UkraiNER: A New Corpus and Annotation Scheme Towards Comprehensive Entity Recognition

Lauriane Aufrant*, Lucie Chasseur*
Inria, France
first.last@inria.fr

Abstract

Named entity recognition as it is traditionally envisioned excludes in practice a significant part of the entities of potential interest for real-word applications: nested, discontinuous, non-named entities. Despite various attempts to broaden their coverage, subsequent annotation schemes have achieved little adoption in the literature and the most restrictive variant of NER remains the default. This is partly due to the complexity of those annotations and their format. In this paper, we introduce a new annotation scheme that offers higher comprehensiveness while preserving simplicity, together with an annotation tool to implement that scheme. We also release the corpus UkraiNER, comprised of 10,000 French sentences in the geopolitical news domain and manually annotated with comprehensive entity recognition. Our baseline experiments on UkraiNER provide a first point of comparison to facilitate future research (82 F1 for comprehensive entity recognition, 87 F1 when focusing on traditional nested NER), as well as various insights on the composition and challenges that this corpus presents for state-of-the-art named entity recognition models.

Keywords: nested NER, annotation scheme, corpus

1. Introduction

Named entity recognition, as the foundational task of Information Extraction, has known many flavours over the last three decades. While it was originally framed as a sequence labelling task, aiming at identifying non-overlapping names (flat NER), real-world needs have led to gradually broadening its scope to identifying all occurrences of names (nested NER), and even pronominal or nominal mentions of entities beyond names only (extended NER).

Figure 1 illustrates the loss of useful information that this evolution mitigates, with numerous entities missed, even for common entity types such as persons, locations and organisations.

To date however, no annotation scheme achieves full comprehensiveness with respect to those entity types. For instance, in Figure 1, none of the NER variants captures "*students*" or "*NATO representative*", thereby missing a significant part of the information conveyed by the sentence.

Besides, the few annotation schemes and associated corpora that have targeted comprehensiveness rely on rich annotations with complex formats, which until now has hindered their wide adoption. Most works in named entity recognition remain today targeted at flat and proper named entities.

Another gap that we aim to address is the relatively limited amount of annotated data for French – at least compared to English.

Our contribution is thus threefold:

She invited their CEO to visit Princeton University, although students were off to see the NATO representative.	
Flat NER:	Princeton University$_{LOC}$, NATO$_{ORG}$
Nested NER:	Princeton$_{LOC}$, Princeton University$_{LOC}$, NATO$_{ORG}$
Extended NER:	She$_{PER}$, their CEO$_{PER}$, Princeton$_{LOC}$, Princeton University$_{LOC}$, NATO$_{ORG}$

Figure 1: Example sentence and resulting entities, for NER variants of increasing comprehensiveness.

- We propose a new annotation scheme for entity recognition, including both a format and annotation guidelines, that yields better comprehensiveness for entity annotation.
- We release an annotation tool to implement those guidelines and visualize annotated corpora.
- We release UkraiNER, a new corpus of 10,000 French sentences in the geopolitical news domain, annotated along that scheme.

The associated material can be accessed online from https://who.paris.inria.fr/Lauriane.Aufrant.

After reviewing existing annotation schemes and corpora (§2), we describe our new annotation scheme (§3), the corresponding annotation tool (§4) and the resulting corpus UkraiNER (§5). To facilitate further research using UkraiNER, we also provide baseline results in §6.

* Equal contributions

LREC-COLING 2024, pages 16941–16952
20-25 May, 2024.

2. Background

2.1. Entity recognition guidelines

Named entity recognition consists in identifying non-overlapping (i.e. flat) mentions of names in text, and categorizing them into predefined types (e.g. person, organization, location). This task was first formalized in the context of the MUC evaluation campaign (Grishman and Sundheim, 1996; Chinchor and Robinson, 1998) and was later consolidated by the CoNLL-2003 shared task on NER (Tjong Kim Sang and De Meulder, 2003).

Over the years, it appeared however that the exclusive focus on names was falling short of the real-world information needs, and the ACE program proposed to extend the recognition of names to a recognition of entities (Doddington et al., 2004). This resulted in the "entity detection and tracking" task, that combined the extraction of entity mentions and the identification of coreference chains. ACE guidelines recognize 3 types of entity mentions (named, nominal and pronominal), including nested mentions, and 7 entity types (persons, organizations, locations, facilities, weapons, vehicles and geo-political entities), each with associated subtypes (e.g. distinguishing air, water and land vehicles). Covering nested mentions enables for instance to recognize both the person and the geo-political entity in "*US Secretary of State*" and to recognize both organisations (the company and its board) in "*the Lufthansa Executive Board*".

More recently, the Quaero project has proposed, specifically for French, new annotation guidelines (Grouin et al., 2011; Rosset et al., 2011) introducing the concept of "extended named entities" to encompass more entity types (e.g. civilizations) and more entity mentions (with a strong focus on nominal mentions and their nesting). The design of that scheme was driven by the ultimate goal of leveraging those entities to build knowledge bases, which is different in spirit from the frequent use case of applying (CoNLL-style) NER to index documents. Indeed, while knowing that a document contains the mention "her university" is of limited interest by itself, detecting it in "*Her university was founded 500 years before Princeton*" is crucial as a first step towards reconstructing the person's curriculum (through coreference resolution, relation extraction, entity linking and a seed knowledge base).

