Open Justice versus Privacy

77 NRP Digital Transformation National Research Programme

UNIVERSITÄT BERN

PD. Dr. Matthias Stürmer Forschungsstelle Digitale Nachhaltigkeit
Mag. rer. publ. Daniel Kettiger Kompetenzzentrum für Public Management (KPM)

Agenda:

I. Teil: Übersicht über das Forschungsprojekt

- Kontext des Projekts: Publikation von Urteilen
- Inhalte, Aufbau und Ablauf

II. Teil: Schwerpunkt technische Realisierung

- Tool zur Verifikation der Anonymisierung von Gerichtsurteilen (De-Anonymisierung)
- Natural Language Processing (NLP) Methoden
- Knowledge Graph Construction

UNIVERSITÄT BERN

Kontext: Publikation von Urteilen (1)

Justizöffentlichkeit
(Art. 30 Abs. 3 BV; Art. 6 Ziff. 1 EMRK; Art. 14 Ziff. 1 UNO-Pakt II):

- Verfahrensöffentlichkeit (Öffentlichkeit der Verhandlung)
- Urteilsöffentlichkeit (Öffentlichkeit der Urteilsverkündung)

UNIVERSITÄT BERN

Kontext: Publikation von Urteilen (2)

Urteilsöffentlichkeit (Urteil BGer 1C_123/2016 vom 21.06.2016)

«[3.5.1] Öffentliche Urteilsverkündung bedeutet, dass am Schluss eines gerichtlichen Verfahrens das Urteil in Anwesenheit der Parteien sowie von Publikum und Medienvertretern verkündet wird. Darüber hinaus dienen weitere Formen der Bekanntmachung dem Verkündungsgebot, wie etwa öffentliche Auflage, **Publikation** in amtlichen Sammlungen oder **Bekanntgabe über das Internet**. Sie sind im Einzelnen anhand von Sinn und Zweck des Verkündungsgebots daraufhin zu beurteilen, ob sie die verfassungsrechtlich gebotene Kenntnisnahme gerichtlicher Urteile erlauben

[3.6] Die weiteren Formen der Bekanntgabe von Urteilen (vgl. E. 3.5.1 a.E.) sind nicht subsidiär, sondern gehören angesichts der Zweckausrichtung gleichwertig zur öffentlichen Verkündung. ….»

UNIVERSITÄT BERN

Kontext: Publikation von Urteilen (3)

- Leuchtturmprojekt des Vereins eJustice.CH zur Zugänglichkeit kantonaler Urteile >> Befragung 2017 zur Urteilspublikation
- Workshop des Vereins eJustice.CH vom 29.01.2019 «Anonymisierung von Urteilen»
- D. Hürlimann/D. Kettiger (Hrsg.): Anonymisierung von Urteilen (2021)
- Die Anonymisierung von Urteilen erfolgt heute in der Schweiz händisch oder teil-automatisiert in einem Word-Dokument.
- Anonymisierungsaufwand pro Urteil: durchschnittlich 45-50 Minuten

UNIVERSITÄT BERN

Inhalte, Aufbau und Ablauf (1)

Zielsetzungen

- Klärung von rechtlichen und sozialwissenschaftlichen Fragen rund um die Anonymisierung von Urteilen
- Erstellen eines Tools zur automatisierten Anonymisierung von Urteilen
- Hilfestellungen für Gerichte zur Anonymisierung von Urteilen

Inhalte, Aufbau und Ablauf (2)

Working Package 1 «Recht»

- Dissertation: Publikation von Gerichtsurteilen im Spannungsfeld zwischen Transparenz und Geheimhaltungsinteressen (Monographie, inkl. Rechtsvergleichung)
- Masterarbeit: Staatshaftung für fehlende bzw. mangelhafte Anonymisierung von Urteilen

Working Package 2 «Informatik»

- Tool zur De-Anonymisierung von Gerichtsurteilen >> Erfahrungen De-Anonymisierung
- Tool zur Anonymisierung von Gerichtsurteilen

UNIVERSITÄT BERN

Inhalte, Aufbau und Ablauf (3)

Working Package 3 «Sozial- und Politikwissenschaften»

- «Stakeholder Opinions»: Wie stehen Peer-Groups und die Bevölkerung zur Anonymisierung von Urteilen?

Working Package 4 «Synthese und Implementation»

- Synthesebericht aus Working Packages 1-3
- evtl. Pilotprojekt Anonymisierungstool
- Wegleitung für Gerichte

UNIVERSITÄT BERN

Inhalte, Aufbau und Ablauf (4)

Beteiligte Institute der Universität Bern

- Kompetenzzentrum für Public Management (KPM): Prof. Dr. Andreas Lienhard, Magda Chodup, Tania Munz, Daniel Kettiger
- Forschungsstelle Digitale Nachhaltigkeit am Institut für Informatik (INF): PD Dr. Matthias Stürmer, Joel Niklaus
- Institut für Wirtschaftsinformatik (IWI): Prof. Dr. Thomas Myrach

