# Domain-specific Entity Spotting: Curation Technologies for Digital Humanities and Text Analytics

Peter Bourgonje, Julián Moreno-Schneider, and Georg Rehm

Language Technology Lab, DFKI
Alt-Moabit 91c, 10559 Berlin, Germany
{peter.bourgonje,julian.moreno_schneider,georg.rehm}@dfki.de
http://digitale-kuratierung.de

## 1 Introduction

The work described in this paper was carried out in the context of task 1 of the CUTE workshop taking place at the conference for Digital Humanities in the DACH-area in Bern in February 2017.[1] The data released for this task consisted of four transcripts of debates as held in the German parliament (on October 28th, 1999, December 16th, 2004, November 15th, 2007 and March 17th, 2011), a series of letters from Goethe's *Die Leiden des jungen Werther*, the section *Zur Theorie des Kunstwerks* from Adorno's *Ästhetische Theorie* and books 3 to 6 from Wolfram von Eschenbach's *Parzival*. These four data sets represent four different domains and display a diversity in language (standard German as spoken today and *Mittelhochdeutsch* from Parzival) and language usage (spoken, i. e., the transcripts of the parliament debates, and written, i. e., the other two corpora). This means that typical entity recognition tools without any specific training data from these four domains are expected to perform with limited quality. The challenge is, thus, either to construct a domain-specific system to deal with the data for the task(s) at hand, or to customise an existing system in order to increase the quality for the given domain. In the context of the Digital Curation Technologies project [1] we are developing a platform that includes entity spotting and, if possible, linking. Because domain-adaptability is an important feature of the platform, we used some of its components for CUTE task 1 and augmented them with a number of domain-specific adaptations and resources. The rest of this abstract is divided into five sections. Section 2 describes the overall approach, rules and procedures that have been applied on all four domains. The other sections describe the domain-specific adaptations.

[1] http://www.creta.uni-stuttgart.de/index.php/de/cute/

## 2 General Methods

The scope of entity spotting is typically limited to detecting words that can be characterised as proper nouns of specific types; words that uniquely refer to a certain entity, such as a person or an organisation. Terms that semantically refer to a single entity, but do not adhere to the usage of proper nouns or otherwise syntactically identifiable constructions are usually excluded from the definition. In this shared task however, the scope of the definition of what an entity actually is, is much broader. It includes the typical proper noun-definition, but also words or word sequences that point to a uniquely identifiable individual, location, organisation, etc. For example, *the president of the European Council* is annotated in the sentence *I now give the word to the president of the European Council*, because it points to a unique individual. Similarly, a phrase indicating a definite location like *her garden* is annotated in the sentence *Heidi and I met the other day in her garden*. Although anaphoric or cataphoric pronouns typically also point to a uniquely identifiable entity mentioned earlier or later in a text, as a general rule, anaphoric pronouns (*he, she, him, her*) are not annotated in the task. This broader definition causes traditional entity spotting methods, that among other things rely on orthographic features, to produce lower quality output. After training a model with the training data provided for the task using a Maximum Entropy approach [2], we got very low F-scores (around 20 at most, and only for a small number of types and domains) and decided to abandon this strategy. Instead, we constructed a rule-based component that relies on several layers (sometimes also referred to as sieves) to spot entity references in the particular domains. All four systems (for the four domains) contained a first layer where we used gazetteer-based name spotting. One of the gazetteers used in this step was extracted from the training data. If some term only appeared as a name in the training data (and not as another type), it was added to the list. Additionally, if the frequency of a term being a name was factor $t$ higher than the frequency of that term not being a name, it was also added to the list. The optimal value for $t$ based on the training data was 5. The resulting list was augmented with domain-specific knowledge from external sources.

## 3 The Test Data Sets

### 3.1 German Parliament Debates

The specific entity types that we annotated for this domain were PER (person), ORG (organisation), LOC (location) and WRK (work, in this case a piece of legislation, typically). In addition to the gazetteer that was extracted from the training data (as described in 2, a list of members of the parliament[3] was

[2] `https://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html#tools.namefind`