Our work is closest to the ACE approach and to the motivation from Quaero, but it seeks a higher level of comprehensiveness while keeping a lightweight scheme, both regarding the format and the content of annotations. Indeed, it is worth mentioning that named entity recognition has known a high diversity of formats, including BIO (and many variants) for flat NER when framed as a sequence labelling task, XML-based formats for fine-grained annotation of nested NER (with inlined annotations in Quaero and standoff annotations in ACE), and tabular standoff annotations in the "brat" format. Today, tabular formats are gaining stronger adoption in many areas of NLP, in part due to their simplicity and flexibility, to repeated CoNLL shared tasks with tabular formats, and to the flourishing of the CoNLL-U format as part of the Universal Dependencies (UD) project (Nivre et al., 2016). We propose therefore to pursue on that track and adopt a tabular format.

Beyond entity recognition, other areas of NLP have noted and accounted for the importance of entities beyond names. For instance, in Open Information Extraction, the Wire57 benchmark (Lechelle et al., 2019) has been observed to contain 10 times more triples when allowing any noun phrase as relation argument, rather than named entities only. Similarly, in coreference resolution (Grobol, 2020), the mention detection component is meant to identify all expressions that refer to an entity, even if not explicitly mentioning it (e.g. through a possessive). This however follows a much broader concept of entity, including also ideas, disciplines or situations for instance (in short, any noun phrase), and without an objective of typing them. While closely related and complementary to our work, we thus do not see coreference resolution guidelines as an answer to the need for comprehensive entity recognition. Rosales-Méndez et al. (2018) asks similar questions on the side of entity linking and further underlines the importance of tracking directly in the annotations the different concepts of entities and entity mentions, for the sake of flexibility.

2.2. Entity recognition corpora

Across the existing guidelines, a number of corpora have been annotated and distributed, each with its pros and cons. We identify here those that have been most widely adopted in the entity recognition community.

For flat NER, the MUC-7 corpus (Chinchor and Robinson, 1998), based on English news articles, has seen decreasing usage in the last two decades, to the benefit of the CoNLL-2003 corpus (Tjong Kim Sang and De Meulder, 2003) of 22k English sentences drawn from Reuters news stories in 1996 (and 19k in German, from the Frankfurter Rundshau newspaper in 1992), which today remains a widely-used benchmark. WikiNER (Nothman et al., 2013) offers more recent, more massive (130k to 250k sentences per language) and more multilingual (9 languages including English and French) data for flat NER, albeit with lower quality (silver-standard). Indeed, it is composed of Wikipedia articles automatically labelled

based on the presence of hyperlinks to Wikipedia pages (entity annotations being manually defined on those Wikipedia pages' titles). WikiNEuRal (Tedeschi et al., 2021) similarly covers 9 languages including French (around 100k sentences per language), with silver-standard annotations.

Nested NER has been mostly driven by the ACE2004 (Alexis Mitchell, Stephanie Strassel, Shudong Huang, Ramez Zakhary, 2004) multilingual corpus (English, Chinese and Arabic, 150k words per language, multi-genre), and the subsequent ACE2005 (Christopher Walker, Stephanie Strassel, Julie Medero, Kazuaki Maeda, 2005). Concurrent efforts have however been done on the Genia corpus (Kim et al., 2003), which is specific to the biomedical domain (18k English sentences) and focuses on term identification (e.g. extraction of protein mentions) but is often viewed as a form of nested NER. In addition, OntoNotes 5.0 (Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, Ann Houston, 2013) contains entity annotations in 3 languages (English, Chinese, Arabic, 1 million words per language, multi-genre) with a rich set of 18 entity types, and accounts for unrestricted mentions in coreference resolution, but for the entity part it remains centred on proper names.

For the specific case of French, the ESTER corpus (Galliano, Sylvain and Geoffrois, Edouard and Gravier, Guillaume and Bonastre, Jean-François and Mostefa, Djamel and Choukri, Khalid, 2006) composed of broadcast news transcripts was the first major corpus for NER (Galliano et al., 2006), and ESTER 2 was incorporated in the Quaero corpus (Grouin, Cyril and Rosset, Sophie and Zweigenbaum, Pierre and Fort, Karën and Galibert, Olivier and Quintard, Ludovic, 2013), that amounts to 1.5 million words. WiNER-fr (Dupont, 2019) has further pursued on that track, with the release of 13k sentences from Wikinews, annotated with guidelines close to Quaero, in brat standoff format. Concurrently, the 12k sentences of the French Treebank (in the news domain) have also been annotated with named entities (Sagot et al., 2012), but with an exclusive focus on proper names, and only flat entities. Ortiz Suárez et al. (2020) have later extended that corpus by realigning it with UD tokenization and a CoNLL-U-based format (with an additional column for flat NER labels), and complementing it with entity linking information. More recently, DWIE-FR (Verdy et al., 2023) has provided additional data for French, but similarly limited to flat named entities.

3. Annotation scheme for Comprehensive Entity Recognition

We propose to build on previous works by introducing a new annotation scheme that combines broad coverage of entities of interest (including entities that are not considered in previous guidelines), with lightweight annotations and a flexible format.
The proposed annotation scheme is based on three main principles:

1. Targeted entities encompass a broad set of entity types (including e.g. products), but labeled with coarse types (e.g. not distinguishing administrations and companies among organizations). The annotation scheme covers 8 entity types: Person, Location, Date, Event, Organization, Group, Product and Other.

