaltet und Korpora dafür entwickelt (für weitere Informationen s. Yadav und Bethard, 2018).

Im Laufe der Forschung zur Eigennamenerkennung entwickelten sich verschiedene Typologien der semantischen Kategorien, die task-, sprach- oder domänspezifischen Bedürfnissen entsprachen. In GermEval 2014 (Benikova et al., 2014b) wurden bestehende Klassen um Eigennamen als Teilausdrücke, die mit einem Nomen ein Kompositum bilden, und um Derivationen von Eigennamen erweitert. Zusätzlich gab es eine zweite Ebene für geschachtelte Eigennamen, die in einen anderen Eigennamen eingebettet waren (vgl. der eingebettete Ortsname ‚Potsdam‘ im Organisationsnamen ‚Universität Potsdam‘). In WNUT 2015 (Baldwin et al., 2015), das sich mit der Eigennamenerkennung in Twitter befasste, wurden Eigennamen in zehn Klassen unterteilt: Personen, Musiker, Firmen, Einrichtungen, Sportgruppen, Produkte, Filme, Fernsehsendungen, Ortsbestimmungen und Anderes. In der biomedizinischen Forschung wurde der Begriff so umdefiniert, dass unter Eigennamenerkennung die Erkennung und Klassifikation von Proteinen, Bakterien, Medikamenten u. Ä. zu verstehen ist.

Die Entwicklung der wissenschaftlichen Ansätze zur Eigennamenerkennung lässt sich mittels Shared-Tasks betrachten. Die besten Ergebnisse in CoNLL 2002 (Tjong Kim Sang, 2002) und CoNLL 2003 (Sang und Meulder, 2003) haben statistische Modelle wie Maximum-Entropie-Modelle und Hidden-Markov-Modelle erzielt. In GermEval 2014 waren Conditional-Random-Fields im Fokus. In den letzten vier Jahren sind Neuro-nale Netze die vorherrschend verwendeten Modelle zur Eigennamenerkennung (Huang et al., 2015; Chiu und Nichols, 2016; Lample et al., 2016; Ma und Hovy, 2016).

### 2.2 Aktuelle Ansätze der Eigennamenerkennung

Seit der Entstehung der Eigennamenerkennung wurden Studien in verschiedenen Sprachen und Domänen durchgeführt und viele Ansätze entwickelt (für weitere Informationen s. Nadeau und Sekine, 2007; Yadav

<sup>1</sup>Die deutsche Übersetzung des Begriffs ‚Named entity‘ ist eine Entität bzw. ein Gegenstand, der mit einem Eigennamen bezeichnet wird. Unter Eigennamen dagegen versteht man ein Zeichen, mit dem eine Entität benannt wird. Wenn weiterhin von der Eigennamenerkennung gesprochen wird, sei damit die Identifikation der Zeichen im Text, die auf die entsprechenden Gegenstände Bezug nehmen, und die Klassifikation der Zeichen und Gegenstände als Ganzes verstanden.

<sup>2</sup>Shared task (dt. gemeinsame Aufgabe) ist eine öffentliche Forschungsaufgabe. Das Ziel besteht darin, ein System zur Lösung der Aufgabe anhand von Methoden und Techniken zur automatischen Verarbeitung von unstrukturierten Daten zu entwickeln. Bewertet werden Systeme nach standardisierten Metriken. Anschließend werden die Ergebnisse verglichen und veröffentlicht.

und Bethard, 2018). Nach dem heutigen Forschungsstand erzielen Ansätze, die sich auf Conditional-Random-Fields und Neuronale Netze stützen, die besten Ergebnisse, sowohl für das Englische als auch für das Deutsche (Finkel et al., 2005; Faruqui und Padó, 2010; Benikova et al., 2014a, 2015; Chiu und Nichols, 2016; Lample et al., 2016; Riedl und Padó, 2018, etc.).

Conditional-Random-Fields gehören zu den statistischen Modellen, die anhand eines eingegebenen Wortes und optional anhand seiner Merkmale, auch Features genannt, sowie externer Quellen eine vordefinierte semantische Kategorie vorhersagen. Zu den Features eines Wortes zählt man solche, die dessen Form, Wortart mit einem Part-Of-Speech-Tag (POS-Tag), Präfixe, Suffixe, Position im Satz o. Ä. beschreiben. Zu den externen Quellen gehören verschiedene Listen (engl. *gazetteers*), die Namen von Personen, Orten, Organisationen bzw. Hinweise für deren Erkennung enthalten (für eine detaillierte Beschreibung der Features und Quellen s. Nadeau und Sekine, 2007). Von besonderem Interesse sind daher manuell erstellte Features und Namenslisten, deren Auswahl die Performanz eines Modells signifikant verbessern kann.

Der bekannteste Ansatz, dem CRF zugrunde liegt, und das daraus entwickelte Tool für die Eigennamenerkennung ist StanfordNER<sup>3</sup> (Finkel et al., 2005). Der Tagger<sup>4</sup> bietet Module für Englisch, Deutsch, Spanisch und Chinesisch an und erkennt Eigennamen (Personen-, Orts-, Organisationsnamen sowie Namen anderer Herkunft) und numerische Ausdrücke (Zeit-, Datums-, Währungs-, Prozentangaben). Wie Finkel et al. (2005) berichten, gehören u. a. ein aktuelles Wort, N-Gramme<sup>5</sup> für Buchstaben, ein POS-Tag, ein vorangegangenes und ein nächstes Wort zu ausgewählten Features des StanfordNERs. Das deutsche Modul, das von Faruqui und Padó (2010) entwickelt wurde, verwendet zusätzlich distributionelle Cluster (Clark, 2003), die die Ähnlichkeit eines Wortes zu einem anderen beschreiben. Wenn semantisch ähnliche Wörter zu derselben Klasse von Eigennamen gehören, ermöglicht dies dem Tagger neue, ungewöhnliche Wörter zu klassifizieren. Das englische Modul des StanfordNERs erzielt einen  $F_1$ -Wert von 86,86 % und das deutsche Modul einen  $F_1$ -Wert von 78,2 % evaluiert auf den entsprechenden Testkorpora CoNLL 2003 (s. die  $F_1$ -Werte der beschriebenen Ansätze in Tabelle 2.1).

Ein anderes Tool, das speziell für die deutsche Sprache entwickelt wurde, ist GermaNER<sup>6</sup> (Benikova et al., 2015). Der Tagger wurde auf dem Korpus NoSta-D Named-Entity (Benikova et al., 2014b), das im Rahmen GermEval 2014 entwickelt wurde, trainiert und evaluiert. Die aktuelle Performanz, wie die Entwickler auf der Github-Seite berichten, entspricht einem  $F_1$ -Wert von 77 %. GermaNER erkennt auch vier Klassen von Eigennamen. Zu den Features, die mit dem CRF verwendet werden, gehören u. a. Präfixe, Suffixe, POS-Tags und thematische Cluster. Zusätzlich bedient man sich eines Features für die Wortähnlichkeit bestehend aus den vier am meisten ähnlichen Wörtern zum aktuellen Wort, die dadurch seine semantische Ähnlichkeit beschreiben. In GermaNER sind auch Namenslisten integriert, um Eigennamen in Texten besser zu erkennen. Einige Features beziehen sich dabei auf ein aktuelles Wort und Wörter aus dem Kontext, also auf die zwei vorangegangenen und die zwei nächsten Wörter. Über Ergebnisse dieses Tools, die auf dem deutschen Testkorpus CoNLL 2003 evaluiert wurden, berichteten Riedl und Padó (2018). GermaNER zeigte im Vergleich zu StanfordNER eine bessere Performanz mit einem  $F_1$ -Wert von 79,37 %.

Gute Ergebnisse für die englische Sprache zeigte der Ansatz von Tkachenko und Simanovsky (2012). Die Autoren testeten sorgfältig den Einfluss von verschiedenen Features auf die Performanz ihres Modells. Dabei wurden diese in interne und externe Features unterteilt. Zu den internen gehörten Features, die eine Information hinsichtlich linguistischer Eigenschaften eines Wortes geben (Form des Wortes, Großschreibung etc.). Zu den externen Features wurden POS-Tags, Quellen, Wort- und Phrasencluster gezählt. Das finale Modell erzielte einen  $F_1$ -Wert von 91,02 %.

Passos et al. (2014) erarbeiteten ein erweitertes Skip-Gram-Modell (Mikolov et al., 2013a,b), welches statt Worteinbettungen<sup>7</sup> (engl. *word embeddings*) Phraseneinbettungen lernte, und setzten dieses mit zwei gestapelten CRFs um. Das erste CRF analysierte eine Eingabe und sagte eine semantische Kategorie vorher. Phraseneinbettungen wurden im zweiten CRF eingesetzt, welches die Vorhersagen des ersten CRFs und diese Phraseneinbettungen bearbeitete. Die Autoren verglichen den Einfluss von Skip-Gramm-Modellen (Mikolov et al., 2013a,b), Browns Clustern (Brown et al., 1992), Phraseneinbettungen und externen Quellen auf die Performanz des Modells. Die Kombination von externen Quellen und Phraseneinbettungen erzielte das beste Ergebnis, also einen  $F_1$ -Wert von 90,9 %.

<sup>3</sup><https://nlp.stanford.edu/software/CRF-NER.shtml>

<sup>4</sup>Der Tagger ist ein System für die automatische Annotation von beliebigen Kategorien.

<sup>5</sup>Als N-Gramme bezeichnet man eine Folge von  $N$  Elementen eines zerlegten Textes. Die Elemente können Wörter, Buchstaben o. Ä. sein.

<sup>6</sup><https://github.com/tudarmstadt-lt/GermaNER>

<sup>7</sup>Worteinbettungen stellen Wörter in einem  $n$ -dimensionalen Raum als reelle Zahlen dar. Dadurch werden diese Wörter im Kontext anderer Wörter als Vektoren mathematisch zusammengefasst.

Wie sich aus den beschriebenen Ansätzen resümieren lässt, basiert die gute Performanz der CRFs auf manuell generierten Features, externen Quellen und semantischen Clustern. Deswegen sind konstruierte Modelle stark sprach- und domänen spezifisch, sodass neue Features gebraucht werden, um neue Modelle an eine andere Sprache sowie Domäne anzupassen. Die Anpassung ist jedoch sehr kosten- und zeitaufwändig. Diese Tatsache hatte eine Wende zur Folge. Collobert et al. (2011) schlugen ein neues universelles Modell für das Part-Of-Speech-Tagging<sup>8</sup>, die Eigennamenerkennung, das Chunking<sup>9</sup> und die Annotation semantischer Rollen vor, das auf Neuronalen Netzen beruhte. Das Modell profitierte von Worteinbettungen, die von großen unannotierten Datenmengen gelernt wurden. Dank diesem Verfahren wurde das entwickelte Modell sprach- und domänenunabhängig.

Heutzutage bedient man sich bei der automatischen Annotation von Eigennamen Varianten des Convolutional-Neural-Networks (CNN) und Long-Short-Term-Memory-Netzes (Huang et al., 2015; Chiu und Nichols, 2016; Lample et al., 2016; Ma und Hovy, 2016), die entweder mit einem CRF oder untereinander kombiniert werden. Dies ermöglicht es, die Vorteile verschiedener Modellen zu vereinen und dementsprechend die Ergebnisse zu verbessern.

Huang et al. (2015) verglichen vier Modelle für die automatische Annotation von Sequenzen (POS-Tagging, Eigennamenerkennung, Chunking). Das waren Variationen und Kombinationen vom LSTM, BLSTM und CRF: ein LSTM, BLSTM, LSTM-CRF und BLSTM-CRF. Insgesamt zeigten das BLSTM und BLSTM-CRF eine bessere Performanz im Vergleich zum LSTM und LSTM-CRF. Um die Ergebnisse in der Eigennamenerkennung zu verbessern, wurden in das BLSTM-CRF Colloberts Worteinbettungen SENNA<sup>10</sup> und externe Quellen integriert. Der F<sub>1</sub>-Wert stieg so auf 90,1 % an.

Chiu und Nichols (2016) beschrieben ein neues Modell für die Eigennamenerkennung, das Informationen aus Wörtern und Buchstaben dank einer Kombination des BLSTMs und des CNNs automatisch entnehmen konnte. Das CNN bearbeitete dabei die Buchstaben eines Wortes und übergab das Resultat an das BLSTM, welches die Wörter im Kontext eines Satzes analysierte. Dabei testeten die Autoren Colloberts Worteinbettungen sowie Features aus Lexika und Großschreibung als ein weiteres Feature. Die Variante mit Worteinbettungen und Features aus dem SENNA- und DBPedia-Lexikon erzielte das beste Ergebnis für das englische Korpus CoNLL 2003, also einen F<sub>1</sub>-Wert von 91,62 %.

Lample et al. (2016) erarbeiteten auch ein Modell basierend auf der Kombination vom BLSTM und CRF. Zusätzlich zu Worteinbettungen wurden Buchstabeneinbettungen verwendet. Das Modell wurde auf dem englischen, deutschen, spanischen und niederländischen Testkorpus aus CoNLL 2002 und 2003 evaluiert. Der F<sub>1</sub>-Wert für das Englische war 90,94 % und der für das Deutsche 78,76 %. Im Vergleich zu dem BLSTM-CRF von Huang et al. war die Performanz des Modells von Lample et al. um 0,84 % besser, was die Bedeutung der Information von Buchstabeneinbettungen untermauerte.

Ma und Hovy (2016) kombinierten das BLSTM, CNN und CRF in einem Modell, um wiederum Informationen aus Wörtern und Buchstaben zu verwenden. Die Architektur war dem Ansatz in Chiu und Nichols (2016) ähnlich. Der Unterschied bestand darin, dass ein CRF im letzten Schritt eingesetzt wurde. Das Modell erhielt einen F<sub>1</sub>-Wert von 91,21 % evaluiert auf dem englischen Testkorpus CoNLL 2003.

Model	Autoren	F <sub>1</sub> -Wert	
		Englisch	Deutsch
<b>Conditional-Random-Fields</b>			
CRF StanfordNER	Finkel et al. (2005)	86,86 %	
	Faruqui und Padó (2010)		78,2 %
CRF	Tkachenko und Simanovsky (2012)	91,02 %	
CRF	Passos et al. (2014)	90,9 %	
CRF GermaNER	Benikova et al. (2015)		<b>79,37 %★</b>
<b>Long-Short-Term-Memory-Netze</b>		Englisch	Deutsch
BLSTM-CRF	Huang et al. (2015)	90,1 %	
BLSTM-CNN	Chiu und Nichols (2016)	<b>91,62 %</b>	
BLSTM-CRF	Lample et al. (2016)	90,94 %	78,76 %
BLSTM-CNN-CRF	Ma und Hovy (2016)	91,21 %	

**Tabelle 2.1.** F<sub>1</sub>-Werte von Modellen für die Eigennamenerkennung, die auf dem englischen und deutschen Testkorpus CoNLL 2003 evaluiert wurden. Das mit ★ gekennzeichnete Ergebnis ist aus Riedl und Padó (2018) entnommen.

Um die Performanz der beschriebenen Ansätze zu vergleichen, sind die Ergebnisse in Tabelle 2.1 zusammengefasst. Das beste Modell für das Englische ist das BLSTM-CNN. Die Werte anderer Modelle liegen nahe

<sup>8</sup>Unter Part-Of-Speech-Tagging (POS-Tagging) versteht man die automatische morphosyntaktische Annotation, somit das Zuordnen von Wörtern und Satzzeichen zu deren Wortarten.

<sup>9</sup>Unter Chunking versteht man die automatische syntaktische Annotation.

<sup>10</sup><https://ronan.collobert.com/senna/>

beieinander und unterscheiden sich lediglich um 1–1,5 %. Eine Ausnahme bildet StanfordNER mit einem um ca. 5 % niedrigeren  $F_1$ -Wert von 86,86 %. Für die deutsche Sprache sind die Ergebnisse aus den drei Ansätzen verfügbar. Die beste Performanz hat GermaNER erzielt. Die Werte des StanfordNERS und des BLSTM-CRFs unterscheiden sich um ca. 1 %. Die beschriebenen Ansätze haben bestätigt, dass CRFs und Kombinationen von BLSTMs die besten Kandidaten sind, um die Eigennamenerkennung in einer neuen Domäne zu testen. Die Tatsache, dass auch für die deutsche Sprache das CRF und das BLSTM zwei Modelle sind, die sehr gute Ergebnisse erzielen, bekräftigen u. a. Riedl und Padó (2018).

### 2.3 Ansätze der Eigennamenerkennung in Rechtstexten

Die Eigennamenerkennung in Rechtstexten ist trotz ihrer hohen Relevanz heutzutage wenig erforscht. Die wissenschaftliche Literatur, die dem betroffenen Thema gewidmet ist, ist hinsichtlich der untersuchten Schwerpunkte, angewandten Methoden und Techniken sowie der entwickelten Klassifikationen und Korpora uneinheitlich. Dadurch ist es unmöglich, ihre Ergebnisse adäquat zu vergleichen. Jedoch leisten erarbeitete Ansätze einen wichtigen Beitrag zur Eigennamenerkennung in Rechtstexten und bilden eine Grundlage für die weitere Forschung.

Die erste Arbeit, in der die Eigennamenerkennung in der juristischen Domäne als ein Begriff explizit zu stande kam, wurde in Dozier et al. (2010) beschrieben. Die Autoren untersuchten die Eigennamenerkennung in amerikanischen Rechtsprechungen, Aussagen unter Eid, Plädoyers u. Ä., die mithilfe der Lookup-Methode basierend auf der Liste der Eigennamen, der kontextuellen Regeln und der statistischen Modelle umgesetzt wurde. Dabei erarbeiteten die Autoren die Tagger für die Identifikation der Zuständigkeitsbereiche (aufgrund einer zuständigen Behörde), der Gerichte, der Richter, des Titels und der Kategorie des jeweiligen Dokumentes. Das beste Ergebnis hat der Zuständigkeitsbereich-Tagger mit einem  $F_1$ -Wert von 92 % erzielt. Die Werte der anderen vier Tagger waren bei ca. 82–85 %.

Cardellino et al. (2017) entwickelten ein Tool für die Erkennung, Klassifizierung und Verlinkung von Eigennamen, wobei für das Training YAGO- und LKIF-Ontologien angepasst wurden und infolgedessen verschie denkörnige Klassen entstanden, die ausgewertet werden konnten. Insgesamt ging es um vier Abstufungen: NER – 2 Klassen, NERC – 6 Klassen, LKIF – 69 Klassen und YAGO – 358 Klassen. Als Systeme wurden eine Support-Vector-Machine, StanfordNER von Finkel et al. (2005) und ein Neuronales Netz verwendet, die auf Wikipedia-Artikeln und Entscheidungen des Europäischen Gerichtes für Menschenrechte trainiert und evaluiert wurden. Für das Neuronale Netz entwarfen die Autoren zusätzlich Worteinbettungen. Das beste Ergebnis für das Korpus, das aus Wikipedia-Artikeln bestand, hat das Neuronale Netz erzielt. Die  $F_1$ -Werte für die NERC- und YAGO-Klassen entsprachen 86 % und 69 %. Für die LKIF-Klassen war StanfordNER mit einem  $F_1$ -Wert von 77 % jedoch besser. Die Performanz der Modelle war auf dem anderen Korpus, das aus Entscheidungen des Europäischen Gerichtes für Menschenrechte bestand, deutlich schlechter. Die  $F_1$ -Werte variierten je nach Modell und je nach Grad der Abstufung. StanfordNER konnte einen maximalen  $F_1$ -Wert von 56 % mit sechs Klassen erzielen.

Glaser et al. (2018) testeten drei Systeme zur Eigennamenerkennung. Das erste System war GermaNER von Benikova et al. (2015), mit dem Personen-, Orts-, Organisationsnamen und Namen anderer Herkunft erkannt wurden. Neben GermaNER wurden weitere Ansätze verwendet, um zusätzliche Entitäten zu finden. Numerische Ausdrücke für Zeit und Währungen wurden mit regelbasierten Ansätzen erkannt und Referenzen mit dem in Landthaler et al. (2016) beschriebenen Ansatz. Das zweite System war DBpedia-Spotlight<sup>11</sup> (Mendes et al., 2011; Daiber et al., 2013). Das Tool wurde für die automatische Annotation der für DBpedia typischen Entitäten entwickelt. Das dritte System, Templatized genannt, wurde von Glaser et al. entworfen. Es konzentrierte sich auf die Eigennamenerkennung in Verträgen, die mithilfe von Vorlagen erzeugt wurden. Dies erlaubte es, Eigennamen wie z. B. Hersteller-, Produktnamen mit regulären Ausdrücken zu extrahieren und zu klassifizieren. Für GermaNER und DBpedia-Spotlight wurde ein manuell annotiertes Korpus erstellt, das aus 500 Entscheidungen des 8. Zivilsenats des Bundesgerichtshofes bestand und Bezug auf Mietrecht hatte. GermaNER sowie DBpedia-Spotlight wurden auf 20 Entscheidungen aus dem erstellten Korpus evaluiert und Templatized wurde auf fünf verschiedenenartigen Verträgen evaluiert. GermaNER und DBpedia-Spotlight erzielten einen  $F_1$ -Wert von 80 % und 87 %. Das Ergebnis der Eigennamenerkennung in Verträgen betrug 92 %.

Hinsichtlich der beschriebenen Ansätze stellt sich zudem die Frage nach einer Klassifikation von Eigennamen in der juristischen Domäne. Neben den typischen Klassen der Personen, Organisationen, Orte und des Verschiedenen, die als Standard in der Eigennamenerkennung gelten, gibt es weitere domänspezifische

<sup>11</sup><https://www.dbpedia-spotlight.org/>

Klassen. In der wissenschaftlichen Literatur, die das betroffene Thema beleuchtet, wird aber wenig Aufmerksamkeit auf die Klassifikation bzw. die Beschreibung einzelner Klassen gelenkt. Dozier et al. (2010) konzentrierten sich auf juristische Namen, nämlich die der Personen (der Richter und der Rechtsanwälte), der Organisationen (der Gerichte) und der Orte, die bestimmte Zuständigkeitsbereiche widerspiegeln. Cardellino et al. (2017) erweiterten zwar die bekannten Klassen der Personen- und Organisationsnamen um Klassen der Akten-, Dokumenten- und Abstraktionsnamen (entspricht der zweiten Abstufung – NERC), es blieb aber unklar, was zu diesen Klassen gehört und wie sie voneinander abgegrenzt wurden. Glaser et al. (2018) fügten hinsichtlich der Rechtstexte die Klasse der Referenz hinzu, die sich nach dem Grad der Ex- und Implizitheit unterschied und vier Typen beinhaltete (Landthaler et al., 2016). Jedoch wurde darunter die Referenz auf Rechtsnormen verstanden, sodass weitere Referenzen (auf Entscheidungen, Vorschriften, Rechtsliteratur etc.) nicht abgedeckt waren.

Die Forschung der Eigennamenerkennung in Rechtstexten wird auch dadurch erschwert, dass keine frei verfügbaren Korpora existieren, weder für das Englische, noch für das Deutsche. Korpora für Zeitungstexte, die im Rahmen von CoNNL 2003 oder GermEval 2014 entwickelt wurden, eignen sich hinsichtlich der Textsorte und der annotierten Klassen wiederum nicht. In diesem Zusammenhang ist der Bedarf eines manuell annotierten Korpus bestehend aus Rechtstexten enorm, welches die Erarbeitung einer Klassifikation juristischer Entitäten und von einheitlichen Annotationsrichtlinien verlangt. Das juristische Korpus ermöglicht es, die Eigennamenerkennung mit aktuellen Verfahren, nämlich mit CRFs und BLSTMs, umzusetzen und ihre Performanz zu analysieren.