Teilprojekt des Nationalen Forschungsprogramms NFP 77 Digitale Transformation

77 NRP Digital Transformation National Research Programme

UNIVERSITÄT BERN

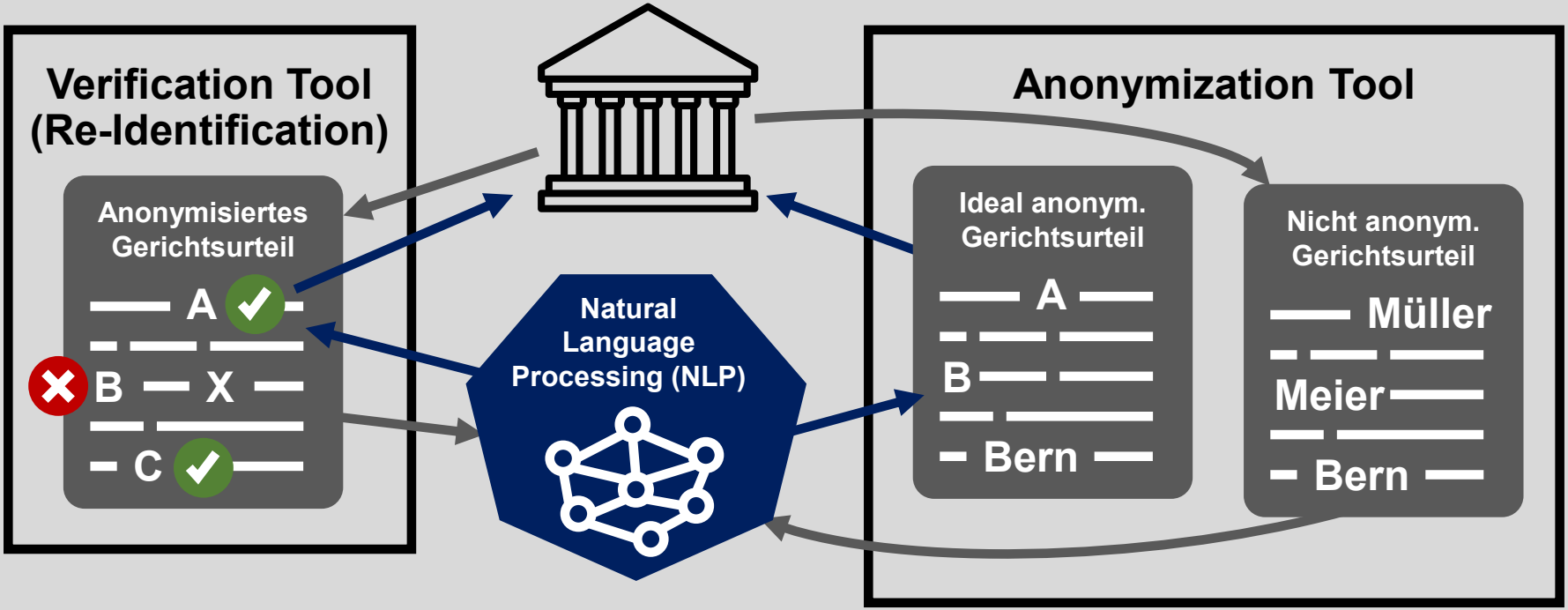
Wir sind nicht die Einzigen …

Projekte mit NLP-basierten Anonymisierungstools

- EU: Multilingual Anonymization for Public Administrations (MAPA)
- Deutschland: Human-in-the-Loop Lernverfahren für verteilte inkrementelle Anonymisierung (HILANO)
- Finnland: ANOPPI-project (Justizministerium)
- Frankreich (2 Projekte):
  - Cour de Cassation/Etalab
  - Conseil d'Etat/Etalab/Éditions Lefebvre Sarrut
- Österreich: Bundesministerium für Justiz

UNIVERSITÄT BERN

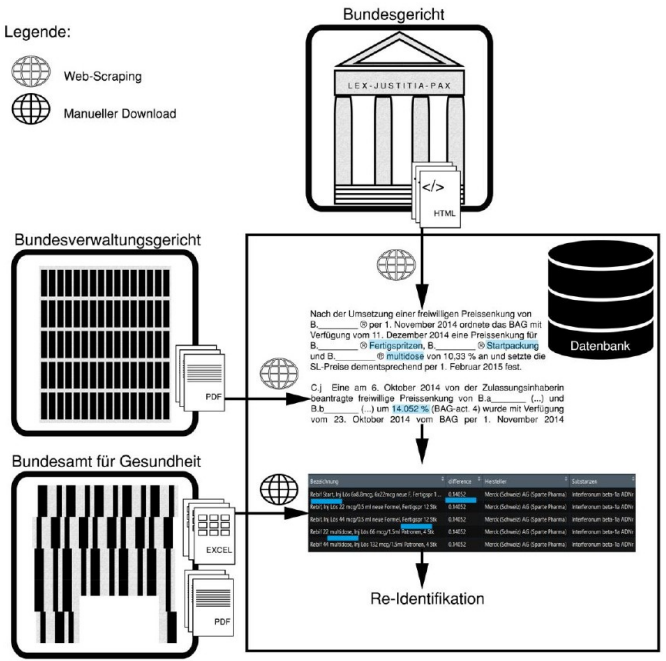
WP2: Verification and Anonymization Tool



