Further with Knowledge Graphs. M. Alam et al. (Eds.)
AKA Verlag and IOS Press, 2021

doi:10.3233/SSW210041

Annotating Entities with Fine-Grained Types in Austrian Court Decisions

Artem Revenko[1][0000−0001−6681−3328], Anna Breit[1][0000−0001−6553−4175], Victor Mireles[1][0000−0003−3264−3687], Julian Moreno-Schneider[2][0000−0003−1418−9935], Christian Sageder[3], and Sotirios Karampatakis[1][0000−0001−7436−7620]

[1] Semantic Web Company GmbH, Austria
{firstname.secondname}@semantic-web.com
[2] DFKI GmbH, Germany
julian.moreno_schneider@dfki.de
[3] Cybly GmbH, Austria
christian.sageder@cybly.tech

Abstract. The usage of Named Entity Recognition tools on domain-specific corpora is often hampered by insufficient training data. We investigate an approach to produce fine-grained named entity annotations of a large corpus of Austrian court decisions from a small manually annotated training data set. We apply a general purpose Named Entity Recognition model to produce annotations of common coarse-grained types. Next, a small sample of these annotations are manually inspected by domain experts to produce an initial fine-grained training data set. To efficiently use the small manually annotated data set we formulate the task of named entity typing as a binary classification task – for each originally annotated occurrence of an entity, and for each fine-grained type we verify if the entity belongs to it. For this purpose we train a transformer-based classifier. We randomly sample 547 predictions and evaluate them manually. The incorrect predictions are used to improve the performance of the classifier – the corrected annotations are added to the training set. The experiments show that re-training with even a very small number (5 or 10) of originally incorrect predictions can significantly improve the classifier performance. We finally train the classifier on all available data and re-annotate the whole data set.

Keywords: Named Entity Recognition · Entity Typing · Legal Corpus · Natural Language Processing

1 Introduction

The ever-increasing amount of unstructured data available in digital form results in a need for technologies that support users in the task of structuring, interpreting or, on a general level, making sense of these data, ideally in an automated way [6, 3]. This is basically the core business of semantic processing, and one of the tasks that has traditionally been very central is Named Entity Recognition (NER). NER is usually an upstream task to concrete use cases such as

text knowledge graph population, information extraction or question answering. Such downstream applications benefit from high-quality NER output, which is why the task of NER is an important and often critical one. At the same time, many NER tools are limited to distinguishing only a relatively small set of entity types, because most of the popular corpora and data sets that are used for training these tools [24, 16] are annotated for the entity types person, location and organisation only. It is especially difficult to find annotated corpora in languages other than English. Producing such a data set is a very expensive task and is completely infeasible in practical applications. In domain-specific corpora it might be even difficult to produce the annotations of the coarse-grained types mentioned above, because the domain-specific use of language and the special terms can easily confuse the general-purpose NER tools resulting in noisy annotations.

In this paper, we tackle the task of classifying the entities recognized by a general-purpose NER tool into fine-grained entity types chosen by a domain expert. We conduct a case study on the corpus of Austrian court decisions collected from The Legal Information System of the Republic of Austria[4] in German language. We consider real industry settings, therefore, we rely on a very small amount of annotated training data that is produced by a domain expert in frames of this work, and we additionally aim at recovering from noisy annotations produced by the general purpose NER tool on the domain-specific corpus.

The task is motivated by the commercial tool `LawThek`[5] that offers its customer a way to access the Austrian legislation and related legal documents. The customers benefit from additional enrichments produced by the system and a domain-specific NER model would provide a further extension that would help the user to better retrieve the relevant documents and identify useful information in those documents.

Open Data All the data, including the original corpus and the manual annotations is publicly available[6]. The code to repeat the experiments is also publicly available[7].

2 Related Work

The most popular NER tools, e. g., SpaCy[8], Stanford CRF-NER[9], and the OpenNLP NameFinder[10], require large amounts of training data and typically distinguish only a small set of entity types. A significant effort has been undertaken

[4] https://www.ris.bka.gv.at
[5] https://lawthek.eu/home
[6] https://doi.org/10.5281/zenodo.4625767
[7] https://github.com/semantic-web-company/austrian_court_decisions
[8] https://spacy.io
[9] https://nlp.stanford.edu/software/CRF-NER.shtml
[10] https://opennlp.apache.org/docs/1.8.3/apidocs/opennlp-uima/opennlp/uima/namefind/NameFinder.html

to create training data sets with more fine-grained entity types, for example two benchmarks: FIGER [1] and OntoNotes [19]. In the work presented here, we investigate a method that significantly reduces the required amount of training data.