## 3 Formale Definition der Sequenzmodelle

Dieses Kapitel leitet in den Begriff der automatischen Annotation mithilfe von Sequenzmodellen ein und beschreibt den Prozess der Eigennamenerkennung aus technischer Perspektive. Danach wird auf die formalen Definitionen von aktuellen Sequenzmodellen eingegangen, die generell für die automatische Annotation und insbesondere für die Eigennamenerkennung verwendet werden. Wie bereits in Kapitel 2 beschrieben, sind dies CRF, LSTM und deren Kombination, bekannt als LSTM-CRF.

### 3.1 Begriffsdefinition und Aufgabenbeschreibung

Automatische Verfahren (einschließlich jener zur Eigennamenerkennung) bearbeiten Daten sequenziell, also Wort für Wort, und ordnen jedem Glied einer Folge ein vordefiniertes Label zu, das einer Kategorie entspricht. Diese Verfahren nennt man Sequenzmodelle (engl. *sequence labeling models*). Sei ein Minimalbeispiel der Satz ‚Angela Merkel besucht Dresden‘ (s. Abbildung 3.1). Die Identifikation und Klassifikation erfolgt mithilfe automatischer Verfahren, nämlich eines Modells, welches den eingegebenen Satz bearbeitet und Labels für jedes Wort ausgibt. Im dargestellten Satz kommen zwei Eigennamen vor. Das ist ein Personename ‚Angela Merkel‘ und ein Ortsname ‚Dresden‘, die zu den Klassen **PER** und **LOC** gehören. Diese Klassen bekommen die zusätzliche Information über die Anfangs-, Innen- und Außenposition eines Wortes (im Englischen als *begin*, *inside*, *out* bezeichnet), die nach IOB2-Schema durch die Zeichen **B**, **I**, **O** kodiert ist (für weitere Schemata s. Sang und Veenstra, 1999). So wird der Personename ‚Angela Merkel‘, der mit ‚Angela‘ beginnt und weiter mit ‚Merkel‘ fortgesetzt wird, mit den Labels **B-PER** und **I-PER** assoziiert. Der Ortsname ‚Dresden‘ besteht aus einem Wort und wird als **B-LOC** gekennzeichnet. Ist ein Wort kein Eigename, wird diesem das Label **O** zugewiesen, so wie mit dem Verb ‚besuchen‘.

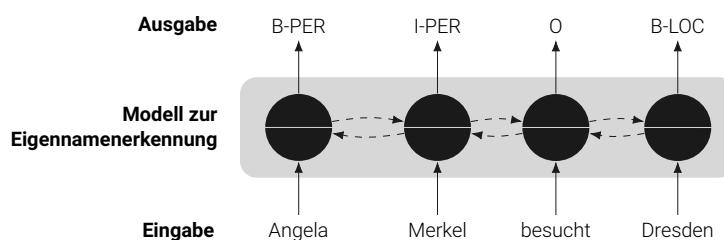


Abbildung 3.1. Schematische Darstellung der Eigennamenerkennung.

Um Sätze automatisch zu kategorisieren, lernt ein Modell Relationen zwischen Wörtern und ihren Labels mithilfe eines manuell annotierten Korpus, das aus diesen Paaren besteht. Diese Methode nennt man überwachtes Lernen (engl. *supervised learning*). Im Lernprozess wird ein Modell so modifiziert, dass es eine möglichst korrekte Vorhersage für jedes Wort trifft.

### 3.2 Formale Definition von CRFs

Conditional-Random-Fields gehören zu den diskriminativen Sequenzmodellen und basieren auf der Idee ungerichteter Graphen, die in Lafferty et al. (2001) beschrieben wurde. CRFs berechnen die bedingte Wahrscheinlichkeit  $p(\bar{y}|\bar{x})$  der möglichen Ausgabesequenz  $\bar{y} = (y_1, \dots, y_n)$ , die vorhergesagt wird, und der gegebenen Eingabesequenz  $\bar{x} = (x_1, \dots, x_n)$ .

Formal ist ein CRF wie folgt definiert: Sei  $\mathcal{X}$  ein Eingabealphabet,  $\mathcal{Y}$  ein Ausgabealphabet, wobei  $\bar{x} \in \mathcal{X}^n$  und  $\bar{y} \in \mathcal{Y}^n$  mit  $n \in \mathbb{N}$  eine Eingabe- und eine Ausgabesequenz sei. Sei  $\vec{\theta} = \{\theta_k\}_{k=1}^K \in \mathbb{R}^K$  ein Parametervektor und  $\mathcal{F} = \{f_k(y_{i-1}, y_i, \bar{x}, i) | f_k \in \mathbb{R}\}_{k=1}^K$  eine Menge von Gewichtsfunktionen mit  $k \in \mathbb{N}$ . Die bedingte Wahrscheinlichkeit  $p(\bar{y}|\bar{x})$  setzt sich aus einem normalisierten Gewicht der Ausgabesequenz  $\bar{y}$  bedingt einer Eingabesequenz  $\bar{x}$  zusammen, wie es in der Formel (3.1) dargestellt wird. Dabei berechnet die

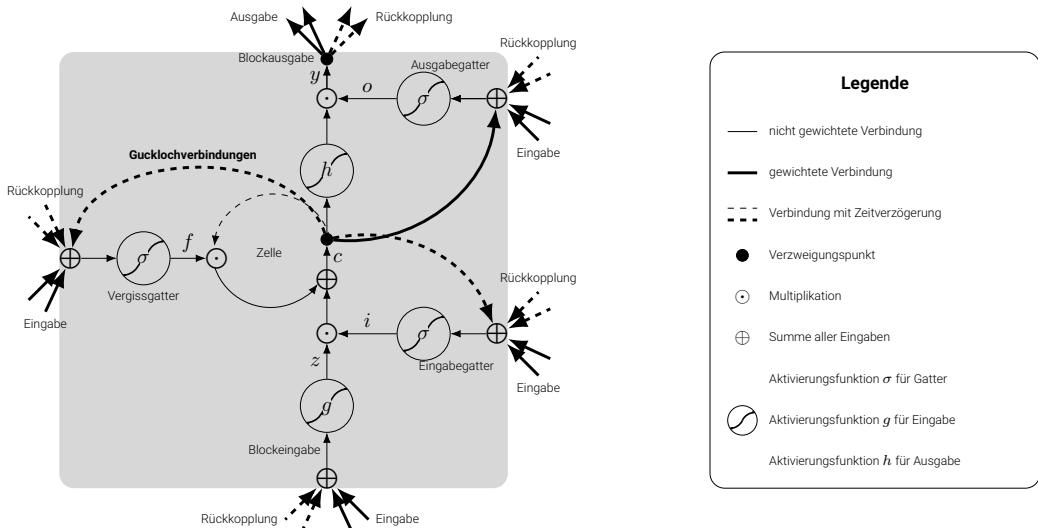
Normalisierungskonstante  $Z$  in der Formel (3.2) die Summe der Gewichte aller möglichen Ausgabesequenzen über dem Alphabet  $\mathcal{Y}$  für eine Eingabesequenz  $\bar{x}$ .

$$p(\bar{y}|\bar{x}) = \frac{1}{Z_{\vec{\theta}}(\bar{x})} \prod_{i=1}^{|\bar{y}|} \exp \left( \vec{\theta} \cdot \mathcal{F}(y_{i-1}, y_i, \bar{x}, i) \right) \quad (3.1)$$

$$Z_{\vec{\theta}}(\bar{x}) = \sum_{\bar{y}' \in \mathcal{Y}^{|\bar{x}|}} \prod_{i=1}^{|\bar{y}'|} \exp \left( \vec{\theta} \cdot \mathcal{F}(y_{i-1}, y_i, \bar{x}, i) \right) \quad (3.2)$$

### 3.3 Formale Definition von LSTMs

Long-Short-Term-Memory-Netze (dt. Langes Kurzzeitgedächtnis) wurden als eine Variante des Rekurrenten Neuronalen Netzes (RNN) entwickelt (Hochreiter und Schmidhuber, 1997), um das Problem des verschwindenden bzw. explodierenden Gradienten zu lösen<sup>1</sup>. Wenn das RNN aus einer Eingabeschicht  $x$ , einer „versteckten“ Verarbeitungsschicht  $h$  und einer Ausgabeschicht  $y$  besteht, hat das LSTM-Netz statt einer Verarbeitungsschicht eine komplexe Speicherzelle. Die Idee dieser neuen Netzwerkschicht besteht darin, den Zugriff auf Informationen aus der Vergangenheit von jedem Zeitpunkt aus zu ermöglichen. Sie ist wie in Abbildung 3.2 dargestellt aufgebaut, wobei es um eine meist bekannte Variante des LSTM-Netzes geht, die in Greff et al. (2017) als Vanilla-LSTM beschrieben wurde (für weitere Informationen s. auch Gers et al., 2000; Gers und Schmidhuber, 2000; Graves und Schmidhuber, 2005). Die Speicherzelle, auch LSTM-Block genannt, setzt sich aus dem Eingabegatter  $i$ , dem Ausgabegatter  $o$ , dem Vergissgatter (*forget gate*)  $f$ , der Blockeingabe  $z$ , der Blockausgabe  $y$ , der Zelle  $c$ , den Aktivierungsfunktionen  $\sigma$ ,  $g$ ,  $h$  und den Gucklöchern (engl. *peepholes*) zusammen.



**Abbildung 3.2.** Aufbau der LSTM-Speicherzelle (Quelle: Überarbeitete Darstellung in Anlehnung an Greff et al., 2017, S. 2)

Der LSTM-Block ist wie folgt formal definiert: Sei  $x^t$  ein Eingabevektor zum Zeitpunkt  $t$ ,  $N$  eine Anzahl der LSTM-Blöcke und  $M$  eine Anzahl der Eingaben. Eine LSTM-Schicht besteht aus den Eingabegewichten  $W_z, W_i, W_f, W_o \in \mathbb{R}^{N \times M}$ , den Rückkopplungsgewichten  $R_z, R_i, R_f, R_o \in \mathbb{R}^{N \times M}$ , den Gucklochgewichten  $p_i, p_f, p_o \in \mathbb{R}^N$  und den Biasgewichten  $b_z, b_i, b_f, b_o \in \mathbb{R}^N$ . Die Gewichte für die Gatter  $i, o, f$ , die

<sup>1</sup>Rekurrente Neuronale Netze werden mit dem Backpropagation-Lernalgorithmus trainiert, der das Gradientenverfahren basierend auf mittleren quadratischen Abweichungen einsetzt. Im Training werden Gewichte in den Schichten so geändert, dass die Möglichkeit der falschen Ausgabesequenz, die vorherzusagen ist, reduziert wird. In anderen Worten wird der Fehler im Gradientenverfahren minimiert, wobei der Gradient aus dem Produkt der partiellen Ableitungen berechnet wird. In mehrschichtigen Netzen führt dies zu inadäquaten Aktualisierungen der Schichten, die weiter von der Ausgabeschicht liegen. Wenn Werte stets kleiner als eins sind, wird das betroffene Gewicht lediglich gering aktualisiert, weil der Fehlerwert in jeder Multiplikation kleiner wird (das Problem des verschwindenden Gradienten, engl. *vanishing gradient problem*). Wenn Werte stets größer als eins sind, wird das Gewicht dagegen zu stark geändert (das Problem des explodierenden Gradienten, engl. *exploding gradient problem*). Siehe Hochreiter (1991); Bengio et al. (1994) für weitere Bemerkungen.

Blockeingabe  $z$ , die Blockausgabe  $y$  und die Zelle  $c$  werden wie folgt berechnet:

$$z^t = g(W_z x^t + R_z y^{t-1} + b_z) \quad (3.3)$$

$$i^t = \sigma(W_i x^t + R_i y^{t-1} + p_i \odot c^{t-1} + b_i) \quad (3.4)$$

$$f^t = \sigma(W_f x^t + R_f y^{t-1} + p_f \odot c^{t-1} + b_f) \quad (3.5)$$

$$c^t = z^t \odot i^t + c^{t-1} \odot f^t \quad (3.6)$$

$$o^t = \sigma(W_o x^t + R_o y^{t-1} + p_o \odot c^t + b_o) \quad (3.7)$$

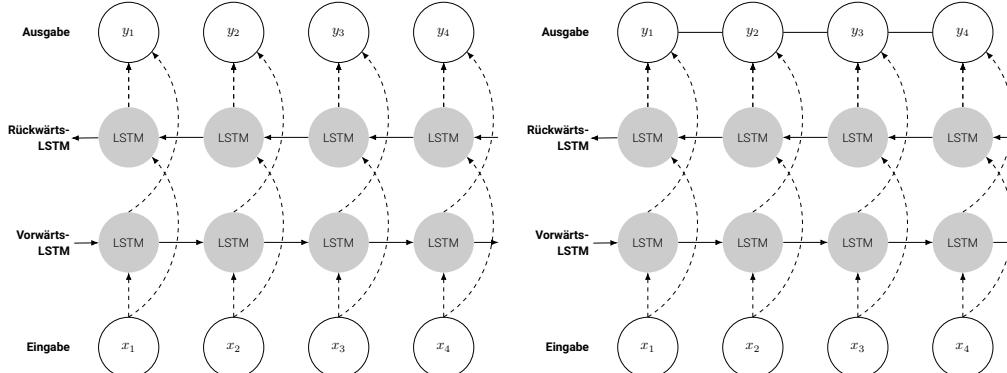
$$y^t = h(c_t) \odot o^t \quad (3.8)$$

Die Aktivierungsfunktion für das Gatter  $\sigma$  wird durch die logistische Funktion und die Aktivierungsfunktionen für die Blockeingabe  $g$  und die Blockausgabe  $h$  durch die Hyperbeltangens-Funktion berechnet:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3.9)$$

$$g(x) = h(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3.10)$$

Auf Grund der Speicherzelle kann das LSTM Informationen aus einem weiten Kontext bearbeiten und lernen. Dabei geht es aber um einen linksseitigen Kontext, der in der Vergangenheit liegt. Um auf Informationen aus der Zukunft zugreifen zu können, wurde das bidirektionale LSTM entwickelt. Es besteht aus zwei Verarbeitungsschichten  $h = \{h_l, h_r\}$  (Graves und Schmidhuber, 2005), die jede Eingabesequenz vorwärts und rückwärts betrachten. Die Vorwärtsverarbeitungsschicht  $h_l$  bearbeitet vergangene Informationen aus einem linksseitigen Kontext und die Rückwärtsverarbeitungsschicht  $h_r$  zukünftige Informationen aus einem rechtsseitigen Kontext.



**Abbildung 3.3.** BLSTM-Modell (links) und BLSTM-CRF-Modell (rechts).

### 3.4 Formale Definition von LSTM-CRFs

LSTMs sowie BLSTMs machen unabhängige Vorhersagen für jede Ausgabe  $y_i$  der Ausgabesequenz  $\bar{y} = (y_1, \dots, y_n)$  aufgrund  $h_i$ . Wenn jedoch zwischen Ausgaben Abhängigkeiten bestehen, reicht es nicht, diese als unabhängig zu betrachten (Lample et al., 2016). Um das Problem zu umgehen, entwickelte man LSTM-CRFs als Modelle. Hierbei wird ein CRF als obere Schicht eingesetzt. In Abbildung 3.3 sind das BLSTM-Modell und das BLSTM-CRF-Modell schematisch dargestellt. Zu beachten ist die Relation zwischen den Ausgaben. Vergleichbare LSTM-Modelle kommen ohne Rückwärtsschicht aus. Die Idee hinter dem LSTM-CRF besteht darin, dass das CRF Gewichte aus einem LSTM nimmt und darauf basierend eine bestmögliche Ausgabesequenz berechnet (Huang et al., 2015; Ma und Hovy, 2016; Lample et al., 2016).

Formal ist das LSTM-CRF nach Lample et al. (2016) wie folgt definiert: Sei  $\mathcal{P}^{n \times k}$  eine vom LSTM ausgegebene Gewichtsmatrix für die Eingabesequenz  $\bar{x} = (x_1, \dots, x_n)$  der Länge  $n$ , wobei  $k$  die Anzahl der Labels ist. Dann entspricht das Element  $\mathcal{P}_{i,j}$  dem Gewicht eines Labels  $j$  für das Wort  $i$  im betroffenen Satz. Sei  $\mathcal{A}^{k+2 \times k+2}$  eine Gewichtsmatrix, die die Übergangsgewichte von einem Zustand  $i$  zu einem anderen Zustand  $j$  widerspiegelt. Die Menge der möglichen Zustände setzt sich aus Labels, dem Anfangs-  $y_{BOS}$  und

dem Endzustand  $y_{EOS}$  zusammen, sodass die Anzahl der Zustände  $k + 2$  ist. Die Gewichte  $s(\bar{x}, \bar{y})$  für die Ausgabesequenz  $\bar{y} = (y_1, \dots, y_n)$  lassen sich wie folgt berechnen:

$$s(\bar{x}, \bar{y}) = \sum_{i=0}^n \mathcal{A}_{y_i, y_{i+1}} + \sum_{i=1}^n \mathcal{P}_{i, y_i} \quad (3.11)$$

Die bedingte Wahrscheinlichkeit  $p(\bar{y}|\bar{x})$  setzt sich wiederum aus einem normalisierten Gewicht der Ausgabesequenz bedingt einer Eingabesequenz zusammen (vgl. mit der Formel (3.1)). Jedoch bedient man sich hierbei der Exponentialfunktion. Im Exponenten im Zähler befindet sich das Gewicht einer betroffenen Ausgabesequenz  $\bar{y}$  für eine Eingabesequenz  $\bar{x}$ . Der andere Exponent im Nenner entsteht aus der Summe der Gewichte aller möglichen Ausgabesequenzen über dem Alphabet  $\mathcal{Y}$  für eine Eingabesequenz  $\bar{x}$ .

$$p(\bar{y}|\bar{x}) = \frac{e^{s(\bar{x}, \bar{y})}}{\sum_{\bar{y}' \in \mathcal{Y}^{|\bar{x}|}} e^{s(\bar{x}, \bar{y}')}} \quad (3.12)$$

Die Architektur der BLSTM-CRFs ist den LSTM-CRFs ähnlich (Huang et al., 2015; Lample et al., 2016; Ma und Hovy, 2016). Das CRF als obere Schicht nimmt in diesem Fall Gewichte aus der Vorwärtsverarbeitungsschicht  $h_l$  und Rückwärtsverarbeitungsschicht  $h_r$ , die die Informationen aus der Vergangenheit und Zukunft bearbeitet haben.

## 4 Juristisches Korpus

Die Sequenzmodelle zur automatischen Annotation von Texten sind wie in Kapitel 3 erläutert überwacht lernende Modelle, die mithilfe des annotierten Korpus Relationen zwischen Wörtern und Labels lernen. Das Problem der Forschung für die Eigennamenerkennung in der juristischen Domäne besteht darin, dass es keine frei verfügbaren annotierten Korpora für Rechtstexte gibt. Im Einklang mit dem Forschungsziel der vorliegenden Arbeit wird ein juristisches Korpus entwickelt. So wird in diesem Kapitel seine Konstruktion behandelt, welche für die Auswahl von gerichtlichen Entscheidungen als Quelltexte argumentiert und die Erarbeitung der Klassen von typischen Entitäten für diese Rechtstexte beschreibt. Schließlich wird über die Korpusgröße, die Anzahl und Verteilung der betroffenen Entitäten sowie über das bereitgestellte Datenformat berichtet.

### 4.1 Rechtstexte als Fachtexte und ihre Besonderheiten

Unter Rechtstexten versteht man Fachtexte, die dem Bereich Recht zugeordnet sind (Simonnæs, 2015). Diese Fachtexte werden von einer zuständigen Institution oder von einem berufstätigen Juristen verfasst (Engberg, 1993) und sind sowohl auf Rechtsexperten als auch auf Bürger bezogen (Sandrini, 1999). Rechtstexte sind durch nationale Rechtsordnungen oder nationale Rechtssysteme geprägt (Wiesmann, 2003), die u. a. Textaufbau, Formulierungsstil, Zitierweise sowie Themenentfaltung bestimmen. Zu den Rechtstexten zählt man beispielsweise Gesetzestexte, Rechtsprechungen, Verwaltungsvorschriften oder Verträge (Kjær, 1992; Engberg, 1993; Busse, 2000; Deutsch, 2017), die im Gegensatz zu Texten aus der Allgemeinsprache durch den Gebrauch juristischer Fachtermini, Nominalisierungen, Passivkonstruktionen, Funktionsverbgefüge, rechts- bzw. linksseitig erweiterter Attribuierung gekennzeichnet sind (vgl. Sander, 2004; Simonnaes, 2012; Simonnæs, 2018).

Rechtstexte sind sowohl nach textinternen als auch nach textexternen Kriterien sehr unterschiedlich, was ihre eindeutige Klassifikation und einheitliche Aufteilung erschwert (Kjær, 1992; Engberg, 1993; Busse, 2000; Deutsch, 2017). Die Heterogenität manifestiert sich wiederum in der sprachlichen Gestaltung verschiedener Rechtstexte. Der Gebrauch von Eigennamen und Zitaten<sup>1</sup> ist hierbei auch betroffen, sodass die Art und die Vielzahl der Entitäten je nach Textsorte deutlich variiert. Daraus stellt sich die Frage nach einer angemessenen Textsorte für das juristische Korpus. Einerseits müssen Quelltexte einer ausgewählten Textsorte der Zielsetzung der Forschungsarbeit entsprechen, indem sie genug verschiedenartige Eigennamen und Zitate enthalten, die automatisch zu erkennen und zu klassifizieren sind. Andererseits müssen die Quelltexte frei zugänglich sein. Wenn man Rechtstexte wie z. B. Gesetze, Rechtsprechungen oder Verwaltungsvorschriften aus elektronischen Datenbanken vergleicht, bieten Rechtsprechungen die beste Möglichkeit. In Gesetzen und Verwaltungsvorschriften ist der Anteil an Personen-, Orts- und Organisationsnamen sehr gering. Rechtsprechungen, die eine eigene Textsorte unter Rechtstexten bilden, schließen dagegen Personen-, Orts-, Organisationsnamen, Verweise auf Gesetze, Rechtsprechungen, Vorschriften u. Ä. ein.

### 4.2 Semantische Klassen juristischer Entitäten

Rechtstexte unterscheiden sich, wie oben kurz angedeutet, sowohl von Texten in anderen Domänen als auch voneinander hinsichtlich textinterner und -externer Merkmale enorm, was einen fundamentalen Einfluss auf die sprachliche und thematische Gestaltung, Zitierweise, Gliederung etc. hat. Dies betrifft auch juristische Entitäten, die in verschiedenen Rechtstexten zum Ausdruck kommen. Typische Eigennamen wie Personen-, Orts- oder Organisationsnamen kommen in Entscheidungen regelmäßig vor, dies betrifft insbesondere den Tatbestand. Sie haben auch Bezüge auf nationale Gesetze, Rechtsverordnungen und auf über nationale Rechtsnormen. Zudem sind Entscheidungen durch Verweise auf andere Entscheidungen gekennzeichnet. Diese Art des Verweises ist z. B. in Gesetzestexten nicht zu finden. Aus diesen Besonderheiten kristallisieren sich zwei Anforderungen an eine Typologie juristischer Entitäten heraus. Erstens müssen entwickelte

<sup>1</sup> Unter dem Begriff „Zitat“ in Rechtstexten wird keine wörtlich oder sinngemäß zitierte Textstelle, sondern der Verweis auf eine Textstelle verstanden. Das Zitat als Argument und auch als Hinweis dient zur Feststellung des Wortlauts eines Rechtstextes, das mittels der zitierten Angaben (z. B. mittels der Angaben des Zitiernamens, der Fundstelle, des Aktenzeichens o. Ä.) gefunden werden kann.

