A German Corpus for Fine-Grained Named Entity Recognition and Relation Extraction of Traffic and Industry Events

Martin Schiersch, Veselina Mironova, Maximilian Schmitt, Philippe Thomas, Aleksandra Gabryszak, Leonhard Hennig
DFKI GmbH
Berlin, Germany
{firstname.lastname}@dfki.de

Abstract

Monitoring mobility- and industry-relevant events is important in areas such as personal travel planning and supply chain management, but extracting events pertaining to specific companies, transit routes and locations from heterogeneous, high-volume text streams remains a significant challenge. This work describes a corpus of German-language documents which has been annotated with fine-grained geo-entities, such as streets, stops and routes, as well as standard named entity types. It has also been annotated with a set of 15 traffic- and industry-related n-ary relations and events, such as accidents, traffic jams, acquisitions, and strikes. The corpus consists of newswire texts, Twitter messages, and traffic reports from radio stations, police and railway companies. It allows for training and evaluating both named entity recognition algorithms that aim for fine-grained typing of geo-entities, as well as n-ary relation extraction systems.

Keywords: Named Entity Recognition, Relation Extraction

1. Introduction

Monitoring relevant news and events is of central importance in many economic and personal decision processes, such as supply chain management (Chae, 2015), market research (Mostafa, 2013), and personal travel planning (Schulz et al., 2013). Social media, news sites, and also more specialized information systems, such as online traffic and public transport information sources, provide valuable streams of text messages that can be used to improve decision making processes (Hennig et al., 2016). For example, a company's sourcing department may wish to monitor world-wide news for disruptive or risk-related events pertaining to their suppliers (e.g. natural disasters, strikes, liquidity risks), while a traveler wants to be informed about traffic events related to her itinerary (e.g. delays, cancellations). To fulfill such information needs, we need to extract events and relations from message streams that mention fine-grained entity types, such as companies, streets, or routes (Yaghoobzadeh and Schütze, 2017; Shimaoka et al., 2017). For example, from the sentence *"Berlin: Rail replacement service between Schichauweg and Priesterweg on route S2"*, we would like to extract a *Rail Replacement Service* event with the arguments *location="S2"* of type *location-route*, and *start-loc="Schichauweg"* respectively *end-loc="Priesterweg"* with types *location-stop*.

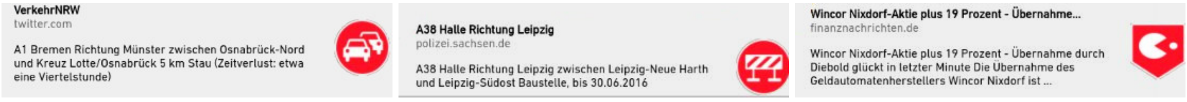
Detecting such relations in textual message streams raises a number of challenges. Social media streams, such as Twitter, are written in a very informal, not always grammatically well-formed style (Osborne et al., 2014), which cannot easily be processed with standard linguistic tools. News sites provide well-formed texts, but their content is very heterogeneous and often hard to separate from non-relevant web page elements. Domain-specific information sources, like traffic reports, on the other hand, are topic-focused, but employ a wide variety of formats, from telegraph style texts to table entries. In addition, existing corpora for German-language Named Entity Recognition (Tjong Kim Sang and De Meulder, 2003; Benikova et al., 2014) are mostly limited to standard entity types, and consist mainly of newswire and Wikipedia texts. These corpora also do not include annotations of events and relations.

In this work, we present a large German-language corpus consisting of documents from three different genres, namely newswire texts, Twitter, and traffic reports from radio stations, police and railway companies (Section 2.). The documents have been annotated with fine-grained geo-entities, as well as standard entity types such as organizations and persons. In addition, the corpus has been annotated with a set of 15 mobility- and industry-related n-ary relation types (Section 3.). Many of these relation and event types, such as accidents, traffic jams, and strike events, are not available in standard knowledge bases and hence cannot be learned in a distantly supervised fashion. The final corpus consists of $2,598$ documents with $22,075$ entity and $1,507$ relation annotations (Section 4.). It allows for training and evaluating both named entity recognition algorithms that aim for fine-grained typing of geolocation entities, as well as for training of n-ary relation extraction systems.