[3] `https://de.wikipedia.org/wiki/Liste_der_Mitglieder_des_Deutschen_Bundestages_(18._Wahlperiode)`

augmented with a list of previous German ministers [4], as these people can be expected to be referred to in the context of parliament debates. The combination of these lists was used to spot complete person names. Subsequently, the list of names was split to arrive at a list of both first names and last names (and, optionally, middle names). This list was filtered for both stop words and some common words in German which could easily result in false positives (family names like *Grund* and *Post*, for example). Because the annotations provided in the training data generally included the full noun phrase (NP), we parsed every sentence using the Stanford LexicalizedParser [2] and extracted the NPs. If a word from the list of names appears in an NP, we annotated the whole NP with a PER annotation. Note that the annotations obtained here are likely to be a superset of the annotations obtained using the full names, but the first layer is still used in the case that the parser did not extract an NP correctly. The procedure was then repeated, using the extracted NPs, to also tag NPs that contain indications that they are about a person. The list used here contained words like *Minister, Kanzler, Kommissar*, etc, where these words were not necessarily surrounded by word boundaries (the set of characters represented by $\backslash b$ in regular expression notation), to include morphologically inflected versions (like *Kanzlerin*).

For LOC-type entities the first pass was similarly done, but with a list of countries [5], augmented with continent names like *Europe*, where the full name of the country was annotated if found. Instead of using the NP-matching method, for locations we annotated all NPs that started with the preposition *in*. Especially in the parliament debates there was a lot of ambiguity regarding entities that can be interpreted as a location, but also as an organisation. In the training data, Turkey was usually marked as an organisation, for example. We did not do anything to resolve this ambiguity, but instead annotated ambiguous candidates with all possible types.

For ORG-type entities, we used the list of organisation names obtained from the training data. In addition, we used a very simple regular expression (*/[A-Z]{2,}*) to detect cases like [*NATO, EU, CDU*].

For WRK-type entities we used the list of WRK-type terms obtained from the training data. In addition, we used the following list in combination with NP-based spotting and annotated the complete NP that contained one of the following words (possibly appearing in a compound): [*vertrag, verträge, charta, verfassung, kriterien*].

### 3.2 Goethe

The entity types that were annotated for this domain were PER, LOC and WRK. Following the annotations provided in the training data, we applied the same NP-based approach as with the previous corpus, but excluded determiners from the annotation, as this resulted in higher F scores on the training set. The

[4] `https://de.wikipedia.org/wiki/Liste_der_deutschen_Bundesminister`
[5] `http://www.laenderdaten.de/laender.aspx`

root forms and all possible inflections of determiners such as [*ein, sein, mein, ihr*] and all articles and their inflected forms ([*der, die, das, den, dem*]) were excluded from being annotated as PER. For the LOC-type entities we followed the same approach, where we first obtained the list of names from the training data to subsequently use the NP-approach and exclude determiners and articles from the annotation. For the WRK-type entities, the same procedure was used.

### 3.3 Adorno

For this domain we annotated entities of the types PER and WRK. Because of the scope of this domain, we augmented the list obtained from the training data with a number of additional lists, including a list of members of the Frankfurt School of social theory [6], and lists of well-known composers, writers and painters collected from several public sources. In line with the training data, we also checked for genitive constructions involving any of the names from the list (basically considering every name if it has the German genitive-suffix *-s*). With this list (and expansion for genitive cases) we used the NP-approach. For WRK-type entities we obtained only the list of entities from the training data and augmented this with the following words: [*gedicht, appasionata, oper, erzählung, sonate, quartett, symphonie, orchesterwerk*]. With the resulting list, the NP-approach was also used for this corpus and entity type.

### 3.4 Parzival

For the Parizval domain we annotated entities of the type PER and LOC. Because this domain consists of Mittelhochdeutsch for which we had no parser available, we could only use the gazetteer-approach for this domain. We obtained the list of names from the training data and worked with that. However, because unlike in Standard German (*Hochdeutsch*) where all nouns are capitalized, in the Mittelhochdeutsch from Parzival, this is not the case, and apart from sentence-initial words, only names are capitalized. So for this domain we annotated every title-cased word that is not sentence-initial as an entity of type PER. For LOC-type entities we resorted to using only the list obtained from the training data.

## 4 Results

The tables below include the results as communicated by the organisers of the task, after submitting our annotated evaluation data. The *bl-ner* system represents the Stanford NER with the newest available German model, *DFKI* represents our system and *IMS* and *IMS2* represent two systems from the organisers of the task.