Klassen diejenigen Entitäten widerspiegeln, die für Entscheidungen typisch sind. Zweitens muss eine Klassifikation die Entitäten betreffen, deren Differenzierung in Entscheidungen von hoher Relevanz ist.

Juristische Entitäten wie Gesetze, Rechtsprechungen, Verträge usw., die für gerichtliche Entscheidungen charakteristisch sind, lassen sich in zwei grundlegende Gruppen unterteilen, nämlich in Bezeichnungen und Verweise. Für Rechtsnormen, also für Gesetze und Rechtsverordnungen, sind Bezeichnungen (sogenannte Namen) Überschriften zu ihren Normtexten, die Auskunft über Rang und Inhalt geben (Bundesministerium der Justiz, 2008, Rn. 321 ff.). Überschriften haben eine einheitliche Form und bestehen in der Regel aus einer Bezeichnung, Kurzbezeichnung und Abkürzung. So lautet z.B. die Überschrift zu der Polstererausbildungsverordnung vom 20. Mai 2014 (BGBl. I S. 539), Verordnung über die Berufsausbildung zum Polsterer und zur Polsterin (Polstererausbildungsverordnung – PolstAusbV). Die Kurzbezeichnung und die Abkürzung stehen in Klammern und sind mit einem Bindestrich getrennt. Die Zitation der Rechtsnormen wird auch fest geregelt, wobei es verschiedene Zitierregeln für Voll- und Kurzzitate gibt (Bundesministerium der Justiz, 2008, Rn. 168 ff.). Die Bezeichnungs- sowie Zitierweise verbindlicher Einzelakte wie Vorschriften oder Verträge ist dagegen nicht einheitlich festgelegt. Insgesamt sind Entscheidungen durch Verweise auf Rechtsnormen, Einzelfallregelungen, Rechtsprechungen und Rechtsliteratur (Rechtskommentare, Gesetzgebungsmaterialien, Lehrbücher, Monographien) gekennzeichnet. Aus dieser Hinsicht wird die Eigennamenerkennung in Rechtstexten und insbesondere in Entscheidungen so umdefiniert, dass man darunter die Identifikation und Klassifikation sowohl der Namen als auch der Zitate versteht.

Grobkörnige Klassen		Feinkörnige Klassen	
Nr.		Nr.	
1	<b>PER</b>	1	Personen
		2	Richter
		3	Anwälte
2	<b>LOC</b>	4	Orte
		5	Länder
		6	Städte
		7	Straßen
		8	Landschaften
3	<b>ORG</b>	9	Organisationen
		10	Unternehmen
		11	Institutionen
		12	Gerichte
		13	Marken
4	<b>NRM</b>	14	Rechtsnormen
		15	Gesetze
		16	Rechtsverordnungen
		17	Europäische Normen
5	<b>REG</b>	18	Einzelfallregelungen
		19	Vorschriften
6	<b>RS</b>	20	Verträge
7	<b>LIT</b>	21	Rechtsprechungen
		22	Rechtsliteratur

Tabelle 4.1. Grob- und feinkörnige Klassen juristischer Entitäten.

Für das juristische Korpus bestehend aus gerichtlichen Entscheidungen wurden insgesamt 19 feinkörnige juristische Klassen erarbeitet, denen sieben grobkörnige zugrunde liegen. Sie sind in Abbildung 4.1 zusammen aufgelistet. Zusätzlich wurden Annotationsrichtlinien für diese Klassen entworfen, die auf den Annotationsrichtlinien von Automatic-Content-Extraction (Linguistic Data Consortium, 2008) und NoSta-D Named-Entity (Benikova et al., 2014b) basieren. Das Grundgerüst der Entitäten bilden die herkömmlichen Klassen **PER**, **LOC** und **ORG**, die sich in feinkörnige Unterklassen gliedern. Hiermit korrelieren die grob- und feinkörnigen Klassifikationen derart, dass in der grobkörnigen Grundklasse der Personen **PER** unter Nummer 1 zwischen den feinkörnigen Unterklassen der Personen **PER** (z.B. der Kläger, Beklagten, Zeugen, Gutachter etc.), der Richter **RR** (in (1)) und Anwälte **AN** (in (10)) unter den Nummern 1 bis 3 unterschieden wird. Zu der Grundklasse der Orte **LOC** gehören die Unterklassen der Länder **LD** (Länder, Staaten und Gliedstaaten), Städte **ST** (Städte, Dörfer und Gemeinden), Straßen **STR** (Straßen, Plätze, Alleen, Stadtbezirke und Sehenswürdigkeiten) und Landschaften **LDS** (Kontinente, Berge, Seen, Flüsse und weitere geographische Objekte), die im Beispiel (2) dargestellt sind.

- (1) Das Ablehnungsgesuch der Beschuldigten vom 1. April 2018 gegen die Vorsitzende Richterin am Bundesgerichtshof **GRT** Sost-Scheible **RR**, die Richterin am Bundesgerichtshof **GRT** Roggenbuck **RR** und die Richter am Bundesgerichtshof **GRT** Cierniak **RR**, Bender **RR** und Dr. Feilcke **RR** wird als unzulässig verworfen.

- (2) Jedoch wird der Verkehr darin naheliegend den Namen eines der bekanntesten Flüsse Deutschlands **LD** erkennen, welcher als Seitenfluss des **Rheins LDS** durch **Oberfranken LDS**, **Unterfranken LDS** und **Südhessen LDS** fließt und bei **Mainz ST** in den **Rhein LDS** mündet.

Die grobkörnige Grundklasse der Organisationen **ORG** wird in öffentliche bzw. gesellschaftliche, staatliche und wirtschaftliche Einrichtungen unterteilt (vgl. die Entitäten in den Beispielen (3, 4, 5)). Zu der feinkörnigen Unterklassse der Organisationen **ORG** gehören gesellschaftliche und öffentliche Einrichtungen wie Parteien, Vereine, Verbände, Zentren, Gemeinschaften, Unionen, Bildungsstätten oder auch Forschungseinrichtungen. In der Unterklassse der Institutionen **INN** werden dagegen staatliche Einrichtungen zusammengefasst, die Aufgaben für die Verwaltung des Staates (insbesondere hinsichtlich der Bürger) erfüllen. Das sind Bundesministerien und -behörden, Landesministerien und -behörden, sonstige Anstalten und Dienststellen. Dazu gehören auch die Verfassungsorgane der Bundesrepublik Deutschland auf der Bundes- und Länderebene: die Bundesregierung, der Bundesrat, der Bundestag, der Gemeinsame Ausschuss, die Landesparlamente und Landesregierungen. Im Unterschied zu **ORG** und **INN** enthält die Unterklassse der Unternehmen **UN** (privat)wirtschaftliche Einrichtungen. Diese beinhaltet Betriebe, Konzerne, Arbeits- und Werkstätten, Firmen u. Ä.

- (3) Der **FC Bayern München ORG** schloss den Beschwerdeführer ... aus dem Verein aus ...  
 (4) Die **Landesregierung Rheinland-Pfalz INN** hat von einer Stellungnahme abgesehen.  
 (5) ... des US-amerikanischen Unternehmens **Apple UN** ...

Eine zentrale Stelle in Entscheidungen nehmen Gerichtsbezeichnungen ein, deswegen werden sie in einer separaten Unterklassse **GRT** zusammengefasst. Das sind Bezeichnungen der Bundes-, Oberlands-, Lands- und Amtsgerichte, wobei sich die Bezeichnungen der Gerichte auf Länderebene aus den Bezeichnungen der ordentlichen Gerichtsbarkeit und ihren Standorten zusammensetzen (vgl. die markierten Entitäten in den Beispielen (6, 10)). Weiterhin werden noch Marken als eine semantische Kategorie ausgesondert, die in Entscheidungen des Bundespatentgerichtes sehr oft thematisiert werden. Sie bilden die Klasse **MRK**, die aus typologischer Sicht der Namensforschung neben der Grundklassse der Organisationen zu den Ergonymen (sogenannten Objektnamen) gehört. Ergonyme bezeichnen vom Menschen geschaffene Objekte, also Gegebenstände bzw. Einrichtungen. Trotz terminologischer und typologischer Ungenauigkeit wird die Klasse **MRK** absichtlich als eine Unterklassse der Organisationen abgestuft (s. Abbildung 4.1). Markennamen können kontextabhängig und mitunter semantisch zweideutig sein, wie z. B. 'Becker' aus dem illustrierten Beispiel (7), der sich aus dem Personennamen entwickelt hat.

- (6) ... wird das Urteil des **Landgerichts Fulda GRT** vom 30. Mai 2017 ... aufgehoben.  
 (7) Vorliegend stehen sich die Widerspruchsmarke **Becker Mining MRK** und die angegriffene Marke **Becker MRK** gegenüber.

Eine grundlegende Besonderheit der veröffentlichten Entscheidungen besteht darin, dass alle persönlichen Angaben aus Datenschutzgründen anonymisiert sind. Dies betrifft Personen-, Orts- und Organisationsnamen. Je nach zuständigem Bundesgericht werden unterschiedliche Muster zur Anonymisierung verwendet. Namen werden entweder durch einen oder zwei Buchstaben, meistens gefolgt von einem Punkt wie in (8) bzw. von Auslassungspunkten wie in (9), oder durch Auslassungspunkte wie in (10) ersetzt.

- (8) ... das Land **B. LD** ...  
 (9) ... unter der Firma **C ... AG UN** ...  
 (10) ... der ebenfalls beim **Bundesgerichtshof GRT** zugelassene Rechtsanwalt **... AN ...**

Neben den in der Eigennamenerkennung typischen Klassen wie **PER**, **LOC**, **ORG** wurden auch für Rechts- texte spezifische juristische Klassen entworfen. Das sind die grobkörnigen Grundklassen der Normen **NRM**, der Einzelfallregelungen **REG**, der Rechtsprechungen **RS** und der Rechtsliteratur **LIT**. Zu den Grundklassen der Normen und der Einzelfallregelungen gehören sowohl Bezeichnungen als auch Zitate, zu der Klasse der

Rechtsprechungen und der Rechtsliteratur jedoch ausschließlich Zitate. Normen werden gemäß ihrer rechtlichen Kraft in die feinkörnigen Unterklassen der Gesetze **GS**, Rechtsverordnungen **VO** und der Europäischen Normen **EUN** unterteilt. Die Unterkategorie **GS** setzt sich aus denjenigen Normen zusammen, die vom Gesetzgeber (dem Bundestag, dem Bundesrat bzw. den Landtagen) beschlossen und bezeichnet werden. Die Unterkategorie **VO** setzt sich aus Normen zusammen, die von der Bundes- oder Landesregierung bzw. von einem Ministerium anhand einer Ermächtigung beschlossen werden. Zu der Unterkategorie **EUN** gehören Normen des Europäischen Primärrechts, des Europäischen Sekundärrechts, der europäischen Organisationen und weitere Übereinkommen und Abkommen.

Im Beispielsatz (11) ist ein Verweis auf das ‚Gesetz über Teilzeitarbeit und befristete Arbeitsverträge‘ und die Bezeichnung ‚Grundgesetz‘ geschildert. Im Vergleich zu der Bezeichnung, die im Fließtext erwähnt wurde, weist das Zitat eine komplexere Struktur auf. Es besteht aus der Referenz auf den bestimmten Gesetzestextabschnitt, der Bezeichnung, der Kurzbezeichnung in Klammern, dem Ausfertigungsdatum, der Fundstelle in Klammern und Angaben zu der letzten Änderung inkl. wiederum deren Fundstelle in Klammern. Wenn dies vorliegt, handelt es sich um ein Vollzitat. Im Beispielsatz (12) ist dagegen ein Kurzzitat dargestellt, das aus Angaben zum entsprechenden Textabschnitt und der abgekürzten Bezeichnung der Rechtsverordnung besteht.

- (11) ... § 14 Absatz 2 Satz 2 des Gesetzes über Teilzeitarbeit und befristete Arbeitsverträge (TzBfG)  
 vom 21. Dezember 2000 (Bundesgesetzblatt Seite 1966), zuletzt geändert durch Gesetz vom 20. Dezember 2011 (Bundesgesetzblatt I Seite 2854) **GS**, ist nach Maßgabe der Gründe mit dem Grundgesetz **GS** vereinbar.

- (12) Mit der Neuregelung in § 35 Abs. 6 StVO **VO** ...

Die Klasse der Einzelfallregelungen **REG** enthält verbindliche Einzelakte, die unterhalb jeder Rechtsnorm stehen. Dazu gehören die Unterklassen der Vorschriften **VS** und Verträge **VT**. Vorschriften sind interne Anordnungen bzw. Anweisungen von einer übergeordneten Behörde an eine nachgeordnete, die deren Tätigkeit regeln. Das sind nebst Verwaltungsvorschriften weiterhin Richtlinien, Runderlasse und Erlasse, die meistens unterschiedlich bezeichnet werden, aber unter Verwaltungsvorschriften zu kategorisieren sind. Im Gegensatz zu Rechtsnormen haben diese Vorschriften keine unmittelbare Außenwirkung auf den Bürger. Zu der Unterkategorie der Verträge **VT** gehören verschiedenartige Verträge, u. a. öffentlich-rechtliche Verträge, Staatsverträge, Tarifverträge. Einige Bezeichnungen und Zitate aus diesen Klassen ähneln hinsichtlich der Form Rechtsnormen (s. die annotierten Entitäten in den Beispielen (13, 14)).

Die letzten zwei Klassen, also die der Rechtsprechungen **RS** und der Rechtsliteratur **LIT**, spalten sich nicht in Unterklassen und sind nach der grob- sowie feinkörnigen Klassifikation identisch. Die Klasse **RS** spiegelt Verweise auf Entscheidungen wider, in welchen stets der Name der amtlichen Entscheidungssammlung, des Bandes und die nummerierte Textstelle zu einer zitierten Aussage genannt werden. Oft erwähnt man hierbei auch das entscheidende Gericht, ggf. den Entscheidungstypen, das Aktenzeichen und das Entscheidungsdatum. Im Beispiel (15) sind Zitate auf Entscheidungen aus dem Bundesverfassungsgericht (BVerfG) und dem Bundessozialgericht (BSG) vorgestellt. Beide unterscheiden sich in der Zitierweise, weil Entscheidungen des BVerfGs nach Seiten und Entscheidungen des BSGs nach Paragraphen, Nummern und Randnummern sortiert sind.

- (13) ... insbesondere durch die Richtlinien zur Bewertung des Grundvermögens – BewRGr – vom 19.

September 1966 (BStBl I, S. 890) **VS**.

- (14) Auf das Arbeitsverhältnis der Parteien fand der Manteltarifvertrag für die Beschäftigten der Mitglieder der TGAOK **VT** (BAT/AOK-Neu **VT**) vom 7. August 2003 Anwendung.

- (15) ... (stRspr; vgl zB BVerfGE 62, 1, 45 **RS**; BVerfGE 119, 96, 179 **RS**; BSG SozR 4 – 2500 § 62 Nr 8 RdNr 20 f **RS**; Hauck/Wiegand, KrV 2016, 1, 4 **LIT**).

Zu der Klasse der Rechtsliteratur **LIT** gehören Zitate der Rechtskommentare, Gesetzgebungsmaterialien, Rechtslehrbücher und -monographien. Ein Rechtskommentar setzt sich wie in (16) aus den Angaben des Autoren- und/oder des Herausgeber-, des Normennamens, des Paragraphen und der Randnummer zusammen. Mehrere Autoren werden mit einem Schrägstrich unterteilt. Lehrbücher und Monographien dagegen

werden wie gewohnt mit dem Autorennamen, dem Titel, der Auflage, dem Erscheinungsjahr und der Seitenangabe zitiert. Zitate der Gesetzgebungsmaterialien bestehen aus dem Titel und der Fundstelle, die mit den Zahlen markiert wird (s. (17)).

(16) ... Klein , in : Maunz / Schmidt-Bleibtreu / Klein / Bethge , BVerfGG , § 19 Rn. 9 **LIT** ...

(17) ... BR-Drs. 756/03 **LIT** ...

Im Laufe der Entwicklung der feinkörnigen Klassen und der Annotation der Entscheidungen wurden die Klassen der Kontinente **KONT** (als Unterklasse der Ortsnamen) sowie der Universitäten **UNI**, der Institute **IS** und der Museen **MUS** (als Unterklassen der Organisationsnamen) ausgesondert. Ihre Häufigkeit lag aber unter dem Mindestwert, der auf 50 gesetzt war. Deswegen wurden die Universitäts-, Instituts- und Museumsnamen mit der feinkörnigen Klasse der Organisationen **ORG** vereinigt. Die Kontinentnamen wurden dementsprechend in die Landschaften **LDS** integriert.

### 4.3 Beschreibung und Eigenschaften des Korpus

Das Korpus *Legal Entity Recognition* (LER) besteht aus 750 Entscheidungen der Jahre 2017–2018, die vom Bundesministerium der Justiz und für Verbraucherschutz auf dem Portal „Rechtsprechung im Internet“<sup>2</sup> veröffentlicht wurden. Die Entscheidungen stammen aus sieben Bundesgerichten: Bundesarbeitsgericht (BAG), Bundesfinanzhof (BFH), Bundesgerichtshof (BGH), Bundespatentgericht (BPatG), Bundessozialgericht (BSG), Bundesverfassungsgericht (BVerfG) und Bundesverwaltungsgericht (BVerwG). Insgesamt wurden ca. 107 Dokumente pro Gericht aus dem bereitgestellten Inhaltsverzeichnis<sup>3</sup> ausgewählt. Die Daten wurden aus Quelltexten im XML-Format<sup>4</sup> (vgl. XML- und PDF-Format am Beispiel eines Beschlusses in den Abbildungen A.1 und A.2 im Anhang) gesammelt, also aus solchen XML-Elementen wie Mitwirkung, Titelzeile, Leitsatz, Tenor, Tatbestand, Entscheidungsgründen, Gründen, abweichender Meinung und sonstigem Titel extrahiert. Dementsprechend wurden diejenigen Angaben gelöscht, die am Anfang des Dokumentes waren (Gerichtsname, Entscheidungsdatum, Aktenzeichen, ECLI, Dokumenttyp, Normen in den Zeilen 6–13 in Abbildung A.1), die zum Verfahrensgang (Zeilen 14–15) gehörten und die interne Merkmale des Dokumentes (Zeilen 5, 16–19, 61–65) widerspiegeln. Der ausschlaggebende Grund dafür war, dass diese Daten schon entsprechend klassifiziert waren. Randnummern zu Paragraphen wie in den Zeilen 37 und 45, die in Quelltexten vorkamen, wurden auch entfernt, weil ihre primäre Funktion darin bestand, das Zitieren von und in Rechtstexten zu präzisieren. Danach wurden die extrahierten Daten durch SoMaJo<sup>5</sup> (Proisl und Uhrig, 2016) in Sätze und Wörter aufgeteilt und in WebAnno<sup>6</sup> (Eckart de Castilho et al., 2016) manuell annotiert.

	BAG	BFH	BGH	BPatG	BSG	BVerfG	BVerwG	insg.
<b>Dokumente</b>	107	107	108	107	107	107	107	<b>750</b>
<b>Token mit Satzzeichen</b>	343.065	276.233	177.835	404.041	302.161	305.889	347.824	<b>2.157.048</b>
<b>Token ohne Satzzeichen</b>	301.584	238.730	154.389	339.994	264.555	263.797	305.306	<b>1.868.355</b>
<b>Sätze</b>	12.791	8.522	5.858	12.016	8.083	9.237	10.216	<b>66.723</b>
<b>Verteilung Entitäten</b>	19,23 %	22,43 %	19,23 %	10,41 %	22,76 %	22,09 %	20,84 %	<b>19,15 %</b>

**Tabelle 4.2.** Korpusgröße: die Anzahl Dokumente, Token mit und ohne Satzzeichen, Sätze und die Verteilung der juristischen Entitäten.

Das Korpus LER ist unter CC BY 4.0<sup>7</sup> lizenziert und zum Download<sup>8</sup> frei verfügbar. Es besteht aus 66.723 Sätzen, die insgesamt 2.157.048 Token mit Satzzeichen und 1.868.355 ohne Satzzeichen beinhalten. Die Größe des gesamten Korpus bzw. der Teilkorpora, die nach den sieben Bundesgerichten unterteilt sind, ist in Abbildung 4.2 dargestellt. Die Größe der Teilkorpora variiert zwischen 5.858 und 12.791 Sätzen und von 177.835 auf 404.041 Token, wobei der Anteil an Entitäten durchschnittlich 19 % der Anzahl der Token entspricht. Eine Ausnahme bildet das Bundespatentgericht mit 10,41 %.