2. Dataset Collection

To create the corpus, we collected a dataset of 3,789,803 tweets, 412,652 RSS feeds, and 860,307 news documents in the time period of Jan 1st, 2016 to March 31st, 2016. We aimed to collect only German-language texts by applying appropriate filter settings when crawling APIs, and by post-processing documents with langid.py (Lui and Baldwin, 2012). Figure 1 gives an example for each type of data source. All web documents, tweets, and RSS documents were transformed into a common Avro-encoded schema,[1] with fields for title, text, URI, and other attributes, as well

[1] avro.apache.org



VerkehrNRW
twitter.com

A1 Bremen Richtung Münster zwischen Osnabrück-Nord und Kreuz Lotte/Osnabrück 5 km Stau (Zeitverlust: etwa eine Viertelstunde)

(a) Example Twitter message

A38 Halle Richtung Leipzig
polizei.sachsen.de

A38 Halle Richtung Leipzig zwischen Leipzig-Neue Harth und Leipzig-Südost Baustelle, bis 30.06.2016

(b) Excerpt of a RSS message

Wincor Nixdorf-Aktie plus 19 Prozent - Übernahme...
finanznachrichten.de

Wincor Nixdorf-Aktie plus 19 Prozent - Übernahme durch Diebold glückt in letzter Minute Die Übernahme des Geldautomatenherstellers Wincor Nixdorf ist ...

(c) Excerpt of a news document

Figure 1: Examples for the three different data sources.

Entity / Concept	**Description**	**Examples**
Location (LOC)	General locations	Bayern, Zugspitze, Norden
Location-City (LOC-CIT)	Municipalities, e.g. cities, towns, villages	Berlin, Berlin-Buch, Hof
Location-Street (LOC-STR)	Named streets, highways, roads	Hauptstrasse, A1
Location-Route (LOC-ROU)	Named (public) transit routes	U1, ICE 557, Nürnberg – Hof
Location-Stop (LOC-STO)	Public transit stops, e.g. train stations, bus stops	S+U Pankow, Berlin-Buch
Organization (ORG)	General organizations	Greenpeace, Borussia Dortmund
Organization-Company (ORG-COM)	The subset of organizations that are businesses	Siemens AG, BMW
Person (PER)	Persons	Angela Merkel
OrgPosition (POS)	A person's position within an organization	CEO, Vizepräsident
Date (DAT)	Point in time, date	1. September 2017, gestern
Time (TIM)	Time of day	8:30, 5 Uhr früh
Duration (DUR)	Time periods	mehr als eine halbe Stunde
Distance (DIS)	Distances with unit	5 Kilometer
Number (NUM)	Other numeric entities, e.g. money, percentages	3%, 4 Millionen Euro
Disaster-Type (DIS-TYP)	Man-made and natural disaster types	Erdbeben, Überschwemmung
Trigger (TRI)	Trigger terms or phrases for events	Stau, Streik, Entlassungen

Table 1: Definition of entity types annotated in the corpus.

as fields for the annotations. From this dataset, we randomly sampled documents from each data source for annotation.

Twitter We use the Twitter Search API[2] to obtain a topically focused streaming sample of tweets. We define the search filter using a list of approximately 150 mobility- and industry-relevant channels and 300 search terms. Channels include e.g. airline companies, traffic information sources, and railway companies. Search terms comprise event-related keywords such as "traffic jam" or "roadworks", but also major highway names, railway route identifiers, and airport codes.

News We retrieve news pages and topically focused web sites using the uberMetrics Search API,[3] which provides an interface to more than 400 million web sources that are crawled on a regular basis. The API allows us to define complex boolean search queries to filter the set of web pages. We employ the same search terms as for Twitter, and limit the language to German. Boilerplate detection is used to remove extraneous contents from the HTML document (Kohlschütter et al., 2010). To speed up the annotation process, we limit each news document to the first 1000 characters, including the title, and discard the remainder of the text. Although this approach may result in the loss of some information, it is well known that in news writing, important information is presented first. The trimming may lead to incomplete final sentences, which annotators were advised to ignore.

RSS Feeds We implemented crawlers for a representative set of approximately 100 German-language RSS feeds that provide traffic and transportation information. Feed sources include federal and state police, radio stations, and air travel sources. The feeds were fetched at regular intervals during the 3-month period.

3. Annotation Guidelines

For the targeted applications of supply chain monitoring and personal travel planning, we are interested in the annotation of fine-grained geo-locations, such as street names and public transport stops, as well as relation (or event) mentions with typically multiple arguments. That is, we are not only interested to know that a given event type occurred, such as a traffic jam, but also, on which road, between which exits, and what the resulting time delay is for drivers. We hence aim to recognize and extract n-ary ACE/ERE-style relations (Doddington et al., 2004; Linguistic Data Consortium, 2015). The annotation guidelines and schema of the ACE Entities V6.6[4] and TimeML[5] served as a basis for annotating standard entity types, such as organizations, persons and dates. The main difference to ACE guidelines is the treatment of geo-political entities (GPE) – we chose to annotate them mainly as locations (LOC), and sometimes as organizations (ORG), in particular for cities, regions, or counties, as the relation types we are interested in typically refer to the location or organization aspects of a potential GPE entity.

