Named Entity Recognition and Linking for Entity Extraction from Italian Civil Judgements

Riccardo Pozzi[1][0000−0002−4954−3837], Riccardo Rubini[1][0009−0001−0955−6721], Christian Bernasconi[1][0009−0005−8655−4446], and Matteo Palmonari[1][0000−0002−1801−5118]

University of Milano-Bicocca, Milan, Italy

Abstract. The extraction of named entities from court judgments is useful in several downstream applications, such as document anonymization and semantic search engines. In this paper, we discuss the application of named entity recognition and linking (NEEL) to extract entities from Italian civil court judgments. To develop and evaluate our work, we use a corpus of 146 manually annotated court judgments. We use a pipeline that combines a transformer-based Named Entity Recognition (NER) component, a transformer-based Named Entity Linking (NEL) component, and a NIL prediction component. While the NEL and NIL prediction components are not fine-tuned on domain-specific data, the NER component is fine-tuned on the annotated corpus. In addition, we compare different masked language modeling (MLM) adaptation strategies to optimize the result and investigate their impact. Results obtained on a 30-document test set reveal satisfactory performance, especially on the NER task, and emphasize challenges to improve NEEL on similar documents. Our code is available on GitHub[1].

Keywords: Named Entity Recognition · Named Entity Linking · NIL Prediction · Italian Civil Court Judgments · Legal · Domain Adaptation.

1 Introduction

Solutions to extract information from legal texts have a long tradition [37] and are attracting even greater interest, also due to the performance boost on many tasks that has been made possible by recent advances in natural language processing (NLP) technologies (see Section 2). The language in legal text and the downstream application may differ significantly depending on the domain, a broad spectrum of specific solutions have been proposed in a variety of domains, e.g., from law to court judgments, from contracts to criminal investigations [3]. For example, many approaches have been proposed to extract legal terminology [5] and some approach has focused on named entities [37]. In this paper, we focus on a specific domain: the extraction of named entities from Italian

[1] https://github.com/rpo19/pozzi_aixia_2023. We are not allowed to publish sensitive data and the NER models trained on sensitive data.

court judgments, and, in particular, from judgments produced in the context of civil trials. The work discussed in this paper is part of activities conducted in two projects developed in cooperation with or funded by DGSIA, the body that manages information systems of the Ministry of Justice, and the Ministry of Justice itself.

Entity extraction is applied to enrich court judgment data to support three main target downstream applications: 1) semantic search, where stakeholders (mainly judges) can search for previous judgments, and use named entities therein to filter out results; 2) anonymization, where finding references, especially to people and organizations, is a prerequisite to anonymize the judgments; 3) calculate advanced statistical analyses, which can use variables that are not found in trial records and metadata, and can be found only in the actual text (e.g., average alimony by district). While additional NLP processing methods may be required for advanced statistical analyses, named entity extraction remains a crucial component for solutions targeting this application.

With entity extraction, we refer to a task that goes a bit beyond NER, as proposed in most of the previous approaches. In fact, for all or some of the above applications, it is valuable not only to find named entity mentions and classify them into a set of known classes, but also to consolidate these mentions into an entity-centric knowledge layer, which supports deeper data integration functionalities and related downstream functionalities. In particular, deeper integration can be achieved by: 1) reconciling the mentions of different entities and linking references to known entities described in background knowledge bases, e.g., Wikipedia entities (named entity linking - NEL); 2) reconciling different mentions of entities within a document (entity clustering). Observe that named entity linking contributes to entity clustering, where mentions with the same link are implicitly clustered together. Another reason to use NEL in the entity extraction process is that there are entities in court judgments that are known because described in background knowledge bases, which makes these links useful. In fact, in these projects, we developed an end-to-end entity extraction pipeline that performs the following tasks: NER; NEL; NIL prediction, which decides whether to link an entity mention to an entity in the KB (the one identified by NEL) or to consider that the correct entity is not in the KB, i.e., if a mention is respectively not NIL, or NIL ("not in lexicon"); NIL clustering, i.e., the task of clustering NIL mentions referring to the same entity. The pipeline is inspired by and shares some components of the approach described in previous work [24].

In this paper, we focus on discussing the performance that our neural algorithms achieve on named entity recognition and linking tasks (NEEL) including NIL prediction. To better illustrate the NEEL process, we provide an example in Figure 1. We leave out of the focus of this paper the NIL clustering part, mainly for reasons of space.