<sup>2</sup><http://www.rechtsprechung-im-internet.de>

<sup>3</sup><http://www.rechtsprechung-im-internet.de/rii-toc.xml>

<sup>4</sup>Nähtere Informationen zum Aufbau der XML-Daten sind in der Document Type Definition zu finden: [www.rechtsprechung-im-internet.de/dtd/v1/rii-dok.dtd](http://www.rechtsprechung-im-internet.de/dtd/v1/rii-dok.dtd)

<sup>5</sup><https://github.com/tsproisl/SoMaJo>

<sup>6</sup><https://webanno.github.io/webanno/>

<sup>7</sup><https://creativecommons.org/licenses/by/4.0/deed.de>

<sup>8</sup><https://github.com/elenanereiss/Legal-Entity-Recognition>

<b>Grobkörnige Klassen</b>		<b>BAG</b>	<b>BFH</b>	<b>BGH</b>	<b>BPatG</b>	<b>BSG</b>	<b>BVerfG</b>	<b>BVerwG</b>	<b>insgesamt</b>
<b>PER</b>	Personen	659	336	1.041	606	262	194	279	<b>3.377</b> <b>6,30 %</b>
<b>LOC</b>	Orte	220	250	189	327	139	518	825	<b>2.468</b> <b>4,60 %</b>
<b>ORG</b>	Organisationen	835	964	506	1.535	934	1.615	1.526	<b>7.915</b> <b>14,76 %</b>
<b>NRM</b>	Rechtsnormen	2.838	3.358	1.302	1.236	3.731	4.078	4.273	<b>20.816</b> <b>38,81 %</b>
<b>REG</b>	Einzelfallregelungen	2.487	92	26	3	413	60	389	<b>3.470</b> <b>6,47 %</b>
<b>RS</b>	Rechtsprechungen	1.984	1.927	1.206	2.057	1.795	1.898	1.713	<b>12.580</b> <b>23,46 %</b>
<b>LIT</b>	Rechtsliteratur	485	472	408	307	532	554	248	<b>3.006</b> <b>5,60 %</b>
<b>Feinkörnige Klassen</b>		<b>BAG</b>	<b>BFH</b>	<b>BGH</b>	<b>BPatG</b>	<b>BSG</b>	<b>BVerfG</b>	<b>BVerwG</b>	<b>insgesamt</b>
<b>PER</b>	Personen	123	318	430	228	251	118	279	<b>1.747</b> <b>3,26 %</b>
<b>AN</b>	Anwälte	2	1	52	38	9	9	0	<b>111</b> <b>0,21 %</b>
<b>RR</b>	Richter	534	17	559	340	2	67	0	<b>1.519</b> <b>2,83 %</b>
<b>LD</b>	Länder	123	130	79	120	86	425	466	<b>1.429</b> <b>2,66 %</b>
<b>ST</b>	Städte	62	67	80	124	43	59	270	<b>705</b> <b>1,31 %</b>
<b>STR</b>	Straßen	7	19	20	17	4	12	57	<b>136</b> <b>0,25 %</b>
<b>LDS</b>	Landschaften	28	34	10	66	6	22	32	<b>198</b> <b>0,37 %</b>
<b>ORG</b>	Organisationen	386	81	70	92	175	198	164	<b>1.166</b> <b>2,17 %</b>
<b>UN</b>	Unternehmen	263	351	61	248	63	55	17	<b>1.058</b> <b>1,97 %</b>
<b>INN</b>	Institutionen	75	66	50	602	137	352	914	<b>2.196</b> <b>4,09 %</b>
<b>GRT</b>	Gerichte	110	464	287	352	559	1.010	430	<b>3.212</b> <b>5,99 %</b>
<b>MRK</b>	Marken	1	2	38	241	0	0	1	<b>283</b> <b>0,53 %</b>
<b>GS</b>	Gesetze	2.565	2.994	1.228	1.056	3.595	3.643	3.439	<b>18.520</b> <b>34,53 %</b>
<b>VO</b>	Verordnungen	112	73	10	28	53	8	513	<b>797</b> <b>1,49 %</b>
<b>EUN</b>	Europäische Normen	161	291	64	152	83	427	321	<b>1.499</b> <b>2,79 %</b>
<b>VS</b>	Vorschriften	144	62	2	2	175	30	192	<b>607</b> <b>1,13 %</b>
<b>VT</b>	Verträge	2.343	30	24	1	238	30	197	<b>2.863</b> <b>5,34 %</b>
<b>RS</b>	Rechtsprechungen	1.984	1.927	1.206	2.057	1.795	1.898	1.713	<b>12.580</b> <b>23,46 %</b>
<b>LIT</b>	Rechtsliteratur	485	472	408	307	532	554	248	<b>3.006</b> <b>5,60 %</b>
		<b>Anzahl Entitäten</b>		<b>53.632</b>	<b>100 %</b>				

**Tabelle 4.3.** Verteilung der grob- und feinkörnigen Klassen in den sieben Bundesgerichten.

Das Korpus verfügt über zwei Varianten für die Klassifikation der juristischen Entitäten, deren Aufteilung die grob- und die feinkörnigen semantischen Kategorien wie oben beschrieben zugrunde liegen (s. Abbildung 4.1). Für die feinkörnige Variante wurden Entscheidungen mit 19 Unterklassen manuell annotiert. Die grobkörnige Variante mit den sieben semantischen Grundklassen wurde automatisch aus den fein annotierten Entitäten erstellt. Insgesamt gibt es 53.632 annotierte Entitäten, die aus 413.082 Token bestehen. Die Verteilung der Entitäten hinsichtlich der zwei Varianten der Klassifikation und der sieben Bundesgerichte ist in Abbildung 4.3 dargestellt. Ihre Anzahl variiert je nach Bundesgericht deutlich. Die typischen Klassen der Eigennamen wie **PER**, **LOC**, **ORG** bilden mit 25,66 % ein Viertel aller Entitäten (also 13.760), wobei mehr als die Hälfte davon Organisationsnamen sind. Im Vergleich dazu ist der Anteil an juristischen Entitäten, welche zu Namen und Zitaten der Rechtsnormen **NRM**, Einzelfallregelungen **REG**, Rechtsprechungen **RS** und Rechtsliteratur **LIT** gehören, dreifach größer – 74,34 % oder 39.872 Entitäten. Am meisten sind die Klassen der nationalen Gesetze **GS** mit 34,53 % und der Rechtsprechungen **RS** mit 23,46 % vertreten. Die anderen juristischen Klassen (der Rechtsverordnungen **VO**, der Europäischen Normen **EUN**, der Vorschriften **VS**, der Verträge **VT** und der Rechtsliteratur **LIT**) kommen in Entscheidungen seltener vor und ihr Anteil beträgt relativ zu allen Entitäten ca. 1–6 %.

Im Vergleich zu den Korpora für Zeitungstexte sind personenbezogene Daten in Entscheidungen anonymisiert, wie bereits in Kapitel 4.2 beschrieben. Zuallererst sind die Namen der Personen, Anwälte und Unternehmen betroffen. Ihre Anzahl beläuft sich auf ca. 80 %, 95 % und 70 %. Mehr als die Hälfte der Stadt- und Straßennamen, also ca. 55 %, sind ebenfalls verändert. Für die Namen der Richter, Länder, Institutionen und Gerichte ist dies aber unüblich. Anonymisiert sind hierbei lediglich ca. 1–5 %. Die Landschaftsbezeichnungen und die Organisationsnamen sind auch betroffen, dort sind 40 % und 15 % der personenbezogenen Daten entsprechend anonymisiert.

Das Korpus steht im CoNLL-2002-Format zur Verfügung. Die Daten sind in zwei Spalten aufgeteilt, die mit einem Leerzeichen getrennt sind. Jedes Wort befindet sich in einer Zeile. Die Satzgrenze ist mit einer leeren Zeile markiert. Die erste Spalte enthält ein Wort und die zweite ein Tag im IOB2-Format. Das Datenformat ist in Abbildung 4.1 illustriert. Der Beispielsatz enthält vier Eigennamen: den Namen des Landesorgans ‚saarländische Landesregierung‘, den des Richters ‚Müller‘, der Ministerpräsident des Saarlandes war, und die zwei Organisationsnamen ‚Evangelische Kirche im Rheinland‘ sowie ‚Evangelische Kirche der Pfalz‘.

Am	O
7.	O
März	O
2006	O
fand	O
ein	O
Treffen	O
der	O
saarländischen	B-INN
Landesregierung	I-INN
unter	O
Vorsitz	O
des	O
Ministerpräsidenten	O
Müller	B-RR
mit	O
Vertretern	O
der	O
Evangelischen	B-ORG
Kirche	I-ORG
im	I-ORG
Rheinland	I-ORG
und	O
der	O
Evangelischen	B-ORG
Kirche	I-ORG
der	I-ORG
Pfalz	I-ORG
statt	O
.	O

**Abbildung 4.1.** Datenformat des juristischen Korpus.

## 5 Evaluation und Ergebnisse

Die Entwicklung des juristischen Korpus ermöglicht es nun, aktuelle Verfahren zur automatischen Annotation in Rechtstexten zu untersuchen. Bevor auf die Ergebnisse der CRF- und BLSTM-Modellfamilien mit 19 und sieben Klassen eingegangen wird und diese verglichen werden, wird über die standardisierten Mikro-Metriken wie Precision, Recall und das  $F_1$ -Maß berichtet, mit denen Ergebnisse eines Mehrklassen-Klassifizierungsproblems bewertet werden. Zugleich wird das Experimentdesign beschrieben, dabei wird insbesondere auf die drei jeweils ausgewählten CRF- sowie BLSTM-Modelle und die verwendete Evaluationsmethode eingegangen. Um Ursachen der festgestellten Gemeinsamkeiten in der Performanz aufzudecken, wird eine Analyse der richtigen und falschen Vorhersagen durchgeführt. Schließlich wird eine Diskussion der gewonnenen Erkenntnisse angefügt.

### 5.1 Evaluationsmetriken der Eigennamenerkennung

Die Ergebnisse der Eigennamenerkennung sind mithilfe der Precision, des Recalls und des  $F_1$ -Maßes zu evaluieren. Diese Metriken spiegeln das Verhältnis der korrekt und inkorrekt klassifizierten Einheiten eines Modells wider, welche mit manuell annotierten Klassen (mit sogenanntem Goldstandard) verglichen werden. Dabei unterscheidet man vier Fälle hinsichtlich der Klassifizierungsergebnisse: richtig positive, falsch positive, richtig negative und falsch negative Ergebnisse. Wenn Eigennamen oder Gattungsnamen korrekt klassifiziert werden, geht es um richtig positive ( $t_p$ ) oder richtig negative Ergebnisse ( $t_n$ ). Werden Gattungsnamen als Eigennamen klassifiziert, werden diese zu den falsch positiven Ergebnissen ( $f_p$ ) zugeordnet. Im Gegenteil geht es um falsch negative Ergebnisse ( $f_n$ ), wenn Eigennamen als Gattungsnamen klassifiziert werden.

Die Precision misst den Anteil der korrekt klassifizierten Eigennamen am gesamten Anteil der als Eigennamen klassifizierten Eigennamen und Gattungsnamen. Der Recall misst dagegen den Anteil der korrekt klassifizierten Eigennamen an der gesamten Menge der echten Eigennamen. Das  $F_1$ -Maß kombiniert die Precision und den Recall mithilfe des gewichteten harmonischen Mittels. Aufgrund dessen, dass das Problem der Eigennamenerkennung zu den Mehrklassen-Klassifizierungsproblemen gehört, werden die Precision, der Recall und das  $F_1$ -Maß als Mikro-Metriken nach der Klassenfrequenz gewichtet. Die Anzahl  $t_p, f_p, f_n$  setzt sich aus der Summe  $t_{pc}, f_{pc}, f_{nc}$  für jede Klasse  $c$  aus der Menge aller definierten Klassen  $C$  zusammen. Demnach sind Evaluationsmetriken wie folgt definiert:

$$\text{Precision} = \frac{\sum_{c \in C} t_{pc}}{\sum_{c \in C} t_{pc} + \sum_{c \in C} f_{pc}} \quad (5.1)$$

$$\text{Recall} = \frac{\sum_{c \in C} t_{pc}}{\sum_{c \in C} t_{pc} + \sum_{c \in C} f_{nc}} \quad (5.2)$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.3)$$

### 5.2 Experimentbeschreibung

Für das Experiment wurden zwei frei verfügbare Tools für die Eigennamenerkennung gewählt, denen zwei Ansätze, nämlich CRFs und BLSTMs, zugrunde liegen. Das sind sklearn-crfsuite<sup>1</sup> und UKPLab-BLSTM<sup>2</sup> (Reimers und Gurevych, 2017b). Sklearn-crfsuite ist ein Wrapper<sup>3</sup> für CRFSuite<sup>4</sup>, der für die Software-Bibliothek Scikit-learn<sup>5</sup> entwickelt wurde. Für die CRF-Modelle wurden folgende Gruppen von Features und Quellen ausgewählt:

<sup>1</sup><https://sklearn-crfsuite.readthedocs.io/en/latest/>

<sup>2</sup><https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf>

<sup>3</sup>Der Wrapper dient zur Verbindung inkompatibler Softwarekomponenten, erfüllt also eine Übersetzungsfunktion.

<sup>4</sup><http://www.chokkan.org/software/crfsuite/>

<sup>5</sup><https://scikit-learn.org>

1. F – Features für das aktuelle Wort sowie für die zwei vorangegangenen und zwei nachfolgenden Wörter (Kontext [-2, -1, 0, 1, 2]), die Groß- und Kleinschreibung, Form, Präfixe sowie Suffixe berücksichtigen;
2. G – Namenslisten der Personen, Länder, Städte, Straßen, Landschaften, Firmen, Gesetze, Rechtsverordnungen und Vorschriften für das aktuelle Wort (also im Kontext [0]);
3. L – die Lookup-Tabelle für die Wortähnlichkeit (auch im Kontext [-2, -1, 0, 1, 2]) wie in Benikova et al. (2015), die die vier ähnlichsten Wörter zum aktuellen Wort enthält.

Für das Experiment mit der CRF-Modellfamilie wurden drei Modelle entworfen, die diese drei Gruppen von Features und Quellen der Reihe nach verketten: (1) CRF-F mit Features (Gruppe 1); (2) CRF-FG mit Features und Namenslisten (Gruppe 1, 2); (3) CRF-FGL mit Features, Namenslisten und der Lookup-Tabelle (Gruppe 1, 2, 3). Dementsprechend spiegeln die Abkürzungen in den Bezeichnungen der CRF-Modelle die betroffenen Gruppen wider. Als Lernalgorithmus wird das L-BFGS-Verfahren mit den L1- und L2-Regularisierungsparametern verwendet, welche auf den Koeffizienten 0,1 eingestellt werden. Die Anzahl maximaler Iterationen für die Optimierung des Algorithmus ist auf 100 eingestellt.

Das UKPLab-BLSTM ist eine Implementierung der Sequenzmodelle von Ubiquitous-Knowledge-Processing-Lab, die auf dem BLSTM beruht. Das Tool kann als ein BLSTM-CRF (Huang et al., 2015), ein BLSTM-CNN-CRF (Ma und Hovy, 2016) oder ein BLSTM-CRF mit Buchstabeneinbettungen (Lample et al., 2016) konfiguriert werden und dient für verschiedene automatische Annotationen (Chunking, POS-Tagging, Eigennamenerkennung usw.). Für das Experiment mit der BLSTM-Modellfamilie wurden auch drei Modelle gewählt: (1) das BLSTM-CRF ohne Buchstabeneinbettungen; (2) das BLSTM-CRF+ mit Buchstabeneinbettungen, die vom BLSTM generiert werden; (3) das BLSTM-CNN-CRF, wiederum mit Buchstabeneinbettungen, die aber vom CNN produziert werden. Dabei wurden solche Hyperparameter verwendet, welche die beste Performanz in der Eigennamenerkennung, so Reimers und Gurevych (2017a), erzielt haben. Die BLSTM-Modelle verfügen über zwei BLSTM-Schichten mit jeweils einer Größe von 100 Einheiten, wobei der Dropout auf 0,25 für beide Schichten eingestellt ist. Die maximale Anzahl der Epochen beträgt 100. Zugleich setzt das Tool vortrainierte Worteinbettungen für das Deutsche von Reimers et al. (2014) ein.

Insgesamt sind zwölf Experimente für zwei Modellfamilien, die jeweils durch drei Modelle vertreten sind, und zwei Klassifikationen (feinkörnig – 19 Klassen, grobkörnig – sieben Klassen) durchzuführen. Um die Leistung der Modelle zuverlässig zu schätzen, wird als Evaluationsmethode das k-fache stratifizierte Kreuzvalidierungsverfahren (engl. *stratified k-fold cross-validation*) eingesetzt. Das Korpus wird satzweise gemischt und in zehn sich gegenseitig ausschließende Teilkorpora ähnlicher Größe aufgeteilt. In einer Iteration wird ein Teil zur Validierung und der Rest zum Training verwendet. Das Verwenden dieser Evaluationsmethode hat als Folge, dass das Testkorpus aus 6.673 Sätzen und das Trainingskorpus aus 60.050 Sätzen besteht. Dabei bleiben die relativen Klassenhäufigkeiten im Trainings- und Testkorpus über die k Iterationen gleich. Insgesamt wird zehnmal iteriert, sodass jeder Teil des Korpus neunmal zum Trainieren und einmal für die Validierung verwendet wird. So verhindert die 10-fache Kreuzvalidierung eine Überanpassung (engl. *overfitting*) während des Trainings. Wie aus Abbildung 4.3 in Kapitel 4.3 zu sehen war, unterscheidet sich die Häufigkeit der semantischen Klassen im Korpus. Die Stratifizierung garantiert, dass im Testkorpus die semantischen Klassen gleich häufig relativ zur Größe des Trainingskorpus vorkommen, was Messfehler im Falle der unausbalancierten Daten ausschließt.

### 5.3 Ergebnisse der CRF-Modelle

Für die CRF-Modellfamilie wurden drei Modelle entwickelt, die in zehn Iterationen auf dem juristischen Korpus trainiert und evaluiert wurden. Das waren ein CRF-F mit Features, ein CRF-FG mit Features und Namenslisten sowie ein CRF-FGL mit Features, Namenslisten und einer Lookup-Tabelle für die Wortähnlichkeit. Wie in Tabelle 5.1 darstellt, hat das CRF-FGL die beste Performanz in der Eigennamen- und Zitaterkennung mit einem  $F_1$ -Wert von 93,23 % erzielt, welcher um 0,18 % höher als mit dem CRF-F und um 0,11 % höher als mit dem CRF-FG war. Die Erkennung der Namen und Zitate in den Klassen war je nach Modell unterschiedlich erfolgreich. Für weitere Informationen s. ihre Konfusionsmatrizen in den Abbildungen A.3, A.5, A.7 im Anhang. So haben die Klassen der Anwälte, Institutionen, Gerichte, Verträge und Rechtsprechungen die höchsten  $F_1$ -Werte mit dem CRF-F erreicht. Mit dem CRF-FG konnten bessere Ergebnisse in den Klassen der Richter, Städte, Vorschriften und Literatur erzielt werden. Dies bedeutet, dass Namenslisten einen positiven Einfluss auf die Erkennung dieser Entitäten ausgeübt haben. Die restlichen Klassen zeigten eine bessere Performanz mit dem CRF-FGL. Wenn man die Werte von den drei CRF-Modellen vergleicht, stellt man fest, dass die Precision-, Recall- und  $F_1$ -Werte minimal gestiegen oder abgesunken sind. Die größte Differenz betrug ca. 3 %. Die Verketzung von Namenslisten und der Lookup-Tabelle hat die Ergebnisse zwar verbessert, aber nicht so signifikant

wie erwartet.

Klasse	CRF-F			CRF-FG			CRF-FGL		
	Prec	Rec	F <sub>1</sub>	Prec	Rec	F <sub>1</sub>	Prec	Rec	F <sub>1</sub>
<b>PER</b>	89,41 %	83,53 %	86,32 %	90,50 %	83,54 %	86,83 %	90,44 %	84,22 %	<b>87,18 %</b>
<b>RR</b>	98,22 %	97,62 %	97,92 %	98,68 %	97,75 %	<b>98,21 %</b>	98,55 %	97,75 %	98,14 %
<b>AN</b>	93,14 %	76,84 %	<b>83,73 %</b>	89,81 %	73,51 %	80,39 %	92,17 %	75,04 %	81,99 %
<b>LD</b>	96,73 %	90,42 %	93,44 %	97,03 %	91,98 %	94,40 %	96,93 %	92,62 %	<b>94,70 %</b>
<b>ST</b>	88,99 %	77,37 %	82,70 %	88,27 %	81,77 %	<b>84,77 %</b>	88,09 %	81,82 %	84,67 %
<b>STR</b>	88,69 %	59,58 %	70,51 %	87,51 %	57,95 %	68,90 %	90,50 %	59,85 %	<b>71,30 %</b>
<b>LDS</b>	94,34 %	61,14 %	73,43 %	92,63 %	64,09 %	75,25 %	93,33 %	65,27 %	<b>76,08 %</b>
<b>ORG</b>	86,82 %	71,25 %	78,20 %	86,71 %	71,95 %	78,56 %	88,84 %	72,72 %	<b>79,89 %</b>
<b>UN</b>	92,77 %	86,04 %	89,21 %	93,00 %	86,18 %	89,39 %	93,54 %	86,85 %	<b>90,01 %</b>
<b>INN</b>	92,74 %	89,49 %	<b>91,07 %</b>	92,88 %	89,20 %	90,98 %	92,51 %	89,47 %	90,96 %
<b>GRT</b>	97,23 %	96,35 %	<b>96,78 %</b>	97,03 %	96,35 %	96,69 %	97,19 %	96,33 %	96,75 %
<b>MRK</b>	85,85 %	56,91 %	67,85 %	90,33 %	56,20 %	68,82 %	88,40 %	58,07 %	<b>69,61 %</b>
<b>GS</b>	96,86 %	96,34 %	96,60 %	97,00 %	96,44 %	96,72 %	97,02 %	96,56 %	<b>96,79 %</b>
<b>VO</b>	91,91 %	82,23 %	86,79 %	91,35 %	82,85 %	86,87 %	91,41 %	83,49 %	<b>87,26 %</b>
<b>EUN</b>	89,37 %	86,07 %	87,67 %	88,91 %	85,49 %	87,14 %	89,41 %	86,21 %	<b>87,76 %</b>
<b>VS</b>	83,83 %	71,38 %	77,00 %	84,34 %	71,03 %	<b>77,02 %</b>	84,42 %	70,66 %	76,85 %
<b>VT</b>	90,66 %	87,72 %	<b>89,15 %</b>	90,18 %	87,42 %	88,76 %	90,53 %	87,67 %	89,06 %
<b>RS</b>	93,35 %	93,39 %	<b>93,37 %</b>	93,22 %	93,34 %	93,28 %	93,21 %	93,29 %	93,25 %
<b>LIT</b>	92,98 %	91,28 %	92,12 %	92,94 %	91,42 %	<b>92,17 %</b>	92,79 %	91,28 %	92,02 %
<b>insg.</b>	94,28 %	91,85 %	93,05 %	94,31 %	91,96 %	93,12 %	94,37 %	92,12 %	<b>93,23 %</b>

**Tabelle 5.1.** Precision, Recall und F<sub>1</sub>-Werte der CRF-Modelle für die feinkörnigen Klassen.

In den Grundkategorien wurden stets die besten Ergebnisse bei der Klasse der Richter **RR** unter den Personennamen, bei der Klasse der Länder **LD** unter den Ortsnamen sowie bei der Klasse der Gerichtsnamen **GRT** unter den Organisationsnamen erzielt. Die F<sub>1</sub>-Werte erreichten maximal ca. 98 %, 95 %, 97 %. Unter den Rechtsnormen ist die Erkennung der Gesetzesnamen und -zitate **GS** mit rund 97 % am besten ausgefallen. Dazu konnte man auch die Klasse der Rechtsprechungen **RS** mit einem F<sub>1</sub>-Wert von ca. 93 % und die Klasse der Rechtsliteratur **LIT** mit einem F<sub>1</sub>-Wert von ca. 92 % mitzählen. Unter den Einzelfallregelungen war die Klasse der Verträge **VT** mit einem F<sub>1</sub>-Wert von 89 % besser.

Zufriedenstellende Ergebnisse mit einem F<sub>1</sub>-Wert höchstens bei 84 % und 77 % erreichten die CRF-Modelle bei den Klassen **AN** und **VS**. Die Markennamenerkennung **MRK** ist dagegen am schwächsten ausgefallen und enthielt somit das größte Verbesserungspotential. Der F<sub>1</sub>-Wert variierte zwischen 68–70 %, wobei der Unterschied zwischen Precision und Recall mindestens 29 % betrug. Der höhere Precision-Wert im Vergleich zum niedrigeren Recall-Wert deutet darauf hin, dass viele Markennamen als solche nicht identifiziert wurden (vgl. die annotierten **B-MRK**, **I-MRK** mit dem vorhergesagten **O** in den Konfusionsmatrizen, die in den Abbildungen A.3, A.5, A.7 im Anhang dargestellt sind). Ein ähnliches Verhältnis ließ sich für die Klassen **STR** und **LDS** feststellen. Die weiteren Klassen (**PER**, **ST**, **UN**, **VO**, **EUN**) konnten gut klassifiziert werden. Ihre F<sub>1</sub>-Werte variierten zwischen 85–91 %.

Klasse	CRF-F			CRF-FG			CRF-FGL		
	Prec	Rec	F <sub>1</sub>	Prec	Rec	F <sub>1</sub>	Prec	Rec	F <sub>1</sub>
<b>PER</b>	94,20 %	89,43 %	91,74 %	94,54 %	89,99 %	<b>92,20 %</b>	94,22 %	90,20 %	92,16 %
<b>LOC</b>	94,60 %	84,55 %	89,26 %	93,89 %	85,48 %	89,45 %	94,33 %	86,45 %	<b>90,18 %</b>
<b>ORG</b>	92,82 %	89,00 %	90,87 %	93,02 %	89,08 %	90,99 %	93,23 %	89,10 %	<b>91,11 %</b>
<b>NRM</b>	96,19 %	95,16 %	95,67 %	96,29 %	95,26 %	95,77 %	96,28 %	95,44 %	<b>95,86 %</b>
<b>REG</b>	89,29 %	84,72 %	86,94 %	89,28 %	84,77 %	<b>86,96 %</b>	88,76 %	84,15 %	86,39 %
<b>RS</b>	93,19 %	93,26 %	93,23 %	93,28 %	93,23 %	<b>93,25 %</b>	93,08 %	93,08 %	93,08 %
<b>LIT</b>	92,72 %	91,15 %	91,92 %	92,99 %	91,14 %	92,06 %	93,11 %	91,13 %	<b>92,11 %</b>
<b>insg.</b>	94,17 %	92,07 %	93,11 %	94,26 %	92,20 %	<b>93,22 %</b>	94,22 %	92,25 %	<b>93,22 %</b>

**Tabelle 5.2.** Precision, Recall und F<sub>1</sub>-Werte der CRF-Modelle für die grobkörnigen Klassen.

Wie aus Tabelle 5.2 zu entnehmen ist, erzielten die CRF-Modelle mit den grobkörnigen Klassen eine vergleichbare Performance. Eine Ausnahme bildeten die besser ausbalancierten Precision- und Recall-Werte (für weitere Informationen s. die Konfusionsmatrizen der CRF-Modelle in den Abbildungen A.4, A.6, A.8 im Anhang). Insgesamt stiegen die F<sub>1</sub>-Werte lediglich mit dem CRF-F minimal an. Das CRF-FG und das CRF-FGL erreichten zusammen das beste Ergebnis für die sieben Klassen mit einem F<sub>1</sub>-Wert von 93,22 %. Jedoch gab es kleinere Unterschiede in den einzelnen Klassen. So wurden Personennamen besser mit dem CRF-FG und Orts- sowie Organisationsnamen besser mit dem CRF-FGL erkannt. Bei den Klassen **REG** und **RS** konnte das CRF-FG das beste Ergebnis erzielen. Mit dem CRF-FGL erhöhten sich die Werte in den Klassen **NRM** und **LIT**.

