

CNER: Concept and Named Entity Recognition

Giuliano Martinelli¹, Francesco Maria Molfese¹, Simone Tedeschi^{1,2}

Alberte Fernández-Castro^{2,3} and Roberto Navigli¹

¹Sapienza NLP Group, Sapienza University of Rome

²Babelscape

³Roma Tre University

{martinelli, molfese, tedeschi, navigli}@diag.uniroma1.it,
alb.fernandezcastro@stud.uniroma3.it

Abstract

Named entities – typically expressed via proper nouns – play a key role in Natural Language Processing, as their identification and comprehension are crucial in tasks such as Relation Extraction, Coreference Resolution and Question Answering, among others. Tasks like these also often entail dealing with concepts – typically represented by common nouns – which, however, have not received as much attention. Indeed, the potential of their identification and understanding remains underexplored, as does the benefit of a synergistic formulation with named entities. To fill this gap, we introduce Concept and Named Entity Recognition (CNER), a new unified task that handles concepts and entities mentioned in unstructured texts seamlessly. We put forward a comprehensive set of categories that can be used to model concepts and named entities jointly, and propose new approaches for the creation of CNER datasets. We evaluate the benefits of performing CNER as a unified task extensively, showing that a CNER model gains up to +5.4 and +8 macro F1 points when compared to specialized named entity and concept recognition systems, respectively. Finally, to encourage the development of CNER systems, we release our datasets and models at <https://github.com/Babelscape/cner>.

1 Introduction

In the age of big data, the extraction of valuable knowledge from unstructured text is crucial for a wide range of applications, from Information Retrieval to Text Mining (Grishman, 2015). Within this context, “nouns and noun phrases have a special status in describing the concepts that people are interested in searching for” (Manning et al., 2008). Consequently, to work at their best, Natural Language Processing (NLP) systems should not only leverage the information about named entities, but also that concerning nominal concepts, as the nature of the two is strongly intertwined. A

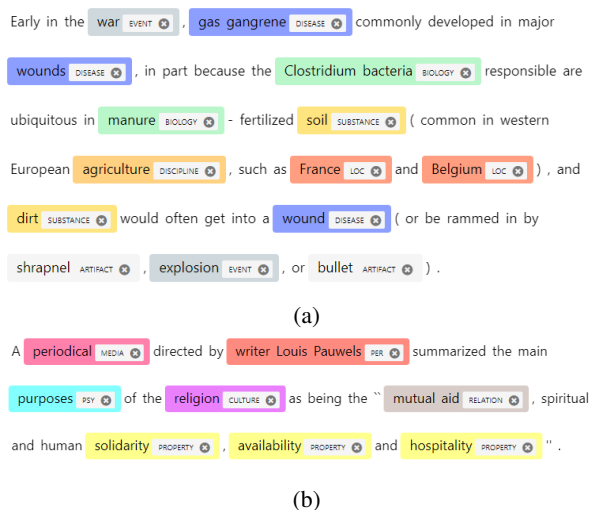


Figure 1: Two examples of CNER annotations where both nominal concepts and named entities are tagged with the proposed categorization.

prototypical task in which such interaction is intrinsic to its very nature is Relation Extraction (Pawar et al., 2017), where the triples extracted from text often connect named entities with concepts, e.g. (Carlsen, plays, chess) or (ChatGPT, is a, chatbot). As a second example, consider Coreference Resolution (Sukthanker et al., 2020), where the named entities mentioned in a text (e.g. Joe Biden) are paired with subsequent expressions referring to them, including common nouns such as *president* or *state leader*. Indeed, many tasks have been shown to benefit from the recognition of named entities (Mollá et al., 2006; Durrett and Klein, 2014; Khosla and Rose, 2020; Tedeschi et al., 2021a; Huguet Cabot et al., 2023), but little attention has been devoted to the identification of concepts.

In this paper, we fill this gap by, first, introducing Concept Recognition (CR), namely, the task of identifying and classifying concepts into predefined semantic types, and, second, harmonizing CR with the well-established task of Named Entity Recognition (NER). As shown in Figure 1, our new

joint formulation enables a denser, richer, and more cohesive semantic annotation. Specifically, we put forward the following innovative contributions:

1. We introduce a novel NLP task, namely, Concept and Named Entity Recognition (CNER), and a unified set of categories which are tailored to semantically annotating both nominal concepts and named entities.
2. We put forward an automatic procedure for the creation of CNER, NER and CR training data, and manually produce a dataset of 2,000 sentences for model evaluation.
3. We study the benefits of performing CNER as a unified task compared to performing NER and CR separately.

Finally, to encourage the development of CNER systems, we release our datasets and models to the research community at <https://github.com/Babelscape/cner>.

2 Related Work

We now review established approaches that deal with the identification and classification of concepts and named entities. Specifically, we observe that the vast majority of works in the literature focus exclusively on either concepts or named entities. This is particularly relevant because, as we will see, the currently available categorizations are not suitable for a unified task that integrates both concept and entity recognition.

2.1 Named Entities

In the last two decades, researchers have devoted significant attention to named entities, with Named Entity Recognition (NER) being a key popular task aimed at locating named entities in unstructured free-form text and then assigning them to predefined semantic types (Nadeau and Sekine, 2007; Li et al., 2022). Although NER datasets were primarily constructed to classify named entities into a limited set of categories, namely PER, ORG, LOC and MISC (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), recent approaches have proposed unwrapping the MISC category – containing miscellaneous entities – into fine-grained categories in order to explicitly identify more specific entity types. Among these, OntoNotes stands out as a well-established resource that identifies 18 categories for named entities (Pradhan et al., 2012).

However, its annotations do not cover instances of several classes, including food, animals, or celestial bodies, among others. This was instead addressed in subsequent works, i.e. Few-NERD (Ding et al., 2021) and MultiCoNER II (Fetahu et al., 2023), which proposed named entity categorizations of 66 and 36 categories, respectively. Although these categorizations are more comprehensive and fine-grained, they tend to focus on excessively detailed semantic types. For instance, the category ISLAND is well-suited for classifying named entities like *Lesbos*, but it proves less suitable for nominal concepts, as only a few, such as *islet*, fall within its scope. Other examples of categories that are not suitable for the classification of concepts are LIBRARY, AIRPORT and HOTEL, *inter alia*.

Ultimately, OntoNotes strikes a good balance in terms of the level of granularity of its categories and, as later discussed in Section 3.2, blends well with a complementary set of abstract semantic types needed for concepts, which includes relations, properties, and feelings, among others.

2.2 Concepts

The closest task to what we have called Concept Recognition is Word Sense Disambiguation (Bevilacqua et al., 2021; Navigli, 2009, WSD), the main difference being that in WSD a word is tagged with one of the senses it can denote in a given inventory (e.g. WordNet), whereas in CR a word is tagged with a more general category that is independent of the word itself. Assigning fine-grained senses to words has been found to be extremely difficult (Izquierdo et al., 2015; Maru et al., 2022), which has led to coarse-grained disambiguation approaches aimed at reducing the number of categories to choose from. Vial et al. (2019) leverage hypernymy relations to reduce WordNet granularity by exploiting its taxonomy graph, releasing a coarser-grained inventory with 39K labels. However, the high number of proposed categories is too specific to be considered as a good candidate categorization for CNER.¹ Izquierdo Beviá et al. (2007), instead, exploit WordNet relations to automatically extract a set of fundamental senses, called Basic Level Concepts, to which all other senses are mapped. However, depending on a threshold, this approach outputs hundreds or thousands of Basic Level Concepts, which are again too fine-grained.

¹Choi et al. (2018) showed that with 10k+ categories a state-of-the-art model struggles to obtain reasonable classification accuracy.

Thanks to their being much closer to the granularity of OntoNotes for entities, WordNet lexicographer files² can, instead, be used as coarse categories for nominal concepts, where each category contains synsets with the same PoS tag and a general semantic type (e.g. NOUN.FEELING). Since this resource is highly reliable and has a strong psycholinguistic grounding, we use it as a starting point for our CNER categorization, as we will discuss in Section 3.2.