Abgrenzung zu Vokinger Mühlematter 2019

bb. «Matching»[52] mit vorinstanzlichen Urteilen

[Rz 31] Die eingeschlossenen Bundesgerichtsurteile wurden a
der direkt über die Referenzangabe in der Regeste bzw. im Urte
Ähnlichkeit Algorithmus[53]) mit dem entsprechenden Vorents
richts verlinkt. Der Cosinus-Ähnlichkeit Algorithmus glieder
einen Vektor, in welchem die Frequenz von jedem Wort im Te
zwei ähnliche Dokumente zu finden, werden deren Vektoren r
tomatisch generierte «Matching» wurde nach Abschluss manu




www.jusletter.ch

Kerstin Noëlle Vokinger / Urs Jakob Mühlematter

Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten(banken)

Eine empirische Analyse anhand von Bundesgerichtsbeschwerden gegen (Preisfestsetzungs-)Verfügungen von Arzneimitteln

Gerichtsurteile werden häufig in anonymisierter Form öffentlich zugänglich gemacht. In der vorliegenden Studie haben wir untersucht, ob es mit der Methodik des «Linkage» – der Verbindung von verschiedenen, öffentlich zugänglichen Daten(banken) – möglich ist, Urteile zu re-identifizieren. Materiell interessierten uns die Fragestellungen, welche pharmazeutischen Unternehmen zwischen 2000 und 2018 in einem Verfahren gegen (Preis-)Verfügungen des BAG vor Bundesgericht involviert und welche Arzneimittel davon betroffen waren. Wir erzielten eine Re-Identifikation in 84% der Fälle. Dies wirft neue Fragen zur Anonymisierung von Daten auf.

Beitragsarten: Beiträge
Rechtsgebiete: Gesundheitsrecht; Öffentliches Recht; Gerichtsorganisation. Gerichtsbarkeit. Verfahren

Zitiervorschlag: Kerstin Noëlle Vokinger / Urs Jakob Mühlematter, Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten(banken), in: Jusletter 2. September 2019

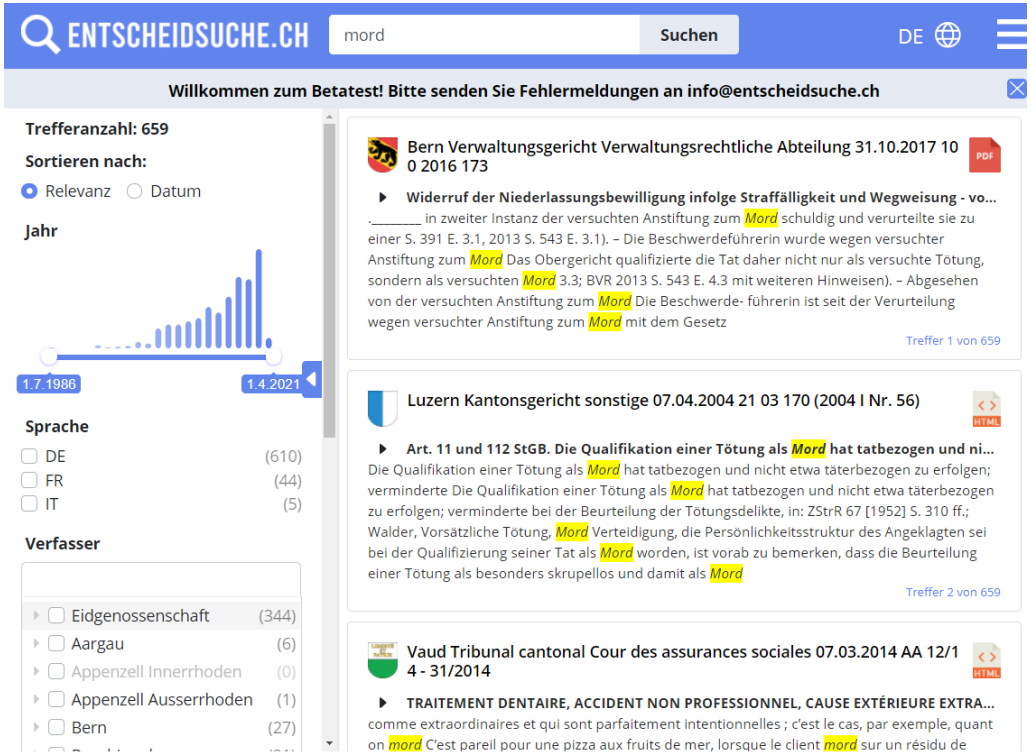
ISSN 1424-7410, jusletter.weblaw.ch, Weblaw AG, info@weblaw.ch, T +41 31 380 57 77

 Vokinger, K. N. and Mühlematter, U. J. (2019). Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten(banken), *Jusletter*