Die drei CRF-Modelle erreichten das beste Ergebnis in der Klasse **NRM** mit einem F<sub>1</sub>-Wert von ca. 96 %. Die

Erkennung in den Klassen der Personen **PER**, der Rechtsprechung **RS** und der Rechtsliteratur **LIT** ist auch sehr gut ausgefallen. Die  $F_1$ -Werte lagen bei ca. 92–93 %. Eine schlechtere Performanz mit einem  $F_1$ -Wert von ca. 89–91% wurde bei den Orts- **LOC** und Organisationsnamen **ORG** festgestellt. Das schlechteste Ergebnis erzielten die CRF-Modelle bei der Klasse der Einzelfallregelungen **REG**. Der  $F_1$ -Wert betrug maximal 87 %.

## 5.4 Ergebnisse der BLSTM-Modelle

Um Eigennamen und Zitate in Entscheidungen zu erkennen, wurden auch drei BLSTM-Modelle auf dem juristischen Korpus getestet: ein BLSTM-CRF, ein BLSTM-CRF+ mit Buchstabeneinbettungen und ein BLSTM-CNN-CRF. Die Abbildung 5.3 zeigt, dass alle drei Modelle gute Ergebnisse erzielt haben. Die  $F_1$ -Werte für die betroffenen feinkörnigen Klassen variieren von 72–99%. Das BLSTM-CRF lieferte einen um 1,7 % niedrigeren  $F_1$ -Wert für alle Klassen, der mit 93,75 % das schlechteste Ergebnis im Vergleich zu den BLSTM-Modellen mit Buchstabeneinbettungen war. Diese Tatsache bestätigt den positiven Einfluss der Buchstabeninformation auf die Performanz eines Modells für die Erkennung semantischer Kategorien. Eine besonders signifikante Verbesserung mit einer Steigerung des  $F_1$ -Wertes um 5–16 % konnte in den Klassen der Organisationen, Unternehmen, Rechtsverordnungen, Vorschriften und Verträge beobachtet werden. Jedoch wurden die Namen der Richter und Anwälte mit dem BLSTM-CRF im Vergleich zu den Modellen mit Buchstabeneinbettungen um ca. 1 % besser erkannt. Interessanterweise erreichten das BLSTM-CRF+ und das BLSTM-CNN-CRF für alle Klassen einen gleichen  $F_1$ -Wert von 95,46 %. Mit dem BLSTM-CRF+ wurden aber Rechtsverordnungen, Europäische Normen, Vorschriften und Verträge besser erkannt und dies sowohl in Bezug auf Precision, Recall und das  $F_1$ -Maß. Mit dem BLSTM-CNN-CRF wurden dagegen die Namen und Zitate von Straßen, Landschaften, Organisationen, Unternehmen und Rechtsprechungen besser identifiziert. Ein besonders großer Unterschied in den  $F_1$ -Werten für das BLSTM-CRF+ und das BLSTM-CNN-CRF ist mit 2,9 % bei Vorschriften und mit 3,6 % bei Straßen festzustellen. In den Abbildungen A.9, A.11, A.13 im Anhang sind die Konfusionsmatrizen für diese Modelle dargestellt, die die Relation zwischen den annotierten und vorhergesagten Klassen besser illustrieren.

Klasse	BLSTM-CRF			BLSTM-CRF+			BLSTM-CNN-CRF		
	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$
<b>PER</b>	89,30 %	91,08 %	90,09 %	90,78 %	92,24 %	<b>91,45 %</b>	90,21 %	92,57 %	91,35 %
<b>RR</b>	98,64 %	99,48 %	<b>99,05 %</b>	98,37 %	99,21 %	98,78 %	98,18 %	99,01 %	98,59 %
<b>AN</b>	94,85 %	84,62 %	<b>88,19 %</b>	86,18 %	90,59 %	87,07 %	88,02 %	87,96 %	87,11 %
<b>LD</b>	94,66 %	95,98 %	95,29 %	96,52 %	96,81 %	<b>96,66 %</b>	95,09 %	97,20 %	96,12 %
<b>ST</b>	81,26 %	86,32 %	83,48 %	82,58 %	89,06 %	<b>85,60 %</b>	83,21 %	87,95 %	85,38 %
<b>STR</b>	81,70 %	75,94 %	78,10 %	81,82 %	75,78 %	77,91 %	86,24 %	78,21 %	<b>81,49 %</b>
<b>LDS</b>	78,54 %	79,08 %	77,57 %	78,50 %	80,20 %	78,25 %	80,93 %	81,80 %	<b>80,90 %</b>
<b>ORG</b>	79,50 %	74,72 %	76,89 %	82,70 %	80,18 %	81,28 %	84,32 %	81,00 %	<b>82,51 %</b>
<b>UN</b>	85,81 %	81,34 %	83,44 %	90,05 %	88,11 %	89,04 %	91,72 %	89,18 %	<b>90,39 %</b>
<b>INN</b>	88,88 %	90,91 %	89,85 %	89,99 %	92,40 %	91,17 %	90,24 %	92,23 %	<b>91,20 %</b>
<b>GRT</b>	97,49 %	98,33 %	97,90 %	97,72 %	98,24 %	<b>97,98 %</b>	97,52 %	98,34 %	97,92 %
<b>MRK</b>	78,34 %	73,11 %	75,17 %	83,04 %	76,25 %	<b>79,17 %</b>	83,48 %	73,62 %	77,79 %
<b>GS</b>	96,59 %	97,01 %	96,80 %	98,34 %	98,51 %	<b>98,42 %</b>	98,44 %	98,38 %	98,41 %
<b>VO</b>	82,63 %	72,61 %	77,08 %	92,29 %	92,96 %	<b>92,58 %</b>	91,00 %	91,09 %	90,98 %
<b>EUN</b>	90,62 %	89,79 %	90,18 %	92,16 %	92,63 %	<b>92,37 %</b>	91,58 %	92,29 %	91,92 %
<b>VS</b>	75,58 %	68,91 %	71,77 %	85,14 %	78,87 %	<b>81,63 %</b>	79,43 %	78,30 %	78,74 %
<b>VT</b>	87,12 %	85,86 %	86,48 %	92,00 %	92,64 %	<b>92,31 %</b>	90,78 %	92,06 %	91,40 %
<b>RS</b>	96,34 %	96,47 %	<b>96,41 %</b>	96,70 %	96,73 %	96,71 %	97,04 %	97,06 %	<b>97,05 %</b>
<b>LIT</b>	93,87 %	93,68 %	93,77 %	94,34 %	93,94 %	94,14 %	94,25 %	94,22 %	<b>94,23 %</b>
<b>insg.</b>	93,80 %	93,70 %	93,75 %	95,36 %	95,57 %	<b>95,46 %</b>	95,34 %	95,58 %	<b>95,46 %</b>

**Tabelle 5.3.** Precision, Recall und  $F_1$ -Werte der BLSTM-Modelle für die feinkörnigen Klassen.

Unter den typischen Klassen der Eigennamen haben die Klassen der Richter **RR**, Länder **LD** und Gerichte **GRT** zu den besten Ergebnissen geführt. Die  $F_1$ -Werte betragen hierbei ca. 99 %, 96 % und 98 %. Das BLSTM-CRF+ und BLSTM-CNN-CRF konnten beide sowohl bei den Klassen der Personen **PER** als auch der Unternehmen **UN** sowie der Institutionen **INN** einen  $F_1$ -Wert von rund 90 % erreichen. Das BLSTM-CRF erzielte dagegen vergleichbare Werte nur für die Klassen der Personen und Institutionen. Der Unterschied in den  $F_1$ -Werten für Unternehmen betrug zwischen dem BLSTM-CRF und den Modellen mit Buchstabeneinbettungen fast 7 %. Ebenfalls hatte das BLSTM-CRF eine schlechtere Performanz in Bezug auf Organisationsnamen, sodass das BLSTM-CRF+ sowie das BLSTM-CNN-CRF einen  $F_1$ -Wert von 81–83 % und das BLSTM-CRF von nur 77 % erzielte. Alle drei Modelle erreichten bei der Erkennung der Anwaltsnamen **AN** ca. 87–88 % und der Stadtnamen **ST** ca. 83–86 %. Weitere Ortsnamen aus den Klassen der Straßen **STR** und Landschaften **LDS**, die in Entscheidungen seltener vorkamen, wurden schlechter erkannt und die Modelle erzielten hier 78–81 %.

Bei der Klasse der Markennamen **MRK** erreichten alle drei BLSTM-Modelle die schlechtesten Ergebnisse im Vergleich zu den anderen typischen Klassen (im Sinne von Personen, Orten, Organisationen). Der  $F_1$ -Wert überschritt nie 80 %. Weiterhin wichen Precision und Recall stärker voneinander ab. Mit dem BLSTM-CNN-CRF betrug der Unterschied fast 10 %. Höhere Precision- und niedrigere Recall-Werte wurden auch bei der Klasse **AN** mit dem BLSTM-CRF und bei der Klasse **STR** mit allen Modellen festgestellt. Im Falle von **ST** war aber eine gegensätzliche Korrelation zu finden. Dies bedeutet, dass mehr Stadtnamen gefunden wurden, was ihrer tatsächlichen Anzahl nicht entsprach.

Die beste Performanz bei den Klassen, die Namen und Zitate beinhalteten, wurde bei der größten Klasse **GS** und bei der zweitgrößten Klasse **RS** erzielt, was einem  $F_1$ -Wert von ca. 98 % und ca. 97 % entspricht. Das Ergebnis in der Klasse der Rechtsliteratur **LIT** war um einiges niedriger und betrug sogar ca. 94 %. Bei der Klasse der Vorschriften **VS** sank der  $F_1$ -Wert mit dem BLSTM-CRF bis auf 72 %, mit dem BLSTM-CRF+ bis auf 82 % und mit dem BLSTM-CNN-CRF bis auf 79 % ab. Das war das schlechteste Ergebnis unter den Rechtsnormen und Einzelfallregelungen. Darüber hinaus erreichte das BLSTM-CRF bei der Klasse der Rechtsverordnungen **VO** 77 %, was sich von den Ergebnissen mit dem BLSTM-CRF+ und dem BLSTM-CNN-CRF drastisch unterschied. Ihre  $F_1$ -Werte beliefen sich auf ca. 93 % und 91 %. Bei der Klasse der Verträge **VT** wurde wiederum ein sehr gutes Ergebnis erzielt, welches mit den BLSTM-Modellen mit Buchstabeneinbettungen ca. 92 % betrug. Mit dem BLSTM-CRF war allerdings der  $F_1$ -Wert um ca. 6 % niedriger. Die Erkennung der Europäischen Normen **EUN** war mit den drei Modellen sehr gut. Ihre  $F_1$ -Werte lagen bei 90–92 %.

Klasse	BLSTM-CRF			BLSTM-CRF+			BLSTM-CNN-CRF		
	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$
<b>PER</b>	94,34 %	95,16 %	94,74 %	94,82 %	96,03 %	<b>95,41 %</b>	94,09 %	96,21 %	95,12 %
<b>LOC</b>	90,85 %	92,59 %	91,68 %	92,60 %	94,05 %	<b>93,31 %</b>	91,74 %	93,45 %	92,57 %
<b>ORG</b>	91,82 %	90,94 %	91,37 %	92,87 %	92,89 %	92,87 %	93,80 %	92,65 %	<b>93,21 %</b>
<b>NRM</b>	97,04 %	96,50 %	96,77 %	97,93 %	98,04 %	<b>97,98 %</b>	97,71 %	97,87 %	97,79 %
<b>REG</b>	86,79 %	84,15 %	85,43 %	90,72 %	90,53 %	<b>90,61 %</b>	90,11 %	90,80 %	90,43 %
<b>RS</b>	96,54 %	96,58 %	96,56 %	96,93 %	97,05 %	<b>96,99 %</b>	96,73 %	96,83 %	96,78 %
<b>LIT</b>	93,78 %	93,91 %	93,84 %	94,23 %	94,62 %	<b>94,42 %</b>	94,24 %	93,80 %	94,02 %
<b>insg.</b>	94,86 %	94,49 %	94,68 %	95,84 %	96,07 %	<b>95,95 %</b>	95,71 %	95,87 %	95,79 %

**Tabelle 5.4.** Precision, Recall und  $F_1$ -Werte der BLSTM-Modelle für die grobkörnigen Klassen.

Im Vergleich zu den feinkörnigen Klassen erzielten die BLSTM-Modelle ein wenig bessere Ergebnisse mit den grobkörnigen Klassen, welche in Tabelle 5.4 dargestellt sind (für weitere Informationen s. die Konfusionsmatrizen in den Abbildungen A.10, A.12, A.14 im Anhang). Die  $F_1$ -Werte stiegen für das BLSTM-CRF um 0,9 %, für das BLSTM-CRF+ um 0,5 % und für das BLSTM-CNN-CRF um 0,3 % an. Die besten Ergebnisse für alle Klassen lieferte jedoch das BLSTM-CRF+ mit 95,95 % und war dadurch um 0,2 % besser als das BLSTM-CNN-CRF und um 1,27 % besser als das BLSTM-CRF. Darüber hinaus hatte das BLSTM-CRF+ die höheren Precision-, Recall sowie  $F_1$ -Werte in den Klassen der Personen, Orte, Rechtsnormen, Rechtsprechungen und Rechtsliteratur. Das BLSTM-CNN-CRF hatte aber einen Vorteil in der Organisationsnamenerkennung, der einem um 0,3 % höheren  $F_1$ -Wert entsprach. Insgesamt lässt sich feststellen, dass Precision und Recall bei allen Klassen und Modellen im Gegensatz zu den feinkörnigen Klassen ausbalancierter waren.

Die besten Ergebnisse konnten bei den Klassen **NRM** und **RS** erreicht werden. Die  $F_1$ -Werte betrugen ca. 98 % und 97 %. Die Personennamenerkennung ist auch gut ausgefallen (nicht zuletzt dank der Richternamen). Der  $F_1$ -Wert lag bei ca. 95 %. Die  $F_1$ -Werte der Orts- und Organisationsnamenerkennung waren wiederum sehr gut und beliefen sich auf 91–93 %. Bei der Klasse der Einzelfallregelungen **REG** erzielten alle BLSTM-Modelle im Vergleich zu den anderen Klassen das schlechteste Ergebnis. Die Werte schwankten zwischen ca. 85–91 %.

## 5.5 Analyse der richtigen und falschen Vorhersagen

Wie aus den Ergebnissen der CRF- und BLSTM-Modelle hervorgeht, wurden Namen und Zitate der bestimmten semantischen Klassen unterschiedlich erfolgreich erkannt. Daraus lässt sich auch die Tendenz feststellen, dass bei bestimmten Klassen stets schlechtere Ergebnisse erzielt wurden, die nicht durch die Modelle bedingt sind. Eine genaue Analyse kann Gesetzmäßigkeiten in Bezug auf falsche Klassifikationen sowie Schwachstellen des Korpus aufdecken, welche Vorschläge für die Performanzsteigerung abzuleiten helfen. Darüber hinaus wird auch kurz über korrekte Vorhersagen spezifische Klassen bzw. Entitäten betreffend diskutiert, die für Entscheidungen typisch waren. Einfachheitshalber werden im Weiteren Fälle besprochen und Sätze durchgeführt, die regelmäßig richtig bzw. falsch klassifiziert wurden. Zu beachten ist, dass alle Namen und Zitate in den Beispielsätzen so markiert sind, wie sie von den CRFs oder BLSTMs vorhergesagt wurden

(für weitere Informationen über die richtig oder falsch vorhergesagten Entitäten, die dieser Analyse zugrunde liegen, s. die Konfusionsmatrizen A.3, A.5, A.7, A.9, A.11, A.13 im Anhang).

Die Klasse der Personen **PER** war in erster Linie von der Anonymisierung betroffen. Das waren die Namen der Kläger, der Angeklagten, der Zeugen und der Experten. Jedoch haben die CRFs sowie BLSTMs gute Ergebnisse erzielt. Dies lässt sich dadurch erklären, dass ein Name der Bezeichnung dieser Person in Bezug auf ihre Rolle im gerichtlichen Prozess („der Angeklagte“, „der Mitangeklagte“, „der Nebenkläger“ o. Ä.) öfters folgte, wie im Beispiel (18) gezeigt ist. Ebenfalls kamen Namen in Kombination mit solchen Worten wie „Frau“, „Herr“, „Dr.“ vor, so wie im Beispiel (19). Zum Teil wurden auch diejenigen Personennamen korrekt vorhergesagt, die durch Auslassungspunkte ersetzt waren (z. B. in Phrasen „Herr ...“, „unter dem Pseudonym ...“, „Kommandeur ...“).

(18) ... dass der Angeklagte M. **PER** der Geschädigten das Telefon entriss ...

(19) Herr **N PER** wurde ... an drei Tagen beschäftigt.

Einige anonymisierte Personennamen wurden jedoch nicht identifiziert und als Gattungsnamen klassifiziert. In anderen Fällen wurden sie als Ortsnamen erkannt, falls davor Präpositionen verwendet wurden (vgl. (20, 21)), die als lokale Präpositionen interpretiert wurden. Falsche Vorhersagen gab es auch unveränderte Namen betreffend, also z. B. Vornamen (Margit in „Margit Weidner“, Siegfried in „Siegfried Buback“) oder Nachnamen im Genitiv („Melzacks“, „Nassauers“). Es lässt sich vermuten, dass diese Namen im Training nicht gelernt wurden.

(20) Später habe er in R. **ST** s Büro eine Wanze installiert.

(21) ... Geschehen endete, als die Großmutter nach H. **ST** rief.

Bei der Klasse der Richternamen **RR** zeigten dagegen die CRFs sowie BLSTMs eines der besten Ergebnisse mit einem  $F_1$ -Wert von mehr als 97 %. Diese präzise Erkennung war mit der Form und dem Inhalt der Entscheidungen, insbesondere mit dem Gebrauch der Richternamen in ähnlichen Kontexten verbunden. Jedes Gericht hat bestimmte Gesetze und Vorgaben, die schriftliche Erfassung der Urteile, der Beschlüsse und der Anordnungen regeln. So hatten Entscheidungen eines Bundesgerichtes bestimmte Ähnlichkeiten hinsichtlich der Form und des Inhalts sowie des Satzbaus und der Wortwahl. In den Entscheidungen, die aus dem Bundesarbeitsgericht und Bundespatentgericht stammen, standen die Namen der Richter, die dabei mitgewirkt hatten, im Rubrum<sup>6</sup> (in (22)). In den Entscheidungen aus dem Bundesgerichtshof waren die Titel und Namen der Richter am Ende des Dokumentes als einzelne Sätze aufgelistet. Selten wurden aber Worte als Richternamen falsch klassifiziert, die in ähnlichen Kontexten vorkamen. Dies betraf Einwortsätze („Sportbrille“, „Softfeil“) sowie kurze Sätze („Potsdam, den 9. März 2012 J.“).

(22) ... durch die Richterin Lachenmayr-Nikolaou **RR** als Vorsitzende, den Richter Paetzold **RR** und den Richter Dr. Himmelmann **RR** beschlossen:

Die Klasse der Anwälte **AN** war die kleinste im juristischen Korpus und fast alle Namen waren anonymisiert. Trotz dieser Tatsachen ist ihre Erkennung mit den BLSTM-Modellen gut ausgefallen. Dies kann man dadurch erklären, dass das Wort „Rechtsanwalt“ oder „Rechtsanwältin“, seltener „Dr.“, in den meisten Fällen einem Anwaltsnamen voranging, wobei diese Gesetzmäßigkeit im Training gelernt wurde. Die meisten Fehler bestanden darin, dass der Name eines Anwalts als der einer Person erkannt wurde.

Im Vergleich zu den anderen Ortsnamen wurde bei der Klasse der Länder **LD** ein um 10–20 % besseres Ergebnis erzielt. Erstens waren Ländernamen im Korpus gut abgedeckt. Zweitens belief sich die Anzahl der Wortformen, genannt Types, auf ca. 100 Einheiten. Mitunter wurden Länder als Städte falsch identifiziert, falls Namen anonymisiert waren oder falls es um Stadtstaaten wie Berlin, Hamburg oder Bremen ging. Bei der Klasse der Städte **ST** konnte ebenfalls ein gutes Ergebnis erzielt werden. Fast die Hälfte aller Stadtnamen folgte der lokalen Präposition („in“, seltener „von“, „nach“) sowie dem Wort „Stadt“. Anonymisierte Stadtnamen kollidierten jedoch mit Personen- und Ländernamen. Dies war der Fall mit „F.“ und „...“ in den Beispielen (23, 24), wobei andere Namen in diesen Sätzen richtig erkannt wurden.

(23) ... mit einer Fahrt von F. **PER** nach J. **ST** und zurück ...

<sup>6</sup> Als Rubrum wird der Urteilskopf bezeichnet. Das Rubrum besteht aus der Bezeichnung der Parteien, der Prozessbevollmächtigten, des Gerichtes sowie der Richter.

(24) ... K. C. **PER** aus B. **ST**, geboren im ... **LD** in Dagestan **LD** ...

Die Erkennung in den Klassen der Straßen **STR** und Landschaften **LDS** als Unterklassen der Ortsnamen war unter diesen am schlechtesten ausgefallen. Dies kann auf ihre Anzahl im Korpus zurückgeführt werden, welche sich auf höchstens 200 Entitäten beschränkte. Trotz der Tatsache, dass ca. 65 % der Straßennamen mit dem Wort ‚Straße‘ bzw. ‚Str.‘ zusammengesetzt waren, wurden einige Namen wie ‚Merowingerstraße‘, ‚Corneliusstraße‘ und ‚Kstraße‘ doch nicht erkannt. Fehlerhafte Vorhersagen gab es auch für Straßennamen, die durch Auslassungspunkte anonymisiert wurden. Was Landschaftsbezeichnungen angeht, gehören dazu per se verschiedene geografische Namen (Meere, Inseln, Flüsse, Seen, Gebirge, Täler, Wälder usw.). Diese Vielfältigkeit konnte das juristische Korpus kaum abdecken.

Bei der Klasse der Gerichtsnamen **GRT** (im Vergleich zu allen Organisationsnamen) erzielten die CRF- und BLSTM-Modelle das beste Ergebnis. Der häufigste Fehler bezog sich auf die falsche Abgrenzung der Gerichtsnamen, die entweder nicht bekannt oder selten im Korpus vorkamen, wie in (25). Darüber hinaus gab es Fehlklassifikationen, in denen Gerichtsnamen als Teil eines Rechtsprechungssitzes in (26) interpretiert und klassifiziert wurden.