Fine-grained NER Several systems addressing the fine-grained NER task have been successfully applied on those data sets. In [28] and [27] authors exploit modern transformer-based language models to learn joint embeddings of words and entities from large entity-annotated corpora. These models allow one to effectively combine the semantic signals retrieved from both words and entities. The authors of K-Adapter [26] also build on top of modern transformer language models adding special adapters that inject multiple kinds of diverse knowledge. These adapters are task-specific and are, therefore, able continuously infuse knowledge, without forgetting.

The mentioned models reach state of the art performance on the mentioned fine-grained data sets. However, these and other similar models (e.g. [23, 15]) rely on significant amounts of data, incorporating external knowledge bases to learn each type of entities. On the contrast, we are interested in learning the basic "concept" of type verification from very small data and seek to find a model that is able to benefit from cross-type interactions.

Entity Linking and Distant Supervision Solving the problem of typing entities can also be addressed by the methods of Entity Linking (EL), in which a diversity of string-matching techniques are employed to relate found mentions of NEs with known entities. Common tools for entity linking include DBPedia Spotlight[4], Entity Fishing[13] or Babelfly[14]; and the number of approaches to EL continues to increase (see [18] for a recent overview). EL, however, assumes the existence of a catalog of entities (preferably containing additional knowledge). In domain-specific settings, such as the one treated here, this is an unreasonable assumption that limit the potential range of use cases. One reason is that no entity catalog is complete and the entities specific to a domain are unlikely to be found in general domain ones. Another reason is that, for some purposes, entities of interest in the text do not correspond to any particular real world entity. For example, *the complainant* or *the buyer* are specific entities within a document, but are not a surface form of any particular real world entity.

The idea of Distant Supervision [7, 21] is to employ EL to create an (entity) annotated corpora. The drawbacks are the inevitable errors produced by the EL tools, especially false negatives. To mitigate this problem, researchers design neural models that are able to cope with such noisy data and/or help to recover from noisy [22, 12]. Yet, the Distant Supervision approach and its extension suffer from the same shortcomings as EL, because these techniques rely on EL and external data sources, whereas our approach is bound only to the domain-specific corpus and domain experts.

3 Data

The current paper presents a use case study of re-tagging entities with fine-grained types selected by the domain expert in a legal corpus in German language. The corpus is downloaded from Legal Information System of the Republic of Austria (RIS) – a computer-assisted information system on Austrian law[11]. The corpus for the current experiments consists of 2500 randomly selected court decisions taken from the category "Judikatur" (= Judicature of the courts), section "Bundesverwaltungsgericht" (= Federal Administrative Court). These documents are not older than 2014 and provided in the original German language. A large part of the decisions concern decisions on asylum procedures. The personal information is anonymised in the documents, see Example 1.

Example 1. The following quote is taken from the document with European Case Law Identifier `ECLI:AT:BVWG:2021:W109.2195466.1.00`. The replacement token *XXXX* mask the real name and the birth date of an individual.

> Gemäß § 8 Abs. 1 Asylgesetz 2005 wird *XXXX* , geb. *XXXX* , StA. AFGHANISTAN, der Status der subsidiär Schutzberechtigten in Bezug auf den Herkunftsstaat Afghanistan zuerkannt.

3.1 Original Named Entity Annotations

The corpus is initially annotated with a NER tool, based on BERT neural networks, which is developed following the work of Kamal Raj.[12] The original approach is adapted to allow for training of a new model on the WikiNer data set [17] – a general purpose data set containing four coarse-grained types. In the initial annotation process the tool annotated 39,324 Persons (`PER`), 215,699 Locations (`LOC`), 183,045 Organizations (`ORG`) and 324,926 Miscellanea (`MISC`). As expected, given the type of documents, court decisions, the `PER` type is the least abundant, while the other three types are present one order of magnitude as many times.

The model is quite confused by the domain-specific usage of language and also special symbols such as anonymization masks and frequent abbreviations. Therefore, we identified many noisy annotations in the original coarse-grained annotations.