[2] dev.twitter.com/rest/public/search

[3] doc.ubermetrics-technologies.com/api-reference/

[4] www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

[5] www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf

For the corpus annotation we use the markup tool Recon (Li et al., 2012), which allows annotating n-ary relations among text elements. Recon provides a graphical user interface that enables users to mark arbitrary text spans as entities, to connect entities to create relations, and to assign semantic roles to argument entities. Each document was annotated by two trained annotators. In cases of disagreement, a third annotator was consulted to reach a final decision. We measured inter-annotator agreement for entity and relation annotations. For entity annotations, we evaluated agreement at the entity level by comparing labels and offsets. A high inter-annotator agreement thus implies that annotators agreed both on the extent of entities and their type. For relation mentions, we measured role and relation type agreement at the level of relation arguments for each annotated relation mention. Similar to entity inter-annotator agreement, arguments were identified based on the underlying concepts/entities and their character offsets. A high inter-annotator agreement hence means that annotators agreed on entity, entity extent, role, and relation type labels. Table 2 lists the inter-annotator agreement values of our corpus. The pairwise kappa agreement is moderate at around 0.58 for entity annotations, which is somewhat lower than the 0.74 reported by Benikova et al. (2014). For relations, pairwise kappa agreement is 0.51.

Type	**Cohen's κ**	**Krippendorf's α**
Entities	0.58	0.57
Relations	0.51	0.45

Table 2: Inter-annotator agreement for entities and relations.

3.1. Entities

Table 1 lists the entity types we currently annotate, and provides a brief explanation of each type. In general, annotators were advised to choose the more specific entity type for a given entity mention (e.g. organization-company instead of organization), unless it was unclear from the context whether the entity mention referred to the specific or the more general type. This is for example the case for traffic jam and accident reports on main highways, where the exits often use the name of the closest city, e.g. *"accident on the A1 between [Bremen] and [Oldenburg]"*. Here, *"Bremen"* and *"Oldenburg"* are ambiguous between the types *location-city* and *location-street*.[6]

Organization-Company covers all commercial organizations, including media and entertainment businesses. It does not cover governmental or religious organizations, but may include sports teams. In general, though, we are interested in companies that provide services to other companies, e.g. in the form of products, parts, components, technologies, or non-physical services.

Location subtypes are of interest to pin-point exact locations for any event of interest, e.g. by a lookup in OpenStreetMap,[7] and to distinguish between location types where necessary. We do not tag locations if they are used as metonyms for organizations or GPEs, as in the case of capital cities denoting the government of a country.

In the case of traffic reports, we also consider directions, including cardinal points, as locations, for example:

(1) *Stau auf der B2 [stadteinwärts]*
(Traffic jam on the B2 [into town])

(2) *Auf der A1 Nähe Münster Stau in [beiden Richtungen]*
(On the A1, near Münster, traffic jam in [both directions])

For traffic-related locations, specifiers are included in the mention extent when required, e.g. *"[Kreuz München-Nord]"* (*"[Cross Munich-North]"*) as well as *"[Dreieck Havelland]"* (*"[Junction Havelland]"*), *"[Anschlussstelle Adlershof]"* (*"[Exit Adlershof]"*), *"[Abzweig nach Basel]"* (*"[Branch to Basel]"*), etc. Similarly, terms like *"Kreis"* (*"county"*) in *"[Kreis Tuttlingen]"* are included to distinguish the county from the city. However, terms like *"Ecke"* (*"corner"*) or *"Kreuzung"* (*"intersection"*) are not included in the extent of *Location-Street* entities, because they are not an integral part of the location's name.

City names that occur in transit routes are labeled as *Location-City* when they are used to indicate the direction of the route, and as *Location-Stop*s in every other case. In the case of highway exits, city names are labeled as *Location*s, since they actually denote the exit (and its geographic position), and not the city. For flight routes, we chose to label city names as *Location-City* unless the reference includes the specific airport used, e.g. *"Heathrow"* or *"MUC"*.

*Location-Route*s are either generic transit lines (e.g. *"S2"*), or a specific instance of this line (*"the next S2 which was supposed to arrive at 19:40"*). In general, they are referred to by letter-number combinations, but sometimes consist of concatenated stop or city names:

(3) *[Günzburg – Mindelheim]: Störung an einem Bahnübergang*
([Günzburg - Mindelheim]: Disruption at a crossing)

(4) *Ersatzverkehr auf der Linie [RE 3] [Stralsund/Schwedt (Oder) - Berlin – Elsterwerda]*
(Rail replacement service on the route [RE 3] [Stralsund/Schwedt (Oder) - Berlin - Elsterwerda])

Common nouns and noun phrases are annotated like proper names as entities of the corresponding type. Most often, they are used to denote a group of entities, e.g.:

(5) *Die [EC-Züge] zwischen Dresden Hbf und Praha hl.n. fallen aus*
(The [EC trains] between Dresden main station and Prague main station are cancelled)