2. All entities in those classes and all mentions of those entities (e.g. all persons, regardless of who they are and how they are mentioned) are annotated. This notably includes nested, non-named and discontinuous entities.

3. Entity mentions that would not be included according to more restrictive annotation schemes are annotated with rich information that tracks the criteria used to include them. For instance, each entity is labelled as named or non-named. Motivation to track this information is to offer the ability to filter the annotations and retain only the named entities when the use case warrants it, or to ensure comparability of experiments across corpora.

This section provides an overview of the annotation scheme, and the full guidelines are hosted and maintained at `https://who.paris.inria.fr/Lauriane.Aufrant/`, to be versioned across future revisions.
Appendix A further illustrates those guidelines with motivating examples that underline the added value in coverage and appropriate modelling of the information conveyed by the text.

3.1. CoNLL-UI format

We extend the CoNLL-U format with new elements to encode entities. Compared to Ortiz Suárez et al. (2020), since we operate on nested NER it is not possible to add all that information as an extra column. We therefore add entity annotations in a standoff format that is inlined as comment above the tokens. The resulting format CoNLL-UI (for "CoNLL-U with Information extraction") is illustrated in Figure 2.
Each entity line is space-separated and contains 5 fields: the hyphen-separated list of token ids that

```
# text = Biden met the European ambassadors last night.
# entity: 1 PERSON named complete main -- > Biden
# entity: 3-4-5 GROUP non-named complete main -- > the European ambassadors
# entity: 6-7 DATE non-named complete main -- > last night
1  Biden       Biden       PROPN  NNP  ...  ...
2  met         meet        VERB   VBD  ...  ...
3  the         the         DET    DT   ...  ...
4  European    european    ADJ    JJ   ...  ...
5  ambassadors ambassador  NOUN   NNS  ...  ...
6  last        last        ADJ    JJ   ...  ...
7  night       night       NOUN   NN   ...  ...
8  .           .           PUNCT  .    ...  ...
```

Figure 2: Example of sentence annotated with entities in the CoNLL-UI format. The new lines added compared to the CoNLL-U format are in **bold**.

are part of the entity, the entity type (see §3.5), a label "named" or "non-named" (§3.3), a label "complete" or "incomplete" (§3.4), and a label "main" or "subdivision" (§3.2). The last part (after $--$ >) displays the tokens of the entity and serves only as convenience for manual inspection of the corpus.

3.2. Nested entities

By contrast with flat NER schemes, our annotation scheme includes entities whose mention is nested within another entity. For instance, the entity "*the President of the lower house of the Parliament*" (Person) contains two other nested entities: "*the lower house of the Parliament*" (Organization) and "*the Parliament*" (Organization).

This can concern entities that are of a different type from the larger entity, or different entities from the same type, but also a nested mention of the same entity. For instance, the entity "*the French President Jacques Chirac*" has two nested entity mentions ("*the French President*" and "*Jacques Chirac*") that refer to the same person. In such case, all three mentions are annotated but the nested ones are marked as a (coreferring) subdivision of the largest one. This marker offers flexibility with respect to the preferred convention: either all mentions, only the subdivisions, or only the largest mention for that entity.

Subdivision entities are only annotated on semantic units that could autonomously denote the entity. For instance, "*the late Jacques Chirac*" has only one nested subdivision ("*Jacques Chirac*", but not "the late"). Punctuation-based appositions are considered as ellipsis of a verbal phrase and therefore annotated as separate entities (not nested within a larger one): the segment "*France's President, Jacques Chirac*" contains three entity mentions ("*France*", "*France's President*", "*Jacques Chirac*") but is not itself annotated as entity.

Apposition of acronyms (e.g. "*the European Union (EU)*") is annotated as one entity with nested subdivisions (here, "*the European Union*" and "*EU*").

3.3. Non-named entities

Following the rationale of (Paris and Suchanek, 2021; Lechelle et al., 2019), entities in this scheme are not limited to named entities, but also include non-named entities.[1] This concerns both entities that do not have a name (e.g. DATE "*earlier*") and entities whose name is not mentioned (e.g. PERSON "*a former senator*").

This choice implies in particular that annotated mentions are more diverse than proper nouns only. These can include any noun phrase, pronoun, or pronominal phrase. Adverbs ("*earlier*"), numerals and symbols (e.g. for DATE or OTHER), as well as free relative clauses (e.g. EVENT in "*what happened yesterday surprised everyone*"), are other valid examples of entity mentions.

By contrast with mention detection as is done in coreference resolution, possessive determiners ("*my*") are not considered as denoting an entity, but a relation with an entity, hence they are not annotated in this scheme. Possessive pronouns ("*mine*") are still annotated as the possessed entity if any, but not as the possessor. Adjectives such as "*French*" in "*the French President*" are similarly excluded (expression of a relation with entity France, rather than of the entity France itself).

Entities are additionally labeled as named or non-named. For persons, mentions containing a first name, last name or nickname ("*Nelson*", "*Nelson Mandela*", "*Mr Mandela*", "*Madiba*") are considered named, whereas functions and titles are non-named ("*the UN Secretary-General*"). For legally recognized organizations, only their official denomination is considered named ("*France's ministry of Defence*" is not named but its official name "*the Ministry of Armed Forces*" is). Absolute dates ("*January 24"*) are considered named, whereas hours ("at *07:00*") and relative dates ("*tomorrow*", "*on Monday*") are non-named. URLs, brands and product models ("*a Leclerc tank*") are named.