UNIVERSITÄT BERN

Datenquelle Gerichtsurteile

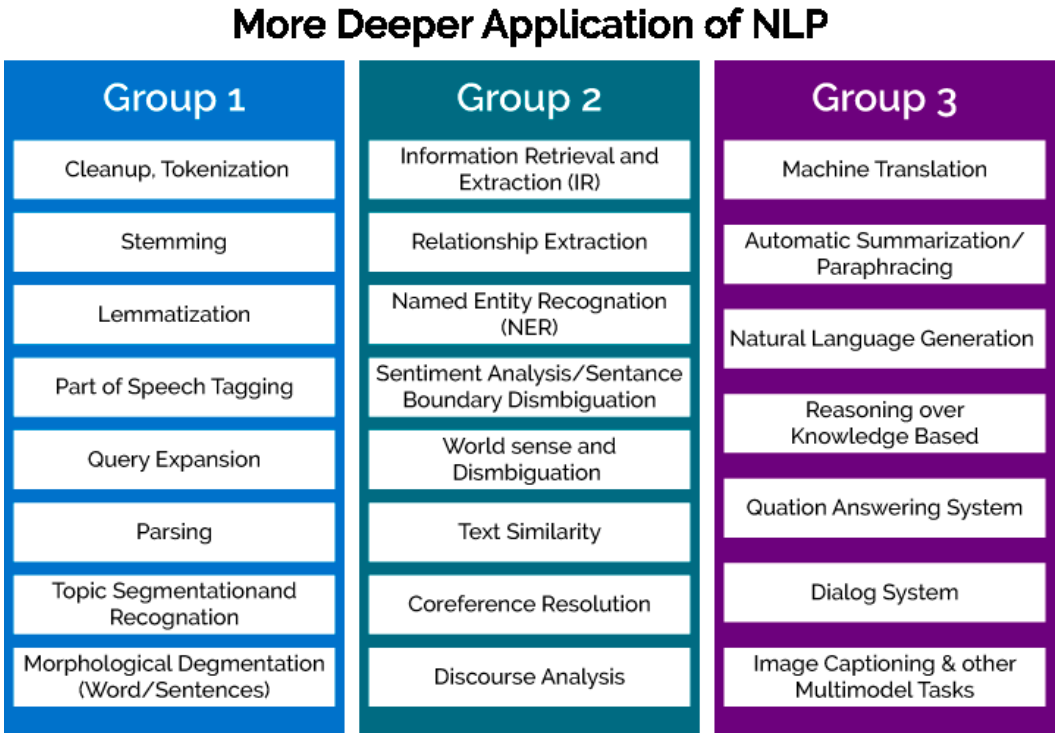
- Kooperation mit Verein **entscheidsuche.ch**
- Täglich aktuelle Gerichtsurteile von allen **Schweiz Gerichten**
- Aktuell rund **540’000 Gerichtsurteile** verfügbar



Forschungs-Methodik NLP

Natural Language Processing (NLP):

- Named Entity Recognition (NER)
- Part-of-speech tagging (POS)
- Information Extraction
- Coreference Resolution etc.



 http://nlpprogress.com http://primo.ai/index.php?title=Natural_Language_Processing_(NLP)

Named Entity Recognition (NER)

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in

UNIVERSITÄT BERN

Annotationen von Gerichtsurteilen

Bundesgericht
Tribunal fédéral
Tribunale federale
Tribunal federal

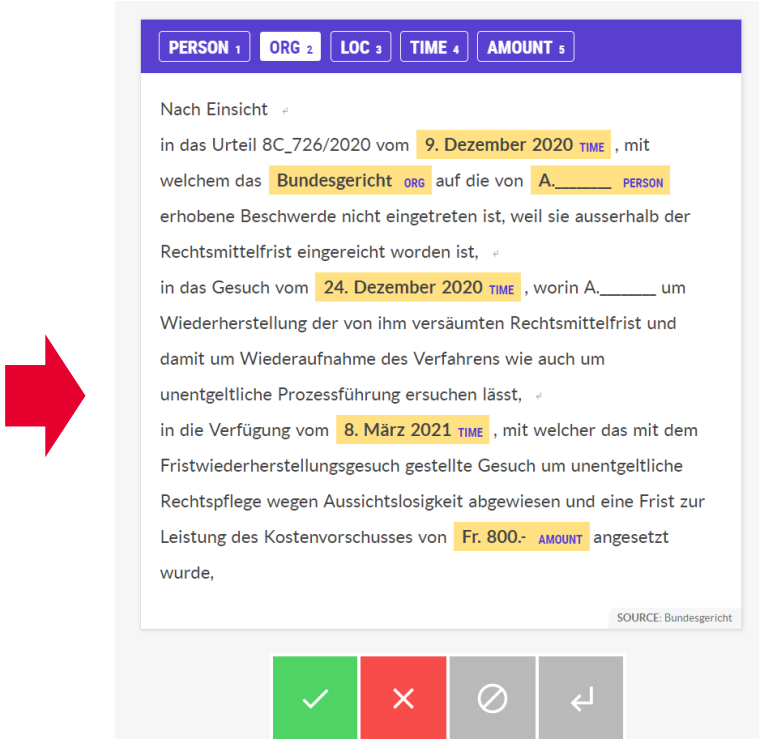
8F_15/2020

Urteil vom 30. März 2021

I. sozialrechtliche Abteilung

Besetzung
Bundesrichter Maillard, Präsident,
Gerichtsschreiber Grünvogel.