3.2 Selection of Named Entity Types of Interest

The original annotations are collected and analysed to group the entities and produce new entity types[13]. These new types are reviewed by the domain expert from the point of view of the targeted functionality of the final application and

[11] https://www.ris.bka.gv.at/UI/Erv/Info.aspx accessed on March 26

[12] https://github.com/kamalkraj/BERT-NER

[13] For the purpose of the current work we omit the details of type induction as the presented classification approach does not depend on the type selection procedure.

9 fine-grained entity types are selected for further analysis, see also Table 1 for the definitions of types.

Gericht to recognize the different courts, therefore, identify the level at which a certain decision was taken. This could be potentially accomplished with the usage of a gazetteer for Austrian courts, however, colloquial usages and international courts would be missed.

Behörde, Administration to be able to see the involvement of different government offices into processes, therefore identifying in the stakeholders from the government.

Verwandtschaft to group physical persons into clusters. Could be further used for information extraction or grouping by kinship.

Land, Staat to identify potential international involvement.

Information, Quelle, Daten to find external information sources that could potentially be interlinked.

Zocken, Spielhallen to group documents w.r.t exact type of criminal activity. This activity was particularly prominent.

Rolle, Gruppe von Personen to identify roles, persons can be assigned too used to for further grouping / information retrieval. E. g. person is a complainant, a buyer, a seller, etc.

Kriegshandlung, Konflikt to group documents w.r.t. exact type of non-criminal activity that could be triggers. In this case, many armed conflicts lead to asylum procedures.

Strasse, Addresse to get more detailed GEO locations down to an exact address which can be checked against an address database.

The new types are chosen from the analysed data and with the idea to provide some additional value for the end user. However, the types appear to have overlaps and do not necessarily cover all the data. For example, the types "Gericht" and "Behörde" are much closer to each other than to "Strasse, Addresse", for example. On the other hand, the type "Rolle, Gruppe" comprises entities of quite different semantics and could be potentially split into two types; the choice is done in favor of this joint type because in the random sample we find many borderline entities such as "complainants" or 'legal representatives". We note that this choice of types makes the classification task more challenging, often an entity might belong to more than one type. We see it necessary to cope with this choice as it was provided by the expert from the point of view of the domain of application itself.

In the following we add an artificial type "Other" that is reserved for 1) original noisy annotations and 2) entities that do not belong to any of the described types.

3.3 Manual Annotations

We annotate a small sample of data manually. For this purpose we choose a few seed entities that unambiguously belong to the chosen fine-grained types and

Table 1. Fine-grained types with definitions and synonyms.

Type	**Definition** **Synonyms**
Gericht	Ort zur gesetzlichen Entscheidung von Rechtsstreitigkeiten *Place where legal disputes are decided*
	Gericht, Gerichtshof, Tribunal *Court, tribunal, court of law*
Behörde, Administration	eine öffentliche Stelle, die die Aufgaben der öffentlichen Verwaltung wahrnimmt *A public body that is involved in public administration*
	Behörde, Administration, Amt *government office*
Verwandtschaft	Zugehörigkeit zur gleichen Familie, gleiche Abstammung *Family or ancestry relations*
	Verwandtschaft, Angehörige, Familie *Relatives, family, kinship*
Land, Staat	unabhängiges politisches Gebilde *Independent political entity*
	Land, Staat *Country, State*
Information, Quelle, Daten	wissenschaftlich auswertbares Primärmaterial *Primary material for scientific research*
	Information, Quelle, Daten *Information, source, data*
Zocken, Spielhallen	Spiel um Geld, bei dem Gewinn und Verlust vom Zufall abhängen *A game where money is waged and whose outcome depends on change*
	Glückspiel, Zocken *Games of chance*
Rolle, Gruppe von Personen	eine Gruppe, deren Mitglieder sich in Kontakt miteinander befinden, gemeinsame Ziele verfolgen und sich als zusammengehörig empfinden *A group whose members are in contact with each other, pursue common goals and feel that they belong together*
	Rolle, Gruppe *Role, group*
Kriegshandlung, Konflikt	Vorgehen gegen einen Gegner oder Feind *Violent actions against an enemy*
	Kriegshandlung, Konflikt, Anschlag, Angriff *War waging, conflict, attack, aggression*
Strasse, Addresse	die genaue örtliche Bezeichnung *An exact description of a location*
	Straße, Addresse, Anschrift *Street, Address*

identify their occurrences in the documents, see Example 2. We then manually verify the correctness. This original manually annotated data set is publicly available, the number of entities is presented in Table 2. There are in total 109 annotated instances for 9 types resulting in ≈12 instances per type on average.