3.4. Discontinuous and incomplete entities

Entities are annotated even if discontinuous, in which case the annotation is positioned on the continuous span that is most salient and the entity is marked as "incomplete". For instance, in "*the French and British ambassadors*", "*the French*" is annotated as an incomplete entity ("*the French ... ambassadors*"), and similarly for "*Prime Minister*" in "*the French President and Prime Minister*". The choice of not annotating all tokens in the entity was driven by UX considerations of the associated annotation tool (see §4), as annotating spans is much faster than individual selection of tokens.

Due to subdivisions (see §3.2) there can be two entities on the same span: for instance "*the Presidents Macron and Putin*" has five nested entities

[1] These entities are sometimes referred to in the literature as *unnamed* entities, but we follow here Paris and Suchanek (2021)'s terminology.

("*the Presidents Macron*", subdivision "*Macron*", "*the Presidents ... Putin*" marked as incomplete on "*Putin*", its complete subdivision "*Putin*", and group "*The Presidents*").

In case of obviously missing tokens (e.g. due to segmentation issues, or typos), the entity is also marked as incomplete. Qualifiers inlined in the entity are considered as part of the entity, hence the entity remains continuous. Lack of determiner (as in "*the French President and British Prime Minister*") is ignored and the entity is considered complete.

3.5. Entity types

Person encompasses all mentions of an individual (a natural person or fictional character), regardless of whether that mention uniquely identifies that individual. For instance, in the sentence '*I think I saw a tall man next to Pete's wife*", the spans "*I*" (twice), "*a tall man*", "*Pete*" and "*Pete's wife*" are all annotated as Person entities.

Location corresponds to places, including any granularity: a room, building, street, city, country, etc. Mentions of relative positioning ("*near Mr Macron*") are annotated as Location entities only when meant to autonomously designate a zone, not when expressing a positioning relationship among entities ("*this happened near Mr Macron*", "*a fishing village, near Marseilles*").

Date includes any expression of time, whether absolute or relative, including both expressions relative to another date ("*before 17:00*") and relative to the context ("*soon*"). Periods of time are also included: "*from February 15 to 19*" is a non-named Date, with nested Date "*February 15*" and incomplete nested Date "*19*" (named). Tokens expressing a relation (e.g. of an event) with that date are not considered as part of the entity (as in "*on Monday*", "*at 07:00*"). Full timestamps including both date and hour are annotated as one entity with two nested entities.

Event includes any public event ("*2024 Summer Olympics*") but also meetings, press conferences, or an official speech for instance. Possible events are only annotated if planned or well-defined in time: for instance, "*Everyone hides since the drone attack*" and "*They hope to launch a drone attack soon*" express an event, but not "*If you go out, be careful in case of drone attack*".

Organization encompasses companies, associations, administrations, or any other structured group or legal entity that is not a Person. This includes entities that have no legal status but present some structure or coordination (not necessarily hierarchical, as is the case with the "Yellow vests" collective), as well as mentions of entities that are supposedly consistent or driven by a common intent or purpose and not focused on individuals. For instance, mentions of a people denote an Organization, whereas mentions of the citizens or residents in a given country belong to Group. An Organization can correspond to a single individual (e.g. for a one-person company), or have members who are not individuals (e.g. a federation). Mentions of countries and administrative regions are typed as organizations when referring to the government or to the underlying legal entity.

Group is any set of individuals or organizations, which does not constitute an organization by itself (not structured). Compared to *pers.coll* in the Quaero scheme, it also includes entities such as "*the victims*" (although non-named) or "*the political parties*" (group of organizations). An enumeration of Persons is annotated as a Group when they act in unison or are regarded as such (e.g. a couple, as in "*Joe and Jill Biden met the King*"). A music band or a one-off collective ("*the demonstrators*") are considered as Organizations even if not legally recognized, but an unstructured aggregate such as "*civil society*" is a Group.

Product corresponds to objects and equipment (e.g. vehicles, weapons, food, artwork), but also software or websites (e.g. social media platform, either referred by its brand or by its address). For a website, the decision between Product and Organization (to refer to the team or company maintaining the website) depends on the context. Documents and laws are also Products.

Other gathers mostly entities related to quantities: mentions of distances, ages, percentages, prices (including non-specific amounts of money, such as "*a loan*"), as well as currencies.

3.6. Span boundaries

Entity spans are annotated to fit constituents as closely as possible. Adpositions before or after the entity are excluded whenever not part of the entity's designation. In particular, in case of adposition-determiner contractions that are undone in the UD tokenization scheme, only the determiner is annotated as part of the entity.

Nominal and clausal modifiers are annotated as part of the entity when they are defining. Some modifiers can however be dropped when this rule would lead to an excessively long entity (e.g. over 20 tokens).

Entities can span across appositions only if not separated by commas or other punctuation (with the exception of parenthesized acronyms).