(6) *Fraport übernimmt [14 griechische Flughäfen]*
(Fraport acquires [14 Greek airports])

Trigger concepts are a generic class of annotations that cover terms or phrases that indicate a specific event type, and that sometimes are required to create at least a binary relation mention within a sentence. For example, given the

[6] In the remainder of this document, '[' and ']' are used to denote the extent of an entity or relation mention

[7] openstreetmap.org

Relation / Event	Definition & Arguments
Accident	Collision of a vehicle with another vehicle, person, or obstruction ⊛location, ⊛trigger, delay, direction, start-loc, end-loc, start-date, end-date, cause (TRI)
Canceled Route	Cancellation of public transport routes ⊛location (LOC-ROU), ⊛trigger, direction, start-loc, end-loc, start-date, end-date, cause (TRI)
Canceled Stop	Cancellation of public transport stops ⊛location (LOC-STO), ⊛trigger, route, direction, start-date, end-date, cause (TRI)
Delay	Delay resulting from remaining traffic disturbances ⊛location, ⊛trigger, delay, direction, start-loc, end-loc, start-date, end-date, cause (TRI)
Disaster	Sudden catastrophe causing great damage to structures or loss of life ⊛type, ⊛location, date, victims (NUM), damage-costs (NUM), trigger
Obstruction	Temporary installation to control traffic ⊛location, ⊛trigger, delay, direction, start-loc, end-loc, start-date, end-date, cause (TRI)
Rail Replacement Service	Replacement of a passenger train by buses or other substitute public transport services ⊛location (LOC-ROU), ⊛trigger, delay, direction, start-loc, end-loc, start-date, end-date, cause (TRI)
Traffic Jam	Line of stationary or very slow-moving traffic ⊛location (LOC-STR), ⊛trigger, delay, jam-length, direction, start-loc, end-loc, start-date, end-date, cause (TRI)
Acquisition	Purchase of one company by another ⊛buyer, ⊛acquired, seller, date, price, trigger
Insolvency	Insolvency of a company ⊛company, ⊛trigger, date, location
Layoffs	Layoffs from companies, including number of people fired. ⊛company, ⊛trigger, date, location, num-laid-off
Merger	Merger of companies that is not a clear buy-up ⊛old (ORG-COM A), old (ORG-COM B), new (ORG-COM), date, trigger
Organization Leadership	Relationship between an organization and its leaders, board members, directors, etc. ⊛organization, ⊛person, position, from, to, trigger
SpinOff	Parent company "splits off" a section as a separate new company ⊛parent (ORG-COM), ⊛child (ORG-COM), location, trigger
Strike	Strike action affecting a company or organization ⊛company, ⊛trigger, date, location, num-striking, striker, union (ORG)

Table 3: Definition of the 15 target relations of the domains *Mobility* and *Industry*. ⊛ denotes the essential arguments of the relation that define the identity of a relation instance. Entity types are abbreviated or omitted in unambiguous cases.

message *"Stau auf der Warschauer Strasse"* (*"Traffic jam on Warschauer street"*), the location *"Warschauer Strasse"* alone is not sufficient to annotate a *Traffic Jam* event, which requires the additional annotation of the trigger *"Stau"* to distinguish it from other traffic-related event types. This reasoning applies for the relations *Insolvency*, *Layoffs* and *Strike* of the *Industry* domain, and for all relations of the *Mobility* domain except for the relation *Disaster*. However, the argument *type* of the relation *Disaster* can be filled only with concepts which can be considered as triggers for this event (earthquake, flood, nuclear accidents, etc.). The majority of mobility-related events we are interested in follow this pattern of *Location* + *Trigger* (or in the case of industry-related events, *Company* + *Trigger*) to distinguish between different event types that are expressed using the same syntactic patterns (see Section 3.2.).

Punctuation characters, such as "-", "/", "#" and "@" are not included in the mention extent unless they occurred inside a multi-token entity, e.g. *"#[Flughafen #Tempelhof]"*. Annotators were advised to make the mention extent as long as required to accurately denote a specific entity. The extent could include adjectives, numerals (*"more than"*, *"a few"*, *"several"*), or numbers, if these were used to denote a specific subset of a set-based named entity mention.

If an entity was referred to by two or more token sequences, e.g. *"Volkswagen (VW)"*, *"A1 Bremen - Hamburg"*, the annotators were advised to annotate two separate entities as in *"[Volkswagen] ([VW])"*.

As a rule, unless required for annotating a relation mention, nested entity mentions were not annotated, e.g. in *"PD Zwickau"* (*"police department Zwickau"*), *"[PD Zwickau]"* was labeled as an *Organization*, but the nested *"Zwickau"* was not labeled as a *Location-City*.