We choose clean, unambiguous entities as the seed entities. This process is tedious for the expert, as it requires to skim through the documents to identify those entities, therefore it is not feasible to produce a large initial data set. Yet, these initial manual annotations are not expected to represent the whole data set, but rather produce a good seed data set for the chosen fine-grained types.

Example 2. The following quote is taken from the document with European Case Law Identifier `ECLI:AT:BVWG:2015:W162.1418315.1.00`. The entity "**Tschetschenien**" is an example of type "Land, Staat".

> Seit 2002 sind in **Tschetschenien** über 2.000 Personen entführt worden, von denen über die Hälfte bis zum heutigen Tage verschwunden bleibt.

Table 2. Statistics of manually annotated and manually verified data sets.

Type	**Annotated**	**Verified**
Gericht	9	47
Behörde, Administration	12	36
Verwandtschaft	12	1
Land, Staat	12	60
Information, Quelle, Daten	12	44
Zocken, Spielhallen	20	15
Rolle, Gruppe von Personen	11	67
Kriegshandlung, Konflikt	15	3
Strasse, Addresse	6	19
Other	-	255
Total	109	547

4 Classifier

We design a classifier that is capable of verifying the type of a given entity. We take into account the lack of training data and, therefore, aim at a robust solution that would be capable to efficiently use some preliminary training to solve the task. Therefore, we focus our attention on the Target Sense Verification (TSV) task [2]. The core task is a binary classification task – given an entity of interest in a context and definitions / hypernyms of an entity's sense decide if the entity is mentioned in the given sense or not. The main challenge of the task is to generalize the ability of verifying the sense of an entity to unseen domains and senses.

Our task setting can be formulated in a similar way, with the difference of verifying the type of an entity instead of its sense – target type verification. Yet the inputs – the target in context, the definition and the synonyms of the target type – are very similar to TSV. Therefore, we reuse the results of the challenge and employ the model from [25][14] that showed the best results in Task 3 of the challenge. The model is based on a transformer model (we use Bert [5][15]), and it marks the input to let the encoder focus on the target and sense/type identifiers.

4.1 General Purpose Fine-tuning

For fine-tuning our model on the proposed task, we chose a learning setup similar to [2]: we created a training set where each instance consists of a target word in a context (e.g. *the **spring** was broken*), and a target sense, indicated by the definition and hypernyms of the target word (e.g., *the season of growth* and *season*) as well as a label indicating if the target word was used in the target sense (in this case, *F*). As it has been shown, that the proposed classifier is to some extent able to transfer intrinsic classification capabilities gained on a general purpose data set into specific domains [25], we generated this training set from German Wiktionary. Herefore, we scraped the entries of nouns for which multiple meanings, definitions, hypernyms and examples were available. We removed senses that were too close (i.e., those that were listed as *[1a]* and *[1b]* instead of *[1]* and *[2]*) and manually cleaned the data set to reduce noise. The final training set consists of 3,564 instances, with 55% positive examples and 45% negative ones.

4.2 Training

We always start from a model tuned on TSV data set. We remind that the classification task is binary and the input is encoded as:

$$\texttt{[CLS]}\ \texttt{T}_1\ \texttt{T}_2\ \ldots\ \texttt{\$}\ \texttt{TARGET\_T}_1\ \texttt{TARGET\_T}_2\ \ldots\ \texttt{\$}\ \texttt{T}_N\ \ldots$$
$$\texttt{[SEP]}\ \texttt{DEF\_T}_1\ \texttt{DEF\_T}_2\ \ldots\ \texttt{\$}\ \texttt{SYN}_1\texttt{\_T}_1\ \texttt{SYN}_1\texttt{\_T}_2\ \ldots\ \texttt{\$}\ \texttt{SYN}_2\texttt{\_T}_1\ \ldots\ \texttt{[SEP]},$$

where $\texttt{T}_i$ stands for ith token of the context, $\texttt{TARGET\_T}_i$ stands for ith token of the target entity, $\texttt{DEF\_T}_i$ – ith token of the definition of the target type, $\texttt{SYN}_j\texttt{\_T}_i$ – ith token of the jth synonym of the target type, and `[CLS], [SEP]` are the special tokens used by the model, "`...`" denotes a continuation of an enumeration, i. e. $\texttt{T}_i$ or $\texttt{DEF\_T}_i$. It is straightforward to generate negative training examples – it is enough to substitute the definition and the synonyms of the correct type with some other type's definition and synonyms. In the preliminary experiments on a

[14] We note that a very similar model is introduced in [8]. However, the former is an extension and is better suited for the task at hand.