Nach Einsicht
in das Urteil 8C_726/2020 vom 9. Dezember 2020, mit welchem das Bundesgericht auf die von A._____ erhobene Beschwerde nicht eingetreten ist, weil sie ausserhalb der Rechtsmittelfrist eingereicht worden ist,
in das Gesuch vom 24. Dezember 2020, worin A._____ um Wiederherstellung der von ihm versäumten Rechtsmittelfrist und damit um Wiederaufnahme des Verfahrens wie auch um unentgeltliche Prozessführung ersuchen lässt,
in die Verfügung vom 8. März 2021, mit welcher das mit dem Fristwiederherstellungsgesuch gestellte Gesuch um unentgeltliche Rechtspflege wegen Aussichtslosigkeit abgewiesen und eine Frist zur Leistung des Kostenvorschusses von Fr. 800.- angesetzt wurde,



NER-Analyse deutscher Rechtstexte

UNIVERSITÄT BERN

… hat bislang nur das Land Mecklenburg-Vorpommern **LD** Gebrauch gemacht.

'So far, only the state of Mecklenburg-Vorpommern has made use of it.'

Dem Haftbefehl liegt eine Entscheidung des Berufungsgerichts in Bukarest **ST** vom 18. Februar 2016 zugrunde …

'The arrest warrant is based on a decision of the Appeal Court in Bucharest of 18 February 2016 …'

Zwar legt der Bezug auf die Grenzwertüberschreitung 2015 insbesondere in der Cornelius-straße **STR** …

'Admittedly, the reference to the exceedance of the 2015 threshold applies in particular to Corneliusstraße …'

… aus der Region um den Fluss Main **LDS** stammen bzw. dort angeboten werden …

'… come from the region around the river Main or are offered there…'

Der FC Bayern München **ORG** schloss den Beschwerdeführer … aus dem Verein aus …

'Bayern Munich closed the complainant … from the club …'

Die Landesregierung Rheinland-Pfalz **INN** hat von einer Stellungnahme abgesehen.

'The state government of Rhineland-Palatinate refrained from commenting.'

… eingeführte Smartphone-Modellreihe des US-amerikanischen Unternehmens Apple **UN** …

'… introduced smartphone model series of the US company Apple …'

Diesen Anspruch hat das LSG Mecklenburg-Vorpommern **GRT** mit Urteil vom 22.2.2017 verneint …

'This claim was rejected by the LSG Mecklenburg-Vorpommern by judgment of 22.2.2017 …'

Vorliegend stehen sich die Widerspruchsmarke Becker Mining **MRK** und die angegriffene Marke Becker **MRK** gegenüber.

'In the present case, the opposing brand Becker Mining and the challenged brand Becker face each other.'

			Classes	#	%
f	1	PER	Person	1,747	3.26
f	2	RR	Judge	1,519	2.83
f	3	AN	Lawyer	111	0.21
c	**1**	**PER**	**Person**	**3,377**	**6.30**
f	4	LD	Country	1,429	2.66
f	5	ST	City	705	1.31
f	6	STR	Street	136	0.25
f	7	LDS	Landscape	198	0.37
c	**2**	**LOC**	**Location**	**2,468**	**4.60**
f	8	ORG	Organization	1,166	2.17
f	9	UN	Company	1,058	1.97
f	10	INN	Institution	2,196	4.09
f	11	GRT	Court	3,212	5.99
f	12	MRK	Brand	283	0.53
c	**3**	**ORG**	**Organization**	**7,915**	**14.76**
f	13	GS	Law	18,520	34.53
f	14	VO	Ordinance	797	1.49
f	15	EUN	EU legal norm	1,499	2.79
c	**4**	**NRM**	**Legal norm**	**20,816**	**38.81**
f	16	VS	Regulation	607	1.13
f	17	VT	Contract	2,863	5.34
c	**5**	**REG**	**Case-by-c. regul.**	**3,470**	**6.47**
f	18				
c	**6**	**RS**	**Court decision**	**12,580**	**23.46**
f	19				
c	**7**	**LIT**	**Legal literature**	**3,006**	**5.60**
Total				**53,632**	**100**

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4478–4485
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

A Dataset of German Legal Documents for Named Entity Recognition

Elena Leitner, Georg Rehm, Julián Moreno-Schneider
DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
{firstname.lastname}@dfki.de

Abstract

We describe a dataset developed for Named Entity Recognition in German federal court decisions. It consists of approx. 67,000 sentences with over 2 million tokens. The resource contains 54,000 manually annotated entities, mapped to 19 fine-grained semantic classes: *person, judge, lawyer, country, city, street, landscape, organization, company, institution, court, brand, law, ordinance, European legal norm, regulation, contract, court decision,* and *legal literature*. The legal documents were, furthermore, automatically annotated with more than 35,000 TimeML-based time expressions. The dataset, which is available under a CC-BY 4.0 license in the CoNNL-2002 format, was developed for training an NER service for German legal documents in the EU project Lynx.