(25) ... der Finanzverfassungsrechtler und ehemalige Richter im Zweiten Senat des **O** Bundesverfassungsgerichts **GRT** Paul Kirchhof **RR** ...

(26) Mit seiner am 1.2.2016 vor dem Sozialgericht (SG) Darmstadt **RS** erhobenen Klage ...

Das nächstbeste Ergebnis wurde bei der Klasse **INN** festgestellt. Häufige Fehler gab es, wenn eine Ortsangabe aus dem Institutionenamen, die eingebettet war, erkannt wurde und ein Institutionenname dagegen nicht. Das war der Fall z. B. mit ‚Bremen‘ in ‚das 62. Kommissariat der Polizei Bremen‘, mit ‚L.‘ in ‚die Ausländerbehörde der Stadt L.‘ u. Ä. In einigen Fällen wurden anonymisierte Institutionenamen als Personennamen falsch klassifiziert (s. (27)). Schließlich betrafen viele Fehlklassifikationen benachbarte Unterklassen, wenn **ORG**, **INN**, **UN** als Eigennamen von anderen Klassen angesehen wurden. So wurde eine Versorgungseinrichtung, also eine Institution des Landes, ‚die Baden-Württembergische Versorgungsanstalt für Ärzte, Zahnärzte und Tierärzte‘ und ein Unternehmen ‚Hausärztliche Vertragsgemeinschaft Aktiengesellschaft‘ als Organisationen erkannt und eine Organisation ‚BVKJ-Service GmbH‘ als ein Unternehmen.

(27) Die Klägerin und Beschwerdeführerin (Klägerin) ist umsatzsteuerrechtlich Organgesellschaft des Organträgers **N** **PER**.

Im Vergleich zu den Werten in den Klassen **UN** und **INN** erzielten alle Modelle bei der Klasse **ORG**, die Namen der öffentlichen Einrichtungen widerspiegeln, das schlechtere Ergebnis. Eine solche Erkennung war nicht durch die Abdeckung im Korpus bzw. die Anzahl der anonymisierten Namen bedingt. Die Anzahl der Entitäten war in diesen drei Klassen annähernd gleich und die Anzahl der anonymisierten Organisationsnamen betrug 15 %. Es gab zwar abgekürzte Bezeichnungen, ihr Anteil war aber nicht groß. Meistens wurden Organisationsnamen, die vermutlich im Lernprozess nicht teilnahmen und unbekannt waren, im Testverfahren als Gattungsnamen erkannt.

(28) Der österreichische Hanreich-Verlag **O** verlegt ein Buch mit dem Titel ‚Pfiffige Rezepte für kleine und große Leute‘ ...

Die semantische Klasse der Markennamen entstand als eine „experimentelle“ Klasse während der Konstruktion des Korpus, insbesondere bei der Annotation der Entscheidungen aus dem Bundespatentgericht, welche das Thema der Patente und Marken zu Diskussion stellten. Sehr häufig kamen Markennamen in einer sogenannten linguistischen Analyse zum Ausdruck, die versuchte, den Sinngehalt einer Wortmarke (meistens in einem Rechtsstreit mit der Widerspruchsmarke) zu erschließen und dadurch die Verwendung dieser Bezeichnung zu rechtfertigen. Dies hatte zur Folge, dass eine Bezeichnung in verschiedenen Kontexten geäußert und analysiert wurde, so z. B. die Wortmarke ‚Adamas‘ als Phantasiewort und Ortsbezeichnung; ‚Limmer‘ von der Wortmarke ‚Limmer Kontor‘ gleichzeitig in Bezug auf einen Stadtteil in Hannover sowie auf einen Ortsteil der Stadt Alfeld, des Marktes Gangkofen und der Gemeinde Hörgertshausen; die Wortmarke ‚Einstein’s Garage‘ und der Wissenschaftler Albert Einstein usw. Das Problem der Erkennung bestand deswegen darin, je nach Kontext korrekt Markennamen von Gattungsnamen sowie von anderen Eigennamen zu unterscheiden. Falsche Vorhersagen betrafen entweder Markennamen wie in (29), die als Gattungsnamen eingeordnet wurden, oder Gattungsnamen wie in (30), die dagegen als Markennamen klassifiziert wurden.

- (29) Bei „iPhone“ **O**, „i Drive“ **O** oder „iCar“ **O** werde auch nicht an ... „Internet“ gedacht.
- (30) Die weiteren Zeichenbestandteile der angegriffenen Marke „Abwehr“ **MRK** und „Spray“ **MRK** sind ... hinsichtlich der meisten Waren und Dienstleistungen in keiner Weise beschreibend.

Die Klasse der Gesetze **GS** war die größte Klasse im juristischen Korpus. Zudem wurden diese Entitäten durch eine einheitliche Form gekennzeichnet und teilten sich in Voll-, Kurzzitate und Bezeichnungen auf. Falsche Vorhersagen wurden getroffen, falls Gesetzeszitate eine abweichende Form hatten. So wurde der Paragraph des ‚Gesetzes über die Versorgungsanstalt für Ärzte, Zahnärzte und Tierärzte (VersAnstG)‘ in (31) nicht erkannt sowie das zwölfte Kapitel des neunten Sozialgesetzbuches in (32). Es gab zudem Schwierigkeiten mit Ausdrücken, die das Wort ‚Gesetz‘ enthielten, wie ‚Bundesgesetz‘, ‚Erfahrungsgesetz‘, ‚Strafgesetz‘, ‚Abfindungsgesetz‘ usw. Diese wurden dann als Gesetzesnamen klassifiziert.

- (31) ... richtet sich die Teilnahme an der Versorgungsanstalt nach § 7 des Gesetzes (gemeint ist das **O** **VersAnstG GS**).
- (32) ... im Sinne des zwölften Kapitels des **O** SGB XII **GS** gilt ...

Falsche Vorhersagen betreffend die Klasse der Rechtsverordnungen **VO** bestanden hauptsächlich darin, dass sie mit Gesetzen oder Verträgen verwechselt wurden, wobei ihre Abgrenzung zu Gattungsnamen korrekt war (vgl. (33, 34)). Es gab auch Probleme mit der Erkennung einiger Namen der Rechtsverordnungen wie z.B. ‚Kfz-Gruppenfreistellungsverordnung 2002‘, ‚Schwerbehindertenausweisverordnung‘, ‚Erschwerniszulagenverordnung‘, welche als Gattungsnamen klassifiziert wurden.

- (33) ... dort zu einer Pflegesatzvereinbarung auf der Grundlage des § 17 Abs 1 Bundespflegesatz-verordnung **GS** ...
- (34) Es ist zutreffend als Wortmarke im Sinne von § 7 MarkenV **VT** angemeldet und erfasst worden ...

In der Klasse der Europäischen Normen **EUN** gab es ebenfalls Fehler, sodass Zitate nicht korrekt von Gattungsnamen abgegrenzt wurden. Dies war wahrscheinlich mit ihrer großen Länge im Vergleich zu anderen Normen verbunden. So wurde die Phrase ‚über die Gemeinschaftsmarke‘ in ‚Verordnung (EG) Nr. 40/94 über die Gemeinschaftsmarke‘ als Teil des Zitats nicht identifiziert. Die Bezeichnung ‚Verordnung ... über Marktmissbrauch (Marktmissbrauchsverordnung) und zur Aufhebung der Richtlinie 2003/6/EG ...‘ wurde dagegen in zwei Teile getrennt, also ‚Verordnung ... über Marktmissbrauch (Marktmissbrauchsverordnung) und ‚Richtlinie 2003/6/EG ...‘. Es ist zu betonen, dass die schließende Klammer dabei nicht identifiziert wurde. Darüber hinaus wurden auch einige Namen der Europäischen Normen wie ‚Abfallrahmenrichtlinie‘, ‚Emissionshandelsrichtlinie‘, ‚UN-Behindertenrechtekonvention‘ usw., die im Fließtext vorkamen, nicht erkannt.

In der Klasse der Vorschriften **VS** wurde das schlechteste Ergebnis unter allen Rechtsnormen und Einzelfallregelungen beobachtet. Diese Tatsache lässt sich so begründen, dass erstens ihre Anzahl unter allen Rechtsnormen und Einzelfallregelungen im Korpus am geringsten war. Ihre Bezeichnungs- bzw. Zitierweise war zweitens uneinheitlich, was der Erkennung der betroffenen Namen und Zitate entgegenstand. Diesbezüglich gab es Schwierigkeiten bei ihrer Identifikation, sodass einige Vorschriften als Institutionen, Gesetze, Verträge und Literatur wie in (35) erkannt bzw. unter Gattungsnamen wie in (36) eingeordnet wurden. Das gleiche Problem betraf die Klasse der Verträge **VT**, welche mit Rechtsnormen kollidierten, wie in (37) gezeigt. Jedoch beruhten falsche Vorhersagen meistens darauf, dass Verträge wie in (37) nicht identifiziert wurden.

- (35) Runderlass (RdErl) des **O** Hessischen Ministeriums der Justiz, für Integration und Europa **INN** ...
- (36) ... „als integraler Bestandteil des Einsatzes“ grundsätzlich verpflichtende, siehe Nr. 2 Satz 1 Zentralerlass B-2640/8 **O** ...
- (37) Weiter heißt es in § 9 Abs. 1 Satz 2 des Altersteilzeitarbeitsvertrags **GS** : ...
- (38) ... Einführung des einheitlichen Entgeltsystems für Arbeiter und Angestellte durch das Entgeltarabkommen **O** (ERA **O**) ...

In der Klasse der Rechtsprechungen **RS** gab es einige Fälle, in denen Zitate falsch abgegrenzt wurden. Dies war wie bei Gesetzen damit verbunden, dass einige Zitate eine andere Form hatten und deswegen nicht korrekt erkannt wurden. Zum Einen wurden Grenzen wie in (39) zu eng und zum Anderen wie in (40) zu weit bestimmt. Im Beispiel (41) wurde dagegen der Name des Gerichtes und das Wort ‚Urteil‘ als Zitat falsch klassifiziert.

- (39) Auch eine Mehrdeutigkeit im Sinne der Link **O** economy-Entscheidung des BGH (GRUR – 2012, 270) **RS** liegt hier nicht vor.
- (40) Das Arbeitsgericht Zwickau **GRT** verurteilte die Beklagte am 22. April 2015 (–9 Ca 146/15–) **RS** ...
- (41) Die Beschwerde gegen die Nichtzulassung der Revision in dem Urteil des 4. Zivilsenats des Oberlandesgerichts Frankfurt am Main **RS** vom 1. Juni 2016 wird auf Kosten des Klägers zurückgewiesen.

Bei der Klasse der Rechtsliteratur **LIT** erzielten die Modelle wiederum ein gutes Ergebnis. Fehler bezogen sich auf die Identifikation der zitierten Gesetzesmaterialien, Entwürfe bzw. Entwurfsbegründungen (s. (42)). Gewisse Schwierigkeiten gab es bei der Erkennung von Autoren, deren Namen mit dem Adelsprädikat ‚von‘ begannen (z. B. ‚von Heintschel-Heinegg‘, ‚v. Schönfeld‘).

- (42) ... die Begründung des Regierungsentwurfs zum **O** Bewertungsänderungsgesetz **GS** von 1965 zu **O** § 27 BewG **GS**, **O** BTDrucks IV/1488, S. 39 **LIT** ...

Aus der durchgeföhrten Analyse in den feinkörnigen Klassen lässt sich resümieren, dass die hohe Abdeckung, die einheitliche Form der Entitäten, der gleiche Kontext sowie die geringe Anzahl an Types die Performance direkt positiv beeinflusst. Wie oben bereits beschrieben, betrifft dies die Klassen der Richter, Länder, Gerichte, Gesetze, Rechtsprechungen und Rechtsliteratur. Andere Klassen haben im Gegensatz ein verscheidenartiges Verbesserungspotenzial, das man teils durch eine Korpusweiterung, teils durch automatische Verfahren realisieren kann. Schlechtere Ergebnisse sind durch die begrenzte Abdeckung und die starke Heterogenität der betroffenen Entitäten in einer Klasse bedingt. Dies bezieht sich in erster Linie auf die Klassen der Straßen, Landschaften, Organisationen, Marken sowie Vorschriften.

## 5.6 Diskussion

Wie die durchgeföhrten Experimente zeigen, haben die BLSTM-Modelle im Vergleich zu den CRF-Modellen eine bessere Performance erzielt, sodass die  $F_1$ -Werte zwischen 93,75–95,46 % für die feinkörnigen Klassen und zwischen 94,68–95,95 % für die grobkörnigen Klassen lagen. Die CRF-Modelle konnten 93,05–93,23 % und 93,11–93,22 % erreichen. Dabei wurden einige Klassen durch größere Unterschiede in der Precision und dem Recall gekennzeichnet, was auf gewisse Schwächen der CRFs in der Erkennung von Namen und Zitaten deutet. Insgesamt haben die CRF-Modelle im Gegensatz zu den BLSTMs ca. 1–10 % niedrigere Werte pro Klasse erzielt. Besonders hat sich die Erkennung der Straßen- und Markennamen mit den BLSTM-Modellen verbessert. Die  $F_1$ -Werte stiegen um mindestens 10 % an. Die Werte in den Klassen der Anwälte, Landschaften und Rechtsverordnungen haben sich auch um die 5 % erhöht.

Die Ergebnisse der CRF-Modelle haben offenbart, dass die Verkettung von Features, Namenslisten und der Lookup-Tabelle eine kleine Steigerung der Werte in bestimmten Klassen gebracht hat. Das beste Ergebnis konnte in den Klassen der Anwälte, Institutionen, Gerichte, Verträge und Rechtsprechungen mit Features festgestellt werden. Die Umsetzung der Namenslisten (für die Erkennung der Personen, Orte, Organisationen und Rechtsnormen) hat einen eindeutigen Einfluss auf das Ergebnis in den Klassen der Richter, Länder, Städte, Landschaften und Marken ausgeübt. Die Straßennamenerkennung hat sich dagegen verschlechtert und die  $F_1$ -Werte sind um 2 % abgesunken. Die weitere Verkettung mit der Lookup-Tabelle für die Wortähnlichkeit konnte positiv zur Erkennung der Personen-, Länder-, Straßen-, Organisations-, Unternehmens-, Markennamen sowie der Landschaftsbezeichnungen weiter beitragen. Davon hat auch die Erkennung der Namen und Zitate von Rechtsnormen profitiert.

Mit der Vergrößerung der feinkörnigen Klassen wurde die erstaunliche Erkenntnis gewonnen, dass die CRF-Modelle gleich gut mit feinkörnigen sowie mit grobkörnigen Klassen umgehen. Der Unterschied für die

maximalen F<sub>1</sub>-Werte betrug lediglich 0,01 %. Vor der Bewertung der Ergebnisse wurde jedoch vermutet und erwartet, dass sich die kleinere Anzahl von semantischen Kategorien positiv auf die Performanz auswirkt. Anscheinend konnten auch die ausgewählten Features und externen Quellen nicht genug zur Performanzsteigerung dieser Modelle beitragen. Ebenso waren z. B. die Namenslisten wenig hilfreich für die Erkennung der anonymisierten Personen-, Orts- und Organisationsnamen. Im Rahmen der vorliegender Arbeit, insbesondere wegen des großen zeitlichen Aufwands, der für das Testen im 10-fachen Kreuzvalidierungsverfahren benötigt wurde, wurde eine begrenzte Anzahl von Features und Quellen gewählt. Um die Performanz der CRF-Modelle signifikant zu verbessern und auf ein vergleichbares Niveau mit den BLSTMs zu bringen, ist es nötig, mehrere Kombinationen von Features und Quellen (z. B. Cluster, POS-Tags usw.) zu testen.

Im Gegensatz zu den CRF-Modellen hat sich die Performanz der unterschiedlichen Kombinationen der BLSTM-Modelle wesentlich voneinander unterschieden. Das beste Ergebnis wurde mit dem BLSTM-CRF-Modell in den Klassen der Richter und Anwälte erzielt. Die Erkennung von anderen Namen und Zitaten hat sich mit der Hinzunahme der Buchstabeneinbettungen verbessert. Besonders haben davon die Klassen der Organisationen, Unternehmen, Marken und Verträge profitiert. Der größte Unterschied zwischen den F<sub>1</sub>-Werten wurde jedoch bei der Erkennung der Rechtsverordnungen und Vorschriften mit dem BLSTM-CRF und BLSTM-CRF+ festgestellt (15,5 % und 9,86 %). Mit diesen Befunden wurde die Behauptung bestätigt, dass Sequenzmodelle dank Buchstabeneinbettungen sehr gut mit neuen, unbekannten Wörtern umgehen können, zu welchen die Bezeichnungen bzw. Zitate der Rechtsverordnungen und Vorschriften gehören. Eine weitere auffallende Erkenntnis war die wesentliche Steigerung der F<sub>1</sub>-Werte um ca. 3 % mit dem BLSTM-CNN-CRF, die in den Klassen der Straßen und Landschaften beobachtet wurde. Unklar bleibt, warum das CNN (im Vergleich zu dem BLSTM) die Informationen aus dem Buchstabenniveau für diese Klassen besser verarbeiten konnte.

Im Falle der BLSTMs hat die Vergrößerung der feinkörnigen Klassen dagegen zu einer Steigerung bzw. einem Ausgleich der Precision-, Recall- und F<sub>1</sub>-Werte beigetragen. Der F<sub>1</sub>-Wert hat sich um 0,5 % im Vergleich zu dem besten Ergebnis in den feinkörnigen Klassen, das mit dem BLSTM-CRF+ und BLSTM-CNN-CRF erreicht wurde, verbessert. Insgesamt wurde ein F<sub>1</sub>-Wert von mehr als 90 % mit allen Klassen erreicht. Hohe, ausgeglichene Precision-, Recall- und F<sub>1</sub>-Werte sind vor allem damit verbunden, dass sich die marginalen Unterklassen (im Sinne der schlechten Abdeckung im Korpus und der schlechten Ergebnisse) mit den verwandten zentralen Unterklassen mittels Mikro-Metriken ausbalancierten. Trotzt der guten Ergebnisse besteht darüber hinaus ein Nachteil der Vergrößerung darin, dass viele Informationen hinsichtlich der betroffenen feinkörnigen semantischen Kategorien verloren gehen.

Dank des Experimentdesigns haben die Forschungsergebnisse und die nachfolgende Analyse gezeigt, dass die beiden Modelfamilien gewisse Gesetzmäßigkeiten in Bezug auf die Performanz besitzen, welche auf das Korpus bzw. die Struktur der Daten zurückzuführen sind. Zuallererst ist dies mit der Größe und der unausgeglichenen Klassenverteilung verbunden. Die Modelle haben stets die besten Ergebnisse in den feinkörnigen Klassen der Richter, Gerichte und Gesetze gezeigt. Ihre F<sub>1</sub>-Werte lagen bei 95 %. Die Eigennamen- und Zitaterkennung war in den Klassen der Länder, Institutionen, Rechtsprechungen und Rechtsliteratur schlechter, aber wiederum sehr gut und betrug einen F<sub>1</sub>-Wert über 90 %. Die Werte in den Klassen der Straßen, Landschaften, Organisationen und Vorschriften waren dagegen stets am niedrigsten und beliefen sich auf 69–80 % mit den CRF-Modellen und auf 72–83 % mit den BLSTM-Modellen. Das schlechteste Ergebnis wurde jedoch bei der Klasse der Marken geliefert. Mit den CRF-Modellen konnte ein maximaler F<sub>1</sub>-Wert von 69,61 % und mit den BLSTM-Modellen ein maximaler F<sub>1</sub>-Wert von 79,17 % erreicht werden. Die Unterschiede in der Erkennung gewisser Eigennamen und Zitate waren mit der Spezifika der Rechtstexte verbunden, insbesondere waren sie durch die Abdeckung im Korpus, die Heterogenität bezüglich der Form sowie durch den Kontext bedingt.

Diese Ergebnisse führen zu dem Schluss, dass die Struktur der Daten, die in den Lernprozess involviert sind, sehr stark die Performanz der Modelle beeinflusst, sodass die technische Verbesserung der automatischen Verfahren wenig sinnvoll ist. Von großer Bedeutung ist in dieser Hinsicht die Erkenntnis, dass das Verbesserungspotenzial für die betroffenen marginalen Klassen in der Korpusweiterung bzw. -optimierung liegt. Die Experimente mit den verschiedenen Modellen haben auch gezeigt, dass es kein allgemeingültiges Modell gibt, mit dem alle Klassen am besten erkannt werden könnten. Ein optimales System für die Eigennamen- und Zitaterkennung könnte sich aus einer Kombination verschiedener Modelle, die bestimmte Klassen vorhersagen, zusammensetzen.

# 6 Zusammenfassung und Ausblick

## 6.1 Zusammenfassung

Die vorliegende Arbeit befasste sich mit der Erkennung der semantischen Konzepte in Rechtstexten. Es ging um die Identifikation und Klassifikation von Eigennamen und Zitaten in gerichtlichen Entscheidungen, welche im Vergleich zu anderen Rechtstexten genug vielfältige Entitäten enthalten. In Bezug auf gewisse Limitierungen der wissenschaftlichen Forschung in diesem Bereich, nämlich das Fehlen von einer einheitlichen, anerkannten Typologie der semantischen Konzepte sowie von frei verfügbaren Korpora, wurden Klassen definiert, die für Rechtstexte und insbesondere für Entscheidungen typisch sind. Das sind für die Eigennamenerkennung übliche Klassen der Personen, Orte sowie Organisationen, die sich in feinkörnige Klassen (Richter, Anwälte, Länder, Städte, Institutionen, Gerichte usw.) unterteilen. Dazu gehören weitere Unterklassen bestehend aus Rechtsnormen, Einzelfallregelungen, Rechtsprechungen und Rechtsliteratur, die für Entscheidungen charakteristisch sind. Weiterhin wurde ein juristisches Korpus entwickelt, in dem gerichtliche Entscheidungen mit betroffenen semantischen Kategorien nach erarbeiteten einheitlichen Annotationsrichtlinien manuell annotiert wurden.

Um das Erkennen der Eigennamen und Zitate mithilfe automatischer Verfahren zu realisieren, wurden die Problemstellung analysiert und wissenschaftliche Ansätze, die dem aktuellen Forschungsstand entsprachen, verglichen. Aus den gewonnenen Erkenntnissen über die Performanz verschiedener Verfahren wurde entschieden, Modelle basierend auf Conditional-Random-Fields und bidirektionalen Long-Short-Term-Memory-Netzen einzusetzen. In den Experimenten wurden jeweils drei CRF- sowie BLSTM-Modelle mit den fein- und grobkörnigen Klassen getestet.

Die BLSTM-Modelle haben eine absolute Überlegenheit bewiesen. Sie konnten gute Ergebnisse sogar mit den Klassen liefern, die im Korpus schlecht abgedeckt waren. Die Varianten der BLSTM-CRFs, die durch das BLSTM bzw. CNN die Buchstabeneinbettung repräsentiert haben, konnten das beste Ergebnis für die 19 feinkörnigen Klassen mit einem  $F_1$ -Wert von 95,46 % erzielen. Mit den sieben grobkörnigen Klassen zeigte das BLSTM-CRF mit Buchstabeneinbettungen die beste Performanz. Der  $F_1$ -Wert belief sich auf 95,95 %. Ein besonders signifikanter Unterschied ließ sich in den Klassen der Straßen und Marken bei dem Vergleich der BLSTMs-Modelle zu den CRF-Modellen beobachten. Anhand der Forschungsergebnisse wurde auch festgestellt, dass die Erkennung in einigen Klassen wesentlich vom BLSTM und CNN, die jeweils die Buchstabeneinbettungen bearbeiten, abhängt. Die Klassen der Marken, Rechtsverordnungen und Vorschriften haben von der Verwendung des CNNs profitiert. Die Straßen- und Landschaftsnamenerkennung hat sich aber mit den Buchstabeneinbettungen aus dem BLSTM verbessert.