[15] It has been demonstrated that domain-specific pre-trained language models such as BioBert [10] or PatentBert [9] can improve the performance on various NLP tasks, however, to our best knowledge no publicly available legal German language model exists at the moment.

different similar data set reported in [11] we identified that the optimal number of negative examples is ≈70% of the available other types, therefore, in our experiments for each positive example we generate 6 negative examples.

In the first experiment with only manually annotated data we use all the data for continuing fine tuning, because the size of the data set is small, see Section 3.3 and in particular Table 2. We train the model for 3 epochs. For the consecutive experiments experiments with manually verified data we use up to 15 instances per type from the verified data with 2 negative examples per positive as the development data set. The model is trained for 7 epochs in each run and the best scoring model (in terms of F_1 score on the development data set) is chosen for further evaluation. The training batch size is set to 8, for the rest the parameters are set either as reported by authors or as the defaults by the training framework `PyTorch` [20].

5 Experiment

We recap that the goal of the experiment is to annotate the originally recognized entities of coarse types with new fine-grained types as defined in Table 1. For this purpose we first fine-tune the model (Section 4) on the German TSV data set (Section 4.1). Then we manually annotate a small sample of data with target fine-grained types (Section 3.3). Further we fine-tune the model on this small manually annotated data set and generate predictions of new entities. We randomly take 547 predictions and manually evaluate the correctness of predictions (Section 5.1). Finally, we combine all the manually annotated entities – the initial manually annotated data set and the verified sample – and fine tune our model on the whole data set. We use this latter model to generate predictions for the complete corpus.

As described in Section 3.1 the original coarse annotations contain many noisy annotations that actually do not belong to any type. Another goal of the experiment is to evaluate how efficiently we can recover from these noisy annotations and correctly reject them.

5.1 Manual Verification

After fine-tuning on the manually annotated data set, we manually verify the results. The trained model is used to automatically generate predictions on the original corpus. Then, a sample of those predictions is taken to manually verify their correctness. We take ≈55 randomly sampled predictions per **predicted** fine-grained type.

For each of the fine-grained type, we create a separate spreadsheet, containing one prediction per row. In detail, on each row, the surface form of the entity, the predicted type, the position of the entity in the context and the full context are given. Six independent reviewers were tasked to examine the correctness of these samples. The review is performed as a binary classification task, by determining if the prediction is correct or not. Additionally, in case of a false prediction, the

reviewers are required to provide a proper classification for the entity according to 9 types described in Table 1. The reviewers are instructed to be tolerant only in the case of incomplete boundaries of the annotations.

Example 3. In the following sample, Eurostat is tagged as of type "Information, Quelle, Daten". Even that the annotation is incomplete, i. e. it does not contain the date reference, this prediction is considered correct.

> (...)4. Qu. 2015 655 35 35 15 565 GESAMT 3.510 345 170 120 2.870 Die Daten werden auf die Endziffern 5 oder 0 auf- bzw. abgerundet. (Eurostat 18.9.2015a; Eurostat 18.9.2015b; Eurostat 10.12.2015; **Eurostat** 3.3.2016b) In erster Instanz für das Asylverfahren in Polen zuständig ist das Office for Foreigners(...)

As the predictions are often incorrect, the resulting verified data set is quite unbalanced, with many wrongly annotated entities that end up in the type "Other". The actual frequencies of the verified fine-grained types are presented in 2.

5.2 Results

In the first run the model is trained on manually annotated data for 3 epochs, the results are presented in Table 3[16]. We note that only one type "Gericht" reaches F_1 score above 0.5. It is remarkable that the performance does not always correlate with the size of the training set for a given types; for example, "Strasse, Addresse" with only 6 training samples reaches F_1 of 0.35 that is significantly higher than F_1 score of "Kriegshandlung, Konflikt" with 15 training samples.