In particular, we present a pipeline that combines a transformer-based named entity recognition (NER) component, a transformer-based Named Entity Linking (NEL) component, and a NIL prediction component. While the NEL and NIL prediction components are not fine-tuned on domain-specific data, the NER component is fine-tuned on an annotated corpus of civil court judgments. In addition, we test different masked language modeling (MLM) [10] adaptation strategies, including adaptation with a larger corpus of civil court judgments from which the annotated corpus has been taken.

The paper is organized as follows: in Section 2 we discuss related work; in Section 3, we present our approach; in Section 4 we present the results of our experimental evaluation; finally, conclusion ends the paper in Section 5.

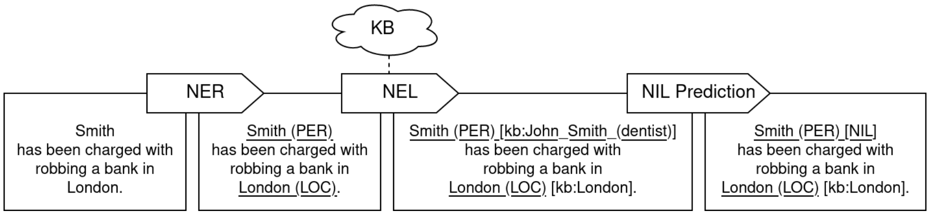
KB
NER
NEL
NIL Prediction
Smith has been charged with robbing a bank in London.
Smith (PER) has been charged with robbing a bank in London (LOC).
Smith (PER) [kb:John_Smith_(dentist)] has been charged with robbing a bank in London (LOC) [kb:London].
Smith (PER) [NIL] has been charged with robbing a bank in London (LOC) [kb:London].

Fig. 1. Overview of NEEL with a NIL (*Smith*) and a ¬NIL mention (*London*). The correct entity for *Smith* is not present in the KB, indeed, NEL provides a wrong candidate, and NIL prediction classifies *Smith* as NIL.

2 Related Work

This section presents an overview of recent advancements in Named Entity Extraction and Linking (NEEL) techniques, including NIL prediction. We start by focusing on NEEL approaches in the legal domain, then we proceed with the status of NEEL for the Italian language. Finally, we briefly highlight recent developments in general-domain NEEL, additionally discussing the advancements achieved in the three subtasks of NEEL: named entity recognition (NER), named entity linking (NEL), and NIL prediction. NER identifies mentions of named entities and classifies them into a predefined set of classes, while NEL links these mentions to corresponding entities in a knowledge base.; 3) NIL prediction determines if the NEL candidate is correct or if the mention refers to an entity that is missing from the KB, i.e., an unlinkable entity mention or NIL ("not in lexicon") mention.

By examining the related works in these areas, we lay the foundation for our research and shed light on existing gaps in the field of NEEL with NIL prediction applied to Italian court judgments.

2.1 NEEL for Legal Documents

Most of the previous work on NEEL for legal documents has focused on the NER task only. The first NER approaches are based on handcrafted rules and statistical models [37], such as conditional random fields (CRFs) [19]. More recent approaches to NER started using BiLSTM-based models combined with CRFs for Brazilian and German legal texts [37]. After the advent of transformers [33] and BERT [10], which obtained impressive performance in multiple NLP tasks, LEGAL-BERT, specialized in the legal domain, has been released [7]. Later, some work compared LEGAL-BERT with previous approaches [16] for NER finding that LEGAL-BERT performance is comparable to simple models (LSTMs, CNNs).

A few approaches have studied NEL in legal texts. One of the first approaches applied NER and NEL on a corpus of judgments of the European Court of Human Rights [4]. As the background KB, they use a legal-specific ontology enriched with YAGO[2] after an alignment procedure. Another approach targets the NEL task only on the EUR-Lex law article dataset [11]. Their NEL system is trained using transfer learning. We study the end-to-end combination of NER, NEL, and NIL prediction, and we use a more recent NEL approach [35] trained on a large Italian Wikipedia corpus without fine-tuning on court judgment data. Another study combined BERT [10] with rule-based techniques for NER and coupled it with an off-the-shelves NEL service to extract entities from court decisions in the Finnish language [29]; NEL is performed with a popularity-based approach. This study is the most similar to ours; however, we focus on the NIL prediction problem, and we use a BERT-based NEL approach; also, in this paper, we discuss only the performance of neural algorithms.