Keywords: Named Entity Recognition, NER, Legal Documents, Legal Domain, Corpus Creation, Corpus Annotation

1. Introduction and Motivation

Just like any other field, the legal domain is facing multiple challenges in the era of digitisation. Document collections are growing at an enormous pace and their complete and deep analysis can only be tackled with the help of assisting technologies. This is where content curation technologies based on text analytics come in Bourgonje et al. (2016). Such domain-specific semantic technologies enable the fast and efficient automated processing of heterogeneous document collections, extracting important information units and metadata such as, among others, named entities, numeric expressions, concepts and topics, time expressions, and text structure. One of the fundamental processing tasks is the identification and categorisation of named entities (Named Entity Recognition, NER). Typically, NER is focused upon the identification of semantic categories such as *person*, *location* and *organization* but, especially in domain-specific applications, other typologies have been developed that correspond to task-, language- or domain-specific needs. With regard to the legal domain, the lack of freely available datasets has been a stumbling block for text analytics research. German newspaper datasets from CoNNL 2003 (Sang and Meulder, 2003) or GermEval 2014 (Benikova et al., 2014) are simply not suitable in terms of domain, text type or semantic categories covered.

The work described in this paper was carried out under the umbrella of the project *Lynx: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe*, a three-year EU-funded project that started in December 2017 (Montiel-Ponsoda et al., 2017).[1] Its objective is the creation of a legal knowledge graph that contains different types of legal and regulatory data (Schneider and Rehm, 2018a; Schneider and Rehm, 2018b; Moreno-Schneider et al., 2020). Lynx aims to help European companies, especially SMEs, that want to become active in new European countries and markets. The project offers compliance-related services that are currently tested and validated in three use cases (UC): (i) UC1 aims to analyse contracts, enriching them with domain-specific semantic information (document structure, entities, temporal expressions, claims, summaries, etc.); (ii) UC2 focuses on compliance services related to geothermal energy operations, where Lynx supports the understanding of regulatory regimes, including norms and standards; (iii) UC3 is a compliance solution in the domain of labour law, where legal provisions, case law, and expert literature are interlinked, analysed, and compared to define legal strategies for legal practice. The Lynx services are developed for several European languages including English, Spanish, and – relevant for this paper – German (Rehm et al., 2019).

Documents in the legal domain contain multiple references to named entities, especially domain-specific named entities, i. e., jurisdictions, legal institutions, etc. Legal documents are unique and differ greatly from newspaper texts. On the one hand, the occurrence of general-domain named entities is relatively rare. On the other hand, in concrete applications, crucial domain-specific entities need to be identified in a reliable way, such as designations of legal norms and references to other legal documents (laws, ordinances, regulations, decisions, etc.). However, most NER solutions operate in the general or news domain, which makes them inapplicable to the analysis of legal documents (Bourgonje et al., 2017; Rehm et al., 2017). Accordingly, there is a great need for an NER-annotated dataset consisting of legal documents, including the corresponding development of a typology of semantic concepts and uniform annotation guidelines. In this paper, we describe the development of a dataset of legal documents, which includes (i) named entities and (ii) temporal expressions.

The remainder of this article is structured as follows. First, Section 2 gives a brief overview of related work. Section 3 describes, in detail, the rationale behind the annotation of the dataset including the different semantic classes annotated. Section 4 describes several characteristics of the dataset, followed by a short evaluation (Section 5) and conclusions as well as future work (Section 6).

2. Related Work

Until now, NER has not received a lot of attention in the legal domain, developed approaches are fragmented and inconsistent with regard to their respective methods, datasets and typologies used. Among the related work, there is