3.7. Textual noise mitigation

In case of apparent typos in the original text, the annotations are decided based on the hypothesized originally intended word. For instance, an entity Person is annotated in "*the French Ministry*

shook his hand", based on the assumption that the intended words were "*the French Minister*".
When using automatic tokenization, in case of incorrect tokenization (e.g. in French, the indefinite determiner "*des*" mistaken for an adposition-determiner contraction "*de+les*"), the annotation is positioned on all tokens that originate from the entity tokens in the untokenized text (hence including the adposition in the previous example).

4. Annotation tool

To implement the new format and guidelines proposed in §3, we additionally release an open source annotation tool.
It is a portable tool based on a lightweight Web server (written in Python with the Django framework, and no other dependency) that can be run either locally or online. The user interface is shown in Figure 3.
The tool takes one or more CoNLL-U files, optionally CoNLL-UI files with partial annotations, and creates or updates the corresponding CoNLL-UI files.
Annotation is done one sentence at a time, with easy navigation across sentences and across files to explore the in-document context of the sentence, or to revise previous annotations based on new information. Navigation utilities also enable to use the tool for pure visualization of an already annotated corpus.
Span selection is done by mouse (pointing to a single token, start then end, or end then start) and labelling is done by key presses. Incorrectly annotated entities can be discarded with a single mouse click (and immediately reannotated with a key press). Ambiguities and identification of difficult sentences to revise later (for instance when detecting a gap in the guidelines) can be traced within the tool and recorded in the annotation file with special markers.
In order to speed up annotation, the tool offers automated suggestions of entities, based on previously made annotations: if the exact same span has already been annotated as entity in the same corpus, it is suggested with the same entity type and markers (if any) and appears with pale display, in which case it can be added to the annotations with a single mouse click.
With these UI utilities and suggestions, it has been measured that a new user of the tool can annotate around 50 sentences per hour, whereas an experienced user can reach 100 to 200 sentences per hour.

5. UkraiNER

As a first corpus following our Comprehensive Entity Recognition guidelines, we release the UkraiNER corpus.[2]
UkraiNER is a corpus of news briefs in French, from the online live feed of the newspaper *Le Monde*, about the Ukrainian situation in February-March 2022. It contains 1,555 news briefs, for a total of 10,604 sentences. Annotations amount to 43,623 entities.

Data collection and pre-processing. The texts have been extracted from open access live feeds on website `lemonde.fr`,[3] corresponding to dates February 19-28 and March 27-30.
Scraping has been applied to retain only the content of the news briefs, including titles and captions, but excluding timestamps (kept however as metadata and added to the document id) and hyperlinks. Questions from readers that are inlined and answered in the live feed are also kept.
Raw texts have been pre-processed by UDPipe 1.2 (Straka and Straková, 2017), using the French GSD model from UD 2.5.

Annotation process. Annotation has been performed using our annotation tool (see §4), mostly by a single annotator. Due to budget limitations it was not possible to hire multiple annotators. Instead, it was chosen to rely on a single person to annotate the corpus (one author of this paper), and to have the other author do counter-annotations on a reduced subset of the corpus, for purposes of quality control.
The main annotator's profile is a native speaker of French between 25 and 30, well-educated (holder of a master's degree) and with a background in computational linguistics. The annotator had little prior knowledge of the geopolitical situation in Ukraine and has reported that a significant part of the corpus required general knowledge and online searches to achieve a sufficient level of understanding to analyze entities in the sentence.
A first version of the guidelines was provided to the annotator to initiate a first series of annotations (first half of February 19), after which an author performed quality control (including counter-annotations, proof-reading of randomly chosen sentences, and automated consistency checks based on scripts and queries) and all issues (including those reported by the annotator) were discussed for adjudication, and update of the guidelines if needed. This process was applied after 100, 200, 500, 1,000, 5,000 and 10,000 sentences were annotated.
After the full corpus was annotated, the main annotator revised previous annotations to account for guidelines' evolutions, for acquired experience

[2]The raw texts are freely accessible online and the annotations are released under CC BY-SA 4.0.

[3]See for instance the February 19 feed. All source URLs are available in the documentation of the corpus.

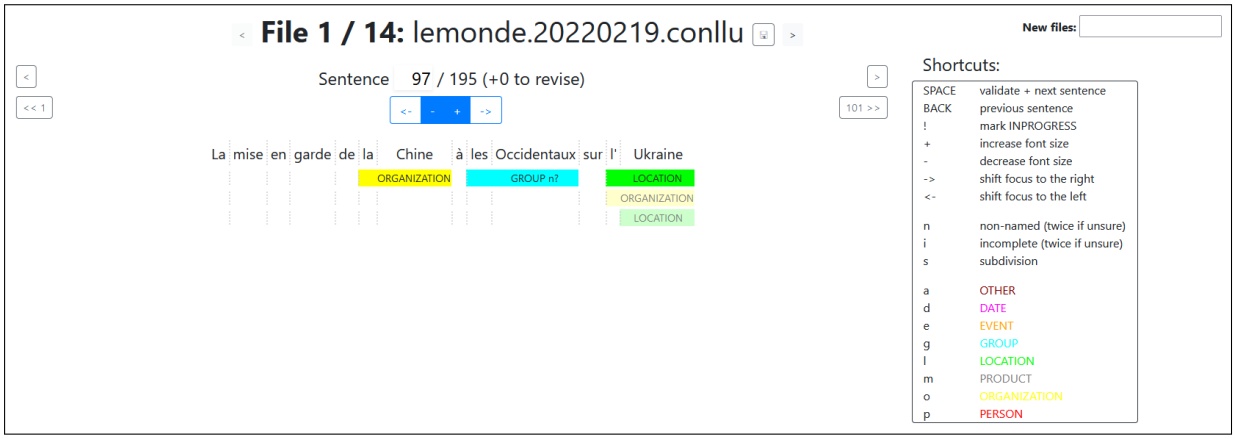

Figure 3: User interface of our annotation tool.

with the annotation tool and the guidelines, and for the initial lack of familiarity with the geopolitical context. In that process, the first 5% annotated sentences have been extensively reannotated (~20% divergences) and the first 20% have been carefully proof-read without further major change.