Some work that studies NEL with NIL prediction is also evaluated on documents that are related to the legal domain [17] (the depositions of the 1641 Irish rebellion[3]). However, to the best of our knowledge, no prior work has investigated end-to-end NEEL considering the NIL prediction problem on recent legal data.

2.2 NEEL in the Italian Language

Italian datasets for NER include multilingual resources [30, 22], and domain-specific datasets, such as [6] in the medical domain. Similarly, Italian NEL datasets comprise multilingual ones, i.e. VoxEL [26] and resources based on micro-posts [2].

Among the ready-to-use NEEL libraries for the Italian language, notable options are SpaCy[4] and Tint [23] for NER and DBpedia Spotlight [8] for both NER and NEL. SpaCy provides pre-trained NER models of different sizes (small, medium and large), but currently does not provide any pre-trained transformer-based model for Italian. Tint performs NER with a combination of CRFs taggers and rule-based systems for dates and money. DBpedia Spotlight is a ready-to-use tool that recognizes and links entity mentions to DBpedia[5].

[2] https://yago-knowledge.org/

[3] http://1641.tcd.ie/

[4] https://spacy.io

[5] https://www.dbpedia.org/

2.3 NEEL for General Domain

NEEL with NIL prediction dates back to the knowledge base population track (TAC-KBP) of the Text Analysis Conference[6] (TAC), which has included the NIL prediction task since 2009 [21]. Work focusing on end-to-end NEEL includes approaches that jointly perform the subtasks [18, 1] and pipeline-based systems [15, 13].

Follows a brief overview of the recent developments of the three subtasks.

NER Recent DL approaches for NER include models based on recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers [28], often combined with CRFs for the final prediction of sequence labels. Several studies highlighted the importance of word embeddings and character embeddings for NER, including non-contextualized embeddings and contextualized embeddings [10]. Indeed, the most effective approaches are based on this latter class of embeddings: the state-of-the-art on CoNLL2003 [31], an important benchmark for English NER, is detained by concatenating character-based, contextualized, and non-contextualized embeddings [34].

NEL Since 2013, representation learning techniques for NEL have been explored to obtain dense representations of mentions and entities and calculate a similarity score (e.g. cosine similarity) to rank linking candidates [12, 36]. The attention mechanism and transformers [33] have played a crucial role in enhancing dense representations, leading to the development of the bi-encoder and cross-encoder paradigms [14, 35], which are widely used for dense-retrieval and candidate reranking, respectively. Recently, promising entity linking paradigms that better leverage the pre-training task of language models are emerging: autoregressive entity linking [9], and extractive entity linking [25].

NIL Prediction NIL entities have been often ignored in the literature of entity linking: among the 38 approaches compared by the survey [28] only 8 considered NIL entities. NIL prediction strategies, several of which derive from the TAC-KBP, include applying a threshold to the entity linking score, representing NIL with an additional class, and using a binary classifier on top of the linking score and additional features [28].

3 Named Entity and Linking Algorithms

As discussed in Section 1, in this paper we focus on presenting our NEEL approach for extracting entities from civil court judgments and evaluating its performance on an annotated dataset. Our approach implements three tasks [24] in a pipeline: NER, NEL, and NIL prediction. For NER, we focus on neural NER

[6] https://tac.nist.gov/

Table 1. Statistics of ICCJ. Number of NIL annotations is indicated in parentheses.

	#Docs	#Ann	PER	ORG	LOC	DATE	MONEY	MISC
Train	102	11940	2997	2761	612	2088	791	2691
Validation	14	1688	308	369	77	350	84	500
Test(#NIL)	30	3006(2539)	722(653)	694(443)	195(58)	555	223	617

algorithms, considering a transformer-based NER module that we fine-tune on an annotated corpus of court judgments. Given that the NER component can use different transformers, we also analyze the impact of domain adaptation, based on masked language modeling (MLM), on downstream performance. Since the classes of entities considered by NER are related to the annotated corpus, we organize this section as follows: we first introduce the Italian Civil Court Judgment Corpus; we discuss the classes used in the NER module; then we provide details about the NER component, the NEL algorithm, and the NIL prediction.

3.1 Italian Civil Court Judgment Corpus and NER classes

The gold standard dataset we use for training and evaluation is composed of 146 annotated judgments derived from a corpus of 900,000 legal judgments, organized as follows: 102 documents as the training set, 14 for validation, and a test set of 30 documents. Unfortunately, we are unable to publish the corpus due to the sensitive nature of the data it contains. However, upon request, we are open to exploring the possibility of sharing it through bilateral agreements. The annotations in the corpus have been performed by two annotators. The inter-annotator agreement (IAA) has been calculated using the F1-measure to assess the coherence between the annotations in terms of both class and span and using Cohen's Kappa, obtaining respectively 80.8% and 66.2%.