2.3 Entities and Concepts

A categorization that spans both concepts and entities has been proposed in the context of the ultra-fine entity typing³ task (Choi et al., 2018, UFET). Here, the objective is to produce a three-level annotation by selecting tags from coarse, fine, and ultra-fine sets of labels consisting of 9, 121 and ~10K types, respectively. Nevertheless, the extreme granularity of the ultra-fine labels makes it difficult for systems to output the right categories, as reported by the authors, while the coarse-grained types are too generic to be considered a good candidate for our fine-grained objective. The fine set of categories comprising 121 tags, instead, not only contains categories that are too specific (e.g. DOCTOR, COACH), but also fails to categorize commonly used concepts, such as relations (e.g. *inferiority*, *brotherhood*), date/times (e.g. *November*, *September 3rd*), feelings (e.g. *love*, *anger*), and psychological features (e.g. *cognition*, *thought*), *inter alia*. Finally, we remark that the proposed ultra-fine entity typing task disregards the identification part. This is in marked contrast to CNER, where each and every span of text denoting concepts or entities has to be identified and tagged.

3 CNER task

We now formalize the Concept and Named Entity Recognition (CNER) task and then introduce our process to obtain CNER categories.

3.1 Task formulation

Formally, CNER is a sequence labeling task whose goal, given a sequence of tokens $t = (t_1, t_2, \dots, t_n)$, is to identify the spans of text S corresponding to nominal concepts and named entities, and categorize each of them into a predefined set of la-

bels $L = \{l_1, l_2, \dots, l_m\}$. Specifically, a text span $s_{ij} \in S$ is defined as a contiguous sequence of tokens $s_{ij} := (t_i, t_{i+1}, \dots, t_j)$, with $1 \leq i \leq j \leq n$.

3.2 Categories

As anticipated in Section 2, we draw upon the robust cognitive foundations of WordNet (Miller, 1995) and OntoNotes (Pradhan et al., 2012) to produce a unified set of categories. Specifically, we leverage the completeness and broad semantic coverage of lexicographer files for nominal concepts, and integrate them with the widely-used semantic categories for named entities available in OntoNotes. Figure 2 shows the resulting 29 CNER categories (inner column) together with their alignment with the OntoNotes and WordNet ones (left and right columns, respectively). In particular, every category in OntoNotes and WordNet has a counterpart in our categorization, whereas the reverse does not hold. Specifically, i) every category highlighted in red has a direct link with a WordNet category (e.g. ANIMAL), ii) every category highlighted in yellow corresponds to an OntoNotes category (e.g. LAW), and iii) every category in a white cell is grounded on both resources (e.g. MEDIA).

To further assess the validity of such categorization and ensure that each instance is assigned to a reasonable category (i.e. semantically appropriate and neither too general nor too specific), we conducted a manual review of the most prominent WordNet synsets, i.e. the top 500 synsets ranked by their number of descendants in the WordNet taxonomy: a group of three human annotators independently reviewed this set of synsets and either assigned a category that is present in the already available categorization or proposed a new one, i.e. a category whose semantics better defines such an instance and that is well distinguished from the others. The result of this step is twofold. First, we obtained 500 synsets tagged with their CNER category (that we refer to as *seed synsets*), which will be employed as a starting point for our automatic annotation procedure, explained in Section 4.1. Second, we identified six new categories, highlighted in green in Figure 2, that complement the WordNet and OntoNotes ones. The practical implication of the latter result is that, for example, biological concepts such as *molecule*, *protein* and *organism* will no longer be tagged with the OBJECT category, but as BIOLOGY. Analogously, terms such as *planet*, *quasar* and *star*, that would all have been included in the OBJECT category, will now

²WordNet synsets are organized into 45 *lexicographer files* based on syntactic categories and logical groupings.

³Entity Typing is often referred to as the second step of NER, aiming at classifying pre-identified entity mentions.

OntoNotes	CNER	WordNet
	ANIMAL	animal
	ASSET	possession
	FEELING	feeling
	FOOD	food
	PART	body
	PLANT	plant
	PROPERTY	shape
		state
		attribute
	PSYCH	cognition
		motive
	RELATION	relation
	SUBSTANCE	substance
PRODUCT	ARTIFACT	object
		artifact
TIME	DATETIME	time
DATE		
	EVENT	act
		event
		phenomenon
		process
NORP	GROUP	group
ORG	ORG	
GPE	LOC	location
LOC		tops
FAC	STRUCT	artifact
CARDINAL	MEASURE	quantity
ORDINAL		
PERCENT		
QUANTITY		
WORK_OF_ART	MEDIA	communication
PERSON	PER	person
LANGUAGE	LANGUAGE	
LAW	LAW	
MONEY	MONEY	
	BIOLOGY	
	CELESTIAL	
	CULTURE	
	DISEASE	
	DISCIPLINE	
	SUPER	

	Grounded on OntoNotes
	Grounded on WordNet
	Grounded on both resources
	New category

Figure 2: Comparison of our CNER categories (center) with the OntoNotes (left) and WordNet ones (right).

be assigned to a specific category named CELESTIAL. The same reasoning can be extended to the new categories CULTURE, DISEASE, DISCIPLINE and SUPER. For completeness, in Appendix C, we provide a textual description along with instance examples for each CNER category.

4 Resources

In the previous section, we described the procedure we used to obtain a new comprehensive categorization specifically designed to jointly capture con-

cepts and named entities. We now provide datasets to enable the training and evaluation of CNER models. Crucially, we note that none of the existing datasets provides full coverage of both concepts and entities, usually expressed by common and proper nouns, respectively.

To fill this gap, we present a methodology for automatically annotating a large corpus of sentences with CNER categories (Section 4.1), which we refer to as CNER_{silver}. Then, we describe the annotation procedure we adopted to produce CNER_{gold}, a manually-curated dataset for model evaluation (Section 4.2). Finally, in Section 4.3 we present a detailed analysis of our two resources compared to the existing ones.

4.1 CNER_{silver}

We propose an automatic approach to the creation of a large-scale training set for the CNER task. Our goal is, first, to disambiguate concepts and entities mentioned within text, and, second, to transform the resulting annotations into CNER categories. We choose the English Wikipedia as our corpus because it offers a large number of heterogeneous texts spanning all domains of knowledge. To ensure high quality, we restrict ourselves to the subset of articles that the Wikipedia community has deemed to be “good” or “featured”⁴ and apply our annotation strategy to these articles.

Importantly, Wikipedia articles contain a few terms that are linked manually to other articles, but many other terms which are left unlinked, leading to an issue of *sparsity*. Therefore, to ensure high density of annotations, we perform three main steps:

- Because the Wikipedia guidelines specify that only the first occurrence of a term should be linked,⁵ we propagate that link to all other occurrences of the term on the same page;
- As most of the mentions linked in Wikipedia articles refer to named entities, we ensure coverage and disambiguation of all common nouns with concepts in BabelNet (Navigli and Ponzetto, 2012; Navigli et al., 2021) through the application of Word Sense Disambiguation;
- All remaining entity mentions are tagged with standard NER.

⁴Wikipedia Good and Featured Articles.

⁵Wikipedia Guidelines.

Once each article has been fully annotated, we transform each tag, be it a Wikipedia hyperlink, a BabelNet concept, i.e. synset, or a CNER category, into its CNER category. In what follows, we delve into the details of our *taxonomy-based tagging* strategy for annotating Wikipedia articles, and discuss the main heuristics for *solving sparsity*.

Taxonomy-based tagging To annotate each Wikipedia hyperlink w_j in an article W_i with a CNER category c , we first retrieve the corresponding BabelNet synset and map it to one of the 500 seed synsets introduced in Section 3.2. Specifically, given that each Wikipedia hyperlink corresponds to a synset node in the BabelNet taxonomy, in order to assign a CNER category to the hyperlink w_j we start from the corresponding BabelNet node n_j and navigate upward along hypernymy edges within the hierarchy. When a seed synset with category c is reached, we assign c to the node n_j . In the case of multiple seed synsets reached at the same distance, we prioritize the most frequent category.