Total duration of the annotation process was 4 weeks (spread over 6 months) for first annotations, and 2 extra weeks for revised annotations and checks. This does not include the time spent for designing and refining the guidelines themselves.

At the end of the annotation process, the author doing quality checks reannotated the last 100 sentences of the corpus and obtained an inter-annotator agreement of 70%. Manual analysis revealed that most of the divergences pertained to span boundaries (e.g. inclusion of the determiners, of adjectival modifiers), as well as differences in interpretation leading to different choices along Group and Organization (e.g. whether "*the Russians*" referred to the government or the citizens, or similar ambiguities between the journalists and the newspaper organization). There was a limited number of omissions on both sides, and overall the named/complete/subdivision labels were consistent.

Entity statistics. Table 1 presents summary statistics on entities found in the UkraiNER corpus. Compared to traditional nested NER, our guidelines yield more than twice as many entities on average, and significantly more for some entity types.

ent. type	# ent.	%	% n	% i	length
all	43,623		44.9	1.3	2.6 $\pm$1.9
PERSON	8,373	19.2	53.2	0.7	2.6 $\pm$2.0
ORG.	10,869	24.9	71.7	0.2	2.6 $\pm$1.8
GROUP	5,978	13.7	**0.3**	**4.8**	2.6 $\pm$1.8
LOC.	8,457	19.4	59.3	1.1	2.3 $\pm$1.8
DATE	4,195	9.6	**39.9**	0.2	2.3 $\pm$1.4
EVENT	2,366	5.4	**1.7**	0.5	**4.1** $\pm$2.4
PRODUCT	3,119	7.2	**16.9**	2.0	2.7 $\pm$1.7
OTHER	266	0.6	**22.2**	2.6	2.8 $\pm$1.3

Table 1: Occurrence count and frequency of each entity type in UkraiNER, with proportions of named (n) and incomplete (i) entities, and average length of entities (in tokens). Subdivision entities have only been found for Persons (10.3% of entities), Organizations (5.8%), Products (2.2%) and Locations (0.5%), and amount for 3.7% of all entities.

6. Baseline experiments

As a first point of comparison for future research based on UkraiNER, we provide evaluation results for nested NER using a strong baseline from the literature. This set of experiments also enables further analysis of the composition and challenges of the corpus (and more generally, of our annotation scheme), including comparisons with flat NER and with narrower guidelines.

6.1. Experimental setup

Entity recognition is evaluated using precision, recall and F1-score (micro-averaged, and macro-averaged without Other), computed over the entity spans with exact match (type-sensitive).

UkraiNER data (10,604 sentences) is split into train (February 22-28 and March 27-29, 9,115 sentences), dev (February 19-21, 715 sentences) and test (March 30, 774 sentences) sets.

In order to offer quantitative insights on the gap between traditional nested NER guidelines and our Comprehensive Entity Recognition guidelines, we conduct experiments with both the full annotations and filtered annotations. In the latter, we discard all entities that are non-named or incomplete, as well as all subdivisions (but retain other nested en-

train→test	Precision	Recall	F1
CER→CER	81.7	83.0	82.3
NNER→CER	88.8	42.4	57.4
NNER→NNER	89.1	85.8	87.4

Table 2: Micro-averaged performance for CNN-NER on UkraiNER, when trained on the full annotations (CER) and on a reduced set (NNER).

tities), which yields guidelines that are close to traditional nested NER.

6.2. Models

Two models are evaluated, one designed for nested entity recognition and one for traditional NER (as a control experiment):

- **CNN-NER** (Yan et al., 2023) is a state-of-the-art model for nested NER. It extracts a feature matrix using BERT and a multi-head biaffine decoder, then predicts entities for each span using a CNN. Experiments are done using the authors' implementation,[4] only modified to support the UkraiNER corpus and the CamemBERT-base model for French (Martin et al., 2020). We train it on UkraiNER using the hyperparameters recommended by Yan et al. (2023) for the Genia experiments. All CNN-NER results are averaged over 5 runs.
- **spaCy** (Honnibal et al., 2020) is an off-the-shelf tool for flat NER. We use version 3.7.1 with model `fr_core_news_lg` (without retraining). As spaCy is trained on WikiNER and WikiNER's entity type set differs from that of UkraiNER, we map PER to Person, ORG to Organization, LOC to Location, and ignore spacy's type MISC and all other UkraiNER types.

Additional experimental results are also reported in Appendix B.