All the documents have NER annotations considering the following classes: Person (*PER*), Organization (*ORG*), Location (*LOC*), Date (*DATE*), Money (*MONEY*), and Miscellaneous (*MISC*), that includes references to court judgments, law articles, court decrees, or any entity not covered by the above classes. In total, the dataset is composed of more than 16,000 annotations and each document counts on average ∼1,900 words. Table 1 reports detailed statistics, including the number of annotations for each class.

The annotations for named entity linking (NEL) and NIL prediction are only available in the test set. These annotations are limited to the classes *PER*, *ORG*, and *LOC*. *DATE*, *MONEY*, and *MISC* mentions have not been annotated for NEL and NIL prediction because they are expected to be processed by rule-based algorithms that we do not cover in this work.

For the remainder of this work, we will refer to our annotated corpus of 146 documents as ICCJ (Italian Civil Court Judgment) and to the 900,000 legal judgment (without annotations) as ICCJ900k.

Table 2. NER backbones details with the applied MLM-adaptations (one model per column). Model names indicate the order of applied adaptations. Legal domain data used for the LGL adaptation vary: *3.7GB legal corpus from the National Jurisprudential Archive; **6.6GB legal corpus composed of civil and criminal cases.

	ITA	ITA+LGL+ICCJ900k	ITA+LGL	LGL+ICCJ900k	LGL
ITA	Y	Y	Y	-	-
LGL	-	Y*	Y*	Y**	Y**
ICCJ900k	-	Y	-	Y	Y

3.2 NER with MLM-adaptation and fine-tuning

We use the library SpaCy-transformers[7] with the SpaCy transition-based parser to leverage contextualized token representations obtained from a transformer [33].

As the backbone transformer for the NER system, we evaluate five different BERT encoders that have been trained with one of three different MLM-adaptations or with a combination of them. They are further described in Table 2 where *ITA* (Italian) denotes the pre-training with MLM on general-domain Italian data, *LGL* (Legal) the MLM-adaptation to legal-domain data, and *ICCJ900k* to our corpus of 900,000 Italian civil court judgments. It is important to note that some models (*LGL* and *LGL+ICCJ900k*) are directly pre-trained on legal domain data using MLM.

As a baseline, we consider the general-domain model *ITA* (available pre-trained on huggingface[8]). Also, *ITA+LGL* [20] and *LGL*[9] are available pre-trained on huggingface, while for the *ICCJ900k* versions we perform the adaptation with MLM. For each backbone, we consider five models with a different random weight initialization.

Finally, each model has been fine-tuned for the NER task on ICCJ training set using the SpaCy library with early stopping on the validation set and AdamW as the optimizer with the initial learning rate set to 5×10^{-5}.

3.3 NEL with BLINK-ITA-bi-encoder

For NEL we use the bi-encoder architecture of BLINK [35]. We initialize the bi-encoder with the weights from Italian BERT-base[10] [27], then we fine-tune them on 9M training samples from Italian Wikipedia hyperlinks, following the original work, for 5 epochs (in the last one we train with hard-negatives instead of random negatives) using AdamW optimizer with the initial learning rate set to 1×10^{-5} and a batch size of 20. As the linking KB, we use $\sim 1.5M$ entities obtained from Italian Wikipedia[11] after filtering out redirects and disambiguation pages.

[7] https://spacy.io/universe/project/spacy-transformers

[8] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[9] https://huggingface.co/dlicari/Italian-Legal-BERT-SC

[10] https://huggingface.co/dbmdz/bert-base-italian-uncased

[11] https://it.wikipedia.org

3.4 NIL Prediction

In the NIL prediction component, we use a *logistic regression classifier* that receives as input features 1) the score of the top-ranked entity given by the NEL system and 2) the difference between the top-ranked entity score and the second-best one [24], and produces an output $p \in [0, 1]$, where 1 means the top-ranked entity is correct for linking the mention, while 0 means the opposite. In this latter case, we consider the mention NIL, assuming that if the correct entity is not in the top-ranked position, then it is not in the KB.