3.2. Relations and Events

We annotated two different sets of relations and events in the corpus, based on the requirements of the project this corpus was developed in. The first group of relations are mobility-related, and include for example *Traffic Jam*s, *Accident*s and *Disaster*s. The second group of relations concerns companies, and includes e.g. *Acquisition*, *Strike* and *Insolvency* events. Table 3 lists all relation types, together with their definitions and arguments. All relations have a set of required (typically two) and a set of optional arguments. For example, the relation *Acquisition* has required arguments *buyer* and *acquired*, and optional arguments *date*, *price*, and *seller*. The following examples illustrate our n-ary relation annotations:

Relation examples of the *Mobility* domain

(7) Accident*: [A8]$_{loc}$ Augsburg Richtung [München]$_{dir}$ - Schwerer [Unfall]$_{tri}$ - kurz vor [Ausfahrt Dasing]$_{sta}$.*

(A8 from Augsburg to Munich - a severe accident - just before the exit Dasing)

(8) Canceled Route*: Wegen des Warnstreiks hat die Lufthansa [mehrere Flüge]$_{loc}$ in [Hamburg]$_{sta}$, [Hannover]$_{sta}$ und weiteren Flughäfen [gestrichen]$_{tri}$.*

(Because of the warning strike, Lufthansa has canceled several flights in Hamburg, Hanover and other airports)

(9) Canceled Stop*: Rinjani macht Ärger: [Flughafen auf Bali]$_{loc}$ wegen [Vulkanausbruch]$_{cau}$ [gesperrt]$_{tri}$.*

(Rinjani causes trouble: Bali airport is closed due to volcanic eruption.)

(10) Delay*: [S-Bahn-Verkehr Stuttgart]$_{loc}$: [Notarzteinsatz]$_{cau}$ in [Feuerbach]$_{sta}$ sorgt für [Verspätungen]$_{tri}$*

(S-Bahn traffic Stuttgart: Emergency medical service in Feuerbach causes delays)

(11) Disaster*: [Mehrere Tote]$_{vic}$ bei erneutem [Erdbeben]$_{typ}$ in [Japan]$_{loc}$*

(Several dead in another earthquake in Japan)

(12) Obstruction*: Wegen [Notarzteinsatzes]$_{cau}$ ist derzeit die [Strecke]$_{loc}$ zwischen [Gerlenhofen]$_{sta}$ und [Senden]$_{end}$ [gesperrt]$_{tri}$.*

(Due to an emergency medical service, the route between Gerlenhofen and Senden is currently closed)

(13) Rail Replacement Service: *[RB59]$_{loc}$: Vom [11.6.]$_{sdat}$ - [3.7.]$_{edat}$ [Schienenersatzverkehr]$_{tri}$ zwischen [Soest]$_{sta}$ und [Holzwickede]$_{end}$ im Spätverkehr.*

(RB59: Rail replacement service from 6/11 until 7/3 between Soest and Holzwickede during evening hours.)

(14) Traffic Jam*: [A40]$_{loc}$ Duisburg Richtung [Venlo]$_{dir}$ zwischen [Neukirchen- Vluyn]$_{sta}$ und [Kempen]$_{end}$ [10 km]$_{len}$ [Stau]$_{tri}$*

(A40 Duisburg - Dortmund between Neukirchen-Vluyn and Kempen 10 km traffic jam)

Relation examples of the *Industry* domain

(15) Acquisition*: Wirecard AG und ihre Tochtergesellschaft [Wirecard Acquiring & Issuing]$_{buy}$ haben den Zahlungsdienstleister [Moip Pagamentos]$_{acq}$ [übernommen]$_{tri}$.*

(Wirecard AG and its subsidiary Wirecard Acquiring & Issuing have acquired the payment service provider Moip Pagamentos.)

(16) Insolvency*: [Imtech]$_{com}$ [Insolvenz]$_{tri}$ gefährdet BER-Eröffnung*

(Imtech insolvency endangers BER opening)

(17) Layoffs*: [Entlassungen]$_{tri}$ bei [Credit Agricole Indosuez]$_{com}$ in [Genf]$_{loca}$*

(Layoffs at Credit Agricole Indosuez in Genf)

(18) Merger*: Der Panzerhersteller [Krauss-Maffei Wegmann]$_{old}$ besiegelt den [Zusammenschluss]$_{tri}$ mit dem französischen Rüstungskonzern [Nexter]$_{old}$.*

(Tank manufacturer Krauss-Maffei Wegmann seals merger with French arms company Nexter.)