Solving sparsity As a result of the previous step, each hyperlink in a Wikipedia article is annotated with a CNER category. However, as previously mentioned, hyperlinks in Wikipedia are sparse (cf. line 1 in Table 1). To address this problem, we apply a surface matching heuristic where, for each link w_i with an associated category c and a surface text t_i , we propagate c to all text spans s in the same document such that $s = t_i \vee s \in \text{syn}(w_i)$, where $\text{syn}(w_i)$ is the set of synonyms of the BabelNet synset corresponding to w_i . However, while this methodology is remarkably effective for named entities, it falls short when it comes to densely annotating concepts. Indeed, after the surface matching heuristic, only 15% of common nouns are tagged, in contrast to 55.3% of proper nouns.

To fill this gap, we complement the above strategy with a state-of-the-art Word Sense Disambiguation model, ESCHER (Barba et al., 2021), to disambiguate each common noun in our dataset. As a result, we also obtain BabelNet synsets for the remaining unlinked common nouns, which we classify through the same taxonomy-based technique presented earlier. Finally, in order to annotate the remaining proper nouns with a CNER category, we use the Stanza NLP toolkit (Qi et al., 2020). This toolkit produces named entity annotations using OntoNotes categories, which are then mapped to CNER categories by exploiting the one-to-one link between OntoNotes and CNER that was presented

Sparsity heuristic	NOUN	PROPN
Wikilinks	7.1	33.2
+ surface matching heuristic	15.0	55.3
+ WSD	92.8	56.3
+ Stanza NER	97.3	98.4

Table 1: Impact of the various modules for solving sparsity in Wikipedia articles. The first row indicates the percentage of common and proper nouns hyperlinked in Wikipedia articles without applying any heuristic.

in Section 3.2. Remarkably, following the above-described process, we are able to annotate 98.4% of proper nouns, and 97.3% of common nouns, hence obtaining extremely dense annotations. This is in contrast to prior work as will be detailed later, in Section 4.3.

Table 1 reports an ablation study highlighting the distinct contributions of the previously described modules. Importantly, for each annotated span, we include explicit information on whether it is a concept (C) or a named entity (NE), which is essential for the training of the specialized NER and CR models, as will be detailed in Section 5.1. Since BabelNet provides this information for its synsets, we include it for each instance that is annotated via taxonomy-based tagging and WSD. The instances annotated using Stanza NER, instead, are all considered to be named entities except for dates and numbers. We then convert the dataset that has been produced employing the BIO tagging scheme.

4.2 CNER_{gold}

To enable a proper and rigorous evaluation of CNER models, we introduce an annotation procedure for constructing CNER_{gold}, a manually curated dataset. The annotation process started by formulating guidelines through a meticulous examination of examples within our CNER_{silver} dataset. However, the task presented multiple challenges related to the annotation of multiword nominal expressions. First, we established that non-compositional expressions such as *white shark* and *bald eagle*, which represent specific concepts, have to be identified as whole spans, while in compositional expressions, such as *black laptop*, only the noun has to be tagged. Second, in the presence of nested named entities, e.g. *Mr. Smith goes to Washington*, we decided to tag the largest available span denoting an entity, rather than tagging e.g. *Mr. Smith* and *Washington* separately. Third, we considered

Dataset	Type	# Sentences	# Tokens	# Spans	# Tagged Tokens	AVG # Spans	Density %
OntoNotes 5.0	Gold	76 714	1 388 973	104 151	190 310	1.36	13.69
UFET _{crowd}	Gold	5878	154 802	5994	17 627	1.01	11.38
UFET _{silver}	Silver	2 272 421	59 500 651	3 121 857	7 016 134	1.37	11.80
CNER _{gold}	Gold	2000	56 843	14 730	22 461	7.37	39.56
CNER _{silver}	Silver	317 590	8 725 846	2 282 800	3 403 952	7.18	39.00

Table 2: Comparison between our newly-introduced CNER resource, OntoNotes and UFET. AVG # Spans is the average number of tagged spans – either concepts or named entities – per sentence. Density % is the percentage of tagged tokens over the total number of tokens.

the cases in which the annotator is uncertain about whether a specific span of text is an entity or not (e.g. *One flew over the cuckoo’s nest*), and what its potential meanings are (e.g. *mamihlapinatapai*). To help annotators disentangle these cases, we allowed them to use external world knowledge, i.e. giving access to web search, Wikipedia, etc., to properly identify and classify spans of text. We point out that this procedure proved to be effective for the annotation task: in a sample of 100 sentences, annotators, on average, utilized external resources for more than one span per sentence.

To start the annotation task, we randomly sampled 2,000 sentences from CNER_{silver} and provided the annotators with a simple interface where each row displayed a specific token and its corresponding PoS tag. The annotators were asked to identify concepts and named entities by means of specific C vs NE tags and label them with the corresponding CNER categories by making their choice via a drop-down menu.

Because the dataset was annotated by a single expert linguist with a robust background in the annotation of lexical-semantic tasks, and an author of this paper, we asked two other annotators to label 10% of the data. We then calculated the inter-annotator agreement on such instances and obtained a Fleiss’ κ score of 89.8, indicating optimal agreement. The overall annotation process took 2 months, with a total of 320 hours spent on the task. This process led to the creation of a corpus of 2,000 sentences and more than 56,000 annotated tokens, \sim 22,000 of which were tagged with a CNER category.

4.3 Dataset Statistics

In Table 2, we provide the statistics of our CNER_{gold} and CNER_{silver} datasets compared to OntoNotes. We observe that our training set comprises a significantly larger number of annotated sentences, spans and tokens. Furthermore, from

the last two columns, we observe a considerably higher annotation density, highlighting that tagging both named entities and nominal concepts leads to a denser and richer annotation compared to focusing on entities only. In Table 2 we also present the statistics of UFET (cf. Section 2.3), but while their formulation encompasses the classification of both named entities and concepts, the annotation density is even lower than that of OntoNotes, as they disregard the identification part.

In addition, the pie charts in Figures 3 and 4 show the category balance in CNER_{silver} and CNER_{gold}, respectively. Finally, in order to evaluate the quality of the automatic annotation procedure (Section 4.1), we compute the agreement over the set of 2,000 sentences reserved for CNER_{gold}. The Cohen’s kappa coefficient over this set of samples is $\kappa = 71.4$, indicating a substantial agreement between the automatic and manual procedures. We provide further dataset statistics in Appendix A.

5 Experimental Setup

We now describe the setup, including datasets (Section 5.1) and model architecture (Section 5.2), used to answer the following research questions:

- **(RQ1)** How does a competitive neural model fare on the CNER task?
- **(RQ2)** Can a CNER system perform on par with, or even better than, specialized NER and CR systems on the respective subtasks?

5.1 Datasets

In our experiments we use several versions of our data:

- CNER_{silver} and CNER_{gold} are, respectively, the silver- and gold-standard datasets introduced in Section 4 covering both nominal concepts and named entities;

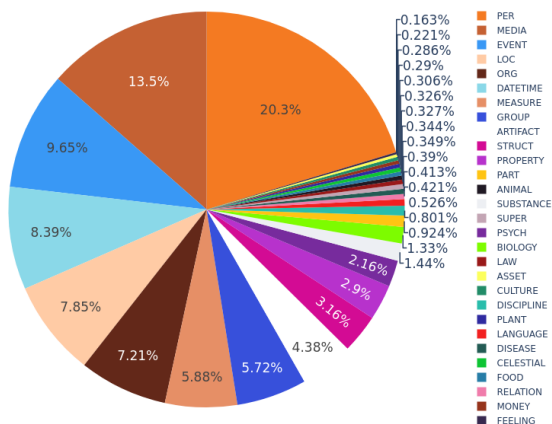


Figure 3: Pie chart showing the CNER_{silver} category distribution.