Die CRF-Modelle haben ausschließlich bei Namen und Zitaten aus den zentralen Klassen eine gute, ausbalancierte Performanz erzielt. Dazu gehörten Gesetze, Rechtsprechungen und Rechtsliteratur, die durch die hohe Abdeckung im Korpus gekennzeichnet waren, sowie Richter, die meistens im gleichen Kontext vorkamen. Der Einsatz der zusätzlichen Listen für die Wortähnlichkeit sowie für die Erkennung der Personen, Orte, Organisationen und Rechtsnormen hat zu einer minimalen Steigerung geführt.

Dank der durchgeführten Analyse wurde anhand von richtig und falsch erkannten Entitäten festgestellt, welche Faktoren die Performanz in der Eigennamen- und Zitaterkennung positiv oder negativ beeinflussen. So bedingt die schlechte Abdeckung im Korpus, die Vielfältigkeit der Eigennamen bzw. eine uneinheitliche Form der Zitate eine unpräzise Erkennung. Dagegen werden Entitäten aus den Klassen, die gut abgedeckt und einheitlich sind, besser identifiziert und klassifiziert. Dies haben die Ergebnisse in den Klassen der Richter, Gerichte, Gesetze und Länder bestätigt, welche mit den CRF- sowie mit den BLSTM-Modellen besonders präzise erkannt wurden.

Daraus lässt sich schlussfolgern, dass die vorliegende Arbeit einen guten Ausgangspunkt für die zukünftige Forschung der Eigennamen- und Zitaterkennung in Rechtstexten bietet. Insbesondere leistet das juristische Korpus einen wichtigen Beitrag, sodass neue Verfahren in der juristischen Domäne problemlos getestet werden können. Außerdem kann das beste der erarbeiteten Modelle von verschiedenen Unternehmen, die sich mit Rechtstexten beschäftigen, zur automatischen Textanalyse praktisch eingesetzt werden.

## 6.2 Ausblick

Einerseits folgen Anknüpfungspunkte für die zukünftige Forschung im Bereich der Eigennamen- und Zitaterkennung in Rechtstexten aus Limitierungen der vorliegenden Arbeit. Zur Performanzverbesserung der getesteten Modelle können sowohl die optimierten Hyperparameter als auch die zusätzlichen Features bzw. externen Quellen beitragen. Für die BLSTM-Modelle wäre auch eine Verbesserung mit einer Entwicklung der Worteinbettungen möglich, die auf deutschen Rechtstexten bzw. Entscheidungen trainiert werden.

Andererseits bieten die gewonnenen Erkenntnisse Anknüpfungspunkte für weitere wissenschaftliche Arbeiten an. Die Messwerte der getesteten Modelle deuten darauf hin, dass die Häufigkeit der Eigennamen und Zitate im Korpus ein aussagekräftiges Merkmal ist. Um die Performanz der Sequenzmodelle zu verbessern, kann man, wie mehrmals angedeutet wurde, das Korpus erweitern und bezüglich der Klassenverteilung ausbalancieren. Dies kann in erster Linie diejenigen Klassen betreffen, deren Vertreter selten vorkommen (wie Anwalts-, Straßen-, Markennamen und Landschaftsbezeichnungen).

Darüber hinaus ermöglicht es die feinkörnige Annotation der semantischen Kategorien, verschiedenartige Manipulationen und Veränderungen durchzuführen, so z. B. die Identifikation auf Eigennamen zu begrenzen oder gewisse Unterklassen auf die eigene Forschung bezogen zu vergrößern bzw. zu löschen. Interessant wäre es auch, die Qualität der annotierten Daten im juristischen Korpus mittels der Bewertung der Übereinstimmung mit anderen Annotatoren (engl. *inter annotator agreement*) zu prüfen.

## Literaturverzeichnis

- Baldwin, T., de Marneffe, M., Han, B., Kim, Y., Ritter, A., und Xu, W. (2015). Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In Xu, W., Han, B., und Ritter, A., Herausgeber, *Proceedings of the Workshop on Noisy User-generated Text, NUT@IJCNLP 2015, Beijing, China, July 31, 2015*, Seiten 126–135. Association for Computational Linguistics.
- Bender, O., Och, F. J., und Ney, H. (2003). Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, Seiten 148–151. Association for Computational Linguistics.
- Bengio, Y., Simard, P. Y., und Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Networks*, 5(2):157–166.
- Benikova, D., Biemann, C., Kisseelew, M., und Padó, S. (2014a). GermEval 2014 Named Entity Recognition Shared Task: Companion Paper. In *Proceedings of the KONVENS GermEval workshop*, Seiten 104–112, Hildesheim, Germany.
- Benikova, D., Biemann, C., und Reznicek, M. (2014b). NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., und Piperidis, S., Herausgeber, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, Seiten 2524–2531. European Language Resources Association (ELRA).
- Benikova, D., Yimam, S. M., Santhanam, P., und Biemann, C. (2015). GermaNER: Free Open German Named Entity Recognition Tool. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015, University of Duisburg-Essen, Germany, 30th September - 2nd October 2015*, Seiten 31–38.
- Brown, P. F., Pietra, V. J. D., de Souza, P. V., Lai, J. C., und Mercer, R. L. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.
- Bundesministerium der Justiz (2008). Bekanntmachung des Handbuchs der Rechtsförmlichkeit. *Bundesanzeiger*, Jahrgang 60(160a):296.
- Busse, D. (2000). Textsorten des Bereichs Rechtswesen und Justiz. *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*, 1:658–675.
- Cardellino, C., Teruel, M., Alemany, L. A., und Villata, S. (2017). A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker. In *Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law, ICAIL '17*, Seiten 9–18, New York, NY, USA. ACM.
- Chiu, J. P. C. und Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *TACL*, 4:357–370.
- Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, Seiten 59–66. Association for Computational Linguistics.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., und Kuksa, P. P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.
- Daiber, J., Jakob, M., Hokamp, C., und Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, Seiten 121–124.
- Deutsch, A. (2017). 5. Schriftlichkeit im Recht: Kommunikations-formen/Textsorten. *Handbuch Sprache im Recht*, 12:91–117.

- Didakowski, J., Geyken, A., und Hanneforth, T. (2006). Eigennamenerkennung mit großen lexikalischen Resourcen. In *Proceedings of KONVENS 2006 (Konferenz zur Verarbeitung natürlicher Sprache) Universität Konstanz*, Seiten 9–14.
- Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., und Wudali, R. (2010). Named Entity Recognition and Resolution in Legal Text. In Francesconi, E., Montemagni, S., Peters, W., und Tiscornia, D., Herausgeber, *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, Band 6036 in *Lecture Notes in Computer Science*, Seiten 27–43. Springer.
- Eckart de Castilho, R., Mújdríca-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., und Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Hinrichs, E. W., Hinrichs, M., und Trippel, T., Herausgeber, *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities, LT4DH@COLING, Osaka, Japan, December 2016*, Seiten 76–84. The COLING 2016 Organizing Committee.
- Engberg, J. (1993). Prinzipien einer Typologisierung juristischer Texte. *Fachsprache: International Journal of Specialized Communication*, 15(1/2):31–38.
- Faruqui, M. und Padó, S. (2010). Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In Pinkal, M., Rehbein, I., im Walde, S. S., und Storrer, A., Herausgeber, *Semantic Approaches in Natural Language Processing: Proceedings of the 10th Conference on Natural Language Processing, KONVENS 2010, September 6-8, 2010, Saarland University, Saarbrücken, Germany*, Seiten 129–133. universaar, Universitätsverlag des Saarlandes / Saarland University Press / Presses universitaires de la Sarre.
- Finkel, J. R., Grenager, T., und Manning, C. D. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Knight, K., Ng, H. T., und Oflazer, K., Herausgeber, *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, Seiten 363–370. The Association for Computer Linguistics.
- Gers, F. A. und Schmidhuber, J. (2000). Recurrent Nets that Time and Count. In *IJCNN* (3), Seiten 189–194.
- Gers, F. A., Schmidhuber, J., und Cummins, F. A. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10):2451–2471.
- Glaser, I., Waltl, B., und Matthes, F. (2018). Named Entity Recognition, Extraction, and Linking in German Legal Contracts. *IRIS: Internationales Rechtsinformatik Symposium*, Seiten 325–334.
- Graves, A. und Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., und Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232.
- Grishman, R. und Sundheim, B. (1996). Message Understanding Conference- 6: A Brief History. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, Seiten 466–471.
- Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. *Diploma, Technische Universität München*, 91(1).
- Hochreiter, S. und Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.
- Huang, Z., Xu, W., und Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR*, abs/1508.01991.
- Jurafsky, D. und Martin, J. H. (2014). *Speech and language processing*, Band 3. Pearson London.
- Kjær, A. (1992). Normbedingte Wortverbindungen in der juristischen Fachsprache (Deutsch als Fremdsprache). *Fremdsprachen Lehren und Lernen*, 21:46–64.
- Klein, D., Smarr, J., Nguyen, H., und Manning, C. D. (2003). Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, Seiten 180–183. Association for Computational Linguistics.

- Lafferty, J. D., McCallum, A., und Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Brodley, C. E. und Danyluk, A. P., Herausgeber, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, Seiten 282–289. Morgan Kaufmann.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., und Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In NAACL HLT 2016, *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, Seiten 260–270.
- Landthaler, J., Waltl, B., und Matthes, F. (2016). Unveiling References in Legal Texts - Implicit versus Explicit Network Structures. *IRIS: Internationales Rechtsinformatik Symposium*, Seiten 71–78.
- Linguistic Data Consortium (2008). ACE (Automatic Content Extraction) English Annotation Guidelines for Entities.
- Ma, X. und Hovy, E. H. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
- Mayfield, J., McNamee, P., und Piatko, C. (2003). Named entity recognition using hundreds of thousands of features. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, Seiten 184–187. Association for Computational Linguistics.
- McCallum, A. und Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, Seiten 188–191. Association for Computational Linguistics.
- Mendes, P. N., Jakob, M., García-Silva, A., und Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. In Ghidini, C., Ngomo, A. N., Lindstaedt, S. N., und Pellegrini, T., Herausgeber, *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011, ACM International Conference Proceeding Series*, Seiten 1–8. ACM.
- Mikolov, T., Chen, K., Corrado, G., und Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., und Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., Bottou, L., Ghahramani, Z., und Weinberger, K. Q., Herausgeber, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, Seiten 3111–3119.
- Nadeau, D. und Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher: John Benjamins Publishing Company.
- Passos, A., Kumar, V., und McCallum, A. (2014). Lexicon Infused Phrase Embeddings for Named Entity Resolution. In Morante, R. und Yih, W., Herausgeber, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, Seiten 78–86. ACL.
- Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J., und Yangarber, R. (2017). The First Cross-Lingual Challenge on Recognition, Normalization, and Matching of Named Entities in Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Seiten 76–85. Association for Computational Linguistics.
- Proisl, T. und Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In Cook, P., Evert, S., Schäfer, R., und Stemle, E., Herausgeber, *Proceedings of the 10th Web as Corpus Workshop, WAC@ACL 2016, Berlin, August 12, 2016*, Seiten 57–62. Association for Computational Linguistics.
- Reimers, N., Eckle-Kohler, J., Schnober, C., Kim, J., und Gurevych, I. (2014). GermEval-2014: Nested Named Entity Recognition with Neural Networks. In Faaß, G. und Ruppenhofer, J., Herausgeber, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, Seiten 117–120. Universitätsverlag Hildesheim.

- Reimers, N. und Gurevych, I. (2017a). Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. *CoRR*, abs/1707.06799.
- Reimers, N. und Gurevych, I. (2017b). Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Seiten 338–348. Association for Computational Linguistics.
- Riedl, M. und Padó, S. (2018). A Named Entity Recognition Shootout for German. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Band 2, Seiten 120–125. Association for Computational Linguistics.
- Sander, G. G. (2004). *Deutsche Rechtssprache: ein Arbeitsbuch*, Band 2578. UTB.
- Sandrini, P. (1999). *Übersetzen von Rechtstexten: Fachkommunikation im Spannungsfeld zwischen Rechtsordnung und Sprache*. Narr.
- Sang, E. F. T. K. und Meulder, F. D. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Daelemans, W. und Osborne, M., Herausgeber, *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, Seiten 142–147. ACL.
- Sang, E. F. T. K. und Veenstra, J. (1999). Representing Text Chunks. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, Seiten 173–179. The Association for Computer Linguistics.
- Simonnaes, I. (2012). *Rechtskommunikation national und international im Spannungsfeld: von Hermeneutik, Kognition und Pragmatik*, Band 103. Frank & Timme GmbH.
- Simonnaes, I. (2015). *Basiswissen deutsches Recht für Übersetzer: mit Übersetzungsübungen und Verständnisfragen*, Band 122. Frank & Timme GmbH.
- Simonnaes, I. (2018). Terminologische Datenbanken als Verstehens-und Formulierungshilfe beim Übersetzen von Rechtstexten. *Parallèles*, 30(1):120–136.
- Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, Seiten 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tkachenko, M. und Simanovsky, A. (2012). Named entity recognition: Exploring features. In Jancsary, J., Herausgeber, *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, Band 5 in *Scientific series of the ÖGAI*, Seiten 118–127. ÖGAI, Wien, Österreich.
- Wiesmann, E. (2003). *Rechtsübersetzung und Hilfsmittel des Übersetzers: Entwurf und prototypische Entwicklung eines elektronischen Hilfsmittels für den Rechtsübersetzer*. Narr.
- Yadav, V. und Bethard, S. (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, Seiten 2145–2158. Association for Computational Linguistics.