(19) Organization Leadership*: [Bernd Hansen]$_{per}$, CEO$_{pos}$ [Hansen Gruppe]$_{com}$*

(Bernd Hansen, CEO Hansen Group)

(20) SpinOff*: [Kölnische Unfall-Versicherungs-Aktiengesellschaft zu #Köln a.Rhein]$_{chi}$, gegr. [1919]$_{dat}$ als [Ableger]$_{tri}$ der [Colonia]$_{par}$*

(Kölnische Unfall-Versicherungs-Aktiengesellschaft zu #Köln a.Rhein, est. in 1919 as a spin-off of the Colonia)

(21) Strike*: Am [Freitag]$_{dat}$ haben die [Amazon]$_{com}$-Mitarbeiter im [Leipziger]$_{loc}$ Versandzentrum des Unternehmens erneut [gestreikt]$_{tri}$.*

(On Friday, Amazon employees in the company's shipping center in Leipzig once more went on strike.)

Some of the relations are semantically related to each other and can occur together even in very short texts such as tweets or RSS feeds. For example, the relation *Traffic Jam* often correlates with *Accident* and *Obstruction* relation mentions. This also applies to the relation *Obstruction* and the event *Disaster*, and *Delay* relations and the events *Canceled Route*, *Canceled Stop* and *Rail Replacement Service*. In the *Industry* domain, we observe that reports of corporate events often include information about leaders of an organization, i.e. a *Organization Leadership* relation is mentioned together with another relation.

The annotators annotated only explicitly expressed relation mentions where all arguments – required and optional – occurred within a single sentence. In cases of multiple occurrences of an argument, they chose the arguments occurring within the shortest overall text span. Future or planned relations, such as potential acquisitions or announced strikes, were also marked up, and labeled with an additional attribute to indicate this status. Negated relation mentions (e.g. a canceled acquisition), or events marking the end of a relation (e.g. *"Traffic jam has dissolved"*) were not annotated.[8] The following examples illustrate the three types of relation mentions:

(22) Factual*: Der Panzerhersteller Krauss-Maffei Wegmann besiegelt den Zusammenschluss mit dem französischen Rüstungskonzern Nexter.*

(Tank manufacturer Krauss-Maffei Wegman seals merger with French armaments group Nexter.)

(23) Potential*: Größer als BASF: US-amerikanische Chemie-Unternehmen DuPont und Dow Chemical planen Mega-Fusion.*

(Larger than BASF: US-American chemical companies DuPont and Dow Chemical are planning mega-merger)

[8] However, the files containing such mentions were marked by renaming them. Fully annotating the negated relation mentions and including them in the corpus remains future work.

	News	Twitter	RSS	Total
Documents	835	1,138	625	2,598
Sentences	5,951	1,842	1,031	8,824
Sentences (avg.)	7.13	1.62	1.65	3.40
Words	113,089	19,558	19,595	152,242
Words (avg.)	135.44	17.19	31.35	58.60

Table 4: Corpus Statistics

	News	Twitter	RSS	Total
Entities	13,500	3,478	5,097	22,075
Entities (avg.)	16.17	3.06	8.16	8.50
Relations	597	454	456	1,507
Relations (avg.)	0.71	0.40	0.73	0.58

Table 5: Annotation Statistics

(24) Negation*: BHF-Bank: Fosun beteuert, keine Fusion mit Hauck & Aufhäuser anzustreben*

(BHF-Bank: Fosun re-affirms not seeking a merger with Hauck & Aufhäuser)

4. Corpus Statistics

This section summarizes the key characteristics of the final corpus. It contains a total of $2,598$ documents with more than $150,000$ words. Table 4 shows a brief summary of all document types, while Table 5 summarizes the annotation statistics per document type. In total, the annotators labeled $22,075$ entities and $1,507$ relation occurrences. Due to their greater length, news documents contain the largest number of entity mentions, significantly more than the other two document types. Twitter documents on average contain fewer relation mentions than RSS and news documents. The overall fraction of documents containing at least a single relation mention is 58%, a rather high figure that can be attributed to the focused retrieval process which was used to create the initial dataset.

Figure 2 shows the distribution of annotated entities and relations across document types. Companies, general locations, and cities are the most frequent entity types in our dataset. Public transit stops, streets, and routes are less frequently mentioned, and occur predominantly in tweets and RSS traffic reports. This is an expected distribution, since major news outlets generally do not report on day-to-day, local traffic events. With regards to relation types, traffic events like *Traffic Jam*s and *Obstructions* occur very frequently. Other event types occur with lower frequency in our annotated data, in particular, the annotators identified only very few instances of *Canceled Stop* and *SpinOff* events.