[6] `https://en.wikipedia.org/wiki/Frankfurt_School`

**Table 1.** Precision for the indicated entity types and domains

| | Corpus | Bundestagsdebatten | Adorno | Parzival | Werther |
|---|---|---|---|---|---|
| PER | bl-ner | 31.58 | 35.34 | 42.86 | 59.09 |
| | dfki | 20.69 | **71.88** | 41.67 | 33.59 |
| | ims | 0.00 | 1.89 | 39.53 | 5.23 |
| | ims2 | **40.62** | 45.45 | **63.24** | **65.00** |
| LOC | bl-ner | 33.82 | 0.00 | 28.57 | 28.21 |
| | dfki | 42.31 | 0.00 | 29.36 | 34.02 |
| | ims | 25.00 | 0.00 | 0.00 | 0.00 |
| | ims2 | **72.22** | 0.00 | **63.16** | **60.80** |
| ORG | bl-ner | 3.70 | 0.00 | 0.00 | 0.00 |
| | dfki | 3.85 | 0.00 | 0.00 | 0.00 |
| | ims | 0.00 | 0.00 | 0.00 | 0.00 |
| | ims2 | **42.28** | 0.00 | 0.00 | 0.00 |
| WRK | bl-ner | 0.00 | 0.00 | 0.00 | 0.00 |
| | dfki | 40.00 | 0.00 | 0.00 | **25.00** |
| | ims | 0.00 | 0.00 | 0.00 | 0.00 |
| | ims2 | **42.86** | 0.00 | 0.00 | 0.00 |

**Table 2.** Recall for the indicated entity types and domains

| | Corpus | Bundestagsdebatten | Adorno | Parzival | Werther |
|---|---|---|---|---|---|
| PER | bl-ner | 0.31 | **2.42** | 0.46 | 5.36 |
| | dfki | 0.31 | 2.37 | 17.79 | **15.99** |
| | ims | 0.00 | 0.26 | 1.75 | 2.27 |
| | ims2 | **0.67** | 0.52 | **21.56** | 12.07 |
| LOC | bl-ner | **3.48** | 0.00 | 0.30 | 1.67 |
| | dfki | 3.33 | 0.00 | 4.85 | **12.42** |
| | ims | 0.15 | 0.00 | 0.00 | 0.00 |
| | ims2 | 1.97 | 0.00 | **10.91** | 11.52 |
| ORG | bl-ner | 0.32 | 0.00 | 0.00 | 0.00 |
| | dfki | 1.94 | 0.00 | 0.00 | 0.00 |
| | ims | 0.00 | 0.00 | 0.00 | 0.00 |
| | ims2 | **37.10** | 0.00 | 0.00 | 0.00 |
| WRK | bl-ner | 0.00 | 0.00 | 0.00 | 0.00 |
| | dfki | **5.41** | 0.00 | 0.00 | **1.35** |
| | ims | 0.00 | 0.00 | 0.00 | 0.00 |
| | ims2 | 4.05 | 0.00 | 0.00 | 0.00 |

## 5 Conclusions

Due to the specific and rather different requirements for this shared task compared to typical NER approaches, the traditional data-driven and statistical approaches only provide limited annotation quality. Although the scores obtained using the methods described show considerable room for improvement, we think that especially in the area of Digital Humanities, rule-based approaches can play an important role in augmenting traditional systems to precision particularly. The specific annotation requirements, including for some types full noun phrases, and hence the included NP-parsing component may be of limited use for future applications of our Digital Curation platform. But both the overall task of augmenting our NER component with a limited set of domain-specific rules and resources and the development of a component dealing with a language with very limited NLP resources available (*Mittelhochdeutsch*) was a very useful exercise for upcoming challenges that we will face in the Digital Humanities domain.

**Acknowledgments.**
The project "Digitale Kuratierungstechnologien (DKT)" is supported by the German Federal Ministry of Education and Research (BMBF), "Unternehmen Region", instrument "Wachstumskern-Potenzial" (no. 03WKP45). More information on the project can be found online at `http://www.digitale-kuratierung.de`.

## References

1. Bourgonje, P., Moreno-Schneider, J., Nehring, J., Rehm, G., Sasaki, F., Srivastava, A.: Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) The Semantic Web: ESWC 2016 Satellite Events (June 2016)
2. Rafferty, A.N., Manning, C.D.: Parsing three german treebanks: Lexicalized and unlexicalized baselines. In: Proceedings of the Workshop on Parsing German. pp. 40–46. PaGe '08, Association for Computational Linguistics, Stroudsburg, PA, USA (2008), `http://dl.acm.org/citation.cfm?id=1621401.1621407`