## A Anhang

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <!DOCTYPE dokument
3   SYSTEM "http://www.rechtsprechung-im-internet.de/dtd/v1/rii-dok.dtd">
4 <dokument>
5   <doknr>WBRE201800396</doknr>
6   <ecli>DE:BVerwG:2018:260418B2B2.18.0</ecli>
7   <gertyp>BVerwG</gertyp>
8   <gerort/>
9   <spruchkoerper>2. Senat</spruchkoerper>
10  <entsch—datum>20180426</entsch—datum>
11  <aktenzeichen>2 B 2/18, 2 B 2/18 (2 C 5/18)</aktenzeichen>
12  <doktyp>Beschluss</doktyp>
13  <norm>§ 13 Abs 1 DiszG BE, §184b Abs 4 S 2 StGB, §132 Abs 2 Nr 1 VwGO</norm>
14  <vorinstanz>vorgehend Oberverwaltungsgericht Berlin—Brandenburg, 28. September 2017, Az: OVG 80 D 4.14, Urteil<br/>vorgehend
   VG Berlin, 18. September 2014, Az: 80 K 2.14 OL, Urteil<br/>
15 </vorinstanz>
16 <region>
17   <abk>DEU</abk>
18   <long>Bundesrepublik Deutschland</long>
19 </region>
20 <mitwirkung/>
21 <titelzeile>
22   <dl class="RspDL">
23     <dt/>
24     <dd>
25       <p>Zulassung der Revision; disziplinare Ahndung des Besitzes kinderpornographischer Schriften bei Lehrern</p>
26     </dd>
27   </dl>
28 </titelzeile>
29 <leitsatz/>
30 <tenor/>
31 <tatbestand/>
32 <entscheidungsgruende/>
33 <gruende>
34   <div>
35     <dl class="RspDL">
36       <dt>
37         <a name="rd_1">1</a>
38       </dt>
39       <dd>
40         <p>Die Beschwerde des Klägers ist begründet. Die Revision wird wegen grundsätzlicher Bedeutung der Rechtssache (§
           41 DiszG BE i.V.m. §69 BDG und §132 Abs. 2 Nr. 1 VwGO) zugelassen.</p>
41       </dd>
42     </dl>
43   <dl class="RspDL">
44     <dt>
45       <a name="rd_2">2</a>
46     </dt>
47     <dd>
48       <p>Das Revisionsverfahren erscheint geeignet, zur weiteren Klärung der Frage beizutragen, wie bei Lehrern das auß
           erdienstliche Dienstvergehen des Besitzes von kinderpornographischen Schriften unter Berücksichtigung des
           sonstigen außerdienstlichen Verhaltens des Lehrers disziplinarrechtlich zu ahnden ist.</p>
49     </dd>
50   </dl>
51   <dl class="RspDL">
52     <dt/>
53     <dd>
54       <p/>
55     </dd>
56   </dl>
57   </div>
58 </gruende>
59 <abwmeinung/>
60 <sonstlt/>
61 <identifier>http://www.rechtsprechung-im-internet.de/jportal/?quelle=jlink&docid=WBRE201800396&psml=bsjrsprod.
   psml&max=true</identifier>
62 <coverage>Deutschland</coverage>
63 <language>deutsch</language>
64 <publisher>BMJV</publisher>
65 <accessRights>public</accessRights>
66 </dokument>
```

**Abbildung A.1.** Quelldaten im XML-Format am Beispiel der Datei WBRE201800396.xml.

Die XML-Elemente sind blau und der Text ist schwarz markiert. Der Beschluss besteht aus einem Titel (Zeilen 21–28) und aus Gründen (Zeilen 33–58). Die Angaben zu einem Gericht, Entscheidungsdatum, Aktenzeichen, ECLI, Entscheidungstyp und zu Normen sind in den Zeilen 6–13 aufgelistet. Der Verfahrensgang findet sich in den Zeilen 14–15.

<b>Gericht:</b>	BVerwG 2. Senat	<b>Normen:</b>	§ 13 Abs 1 DiszG BE, § 184b Abs 4 S 2 StGB, § 132 Abs 2 Nr 1 VwGO
<b>Entscheidungs-</b>	26.04.2018		
<b>datum:</b>			
<b>Aktenzeichen:</b>	2 B 2/18, 2 B 2/18 (2 C 5/18)		
<b>ECLI:</b>	ECLI:DE:BVerwG:2018:260418B2B2.18.0		
<b>Dokumenttyp:</b>	Beschluss		

### Zulassung der Revision; disziplinare Ahndung des Besitzes kinderpornographischer Schriften bei Lehrern

#### Verfahrensgang

vorgehend Oberverwaltungsgericht Berlin-Brandenburg, 28. September 2017, Az: OVG 80 D 4.14, Urteil  
vorgehend VG Berlin, 18. September 2014, Az: 80 K 2.14 OL, Urteil

#### Gründe

- 1 Die Beschwerde des Klägers ist begründet. Die Revision wird wegen grundsätzlicher Bedeutung der Rechtssache (§ 41 DiszG BE i.V.m. § 69 BDG und § 132 Abs. 2 Nr. 1 VwGO) zugelassen.
- 2 Das Revisionsverfahren erscheint geeignet, zur weiteren Klärung der Frage beizutragen, wie bei Lehrern das außerdienstliche Dienstvergehen des Besitzes von kinderpornographischen Schriften unter Berücksichtigung des sonstigen außerdienstlichen Verhaltens des Lehrers disziplinarrechtlich zu ahnden ist.

© Ein Service des Bundesministeriums der Justiz und für Verbraucherschutz  
in Zusammenarbeit mit der juris GmbH - [www.rechtsprechung-im-internet.de](http://www.rechtsprechung-im-internet.de)

**Abbildung A.2.** Quelldaten im PDF-Format am Beispiel der Datei WBRE201800396.xml.

Die Bezeichnungen der Textabschnitte, die den XML-Elementen in Abbildung A.1 entsprechen, sind fett markiert. Der Titel ist zentriert und auch fett markiert. Die Angaben im Kasten und der Verfahrensgang korrespondieren mit den Zeilen 6–15 in Abbildung A.1.

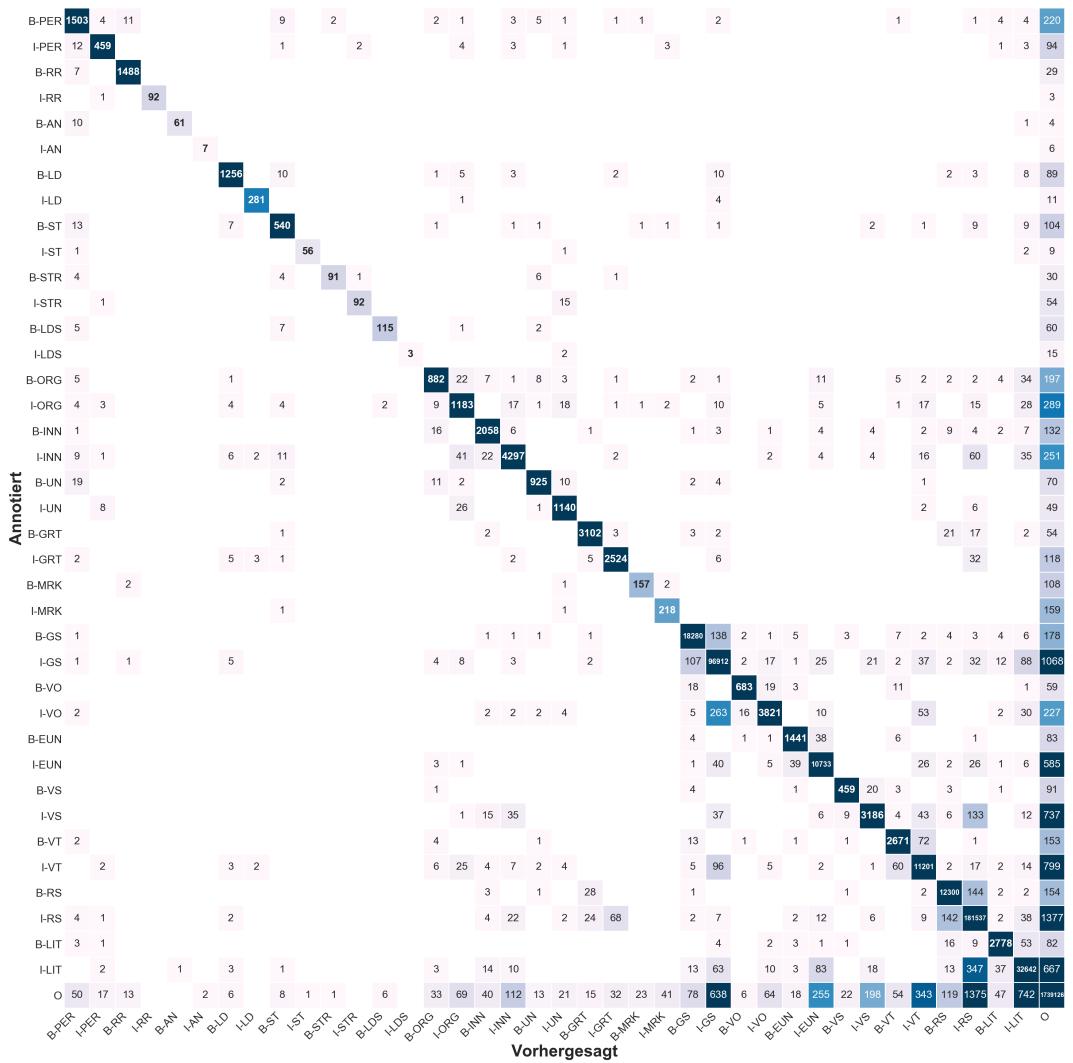


Abbildung A.3. Konfusionsmatrix des CRF-Fs für die feinkörnigen Klassen.

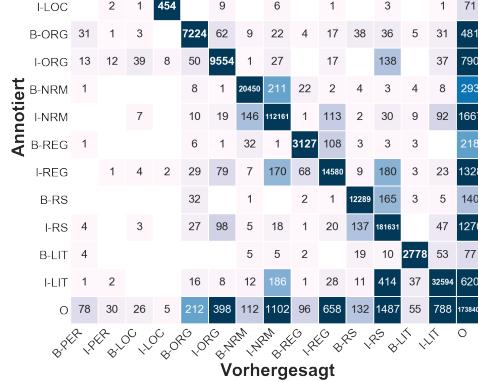


Abbildung A.4. Konfusionsmatrix des CRF-Fs für die grobkörnigen Klassen.

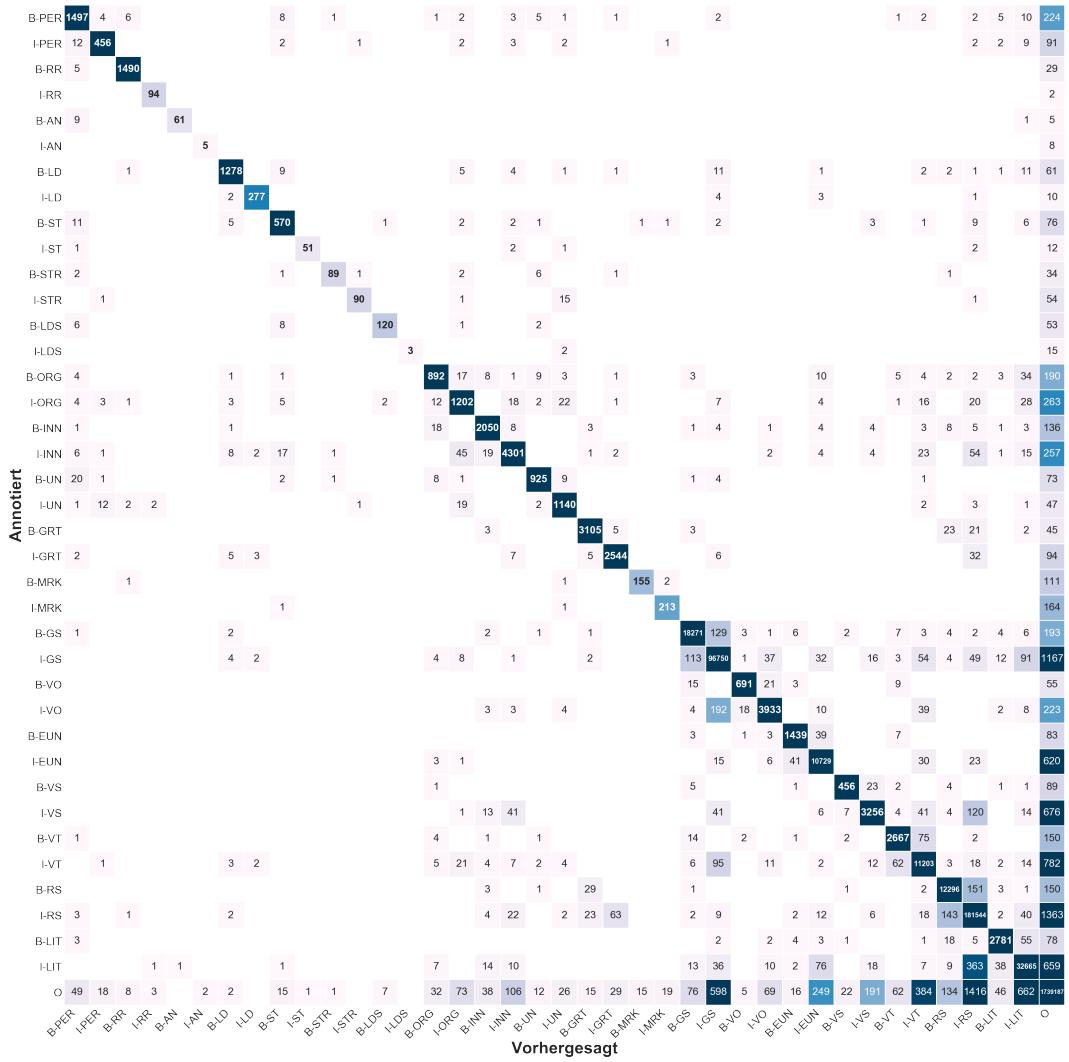


Abbildung A.5. Konfusionsmatrix des CRF-FGs für die feinkörnigen Klassen.

	B-PER	I-PER	B-LOC	I-LOC	B-ORG	I-ORG	B-NRM	I-NRM	B-REG	I-REG	B-RS	I-RS	B-LIT	I-LIT	O	
B-PER	3051	9	17		10	10	1	1	1	1	1	4	8	262		
I-PER	13	559	3	1	2	19					2	1	3	89		
B-LOC	24	1	2074		11	26		16		11	3	15	10	216		
I-LOC	1	3	4	450		9	4			4		3	70			
B-ORG	29	1	2		7230	58	9	17	4	19	32	35	5	38	485	
I-ORG	13	12	44	7	50	9558		31		32	120	48	771			
B-NRM	2				8		20449	202	20	2	4	3	4	9	304	
I-NRM	1	7			11	18	154	11296	115	2	37	10	101	1705		
B-REG	1				7	1	29	1	3128	100	3	5	3		225	
I-REG	1	4	2		26	73	5	139	74	14528	7	173	3	32	1416	
B-RS					32		2	1	2	3	12281	161	1	1	153	
I-RS	3	4	2	28	94	5	29	1	20	143	181562	16	1354			
B-LIT	2					4	5	1		18	9	2781	61	72		
I-LIT	1	1			15	4	12	168	29	11	382	34	3268	587		
O	73	29	31	5	198	368	112	1095	96	707	127	1395	48	767	173837	

Abbildung A.6. Konfusionsmatrix des CRF-FGs für die grobkörnigen Klassen.

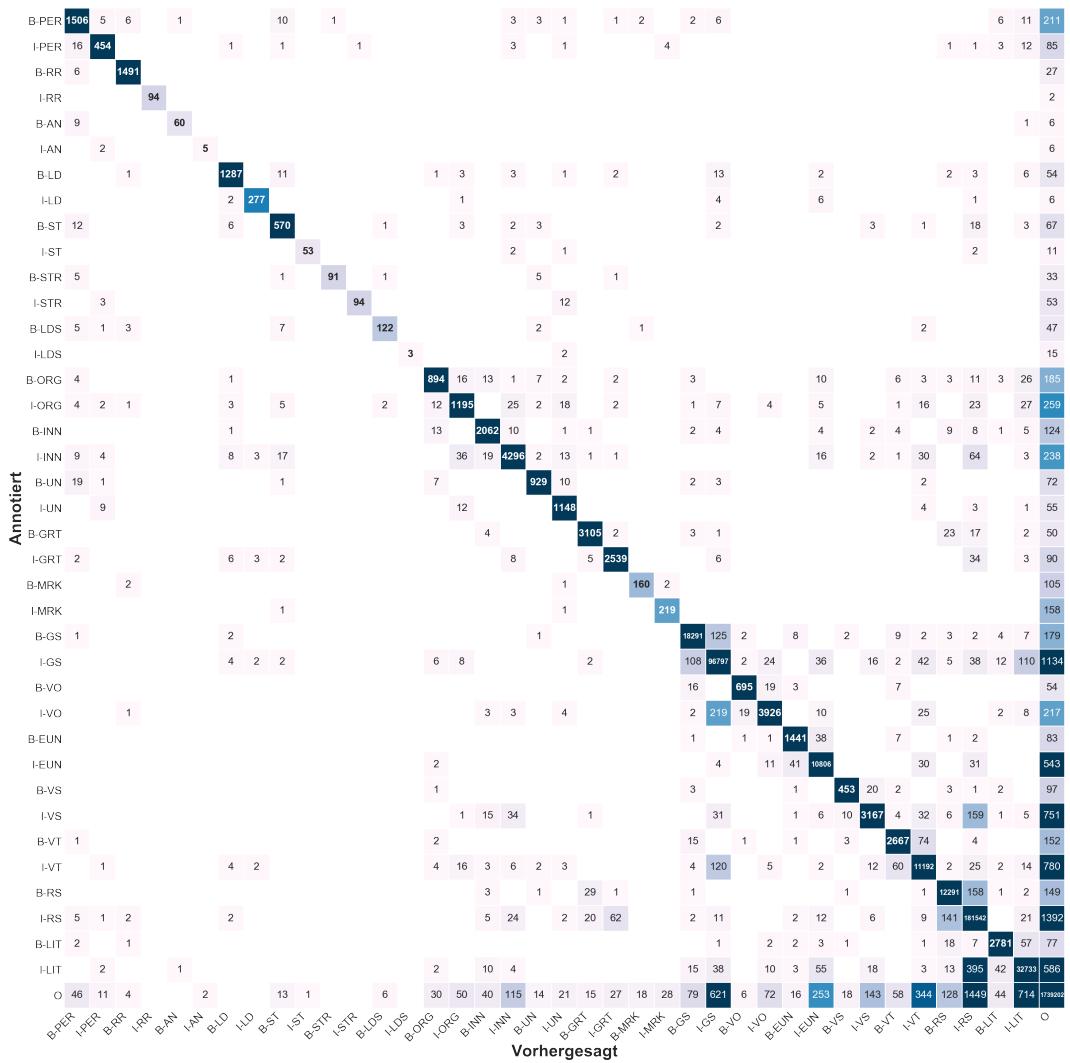


Abbildung A.7. Konfusionsmatrix des CRF-FGLs für die feinkörnigen Klassen.

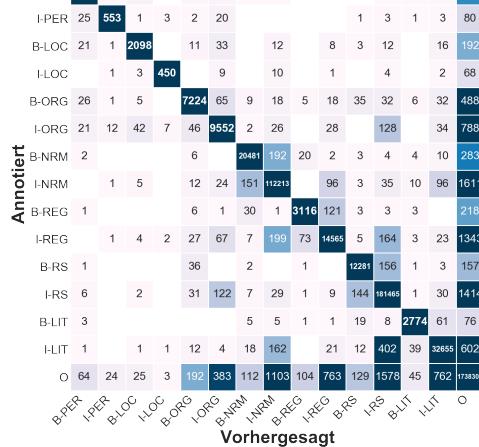
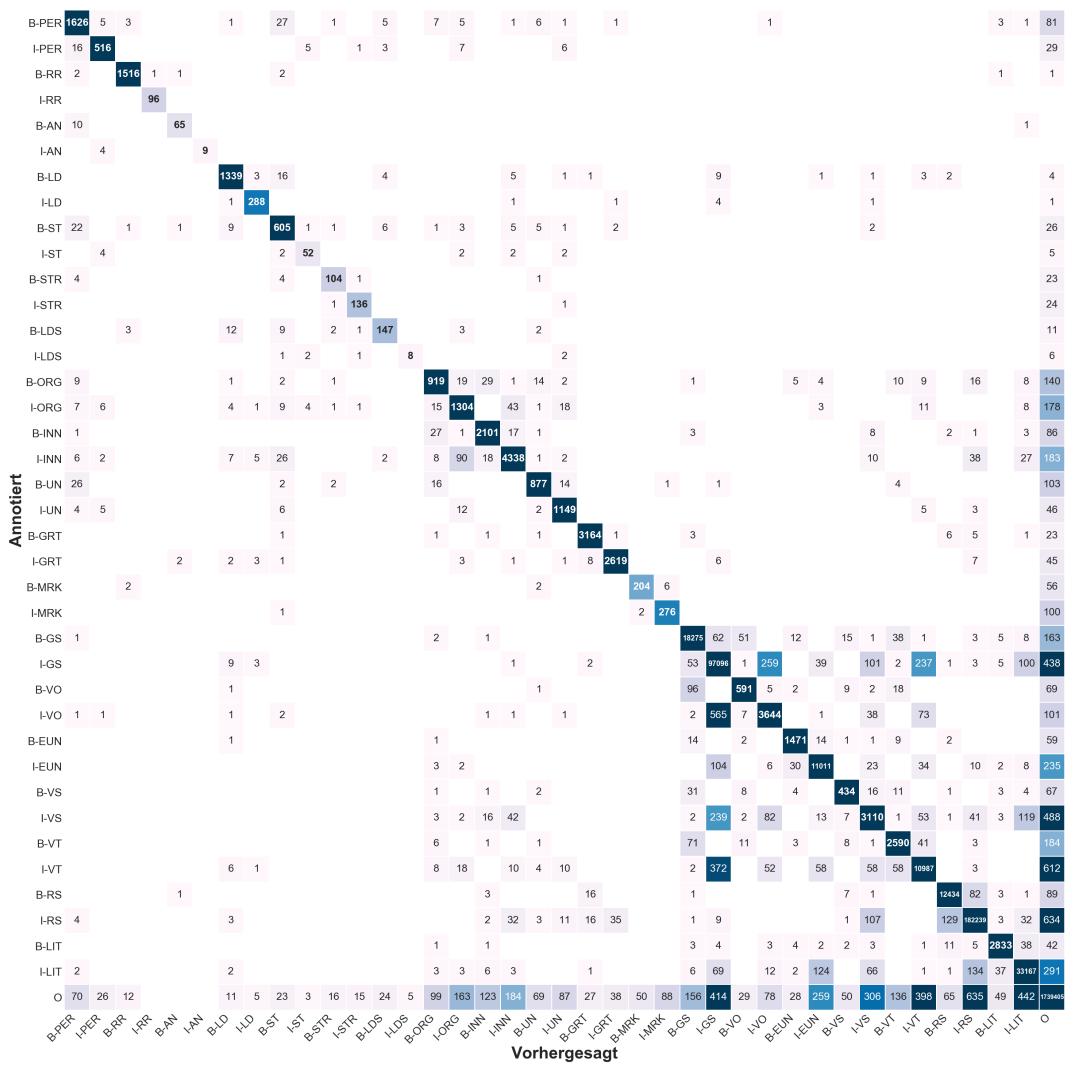
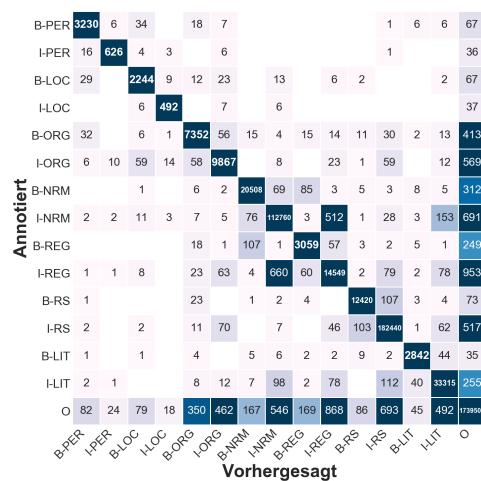


Abbildung A.8. Konfusionsmatrix des CRF-FGLs für die grobkörnigen Klassen.



**Abbildung A.9.** Konfusionsmatrix des BLSTM-CRFs für die feinkörnigen Klassen.



**Abbildung A.10.** Konfusionsmatrix des BLSTM-CRFs für die grobkörnigen Klassen.

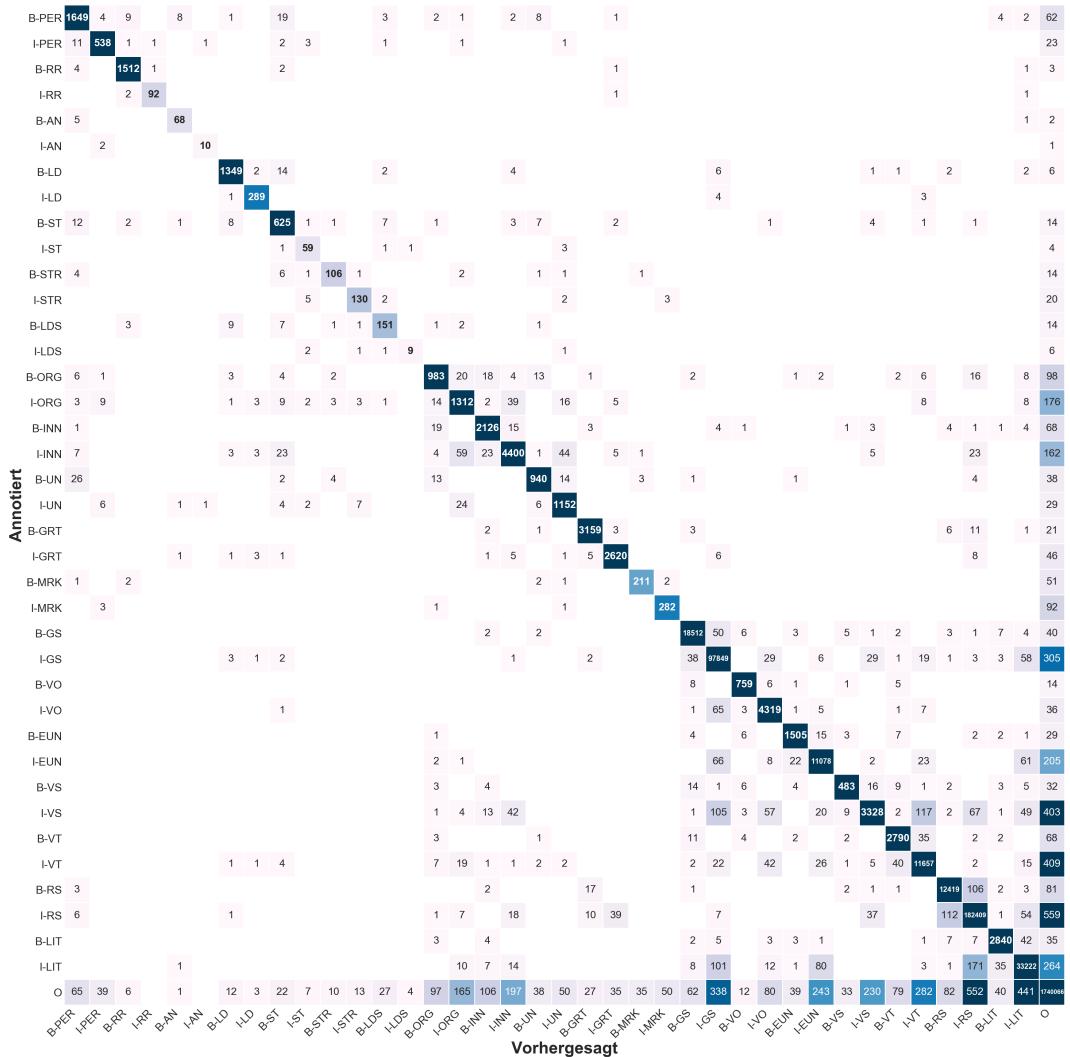


Abbildung A.11. Konfusionsmatrix des BLSTM-CRFs+ für die feinkörnigen Klassen.

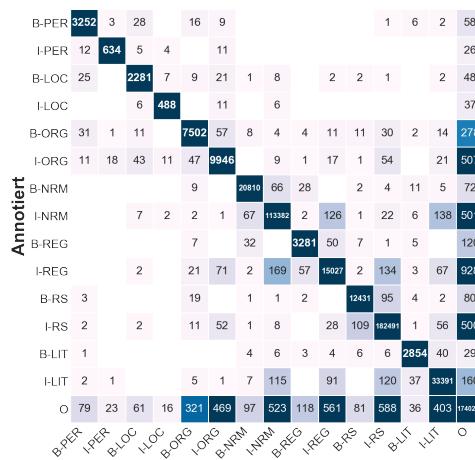


Abbildung A.12. Konfusionsmatrix des BLSTM-CRFs+ für die grobkörnigen Klassen.

	Vorhergesagt																																							
Annotiert	B-PER	I-PER	B-RR	I-RR	B-AN	I-AN	B-LD	I-LD	B-ST	I-ST	B-STR	I-STR	B-LDS	I-LDS	B-ORG	I-ORG	B-INN	I-INN	B-UN	I-UN	B-GRT	I-GRT	B-MRK	I-MRK	B-GS	I-GS	B-VO	I-VO	B-EUN	I-EUN	B-VS	I-VS	B-VT	I-VT	B-RS	I-RS	B-LIT	I-LIT	O	
B-PER	1663	3	10		3	4	19		3	2	1		2	7	1		2	1														4	1	49						
I-PER	12	531	2		1														5														2	28						
B-RR	7		1509	1																2														3						
I-RR		2	94																																					
B-AN	6		68																														1	1						
I-AN		4	1		8																																			
B-LD				1354	2	8											1	6															1	9						
I-LD					2	285											3	1																2						
B-ST	15	3	2	10	614	3			4	1	3	7	5	1	2																	16								
I-ST	2			59							1		3																				2							
B-STR	4			4	108	1				2	1	2																				14								
I-STR		5		127	1					2		2																				25								
B-LDS	1	3		8	9	1	153				1		1																		12									
I-LDS	1			2			1	7										1													8									
B-ORG	11			3	4			992	21	16	3	11	1							2											91									
I-ORG	2	12		2	2	8	5	1	1	10	1316	3	29	1	18					4										188										
B-INN										14	1	2123	22	1	1					2											67									
I-INN	1			5	1	31	3		2	3	42	21	4410	2	28					4	1	1	19						1	37	151									
B-UN	26			4					14	1	949	8			1															5	39									
I-UN	5			6	1				21	4	1154																				36									
B-GRT								2			3160	2							3												9	8	21							
I-GRT		2	3	3	2						4								7	2616											16		39							
B-MRK	2											2				202	3															61								
I-MRK																	269															109								
B-GS	1								1	3	2								18468	46	10	8	1	10	1				1	1	10	11	64							
I-GS	1			3	2	2				2	9		2					39	87619		37	65	53						1	9	3	128	375							
B-VO									1									6	746	7		2	12									20								
I-VO		1	1															1	45	3	4289		14	42								43								
B-EUN												1	2					3	9		1503	18	2	10							1	26								
I-EUN													1					1	27	36	28	11143		45							24	162								
B-VS													1	2					5	1	4	3	2	491	18	5	2	1				46								
I-VS													1	13	50				83	1	24	14	7	3549		50						4	3	97						
B-VT	1												6						15	5	4	5	2	2772	35								75							
I-VT	2				4	1							6	20					3	60	54	34	1	18	49	11599		1	14		427									
B-RS		1												3																		12437								
I-RS	2													23	1	2	14	43		1	7									111	162356									
B-LIT		1												2																		83								
I-LIT	1	1												5	4																	635								
O	68	33	6	1	1	16	1	18	7	13	12	24	2	89	127	106	197	27	31	30	32	36	54	68	261	18	91	27	189	49	382	109	332	68	487	39	404	1746133		

Abbildung A.13. Konfusionsmatrix des BLSTM-CNN-CRFs für die feinkörnigen Klassen.

	Vorhergesagt																												
Annotiert	B-PER	I-PER	B-LOC	I-LOC	B-ORG	I-ORG	B-NRM	I-NRM	B-REG	I-REG	B-RS	I-RS	B-LIT	I-LIT	O														
B-PER	3258	4	27		15	6													5	2	58								
I-PER	14	642	1	1		9																						25	
B-LOC	26		2271	6	8	23			7		6	2	2															54	
I-LOC	6		487		11		6												1	1	36								
B-ORG	50		7	7474	39	13	6	4	6	12	32	3	11															307	
I-ORG	12	21	40	6	62	9848			22		24		35							12	604								
B-NRM	1			5			20765	72	34	2	2	1	6	9	109														
I-NRM	9	3	4	1	66	11335	1	226	1	9	2	64																436	
B-REG				9	1	37	1	3272	59	5	3	4	5															107	
I-REG	1			15	45	6	246	51	1533	101	3	94															564		
B-RS	2		1		20		1		2	1	12430	87																91	
I-RS	6		2	13	62	2	11	1	42	120	182341																635		
B-LIT	2				2	10	11	3	2	3	7	2837	42	34															
I-LIT	2	1	2		7	11	11	255	1	79	1	102	38	33152	268														
O	78	17	26	236	428	132	682	160	918	76	540	42	538	1739632															

Abbildung A.14. Konfusionsmatrix des BLSTM-CNN-CRFs für die grobkörnigen Klassen.

## **Selbstständigkeitserklärung**

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

---

Potsdam, 25. Februar 2019  
Elena Leitner