4 Experimental Evaluation

In order to evaluate the NEEL pipeline, we study the overall effectiveness of the pipeline and each component separately. By doing so, we have been able to independently study the behavior of the NER, the NEL, and the NIL prediction systems, and finally of NEL with NIL prediction combined. We would like to remind the reader that the NEL and NIL prediction components are applied only to mentions classified as *PER*, *ORG*, and *LOC* by the NER component. In our evaluation, we focus especially on the following objectives:

1. investigating the performance with different backbone transformers and the impact of different MLM adaptation strategies on the NER performance;
2. investigating which classes of entities are more challenging;
3. investigating the performance of NEL and NIL prediction in a best-case scenario, independently from the NER component;
4. investigating the performance of the end-to-end NEEL pipeline;
5. to discuss the main challenges.

4.1 Evaluation Settings and Measures

Please note that the ICCJ training set is only used to fine-tune the NER component. All results refer to the ICCJ test set.

NER NER is evaluated using strong and partial matching measures, which is a quite common practice in evaluating NER approaches [32]. Both measures rigorously require that the predicted class matches the gold standard one, while they differ with respect to span detection: the former measure considers an annotation correct when the predicted boundaries perfectly match the gold standard, the latter when there is an overlap between them. Considering both measures is useful also for two other reasons: 1) we can investigate to what extent some correct annotations are returned by the algorithms, even when the span of the mention is not perfectly identified; 2) by considering the gap between strong and partial matching measures in a per-class performance analysis, we can investigate which classes are more affected by boundary identification issues. For each of the measures, as in a multiclass classification problem, we calculate *precision*, *recall*, and *F1-measure*, micro and macro-averaged on the class, and separately for each class.

Comparison of Backbone Transformers for NER The comparison of backbone transformers for NER considers five different random weight initializations for each backbone. We calculate the mean and the standard deviation of the micro *precision*, *recall*, and *F1-measure* of the five initializations for each transformer. We identify the top-performing model based on its *F1-measure* and utilize it as the foundation of the NER component for subsequent evaluations.

NEL We evaluate the NEL component in terms of *accuracy* and *recall@100*, similarly to recent work [35], on the ICCJ test set, additionally comparing to the news-based benchmark VoxEL [26].

It is important to remind that the NEL evaluation and the following ones (NIL prediction, and end-to-end NEEL) exclusively focus on the classes *PER*, *ORG*, and *LOC*.

NIL Prediction The NIL prediction component is evaluated as a binary classifier with *precision*, *recall*, and *F1-measure* calculated for both classes (NIL and ¬NIL). It is important to emphasize that, to evaluate the NIL prediction independently from NEL errors, we consider it correct when the NIL prediction classifies as NIL the mentions incorrectly linked by the NEL component. We attribute a positive value to this behavior as it showcases the ability of the NIL prediction to effectively identify NEL errors, thereby mitigating their impact.

NEL & NIL Prediction We evaluate NEL with NIL prediction independently from NER errors by calculating 1) the *recall* of the *mentions to link*, 2) the *recall* of *NIL mentions*, and 3) the *accuracy* of all the mentions.

NEEL end-to-end Finally, we evaluate the end-to-end NEEL using strong and partial matching measures; in this case, an annotation is considered correct when 1) the predicted class matches the gold standard, 2) the span matches the gold standard according to the measure, and 3) the mention is linked to the correct entity (if ¬NIL) or correctly identified as NIL. We calculate *precision*, *recall*, *F1-measure* micro and macro-averaged for each class, exactly as in the NER evaluation.

4.2 Results

Comparison of backbone transformers Table 3 shows the results for the comparison of the 5 backbone transformers for NER. Based on the sample mean, the encoder that gives the best results is ITA+LGL+ICCJ900k, while the worst one is LGL+ICCJ900k.

In order to properly analyze the presence of statistical differences based on the choice of the backbone transformer, we conducted an analysis of variance (ANOVA) test on the *F1-measure*. The results reveal a highly significant difference (with significance level $\alpha = 0.05$). To further investigate the pairwise differences, we conducted a Tukey's HSD test with a significance level of $\alpha = 0.05$. We

Table 3. Comparison of the backbone transformers (one per row) for NER on ICCJ test. Using strong matching we calculate mean (± std) on 5 random initializations.