Overall accuracy and F_1 are below 0.3. These low scores can be explained by a significantly different real verified data as compared to the clean seed instances used for the manual annotations. Therefore, we do further runs of the experiment taking a few samples of the manually verified types. In these runs we extend the training set by 5 and 10 instances, respectively, of the *incorrectly labelled* manually verified data set. We add those instances with the corrected tag and, in the preparation of the training set, we generate negative examples for them as described in Section 4.2. We only do it for those types that have at least 20 incorrectly labelled instances, namely for "Information, Quelle, Daten", "Rolle, Gruppe von Personen", "Behörde, Administration", "Gericht", "Land, Staat". We also add 5 and 10 instances from the type "Other" to generate negative examples for training. Therefore, we use roughly 5% and 10% of the manually verified data set, respectively, to extend the training set. For both settings 3 runs were performed with randomly chosen 5 and 10 instances, average results are reported.

The results of training on the data set extend by 5 additional instances are presented in Table 4. We note that now for 5 types the F_1 scores are 0.5 and

[16] We used the functionality of `scikit-learn` (https://scikit-learn.org/stable/modules/model_evaluation.html#classification-report visited on 02.04.2021) to produce the classification report.

Table 3. Results of the evaluation of the model trained on manually annotated data set. Overall accuracy is .27. The values above 0.5 are in **bold**, the maximum value per column is in *italics*.

	precision	recall	F_1 score	support
Gericht	**.51**	**.55**	**.53**	47
Behörde, Administration	.06	.08	.07	36
Verwandtschaft	0	0	0	1
Land, Staat	.43	.43	.43	60
Information, Quelle, Daten	.25	.30	.27	44
Zocken, Spielhallen	.28	***1.0***	.44	15
Rolle, Gruppe von Personen	.17	.13	.15	67
Kriegshandlung, Konflikt	.05	***1.0***	.09	3
Strasse, Addresse	.24	**.63**	.35	19
Other	***.66***	.16	.26	255
macro avg	.27	.43	.26	547
micro avg	.46	.27	.28	547

above. Remarkably, for the type "Zocken, Spielhallen" the results have grown significantly, though no additional training instances were added. On the other hand, the type "Strasse, Addresse" is now much more often misclassified as false negative, therefore recall and F_1 are much lower. This demonstrates the cross-type interactions in our model, i. e. the model seems to be able to learn the idea of type verification in general and not fit to the specific training data at hand.

Overall accuracy and F_1 scores grow by $\approx$0.2, which can be considered a very significant growth. This demonstrates that with the current model and training routine even a very small amount of real annotated data can have a significant impact on classification results.

We are further interested if adding more data can still have a significant impact and proceed with 10 additional instances per type for the populated types (with support more than 20), the results are presented in Table 5. We observe further grows of scores, now 6 types have F_1 scores of 0.5 and above. Though the absolute values now demonstrate a more moderate growth of not more than 0.1 for accuracy and average F_1, the variance has significantly decreased for most scores. This might due to the fact that the verified data set is very challenging for classification, including noisy and arguable entities. Therefore, we slowly see the "saturation" of scores, i. e. some entities are outliers in its types and can either not be classified well or corrupt the model if added as training instances. However, as the variance decreases, with 10 added instances we observe less impact from those noisy instances. We also note further cross-type learning as, for example, for "Strasse, Addresse" and "Zocken, Spielhallen" the scores keep growing without any further training instances of this type.

We see that using the extended data set we also manage to train the model to recover from noise to a certain extent. In the latter experiment the F_1 score for the type "Other" is above 0.5. Yet, in both extended experiments we observe that often the precision for this type is higher than recall.

Table 4. Averaged results of the evaluation of the model trained on manually annotated with 5 additional instances for each type with support higher than 20. Best models after epochs 4, 3, 3. Overall accuracy is **.50** ± .04. The values above 0.5 are in **bold**, the maximum value per column is in *italics*.

	precision	recall	F_1 score	support
Gericht	**.69** ± .09	***.91*** ± .04	***.77*** ± .04	47
Behörde, Administration	.25 ± .04	**.67** ± .03	.36 ± .04	36
Verwandtschaft	0	0	0	1
Land, Staat	**.70** ± .08	**.85** ± .03	**.76** ± .05	60
Information, Quelle, Daten	.28 ± .05	**.75** ± .10	.40 ± .03	44
Zocken, Spielhallen	***.83*** ± .17	.45 ± .10	**.54** ± .06	15
Rolle, Gruppe von Personen	**.74** ± .05	**.64** ± .07	**.68** ± .02	67
Kriegshandlung, Konflikt	**.80** ± .20	**.56** ± .10	**.60** ± .10	3
Strasse, Addresse	.42 ± .20	.09 ± .05	.14 ± .07	19
Other	**.66** ± .01	.26 ± .10	.35 ± .10	255
macro avg	**.55** ± .03	**.55** ± .03	.47 ± .03	547
micro avg	**.62** ± .01	**.50** ± .04	.48 ± .06	547