6.3. Results

Evaluation results for CNN-NER are reported in Table 2 (CER→CER). When training CNN-NER on filtered annotations, results show that Comprehensive Named Entity Recognition is a more challenging task than traditional nested NER (NNER→NNER yields higher F1), and that traditional nested NER models fail to identify more than half of UkraiNER's entities (NNER→CER has low recall), which is fully consistent with the distribution of entities in UkraiNER (see Table 1). The drop in precision between NNER→CER and CER→CER additionally suggests that the higher diversity of entity mentions in CER leads the model to overgenerate entities when trained on CER.

[4] https://github.com/yhcc/CNN_Nested_NER

	Precision	Recall	F1
PERSON	94.8	95.4	95.1
ORGANIZATION	87.3	86.3	86.8
GROUP	68.5	76.0	72.0
LOCATION	83.4	84.8	84.1
DATE	72.3	80.7	76.3
EVENT	57.4	50.4	53.6
PRODUCT	62.1	63.3	62.7
macro	75.1	76.7	75.8

Table 3: Fine-grained and macro-averaged performance for CNN-NER on UkraiNER (using the full annotations). Type OTHER is not considered here due to its scarcity in train data and its absence from test data.

	Precision	Recall	F1
PERSON	99.4	97.5	98.4
ORGANIZATION	90.0	85.0	87.4
LOCATION	76.7	81.6	79.0
DATE	96.4	81.4	88.2
PRODUCT	60.5	55.8	57.5
macro	84.6	80.2	82.1

Table 4: Fine-grained and macro-averaged performance for CNN-NER on UkraiNER (using filtered annotations). Types GROUP, EVENT and OTHER are not considered here due to their scarcity in filtered annotations.

Measuring the CER→CER recall separately on named entities (96.5) and non-named entities (68.6) additionally reveals that non-named entities are much more challenging to detect, which can again be linked with their diversity.

While performance is not comparable across datasets, across domains and across entity type sets, it remains interesting to observe that the CER→CER F1 is on par with Yan et al. (2023)'s results on the Genia dataset (~81 F1), which appears meaningful considering that part of Genia's entities (molecule names) are nominal and can thus present challenges that are similar to non-named entities, whereas the NNER→NNER F1 is closer to their results on ACE2004 and ACE2005 (~87 F1) which are much more focused on proper nouns.

Table 3 reports fine-grained results for CNN-NER, per entity type. In line with usual results in NER and nested NER, PERSON (which has the lowest diversity of mentions) is the easiest type to recognize, with ORGANIZATION and LOCATION (the two most frequent types) also well recog-

nized, while other entity types are more challenging. Macro-averaged performance is also reported for future comparison. Similar measures are performed in Table 4 for the NNER→NNER setting.
For the spaCy experiments, only the filtered annotations are evaluated on. Performance for PERSON remains high (R=82.1, P=85.2, F1=83.6), while it drops for LOCATION (R=60.0, P=37.5, F=46.2) and almost no ORGANIZATION is correctly recognized. These results are meaningful considering that spaCy performs only flat NER, and person names are rarely nested. Besides, manual analysis reveals that errors on flat locations and organizations can often be linked to annotation divergences between UkraiNER and WikiNER, such as different span boundaries (e.g. for determiners) and different criteria for considering country names as organizations or locations.

7. Conclusion

We have introduced a new annotation scheme for comprehensive entity recognition, including entities that are nested, discontinuous, and non-named: all entities of a given type are annotated, regardless of the nature of their mention.
As a first application of those guidelines, we release UkraiNER, a corpus of 10,000 French sentences annotated in comprehensive entity recognition, using our own annotation tool, which we release as well.
In addition, we have conducted a series of baseline experiments and further analysis, in order to facilitate future research using UkraiNER, but also provide some quantitative insights on the composition of the corpus and its most salient challenges.
An interesting track for future work would be to adapt and extend our guidelines to be applicable to user-generated content with a full account of their peculiarities, including unreliable punctuation and grammar, or even tweets. Such extension would indeed enable information extraction from social media, and thereby support numerous real-world use cases. As the primary motivation for comprehensiveness is to extract entities as a first step towards knowledge base extraction, another meaningful extension of the annotation scheme would be to complement it with other layers of information extraction built on top of those entities (e.g. coreference resolution, entity linking, relation extraction). This work would include making our comprehensive guidelines interoperable with related ideas that have emerged in the literature for those other tasks, such as (Rosales-Méndez, 2021) for entity linking.

8. Ethical Considerations

The annotator is a full-time employee with a multi-year contract and a salary that accounts for qualification.
The most salient ethical considerations for this work pertain to the topic addressed in the UkraiNER data (armed conflict in Ukraine). Indeed it can be morally difficult for uninformed users of the corpus to read some of the news briefs (e.g. mentions of deaths), in particular if personally affected by the situation in Ukraine or having relatives on either side of the conflict.
However, it is noteworthy that UkraiNER contains only public texts from a broad-audience newspaper, written by journalists, and it does not hold directly offending content.
The annotator has reported during the work the moral impact of working extensively on armed conflict data, but also satisfaction in gaining a better understanding of contemporary events.

9. Limitations

The main limitation of our annotation scheme, annotation tool and corpus is the limited handling of discontinuous entities, which are identified as such but not entirely annotated (not all tokens). As discontinuous entity recognition has received increased interest in the recent years, this is an important aspect to consider in future versions of the annotation scheme.
More specifically for UkraiNER, another important limitation of the corpus is the reliance on a single annotator, which renders the annotations prone to biases but has been mitigated by increasing the quality check efforts.