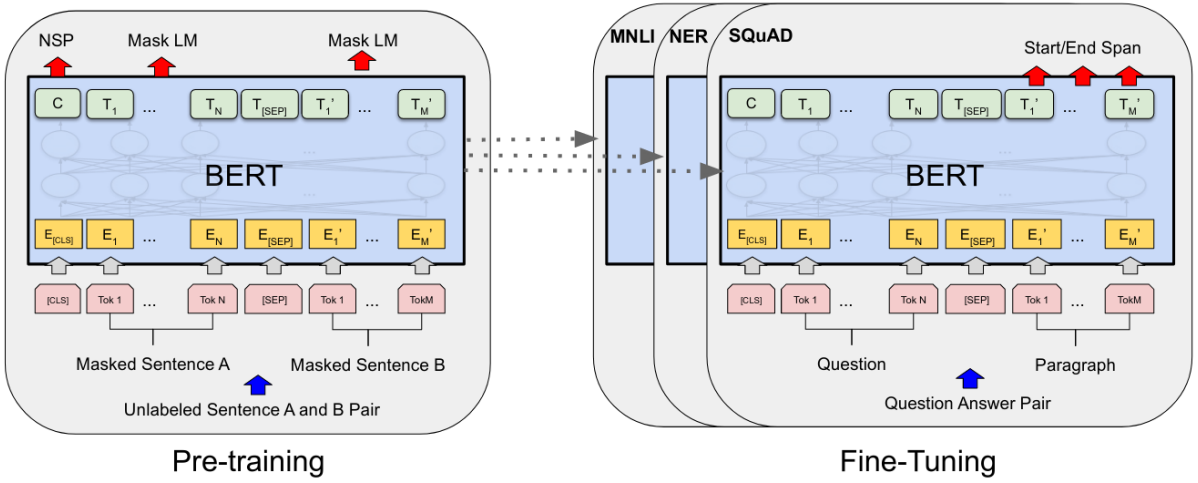
[1] http://www.lynx-project.eu

 Leitner, E., Rehm, G., Moreno-Schneider, J. (2020). A Dataset of German Legal Documents for Named Entity Recognition, https://www.aclweb.org/anthology/2020.lrec-1.551.pdf

UNIVERSITÄT BERN

Google Open Source NLP Technologie

BERT: Bidirectional Encoder Representations from Transformers



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

arXiv:1810.04805v2 [cs.CL] 24 May 2019

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

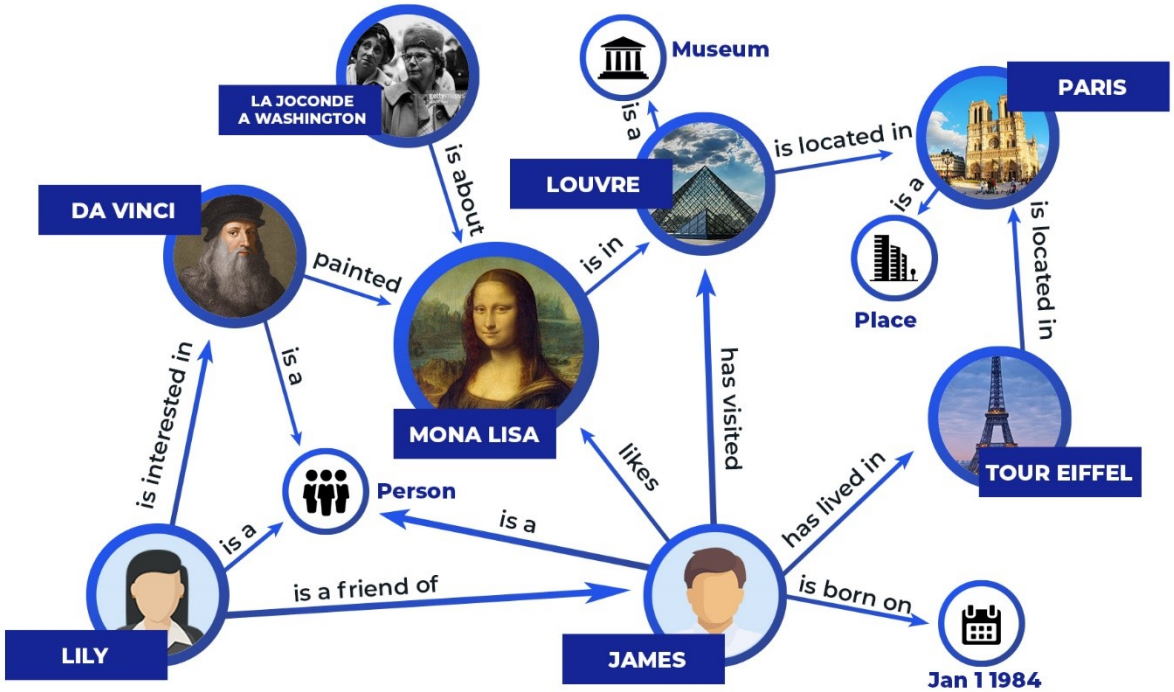
There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, http://arxiv.org/abs/1810.04805