Table 5. Averaged results of the evaluation of the model trained on manually annotated with 10 additional instances for each type with support higher than 20. Best models after epochs 6, 5, 7. Overall accuracy is **.59** ± .02. The values above 0.5 are in **bold**, the maximum value per column is in *italics*.

	precision	recall	F_1 score	support
Gericht	**.63** ± .08	***.96*** ± .01	***.75*** ± .05	47
Behörde, Administration	.40 ± .04	**.60** ± .04	.47 ± .03	36
Verwandtschaft	0	0	0	1
Land, Staat	**.69** ± .05	**.88** ± .01	**.77** ± .02	60
Information, Quelle, Daten	.34 ± .02	**.71** ± .07	.46 ± .01	44
Zocken, Spielhallen	**.82** ± .13	.49 ± .06	**.61** ± .08	15
Rolle, Gruppe von Personen	**.62** ± .06	**.84** ± .01	**.71** ± .04	67
Kriegshandlung, Konflikt	***.83*** ± .16	**.67** ± .00	**.72** ± .08	3
Strasse, Addresse	**.64** ± .05	.35 ± .06	.44 ± .03	19
Other	**.76** ± .02	.39 ± .06	**.51** ± .06	255
macro avg	**.57** ± .03	**.59** ± .01	**.54** ± .02	547
micro avg	**.66** ± .02	**.59** ± .02	**.58** ± .03	547

Analysis of errors Some errors are listed in Table 6. For the first row the entity **Art** (stands for "article") is annotated as "Other" because the phrase is incomplete. However, the model still classifies it as "Information, Quelle, Daten". In the second row we see the inverse error. These errors are due to incomplete original annotations, therefore we think some heuristics to extend the annotations to complete entities could be useful. In the third and fourth rows we see examples of entities that could be classified into more than one type, however, the manual annotators had to choose only one type for their input.

Table 6. Table of some prediction errors by a model trained on manually annotated with 10 additional instances for each type with support higher than 20.

Context with **target**	True type	Predicted type
ihrer Religion, ihrer Nationalität, ihrer Zugehörigkeit zu einer bestimmten sozialen Gruppe oder ihrer politischen Ansichten bedroht wäre (**Art**. 33 Z 1 GFK), es sei denn, es bestehe eine innerstaatliche Fluchtalternative (§ 11 AsylG 2005).	Other	Information, Quelle, Daten
Gemäß § 9 Abs. 2 BFA-VG sind bei der Beurteilung des Privat- und Familienlebens im Sinne des Art. **8 EMRK** insbesondere folgende Punkte zu berücksichtigen	Information, Quelle, Daten	Other
Zwar hat auch die **somalische Polizei** eine eigene Anti-Terror-Einheit gegründet, trotzdem ist die NISA bei der Reaktion auf Terrorangriffe in Mogadischu hauptverantwortlich (EASO 2.2016).	Rolle, Gruppe von Personen	Behörde, Administration
Quellen: - **AA** - Auswärtiges Amt (1.4.2015b): Russische Föderation - Reise- und Sicherheitshinweise	Behörde, Administration	Information, Quelle, Daten

6 Conclusion

We aim at producing an annotated fine-grained domain-specific data set for training an NER tool, while attempting to reduce the high cost of manual annotations produced by domain experts. In particular, we address a realistic case of 1) having to cope with the choice of fine-grained types produced by domain experts and 2) small golden training data set. Therefore, we formulate the (named) entity typing task as a binary classification task and explore the cross-type learning of the model. We exploit a binary classifier that has shown good results on a similar binary task of sense verification.

The experiments demonstrate that the initial clean and manually annotated data set might not be enough to achieve good classification results. However, adding even a small amount of randomly sampled incorrectly classified entities to the training set might significantly improve the performance. Moreover, we

observe that the performance of the model increases even for those types where no additional instances are added, therefore exploiting cross-type learning effect. We also observe the models increasing ability to recover from original incorrect (noisy) annotations produced by the general purpose NER model.

Finally, we train the model on all the available manual data and (re-)annotate the whole original corpus.

Acknowledgments

This work has been partially funded by the project LYNX, which has received funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 780602, see http://www.lynx-project.eu.