	Precision	Recall	F1 Score
ITA	81.96($\pm$0.76)	83.77($\pm$1.39)	82.76($\pm$0.63)
ITA+LGL+ICCJ900k	**82.08($\pm$0.87)**	**84.69($\pm$0.52)**	**83.36($\pm$0.41)**
ITA+LGL	81.11($\pm$1.00)	83.57($\pm$1.04)	82.41($\pm$0.55)
LGL+ICCJ900k	80.87($\pm$0.73)	82.62($\pm$1.55)	81.72($\pm$0.52)
LGL	79.90($\pm$1.05)	82.62($\pm$1.36)	81.23($\pm$0.47)

observe that *ITA*, the pre-training on general-domain Italian data, has a positive impact on performance: the models *ITA+LGL+ICCJ900k* and *ITA+LGL* tend to perform better than those trained from scratch on domain-specific data (*LGL* and *LGL+ICCJ900k*).

Surprisingly, the findings suggest that employing a domain-specific legal BERT does not result in a substantial enhancement in NER performance compared to a generic Italian BERT. This observation extends to the adaptation to the corpus of judgments (ICCJ900k) as well. Furthermore, we emphasize that the use of a pre-trained generic Italian BERT significantly reduces the effort required for adaptation in terms of time, costs, and environmental imprint.

NER The evaluation results for the NER component, as shown in Table 4, are promising. All the strong matching measures exceed 80%, and all the partial matching measures surpass 90%, indicating overall proficiency in NER recognition. The classes *MONEY* and *PER* achieve high recognition rates, surpassing 90% with the strong matching measure. However, the performance for *MISC* is lower compared to other types. This discrepancy may be attributed to the intrinsic heterogeneity of the MISC class. Additionally, *MISC* exhibits the largest disparity between strong matching performance and partial matching performance. A significant difference (approximately 12%) between strong and partial matching outcomes also affects the class *ORG*, highlighting the difficulty in precisely detecting the boundaries of organization mentions.

We also consider the successful results achieved by the NER component indicative of the high quality of our annotated corpus ICCJ.

NEL and NIL Prediction Table 5 reveals that the NEL and NIL prediction components do not exhibit the same level of effectiveness as the NER component. The independent evaluation of the NEL component ($NEL_{\perp}$) demonstrates a lower *accuracy* (73.52%) but achieves a *recall@100* of 90.81%, suggesting that the integration of a re-ranking system could potentially enhance our results. Additionally, the comparison with the outcomes obtained with the news-based VoxEL benchmark [26] further underscores the challenges presented by the ICCJ corpus. We also remind you that the NEL component has not been fine-tuned on

Table 4. NER evaluation with strong and partial matching on ICCJ test.

	Strong Match			**Partial Match**		
	Prec	Recall	F1	Prec	Recall	F1
DATE	83.49	80.18	81.80	92.12	87.84	89.93
LOC	86.34	84.62	85.49	94.24	92.31	93.26
MONEY	96.19	90.58	93.30	99.52	93.72	96.54
ORG	76.58	80.12	78.31	89.12	92.83	90.93
PER	90.37	91.00	90.68	95.77	95.43	95.10
MISC	73.97	70.02	71.94	91.27	85.14	88.10
Macro by Class	84.50	82.75	83.59	93.51	91.21	92.31
Micro	82.70	81.74	82.22	92.53	90.97	91.74

Table 5. NEL and NIL Prediction evaluation on ICCJ test. $\text{NEL}_\perp$ and NIL $\text{Pred}_\perp$ are independent from other tasks. NEL & NIL $\text{Pred}_\perp$ evaluate the two tasks independently from NER. *$\text{NEL}_\perp$ also reports results on VoxEL [26] for comparison.

$\text{NEL}_\perp$			**NIL $\text{Pred}_\perp$**				**NEL & NIL $\text{Pred}_\perp$**	
	Acc	Rec@100		Prec	Rec	F1	Link_{Rec}	52.95
ICCJ	73.52	90.81	NIL	92.15	86.51	89.24	NIL_{Rec}	86.31
sVoxEL-it*	88.89	96.83	¬NIL	58.45	72.02	64.53	Overall_{Acc}	76.85

ICCJ, and that the utilized knowledge base has not been restricted to domain-related entities. These two factors represent possibilities for enhancing this component.

The NIL prediction classifier ($\text{NIL}_\perp$) is effective in recognizing the NIL class, while it suffers with $\neg NIL$ mentions: the low precision of 58.45% highlights that several NIL mentions are wrongly predicted as $\neg NIL$.