4.1. Baseline NER and RE experiments

We conducted a series of experiments to report initial performance figures on the presented corpus for the tasks of named entity recognition and relation extraction. We use the Stanford CoreNLP tools (Manning et al., 2014) for training a NER classifier, and a dependency pattern based model for relation extraction. The relation extraction algorithm, DARE, learns minimal dependency subgraphs that connect all relation arguments, and is described in (Xu et al., 2007; Krause et al., 2012). We did not perform any filtering of the extracted dependency patterns, i.e. we include all learned patterns, even ambiguous or low-frequency ones, in the model.[9]

We randomly split the full dataset into 50% training and 50% test data. The NER model was trained using the standard feature configuration employed by Stanford CoreNLP for NER.[10] The RE models were trained with and without using gold standard NE annotations.

For NER, we evaluated the model's performance both at the token level and at the concept level. The results are shown in Table 6. We see that the overall token-level F1 score is close to 0.85, a respectable figure given the confusability of location subtypes such as streets, stops, and routes in our dataset. For *organization* and *organization-company* entities, the average token-level F1 score is lower at approximately 0.78, but for *location* and its subtypes, it lies between $0.85 - 0.92$ (not shown).

Evaluation Type	Precision	Recall	F1
CRF (token)	0.8966	0.8024	0.8469
CRF (concept)	0.7984	0.6797	0.7343

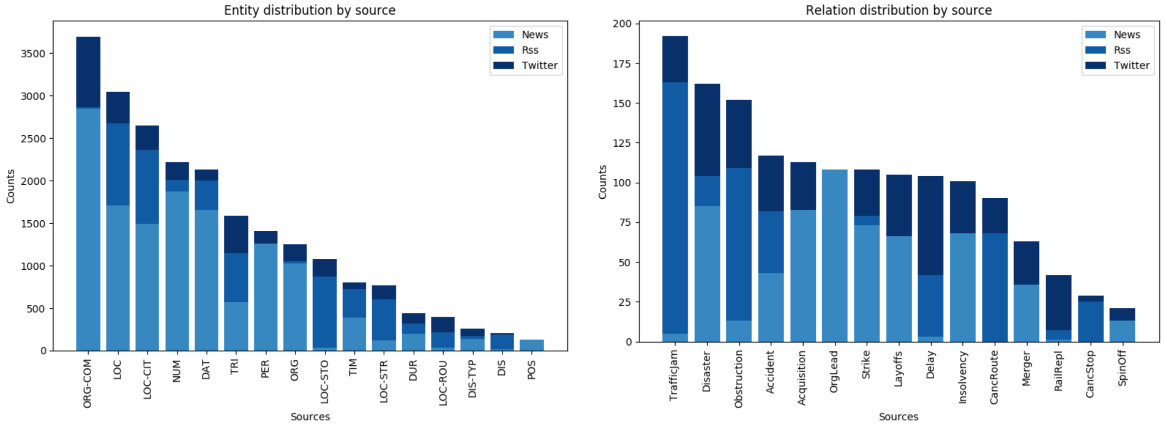
Table 6: Performance of a standard CRF-based NER classifier on the presented dataset.

For relation extraction, the models were evaluated at the mention level, by comparing predicted relation mentions with gold relation mentions. Since our dataset contains n-ary relations with optional and required arguments, we chose a soft matching strategy that counts a predicted relation mention as correct if all predicted arguments also occur in the corresponding gold relation mention, and if all required arguments have been correctly predicted, based on their role, underlying entity, and character offsets / extent. Optional arguments from the gold relation mention that are not contained in the predicted relation mention do not count as errors. In other words, we count a predicted relation mention as correct if it contains all required arguments and is subsumed by or equal to the gold relation mention.

Table 7 shows the results of two RE evaluation runs, once with gold-standard NE annotations, and once without any gold annotations. As can be expected, the performance of the RE models using gold-standard NE annotations is significantly higher than that of the models using the trained NER classifier. The dependency-based DARE model achieves an F1 score of 0.28 using gold-standard NEs, and is biased toward high-precision patterns, at the expense of recall.

[9] Obviously, properly filtering patterns may significantly improve performance, but state-of-the-art RE performance is not the goal of this study.

[10] See nlp.stanford.edu/software/crf-faq.html



(a) Distribution of annotated entities

(b) Distribution of annotated relations

Figure 2: Entity and relation type distribution across document source types

Model	Precision	Recall	F1
DARE (CRF NE)	0.4670	0.1308	0.2043
DARE (Gold NE)	0.5274	0.1923	0.2818

Table 7: Performance of a dependency pattern based RE model on the presented dataset.

5. Related Work

There are very few available NER and RE datasets for German. Most noteworthy are the NER dataset presented by Benikova et al. (2014), and the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003). Both datasets contain annotations only for the three standard entity types, PER, ORG and LOC, as well as MISC/OTHER. The dataset by Benikova et al. includes nested annotations, whereas in our corpus, nested annotations are only annotated if they are required for a relation mention. Both datasets use news articles as their data source, with Benikova et al.'s dataset including Wikipedia texts. In contrast, our dataset also contains Twitter texts, as well as telegraphese-style reports from official traffic channels, which allows for text genre-specific evaluation of NER approaches.