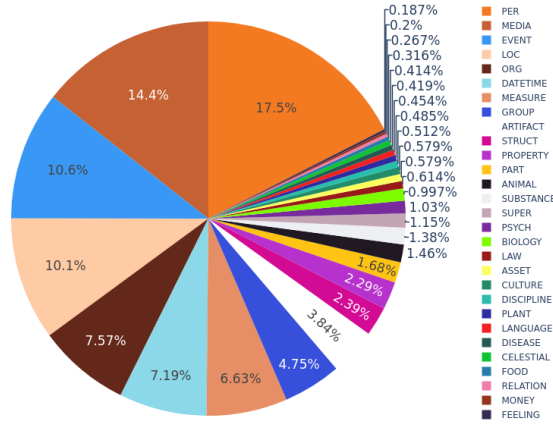


Figure 4: Pie chart showing the CNER_{gold} category distribution.

- NER_{silver} and NER_{gold} are the same datasets as above, in which only entity-related annotations are retained (i.e. concepts are mapped to the O class);
- CR_{silver} and CR_{gold} are the same datasets as above, in which only concept-related annotations are retained (i.e. named entities are mapped to the O class);

Based on these datasets, we train and evaluate the same model architecture under the following four different settings: i) we train on CNER_{silver} and test on CNER_{gold}, ii) we train on CNER_{silver} and test on NER_{gold} and CR_{gold}, iii) we train on NER_{silver} and test on NER_{gold}, and iv) we train on CR_{silver} and test on CR_{gold}.⁶ Differently from setting (i) in which a distinction between concepts and named entities is not required, in setting (ii) we train a CNER model on CNER_{silver} which, for each predicted span, provides the type (C or NE) in addition to the CNER category (e.g. for the token *doctor* the model outputs B-C-PERSON). This allows us to retain only entity- or concept-related annotations when evaluating on NER_{gold} and CR_{gold}, respectively, enabling a fair evaluation and ensuring comparability between the CNER model and the specialized NER and CR systems.⁷ In fact, an unfiltered evaluation of CNER predictions would lead to uninformative results, since a CNER model would naturally provide dense annotations regarding all the nominal expressions in a sentence, independently of their type (C or NE).

⁶We split the gold data into half for validation, half for testing (we refer to the latter as CNER_{gold}^{test}, NER_{gold}^{test} and CR_{gold}^{test}).

⁷We highlight that NER systems implicitly learn to identify NEs, in the same manner as CR does with concepts.

As evaluation metrics, we adopt the macro F1 score at the token level and the span-based F1 score. Additionally, we also report token-level micro F1 scores, because – differently from NER – the O category is much less frequent, given the high density of CNER annotations (cf. last column of Table 2).

5.2 Architecture

For our experiments, we opted for an architecture that strikes a balance between accuracy and simplicity, offering the robustness required for our task while being efficient and easy to be fine tuned. Specifically, we employ a pretrained transformer encoder, namely DeBERTa-v3 base (He et al., 2023), and train a classification head on top of it to predict the category for each token in a sentence. The classification head consists of two linear layers and a normalization layer, with a dropout of 0.1 and the GeLU activation function. Since this architecture has been proven to achieve competitive performance compared to the current state of the art in standard NER benchmarks (He et al., 2023), we believe that it constitutes a strong baseline for the CNER task.

Our systems are developed using the PyTorch Lightning framework⁸ and the HuggingFace models library.⁹ We train them on a single RTX 4090 Ti, with a patience parameter of 20 and a validation step every 30% of the total number of steps per epoch, resulting in 4 epochs and approximately 2 hours of training time for each model. We use the RAdam optimizer with a learning rate of 5×10^{-6} , and a linear scheduler with a warm-up of 10% of the total steps. We adopt the same architectural

⁸<https://www.pytorchlightning.ai/>

⁹<https://huggingface.co/docs/transformers/>

Train \ Test	NER ^{test} _{gold}			CR ^{test} _{gold}		
	Micro	Macro	Span	Micro	Macro	Span
(ii) CNER _{silver}	94.07	39.65	69.15	91.41	52.47	63.77
(iii) NER _{silver}	93.59	34.28	66.73	—	—	—
(iv) CR _{silver}	—	—	—	89.71	44.48	59.66

Table 3: Micro, Macro and Span F1 scores (%). (ii), (iii) and (iv) refer to the settings described in Section 5.1.

setup and hyperparameters to train the CNER, NER and CR models. All models are trained to minimize the cross-entropy loss function. We select the best model based on its macro F1 score on the validation set.

6 Results

RQ1 The proposed architecture achieves 87.20 micro F1, 59.38 macro F1 and 66.72 span F1 score points, when asked to identify and classify both named entities and concepts with our 29-category tagset (cf. setting (i) in Section 5.1). In Figure 5, we provide the individual span F1 scores for each category in our benchmark. In general, by cross-referencing the category scores with the distribution in Figure 3, we observe that performance has a moderate correlation with the number of training instances (Pearson correlation coefficient $\rho = 0.4$).

From a closer perspective, our system is affected by two main factors. First, when evaluated only on the identification of spans, the system achieves 86.83 span F1 score points, setting an upper bound to the overall system performance. Indeed, the model struggles to identify spans like *Triple J hottest 100 of all time in Australia* and *A full day’s Work* in their entirety. Second, the confusion matrix in Appendix B shows that the model sometimes mistakes categories that have an intrinsic level of ambiguity (e.g. ANIMAL or PLANT with FOOD). In particular, terms that can be associated with multiple categories (e.g. *chicken* and *eggplant*) are often difficult to be classified based on the limited sentence context. Additionally, specific terms, often contained in Wikipedia articles, like *Synapsids Ophiacodon* or *metal umlaut*, require technical expertise to be properly classified. We report a detailed error analysis in Appendix B.

RQ2 We summarize the results of our experiments in Table 3. Our findings clearly illustrate the significant synergistic benefits of training a single model to simultaneously identify and classify concepts and named entities, compared to restricting

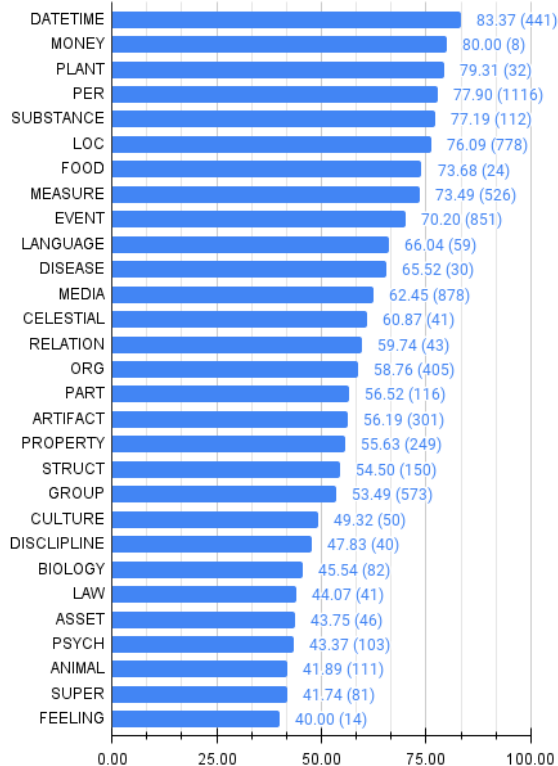


Figure 5: Bar chart with span F1 score (%) for each CNER category sorted in descending order. The support for each category is in parentheses.

the same model to either concepts or named entities. The CNER model exhibits a notable increase of ~ 0.5 micro, ~ 5.4 macro and ~ 2.4 span F1 points over the NER model. A similar improvement is achieved over the CR model with a ~ 1.7 micro, ~ 8.0 macro, and ~ 4.1 span F1 point increase.

These findings underline the advantages of our proposed CNER approach in improving the accuracy and effectiveness of both concept and named entity recognition tasks. In Section 7, we provide a qualitative analysis of our results, and reserve a detailed error analysis for Appendix B.