During the evaluation of *NEL with NIL prediction*$_\perp$, we notice the *overall accuracy* is acceptable (76.85%) and the *recall* on the NIL mentions is satisfactory at 86.31%. However, we observe that the performance of ¬NIL mentions (Link_{Rec}), which should have been linked to the knowledge base (KB), is not up to the desired standard. The errors for this measure include both mentions linked to incorrect entities and mentions inaccurately identified as NIL. After the NIL prediction, indeed, only 52.95% of the ¬NIL mentions are correctly classified, whereas the accuracy of $\text{NEL}_\perp$ stands at 73.52%. This substantial 20% decline in performance can be attributed to the false-NIL predictions.

For these reasons, we consider the NIL prediction to be the most significant challenge in NEEL. It is important to further study and improve this component in order to enhance the overall performance and reliability of NEEL systems.

NEEL end-to-end Lastly, Table 6 presents the comprehensive results for the end-to-end NEEL task. *PER* and *LOC* exhibit similar satisfactory performance levels. On the other hand, *ORG* entities appear to be more challenging.

Table 6. NEEL end-to-end evaluation of PER, LOC, ORG mentions on ICCJ test.

	Strong Match			**Partial Match**		
	Prec	Recall	F1	Prec	Recall	F1
LOC	75.92	74.36	75.13	80.10	78.46	79.27
ORG	51.10	53.46	52.25	60.61	63.22	61.88
PER	76.89	77.42	77.16	80.19	80.86	80.52
Macro by Class	67.97	68.41	68.18	73.63	74.18	73.89
Micro	65.39	66.73	66.05	71.53	72.95	72.24

Furthermore, the difference between strong and partial matching is limited for *PER* and *LOC*, but significant for *ORG*, confirming the difficulty in accurately detecting boundaries for *ORG* entities previously observed in the NER results. Additionally, the relatively modest overall difference of 6% between partial and strong matching, along with the disparity with NER-only results (72.24% vs 91.74%), highlights that the NEL and NIL prediction components are responsible for the majority of errors. This observation, combined with the fact that we fine-tuned only the NER component, suggests that fine-tuning the NEL and NIL prediction components on the data could potentially enhance the overall performance of the end-to-end NEEL system.

5 Conclusion

In this paper, we have presented the application of a NEEL pipeline to Italian civil court judgments and an evaluation of its performance. The experimental evaluation conducted on 30 annotated judgments suggests that the performance of our NEEL pipeline is encouraging, especially the performance of the NER component, and emphasizes some remaining challenges. Quite surprisingly, the gap in performance between models that use domain-specific transformers, adapted with masked language modeling, and those that use transformers trained on generic Italian text is quite limited and not statistically significant. The challenges concern especially the NEL and NIL prediction components, which so far we have not customized for or fine-tuned on domain-specific data. Fine-tuning these algorithms using limited data is a challenge that we plan to address in the future. Moreover, we plan to investigate strategies to support human-in-the-loop NEEL, by improving the extraction quality and minimizing the user effort during the annotation and validation phases. Finally, a prospective scenario for future development involves jointly performing NEL and NIL prediction within a unified module, as recent research indicates that consolidating multiple pipeline tasks in a single module can significantly reduce error propagation [18]. Despite the remaining challenges, we believe that the evidence discussed in the paper suggests that, with further improvements, end-to-end NEEL pipelines could be effectively applied to court judgments to disclose a variety of downstream applications.

Acknowledgements

This research has been partially funded by Cini in the context of the Italian project Datalake@Giustizia and by the project PON Next Generation UPP promoted by the Italian Ministry of Justice.

References

1. Ayoola, T., Tyagi, S., Fisher, J., Christodoulopoulos, C., Pierleoni, A.: ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track. Association for Computational Linguistics (Jul 2022)
2. Basile, P., Caputo, A., Gentile, A.L., Rizzo, G.: Overview of the evalita 2016 named entity recognition and linking in italian tweets (neel-it) task. In: of the Final Workshop. vol. 7 (2016)
3. Batini, C., Bellandi, V., Ceravolo, P., Moiraghi, F., Palmonari, M., Siccardi, S.: Semantic data integration for investigations: Lessons learned and open challenges. In: 2021 IEEE International Conference on Smart Data Services (SMDS) (2021)
4. Cardellino, C., Teruel, M., Alemany, L.A., Villata, S.: A low-cost, high-coverage legal named entity recognizer, classifier and linker. In: Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law. ICAIL '17, Association for Computing Machinery (2017)
5. Castano, S., Falduti, M., Ferrara, A., Montanelli, S.: A knowledge-centered framework for exploration and retrieval of legal documents. Information Systems **106** (2022)
6. Catelli, R., Gargiulo, F., Casola, V., De Pietro, G., Fujita, H., Esposito, M.: Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. Applied Soft Computing **97** (2020)
7. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics (Nov 2020)
8. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems. I-SEMANTICS '13, Association for Computing Machinery (2013)
9. De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., Petroni, F.: Multilingual autoregressive entity linking. Transactions of the Association for Computational Linguistics **10** (2022)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics (Jun 2019)
11. Elnaggar, A., Otto, R., Matthes, F.: Deep learning for named-entity linking with transfer learning for legal documents. In: Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference. AICCC '18, Association for Computing Machinery (2018)