The ACE datasets (Doddington et al., 2004; Linguistic Data Consortium, 2015) are similar to the dataset presented in this paper in that they include both NE and RE annotations. The various ACE datasets developed over the years consider a wide range of entity types, such as PER, ORG, LOC, GPE and FAC. Similarly, a range of different relation types are annotated in these datasets, including geographical, social and business relationships. However, all relations definitions are limited to binary relations, whereas our corpus contains n-ary relation mentions. None of the ACE datasets cover German-language documents.

Other well-known English relation extraction datasets include the corpora prepared for the TAC-KBP challenges (Ji et al., 2011; Surdeanu, 2013), the SemEval-2010 Task 8 dataset (Hendrickx et al., 2010), and the TACRED dataset by (Zhang et al., 2017).

6. Conclusion

We presented a corpus of German Twitter, news and traffic report texts that has been annotated with fine-grained geo-entities as well as a set of mobility- and industry-related events. Many of the event types annotated in the corpus are not available in standard knowledge bases, such as accidents, traffic jams, and strike events. We make the corpus and the guidelines available to the community at `https://dfki-lt-re-group.bitbucket.io/smartdata-corpus`. The dataset is distributed in an AVRO-based compact binary format, along with the corresponding schema and reader tools.

Acknowledgments

This research was partially supported by the German Federal Ministry of Economics and Energy (BMWi) through the projects SD4M (01MD15007B) and SDW (01MD15010A), by the German Federal Ministry of Transport and Digital Infrastructure through the project DAYSTREAM (19F2031A), and by the German Federal Ministry of Education and Research (BMBF) through the project BBDC (01IS14013E).

7. Bibliographical References

Benikova, D., Biemann, C., and Reznicek, M. (2014). NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1251.

Chae, B. K. (2015). Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, 165(Supplement C):247 – 259.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic

Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *Proc. of LREC*.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July. Association for Computational Linguistics.

Hennig, L., Thomas, P., Ai, R., Kirschnick, J., Wang, H., Pannier, J., Zimmermann, N., Schmeier, S., Xu, F., Ostwald, J., and Uszkoreit, H. (2016). Real-time discovery and geospatial visualization of mobility and industry events from large-scale, heterogeneous data streams. In *Proceedings of ACL-2016 System Demonstrations*, pages 37–42, Berlin, Germany, August. Association for Computational Linguistics.

Ji, H., Grishman, R., and Dang, H. T. (2011). Overview of the TAC 2011 Knowledge Base Population Track. In *Proc. of the 4th Text Analysis Conference*.

Kohlschütter, C., Fankhauser, P., Nejdl, W., Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate Detection Using Shallow Text Features. In *Proc. of WSDM*, pages 441–450.

Krause, S., Li, H., Uszkoreit, H., and Xu, F. (2012). Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web. In *Proc. of ISWC*, pages 263–278.

Li, H., Cheng, X., Adson, K., Kirshboim, T., and Xu, F. (2012). Annotating opinions in german political news. In *8th ELRA Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 5.

Linguistic Data Consortium. (2015). Rich ERE annotation guidelines overview. `http://cairo.lti.cs.cmu.edu/kbp/2015/event/summary_rich_ere_v4.1.pdf`.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proc. of ACL: System Demonstrations*, pages 25–30.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241 – 4251.

Osborne, M., Moran, S., McCreadie, R., Von Lunen, A., Sykora, M., Cano, E., Ireson, N., Macdonald, C., Ounis, I., He, Y., Jackson, T., Ciravegna, F., and O'Brien, A. (2014). Real-time detection, tracking, and monitoring of automatically discovered events in social media. In *Proc. of ACL: System Demonstrations*, pages 37–42.

Schulz, A., Ristoski, P., and Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs. In *Extended Semantic Web Conference*, pages 22–33. Springer, Berlin, Heidelberg.

Shimaoka, S., Stenetorp, P., Inui, K., and Riedel, S. (2017). Neural Architectures for Fine-grained Entity Type Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1271–1280, Valencia, Spain, April. Association for Computational Linguistics.

Surdeanu, M. (2013). Overview of the TAC 2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling. In *Proceedings of the Text Analysis Conference*.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans et al., editors, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Xu, F., Uszkoreit, H., and Li, H. (2007). A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity. In *Proc. of ACL*, pages 584–591.

Yaghoobzadeh, Y. and Schütze, H. (2017). Multi-level Representations for Fine-Grained Typing of Knowledge Base Entities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 578–589, Valencia, Spain, April. Association for Computational Linguistics.

Zhang, Y., Zhong, V., Chen, D., Angeli, G., and Manning, C. D. (2017). Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September. Association for Computational Linguistics.