7 Qualitative Analysis

To better understand the behaviour of CNER, NER and CR models, we conduct a qualitative error analysis on CNER^{test}_{gold} and report some representative examples in Table 4.

In the first example, we can immediately appreciate the higher annotation density of both concepts and named entities produced by the CNER model compared to the specialized NER and CR alternatives. We also report, in the second example, an instance in which the CNER model correctly

Type	Prediction
Example 1	Gold The [first] _{MEASURE} [town] _{LOC} to fall into the [Lombards] _{GROUP} ’ [hands] _{PART} was [Forum Iulii] _{LOC} ([Cividale del Friuli] _{LOC}), the [seat] _{LOC} of the local [magister militum] _{PER} .
	NER The first town to fall into the [Lombards] _{GROUP} ’ hands was [Forum Iulii] _{STRUCT} ([Cividale del Friuli] _{LOC}), the seat of the local magister militum.
	CR The [first] _{MEASURE} [town] _{LOC} to fall into the Lombards’ [hands] _{PART} was Forum Iulii (Cividale del Friuli), the [seat] _{LOC} of the local [magister] _{PER} militum
	CNER The [first] _{MEASURE} [town] _{LOC} to fall into the [Lombards] _{GROUP} ’ [hands] _{PART} was [Forum Iulii] _{STRUCT} ([Cividale del Friuli] _{LOC}), the [seat] _{LOC} of the local [magister] _{PER} militum
Example 2	Gold [Tom McCarthy] _{PER} highlighted the prominent [role] _{PROPERTY} of [tobacco] _{PLANT} in the [story] _{MEDIA} , drawing on the [ideas] _{PSYCH} of [philosopher Jacques Derrida] _{PER} to suggest the potential [symbolism] _{MEDIA} of this.
	NER [Tom McCarthy] _{PER} highlighted the prominent role of tobacco in the story, drawing on the ideas of [philosopher Jacques Derrida] _{PER} to suggest the potential symbolism of this.
	CR Tom McCarthy highlighted the prominent [role] _{EVENT} of [tobacco] _{SUBSTANCE} in the [story] _{MEDIA} , drawing on the [ideas] _{PSYCH} of Jacques Derrida to suggest the potential [symbolism] _{PSYCH} of this.
	CNER [Tom McCarthy] _{PER} highlighted the prominent [role] _{PROPERTY} of [tobacco] _{FOOD} in the [story] _{MEDIA} , drawing on the [ideas] _{PSYCH} of [philosopher Jacques Derrida] _{PER} to suggest the potential [symbolism] _{PSYCH} of this.
Example 3	Gold [John] _{SUPER} has new [weapons] _{ARTIFACT} , including [holy water] _{SUBSTANCE} , [bait] _{BIOLOGY} for the [undead] _{SUPER} , and a [blunderbuss] _{ARTIFACT} that uses [zombie] _{SUPER} [parts] _{ARTIFACT} as [ammunition] _{ARTIFACT} .
	NER [John] _{PER} has new weapons, including holy water, bait for the undead, and a blunderbuss that uses zombie parts as ammunition.
	CR John has new [weapons] _{ARTIFACT} , including [holy water] _{SUBSTANCE} , [bait] _{SUBSTANCE} for the [undead] _{BIOLOGY} , and a blunderbuss that uses [zombie] _{BIOLOGY} [parts] _{ARTIFACT} as [ammunition] _{ARTIFACT} .
	CNER [John] _{PER} has new [weapons] _{ARTIFACT} , including [holy water] _{SUBSTANCE} , [bait] _{SUBSTANCE} for the undead, and a blunderbuss that uses [zombie] _{SUPER} [parts] _{ARTIFACT} as [ammunition] _{ARTIFACT} .

Table 4: Examples of annotations from the inference of our CNER, NER and CR models over the test split of our CNER_{gold} dataset. Correct predictions are highlighted in green, while wrong predictions are highlighted in red.

predicts the label PROPERTY for the concept *role* while CR misclassifies it as an EVENT. Furthermore, in the same sentence, we have an example of an instance for which there is one correct label, but, based on the given context, other tags could be associated with it. Specifically, both the CNER and CR models misclassify *tobacco* by tagging it with FOOD and SUBSTANCE, respectively, which are less appropriate, but both plausible annotations within the given context.

Finally, in the third example, we present a sentence in which our system shows better generalization capabilities: while the named entity *John* is misclassified as PER by both the NER and CNER models, the CNER system correctly assigns the label SUPER to the concept *zombie*, which is misclassified by the CR model with the tag BIOLOGY.

8 Conclusions

In this paper, we presented the novel task of Concept and Named Entity Recognition (CNER) and introduced a comprehensive set of categories specifically tailored to encompassing the annotation of both proper and common nouns in a unified frame-

work. To fulfill the need for specific resources in our newly introduced task, we proposed an automatic procedure that enabled the creation of the first large-scale training corpus for the CNER task. Our dataset consists of more than 300k annotated sentences, with dense coverage of both proper and common nouns obtained by devising several heuristics for solving sparsity. Moreover, we carried out a manual annotation of 2,000 sentences and used the resulting data for validation and testing.

Our experiments showed that a competitive pre-trained language model was able to successfully learn the CNER task, achieving 87.20 micro, 59.38 macro and 66.72 span F1 score points. Additionally, we also compared the performance of CNER as a joint task rather than separately identifying and classifying nominal concepts and named entities, reporting a remarkable increase in performance on both tasks.

Finally, to encourage the use and development of CNER systems, we publicly release our data and models to the research community and leave to future work the experiments on possible downstream task applications.



Limitations

The current implementation lacks extension to multiple languages, posing a potential limitation to the broader applicability of the proposed task. Nevertheless, we note that our approach is inherently language agnostic, as Wikipedia hyperlinks are linked to BabelNet synsets. Multilingual data for both NER (Tedeschi et al., 2021b; Tedeschi and Navigli, 2022) and WSD (Pasini et al., 2021) are available, but – as regards named entities – an adaptation to our set of categories would be required. We leave this to future work.

Furthermore, although the proposed tagging procedure demonstrates substantial agreement with manual annotation, as indicated by Cohen’s $\kappa = 71.4$, we remark that the $CNER_{silver}$ dataset is produced automatically, hence it may contain errors. The lack of a large-scale manually annotated gold corpus could represent a limitation to obtaining more accurate CNER models, something that we or the community can address in a future large-scale validation effort.

Finally, the results that our CNER model can achieve on $CNER_{gold}$ are bounded. In particular, during the manual annotation process, humans often had to resort to accessing external world knowledge in order to solve the intrinsic ambiguity of numerous spans. A possible solution to this limitation could be to adopt models that rely not only on the input text, but that are also able to exploit external knowledge for their predictions, as is already done by knowledge-augmented pretrained language models.

Acknowledgements

We gratefully acknowledge the support of the PNRR MUR project   PE0000013-FAIR.

Roberto Navigli also gratefully acknowledges the support of the CREATIVE project (CRoss-modal understanding and gENERATIOn of Visual and tEXtual content), which is funded by the MUR Progetti di Rilevante Interesse Nazionale programme (PRIN 2020). This work has been carried out while Giuliano Martinelli, Francesco Maria Molfese and Simone Tedeschi were enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

References

- Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. [ESC: Redesigning WSD with extractive sense comprehension](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.
- Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. [Recent trends in word sense disambiguation: A survey](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. [Ultra-fine entity typing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.
- Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. [Few-NERD: A few-shot named entity recognition dataset](#). 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)*, pages 3198–3213, Online. Association for Computational Linguistics.
- Greg Durrett and Dan Klein. 2014. [A joint model for entity analysis: Coreference, typing, and linking](#). *Transactions of the Association for Computational Linguistics*, 2:477–490.
- Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. [SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition \(MultiCoNER 2\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Ralph Grishman. 2015. Information extraction. *IEEE Intelligent Systems*, 30(5):8–15.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#).
- Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. [Red^{fm}: a filtered and multilingual relation extraction dataset](#). In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.
- Rubén Izquierdo, Armando Suárez, and German Rigau. 2015. Word vs. class-based word sense disambiguation. *J. Artif. Int. Res.*, 54(1):83–122.