12. He, Z., Liu, S., Li, M., Zhou, M., Zhang, L., Wang, H.: Learning entity representation for entity disambiguation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics (Aug 2013)
13. Heist, N., Paulheim, H.: Nastylinker: Nil-aware scalable transformer-based entity linker. In: The Semantic Web. Springer Nature Switzerland (2023)
14. Humeau, S., Shuster, K., Lachaux, M.A., Weston, J.: Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In: International Conference on Learning Representations (2019)
15. Kassner, N., Petroni, F., Plekhanov, M., Riedel, S., Cancedda, N.: EDIN: An end-to-end benchmark and pipeline for unknown entity discovery and indexing. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (Dec 2022)
16. Keshavarz, H., Vagena, Z., Kouki, P., Fountalis, I., Mabrouki, M., Belaweid, A., Vasiloglou, N.: Named entity recognition in long documents: An end-to-end case study in the legal domain. In: 2022 IEEE International Conference on Big Data (Big Data) (2022)
17. Klie, J.C., Eckart de Castilho, R., Gurevych, I.: From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (Jul 2020)
18. Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. Association for Computational Linguistics (Oct 2018)
19. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
20. Licari, D., Comandè, G.: ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management. CEUR Workshop Proceedings, vol. 3256. CEUR (Sep 2022)
21. McNamee, P., Dang, H.T.: Overview of the tac 2009 knowledge base population track. In: Second text analysis conference (TAC 2009). vol. 2 (2009)
22. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from wikipedia. Artificial Intelligence **194** (2013)
23. Palmero Aprosio, A., Moretti, G.: Tint 2.0: an all-inclusive suite for nlp in italian. In: Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it. vol. 10 (2018)
24. Pozzi, R., Moiraghi, F., Lodi, F., Palmonari, M.: Evaluation of incremental entity extraction with background knowledge and entity linking. In: Proceedings of the 11th International Joint Conference on Knowledge Graphs. IJCKG '22, Association for Computing Machinery (2023)
25. Procopio, L., Conia, S., Barba, E., Navigli, R.: Entity disambiguation with entity definitions. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics (May 2023)
26. Rosales-Méndez, H., Hogan, A., Poblete, B.: Voxel: a benchmark dataset for multilingual entity linking. In: The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17. Springer (2018)

27. Schweter, S.: Italian bert and electra models. Zenodo (Nov 2020)
28. Sevgili, O., Shelmanov, A., Arkhipov, M.V., Panchenko, A., Biemann, C.: Neural entity linking: A survey of models based on deep learning. Semantic Web **13** (2020)
29. Tamper, M., Oksanen, A., Tuominen, J., Hietanen, A., Hyvönen, E.: Automatic annotation service appi: Named entity linking in legal domain. In: The Semantic Web: ESWC 2020 Satellite Events. Springer International Publishing (2020)
30. Tedeschi, S., Navigli, R.: MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In: Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics (Jul 2022)
31. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (2003)
32. Tsai, R.T.H., Wu, S.H., Chou, W.C., Lin, Y.C., He, D., Hsiang, J., Sung, T.Y., Hsu, W.L.: Various criteria in the evaluation of biomedical named entity recognition. BMC bioinformatics **7** (2006)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
34. Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., Tu, K.: Automated concatenation of embeddings for structured prediction. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Aug 2021)
35. Wu, L., Petroni, F., Josifoski, M., Riedel, S., Zettlemoyer, L.: Scalable zero-shot entity linking with dense entity retrieval. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (Nov 2020)
36. Yamada, I., Shindo, H., Takeda, H., Takefuji, Y.: Joint learning of the embedding of words and entities for named entity disambiguation. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Association for Computational Linguistics (Aug 2016)
37. Çetindağ, C., Yazıcıoğlu, B., Koç, A.: Named-entity recognition in turkish legal texts. Natural Language Engineering **29** (2023)