Knowledge Graph

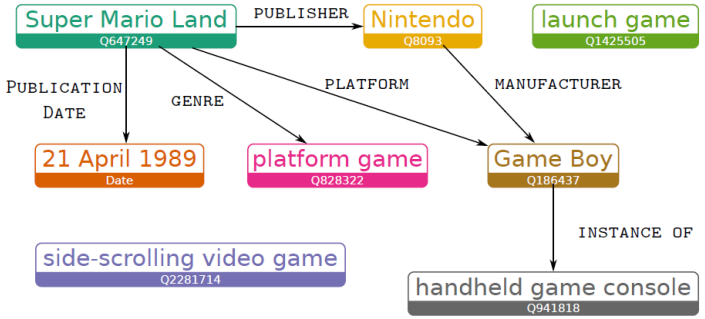


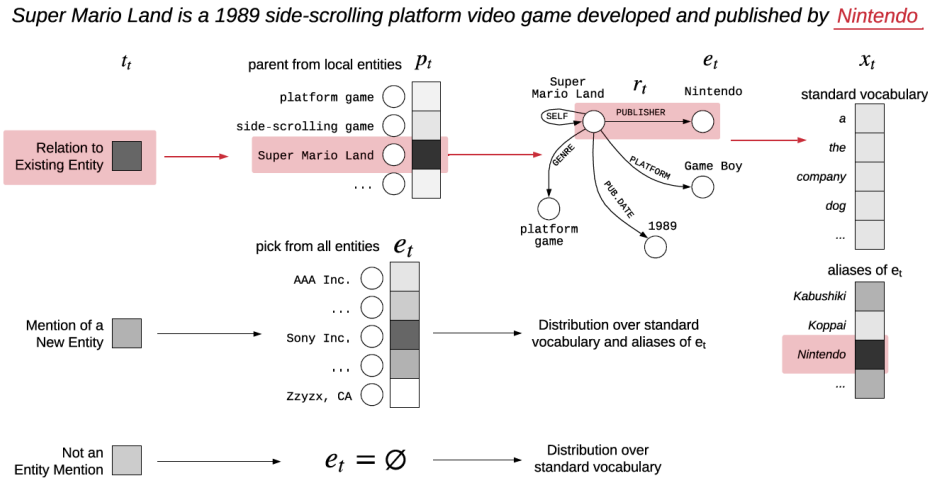
 https://yashuseth.blog/2019/10/08/introduction-question-answering-knowledge-graphs-kgqa/

UNIVERSITÄT BERN

Knowledge Graph aus Text generieren

[*Super Mario Land*] is a [*1989*] [*side-scrolling*] [*platform video game*] developed and published by [*Nintendo*] as a [*launch title*] for their [*Game Boy*] [*handheld game console*].



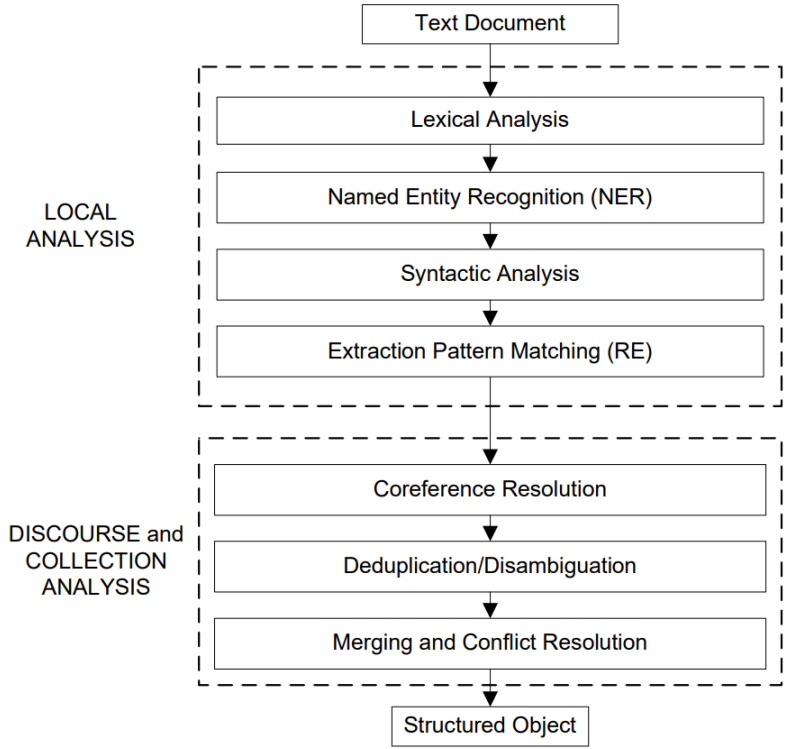


Logan, R., Liu, N. F., Peters, M. E., Gardner, M. and Singh, S. (2019). Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling https://www.aclweb.org/anthology/P19-1598

UNIVERSITÄT BERN

Nächste Schritte 2021/2022

1. **Information Extraction** aus Gerichtsurteilen (Named Entity Recognition mit BERT etc.) → Knowledge Graph
2. **Crawling von Informationen** aus Mediendatenbanken, Social Media Plattformen etc.
3. **Linking von Knowledge Graph** eines Gerichtsurteils mit strukturierten Informationen aus gecrawlten Daten



Agichtein, E. (2005). Scaling information extraction to large document collections https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4024&rep=rep1&type=pdf

UNIVERSITÄT BERN

Herzlichen Dank für Ihre Aufmerksamkeit!

Kontakt:

PD Dr. Matthias Stürmer

Forschungsstelle Digitale Nachhaltigkeit am Institut für Informatik der Universität Bern

Tel. (direkt) +41 76 368 81 65; eMail matthias.stuermer@inf.unibe.ch

Daniel Kettiger

Kompetenzzentrum für Public Management, Universität Bern

Tel. (direkt) +41 33 223 79 25; eMail daniel.kettiger@kpm.unibe.ch