- Rubén Izquierdo Beviá, Armando Suárez Cueto, German Rigau Claramunt, et al. 2007. Exploring the automatic selection of basic level concepts.
- Sopan Khosla and Carolyn Rose. 2020. [Using type information to improve entity coreference resolution](#). In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 20–31, Online. Association for Computational Linguistics.
- Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. [A survey on deep learning for named entity recognition](#). *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.
- Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. [Nibbling at the hard core of Word Sense Disambiguation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.
- George A. Miller. 1995. [Wordnet: A lexical database for english](#). *Commun. ACM*, 38(11):39–41.
- Diego Mollá, Menno van Zaanen, and Daniel Smith. 2006. [Named entity recognition for question answering](#). In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 51–58, Sydney, Australia.
- David Nadeau and Satoshi Sekine. 2007. [A survey of named entity recognition and classification](#). *Linguistic Investigations*, 30(1):3–26.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.
- Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. [Ten years of BabelNet: A survey](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Roberto Navigli and Simone Paolo Ponzetto. 2012. [BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network](#). *Artificial Intelligence*, 193:217–250.
- Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. [Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.
- Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. 2017. [Relation extraction : A survey](#).
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. [Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes](#). In *Joint conference on EMNLP and CoNLL-shared task*, pages 1–40.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the conll-2003 shared task: Language-independent named entity recognition](#).
- Rhea Sukthankar, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. [Anaphora and coreference resolution: A review](#). *Information Fusion*, 59:139–162.
- Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. 2021a. [Named Entity Recognition for Entity Linking: What works and what’s next](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2584–2596, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021b. [WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Simone Tedeschi and Roberto Navigli. 2022. [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*, pages 801–812, Seattle, United States. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang. 2002. [Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition](#). In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [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*, pages 142–147.
- Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. [Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation](#). In *Proceedings of the 10th Global Wordnet Conference*, pages 108–117, Wrocław, Poland. Global Wordnet Association.

A Additional Dataset Statistics

In Table 6 we highlight the overall proportion of concepts (C) and named entities (NE) in our $CNER_{silver}$ and $CNER_{gold}$ datasets, respectively. Additionally, in Figure 6, we provide the distribution of concept and named entities for each individual category in our datasets.

B Additional Results

RQ1 (CNER system) In Figure 5, we presented the results of our fine-tuned CNER system evaluated on the $CNER_{gold}^{test}$ test set. For completeness, and to better interpret the obtained results, in Figure 7, we present a confusion matrix of the system’s outputs. The matrix shows that, in general, the categories exhibit well-defined boundaries, with some notable exceptions. Specifically, the categories PROPERTY, PSYCH, and FEELING are frequently confused due to their intertwined semantics. Additionally, instances belonging to the LAW category, such as *peace treaty*, can occasionally overlap with instances from the MEDIA category, which encompasses general written documents conveying information. Furthermore, the categories ANIM and BIO are prone to misclassification, primarily because distinguishing between animals and biological entities can be challenging (e.g. *chep-halopods*). Finally, DATETIME and MEASURE are often confused due to the difficulties of discerning time periods when used to explicitly measure time, as opposed to identifying specific periods in time.

RQ2 (CNER vs NER & CR) In Table 3 we showed the benefits of performing CNER as a unified task rather than identifying concepts and named entities separately. In Figure 8, we illustrate the category-wise impact of the CNER system when compared to NER and CR systems, highlighting its positive and negative contributions. Notably, the CNER system consistently exhibits a positive influence over both named entities and concepts. Specifically, categories such as DISCIPLINE, FOOD, and CULTURE witness particularly positive contributions, with an increase of up to 40 span-based F1 score points. In some categories, the CNER system provides positive contributions only for named entities, as observed in GROUP and MEDIA, or exclusively for concepts, as evidenced in SUPER and LAW.

Evaluation on Standard NER Benchmark In order to further evaluate our $CNER_{silver}$ data,

Training Strategy	LOC	PER	ORG
(1) NER_{silver}	94.2	96.2	54.7
(2) CoNLL-2003	96.7	98.3	98.2
(3) $NER_{silver}(F)$	96.8	98.0	98.0

Table 5: Classification accuracy (%) of the same architecture trained on three different corpus and tested on the CoNLL2003 benchmark. $NER_{silver}(F)$ is the model pre-trained on NER_{silver} and finetuned on the CoNLL03 training set.

Dataset	NE	C
$CNER_{silver}$	48.3%	51.7%
$CNER_{gold}$	43.6%	56.4%

Table 6: Distribution of named entities and concepts.

we present an out-of-domain comparative analysis with a standard NER benchmark, namely CoNLL-2003 (Sang and Meulder, 2003). In particular, the objective of this analysis is to asses if a model, trained on the NER_{silver} data, is able to attain performance that is comparable to that of directly training on the CoNLL-2003 training corpus. Comparing the two models presents several challenges. First, our span annotation guidelines are different from those of a traditional NER dataset, due to the inclusion of concepts (e.g. "*president George Washington*" is a whole named entity span, while CoNLL03 would only tag "*George Washington*"). For this reason, in order to compare the two models we decided to evaluate classification accuracy only, without evaluating the span extraction. Another significant problem are the differences in the categorizations: while some of our 29 categories are directly associated with PER, ORG and LOC, the MISC category is not comprehensive because the CoNLL03 test set lacks annotations for the instances of 12 of our categories. For this reason, Table 5 shows classification accuracy on PER, ORG and LOC of three training strategies of the same architecture presented in Section 5.2:

1. The model is trained on NER_{silver} and its outputs are mapped to the three standard NER categories.
2. The model is trained on the CoNLL03 dataset.
3. The model is trained on NER_{silver} with an additional fine-tuning on the CoNLL03 dataset.