References

1. Abhishek, A., Taneja, S.B., Malik, G., Anand, A., Awekar, A.: Fine-grained entity recognition with reduced false negatives and large type coverage (2019)
2. Breit, A., Revenko, A., Rezaee, K., Pilehvar, M.T., Camacho-Collados, J.: WiC-TSV: An evaluation benchmark for target sense verification of words in context. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1635–1645. Association for Computational Linguistics, Online (Apr 2021)
3. Castleberry, A., Nolen, A.: Thematic analysis of qualitative research data: Is it as easy as it sounds? Curr Pharm Teach Learn **10**(6), 807–815 (2018)
4. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th I-Semantics (2013)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv:1810.04805 (Oct 2018)
6. Fernández-Macías, E.: Automation, digitalisation and platforms: Implications for work and employment (2018)
7. Fries, J., Wu, S., Ratner, A., Ré, C.: SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data arXiv:1704.06360 (Apr 2017)
8. Huang, L., Sun, C., Huang, X.: GlossBERT: BERT for word sense disambiguation with gloss knowledge. In: Proceedings of EMNLP-IJCNLP 2019. pp. 3507–3512. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
9. Lee, J.S., Hsiang, J.: PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model arXiv:1906.02124 (May 2019)
10. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (09 2019)
11. Leitner, E., Rehm, G., Moreno-Schneider, J.: Fine-grained Named Entity Recognition in Legal Documents. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., Sure-Vetter, Y. (eds.) Proceedings of SEMANTiCS 2019. pp. 272–287. No. 11702 in LNCS, Springer, Karlsruhe, Germany (9 2019)
12. Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., Zhang, C.: BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision arXiv:2006.15509 (Jun 2020)

13. Lopez, P.: entity-fishing. https://github.com/kermitt2/entity-fishing (2016–2020)
14. Moro, A., Raganato, A., Navigli, R.: Entity Linking meets Word Sense Disambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics (TACL) **2**, 231–244 (2014)
15. Murty, S., Verga, P., Vilnis, L., Radovanovic, I., McCallum, A.: Hierarchical losses and new resources for fine-grained entity typing and linking. In: Proceedings of the 56th ACL: Volume 1. pp. 97–109. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
16. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from wikipedia. Artif. Intell. **194**, 151–175 (Jan 2013)
17. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning Multilingual Named Entity Recognition from Wikipedia. Artif. Intell. **194**, 151–175 (Jan 2013)
18. Oliveira, I.L., Fileto, R., Speck, R., Garcia, L.P., Moussallem, D., Lehmann, J.: Towards holistic entity linking: Survey and directions. Information Systems **95**, 101624 (2021)
19. Ontonotes release 5.0. https://catalog.ldc.upenn.edu/LDC2013T19, accessed: 2021-03-21
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)
21. Ren, X., El-Kishky, A., Wang, C., Tao, F., Voss, C.R., Han, J.: Clustype: Effective entity recognition and typing by relation phrase-based clustering. In: Proceedings of the 21th ACM SIGKDD. p. 995–1004. KDD '15, Association for Computing Machinery, New York, NY, USA (2015)
22. Shang, J., Liu, L., Ren, X., Gu, X., Ren, T., Han, J.: Learning Named Entity Tagger using Domain-Specific Dictionary arXiv:1809.03599 (Sep 2018)
23. Shimaoka, S., Stenetorp, P., Inui, K., Riedel, S.: Neural architectures for fine-grained entity type classification. In: Proceedings of the 15th ACL: Volume 1. pp. 1271–1280. Association for Computational Linguistics, Valencia, Spain (Apr 2017)
24. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of HLT-NAACL 2003: Volume 4. p. 142–147. CONLL '03, Association for Computational Linguistics, USA (2003)
25. Vandenbussche, P.Y., Scerri, A., Daniel, Ron, J.: Word sense disambiguation with transformer models. In: Proceedings of SemDeep-6. Association for Computational Linguistics (to appear)
26. Wang, R., Tang, D., Duan, N., Wei, Z., Huang, X., Cao, C., Jiang, D., Zhou, M., et al.: K-adapter: Infusing knowledge into pre-trained models with adapters. arXiv:2002.01808 (2020)
27. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: Luke: Deep contextualized entity representations with entity-aware self-attention. In: EMNLP (2020)
28. Yamada, I., Shindo, H., Takefuji, Y.: Representation learning of entities and documents from knowledge base descriptions. arXiv:1806.02960 (2018)