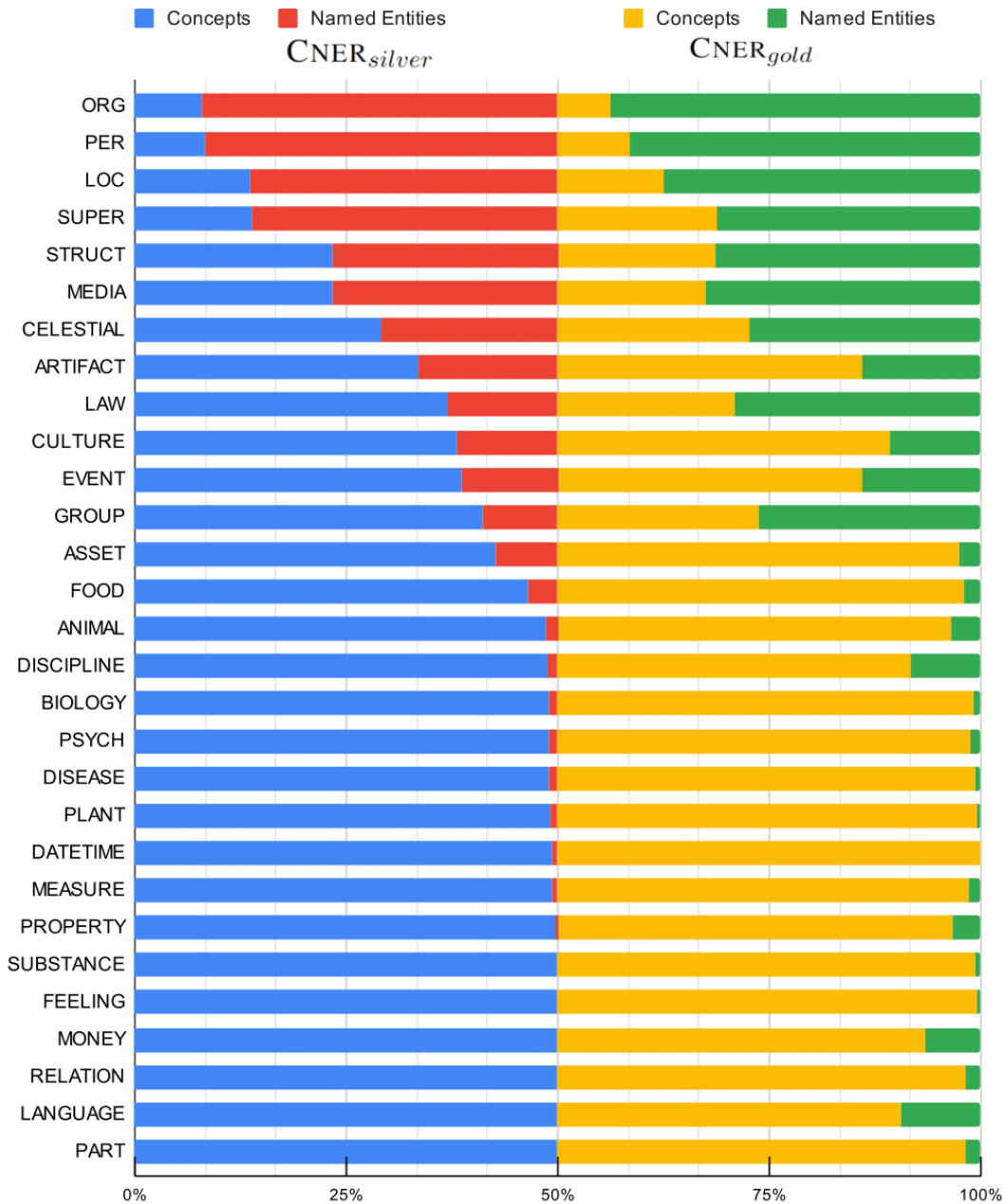


Figure 6: Bar chart with the distribution of named entities and concepts for every CNER category in the CNER_{gold} and CNER_{silver} resources.

Despite the different domains present in NER_{silver} and CoNLL03, namely Wikipedia texts and news articles, respectively, the resulting models obtain comparable performance in labelling PER and LOC instances. However, the same does not hold for the ORG instances. Interestingly, we noticed that the low scores of the model trained solely on our data are strongly influenced by the large number of sports articles available in the CoNLL03 test set. In particular, many of the organizations correspond to sports teams whose names often match the name of the corresponding

city. Nevertheless, fine-tuning the system on the CoNLL03 training set solves this problem.

C CNER Categories

In Table 7, we provide the full list of CNER categories, along with their textual descriptions and instance examples for both named entities and concepts.

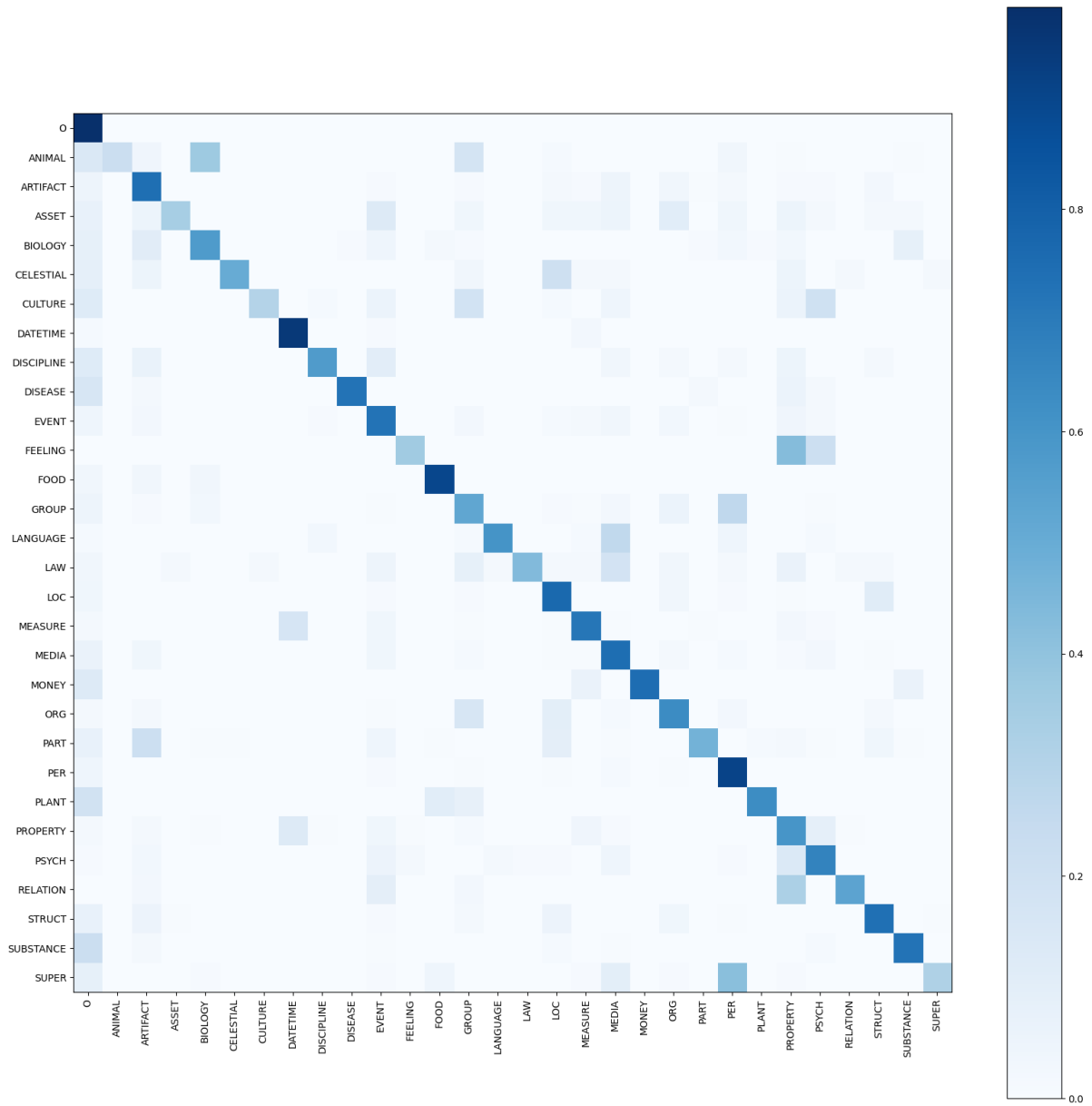


Figure 7: Bar chart with Span F1 score of the different categories.

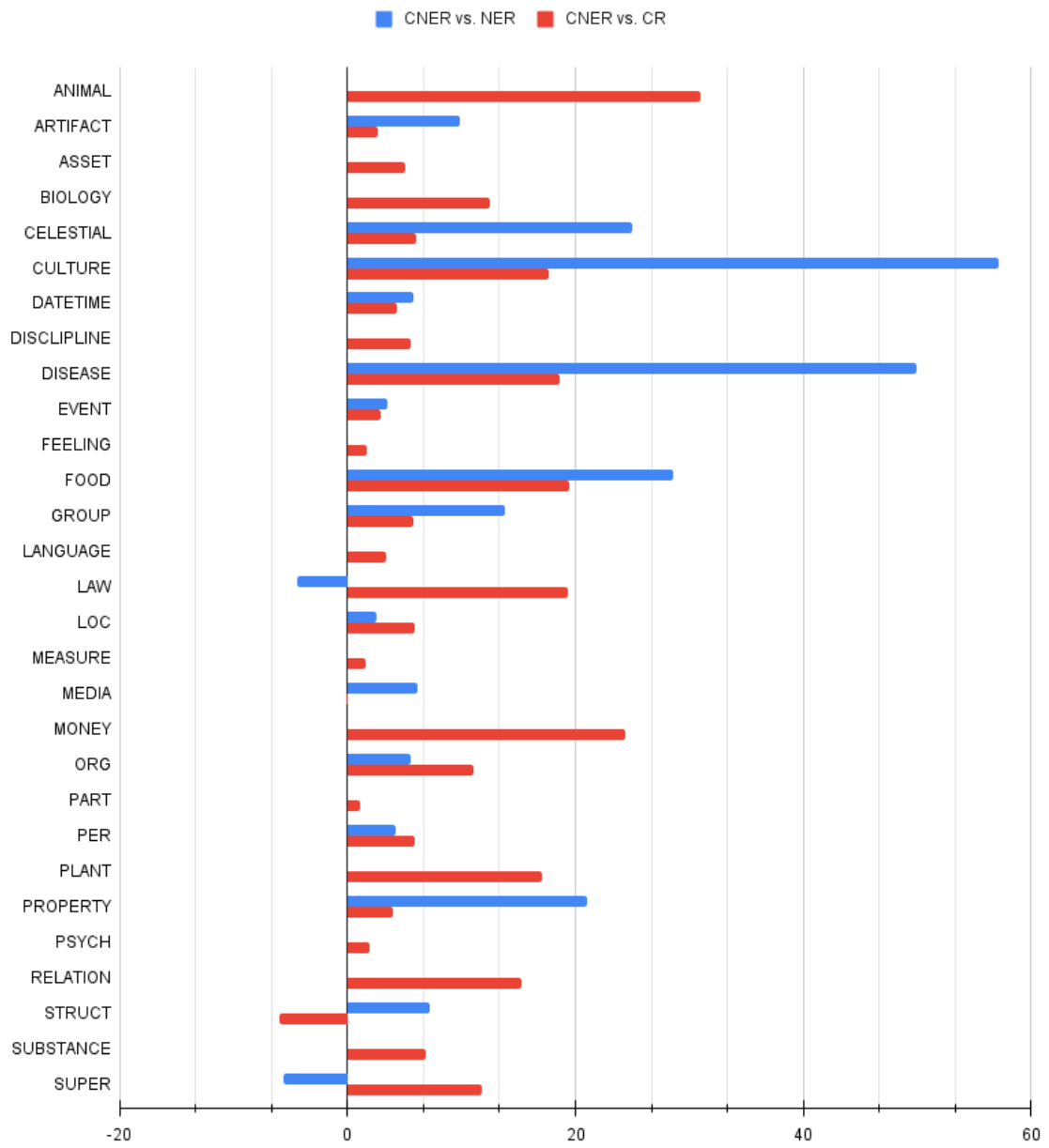


Figure 8: Bar chart with Span F1 score of the different categories.

Category	Description	Examples
ANIMAL	<i>Living beings (excluding humans) with the ability to move and perceive their surroundings.</i>	<i>dog, cat, mammal, carnivore, brown bear, African Wild Dog, Great White Shark,</i>
ARTIFACT	<i>All the objects, artifacts, tools, products and items</i>	<i>vehicle, software, mouse, data stream, Windows XP, Fiat Panda</i>
ASSET	<i>Assets, resources, or possessions with economic or intrinsic value.</i>	<i>capital, stock, wealth, resource, phone bill, Federal Perkins Loan, Investment in Russia</i>
BIOLOGY	<i>Biological entities, including living organisms, cells, or biological components</i>	<i>protein, cell, living organism, lipid, Herpes Simplex Virus, Escherichia Coli</i>
CELESTIAL	<i>celestial bodies as Planets, stars, asteroids, galaxies and other astronomical objects.</i>	<i>comet, nebulae, Sun, Neptune, Asteroid 187 Lamberta, Proxima Centauri</i>
CULTURE	<i>Cultural aspects, traditions, customs, and practices associated with specific groups or societies.</i>	<i>religion, feminism, socialism, capitalism, anarchism, doctrine, cult, Islam, Buddhism</i>
DATETIME	<i>Dates and times</i>	<i>18 March, Saturday, 1979, the evening of 19 November, 15:30 am</i>
DISEASE	<i>medical conditions, illnesses, disorders, and health-related issues affecting living organisms.</i>	<i>infection, allergy, metastasis, complication, acne, Alzheimer's Disease, Cystic Fibrosis</i>
DISCIPLINE	<i>specific fields of study, knowledge, or expertise. It includes academic disciplines, areas of research, and professional domains.</i>	<i>discipline, sport, football, computer science, anatomy, long jump.</i>
EVENT	<i>Events, phenomenon or activities that occur at specific times or places. It includes both significant and everyday occurrences</i>	<i>crime, professorship, temperature change, 2003 Wimbledon Championships, Cannes Film Festival.</i>
FEELING	<i>Emotions, sensations, and subjective experiences related to human or animal consciousness.</i>	<i>affection, attachment, agitation, craving, urge, temptation.</i>
FOOD	<i>edible items, dishes, beverages, and culinary products that are consumed for nourishment or enjoyment</i>	<i>beverage, dish, pork, lasagna, Carbonara, Sangiovese, Cheddar Beer Fondue, Pizza Margherita.</i>
GROUP	<i>group of people or animals</i>	<i>staff, social group, panel, militia, community, trio, duo, family, genealogy, alliance, nationality, peoples</i>
LANGUAGE	<i>individual language-related items, such as words, phrases, or idiomatic expressions</i>	<i>discourse, context, lexeme, morpheme, appellation, eponym, nickname, vowel, syllable, headword</i>
LAW	<i>legal principles, regulations, and rules governing society and various aspects of life</i>	<i>law, civil law, administrative law, martial law, shariah, ordinance, civil right, Magna Carta, Islamic Law</i>
LOC	<i>geographical locations, such as villages, towns, cities, regions, countries, continents, landmarks, or natural features</i>	<i>space, surface, street, road, town, Rome, Lake Paiku, Mississippi River.</i>
MEASURE	<i>units of measurement and quantification used to determine the size, quantity, or quality of various objects or phenomena.</i>	<i>day, microsecond, millisecond, two, 35, 45%, first, temperature, length,</i>
MEDIA	<i>various forms of communication and entertainment media, such as newspapers, television shows, movies, social media or digital content.</i>	<i>soundtrack, report, publication, language, English, Forbes, American Psycho</i>
MONEY	<i>monetary units, currencies, and financial values used in different contexts</i>	<i>monetary unit, dollar, 15 euros, 1116 CHF</i>
ORG	<i>organizations, institutions, and companies involved in diverse sectors or activities</i>	<i>Industry, commercial enterprise, San Francisco Giants, Google, Democratic Party.</i>
PART	<i>individual components or sections of larger entities or objects</i>	<i>finger, chin, head, tail, femur, airplane wing, airplane's wings, flower's stem</i>
PER	<i>individuals or persons, including real people and historical figures</i>	<i>doctor, historian, professor, musician, Ray Charles, Jessica Alba</i>
PLANT	<i>Types of trees, flowers, and other plants, including their scientific names.</i>	<i>grass, peach tree, Forsythia, Artemisia Maritima.</i>
PROPERTY	<i>properties or attributes of objects, entities, or concepts</i>	<i>thickness, height, dimension, shape, age</i>
PSYCH	<i>psychological concepts, mental states, and phenomena related to the human mind and behavior</i>	<i>psychological feature, cognition, attention, necessity</i>
RELATION	<i>relationships, connections, and associations between entities or concepts</i>	<i>apport, competition, comparison, bridge, relatedness, parentage, function, parity, transitivity</i>
STRUCT	<i>physical structures, including buildings, architectural designs, and engineered constructions made by humankind</i>	<i>shelter, gravestone, refuge, tent, loft, San Peter's Church, Golden Bridge</i>
SUBSTANCE	<i>chemical substances</i>	<i>acid, bactericide, carbonyl, explosive, fertilizer, Zyclon B</i>
SUPER	<i>Mythological and religious entities.</i>	<i>Apollo, Persephone, Aphrodite, Saint Peter, Pope Gregory I, Hercules.</i>

Table 7: Label, description, and instance examples of each of our CNER categories.