10. Acknowledgements

This work has been partly funded by the French Defence Innovation Agency.

11. Bibliographical References

N. Chinchor and P. Robinson. 1998. Appendix E: MUC-7 named entity task definition (version 3.5). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Yoann Dupont. 2019. Un corpus libre, évolutif et versionné en entités nommées du français (a free, evolving and versioned french named

entity recognition corpus). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 437–446, Toulouse, France. ATALA.

Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-François Bonastre, Djamel Mostefa, and Khalid Choukri. 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*, pages 139–142. Citeseer.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Loïc Grobol. 2020. *Coreference resolution for spoken French*. Ph.D. thesis, Université Sorbonne Nouvelle-Paris 3.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

William Lechelle, Fabrizio Gotti, and Phillippe Langlais. 2019. WiRe57 : A fine-grained benchmark for open information extraction. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 6–15, Florence, Italy. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. 2020. Establishing a new state-of-the-art for French named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4631–4638, Marseille, France. European Language Resources Association.

Pierre-Henri Paris and Fabian Suchanek. 2021. Non-named entities–the silent majority. In *The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers 18*, pages 131–135. Springer.

Henry Rosales-Méndez. 2021. *Towards a fine-grained entity linking approach*. Ph.D. thesis, PhD thesis, Universidad de Chile.

Henry Rosales-Méndez, Bárbara Poblete Labra, and Aidan Hogan. 2018. What should entity linking link?

Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. 2011. *Entités nommées structurées: guide d'annotation Quaero [in French]*. LIMSI-Centre national de la recherche scientifique.

Benoît Sagot, Marion Richard, and Rosa Stern. 2012. Annotation référentielle du corpus arboré de Paris 7 en entités nommées (referential named entity annotation of the Paris 7 French TreeBank) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 535–542, Grenoble, France. ATALA/AFCP.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. 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 2782–2794, Online. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. 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.

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.

Sylvain Verdy, Maxime Prieur, Guillaume Gadek, and Cédric Lopez. 2023. DWIE-FR : Un nouveau jeu de données en français annoté en entités nommées. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts*, pages 63–72, Paris, France. ATALA.

Hang Yan, Yu Sun, Xiaonan Li, and Xipeng Qiu. 2023. An embarrassingly easy but strong baseline for nested named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1442–1452, Toronto, Canada. Association for Computational Linguistics.

Zheng Yuan, Chuanqi Tan, Songfang Huang, and Fei Huang. 2022. Fusing heterogeneous factors with triaffine mechanism for nested named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3174–3186, Dublin, Ireland. Association for Computational Linguistics.

12. Language Resource References

Alexis Mitchell, Stephanie Strassel, Shudong Huang, Ramez Zakhary. 2004. *ACE 2004 Multilingual Training Corpus*. Linguistic Data Consortium, LDC corpora, ISLRN 789-870-824-708-5.

Christopher Walker, Stephanie Strassel, Julie Medero, Kazuaki Maeda. 2005. *ACE 2005 Multilingual Training Corpus*. Linguistic Data Consortium, LDC corpora, ISLRN 458-031-085-383-4.

Galliano, Sylvain and Geoffrois, Edouard and Gravier, Guillaume and Bonastre, Jean-François and Mostefa, Djamel and Choukri, Khalid. 2006. *ESTER Evaluation Package*. ELRA, ISLRN 110-079-844-983-7.

Grouin, Cyril and Rosset, Sophie and Zweigenbaum, Pierre and Fort, Karën and Galibert, Olivier and Quintard, Ludovic. 2013. *Quaero Broadcast News Extended Named Entity corpus*. ELRA, ISLRN 074-668-446-920-0.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, Ann Houston. 2013. *OntoNotes Release 5.0*. Linguistic Data Consortium, ISLRN 458-031-085-383-4.

A. Appendix: Examples of improved comprehensiveness

In the example of Figure 1, our proposed guidelines for comprehensive entity recognition mandate the extraction of the entity "*students*", whereas traditional nested NER would not, both due to not covering entity type Group, and to the mention being non-named. However, this nominal mention is not about any group of students, but that particular one. Having first recognized the entity thus makes it possible to later extract the size of that group, its location, members (e.g. modelled as a Group-Person relationship), events where that particular group is an agent, etc. This motivation is also independent of the practical form of that entity mention, and extracting "*students*" is not less relevant than if the mention were "*class of '21*" instead. The case of that entity illustrates how our proposed guidelines lay the foundations for more comprehensive information extraction.

In the following, we further illustrate, through example sentences inspired from the UkraiNER texts, the practical benefits of the CER guidelines on the comprehensiveness of the conveyed information – or at least the ability to faithfully convey

the information embedded by the journalist in the text.

For instance, the sentence "*The Ukrainian President was in Munich to meet Western leaders*" yields a similar configuration to "*students*" above: the Group entity first needs to be extracted before considering extracting further information on those leaders (their names, countries, possibly the relationship that links those particular countries, that those individuals were present in Munich on that day, etc.).

In "*The Ukrainian army has announced yesterday the death of a second soldier*", the Event entity and its nested Person entity provide the key information that there was not only one death, but also another one prior to that event. In addition, detecting that Person entity (even if non-named) enables to start collecting information in search for that person's name (and the other one's). Similar motivation applies to Groups, as in "*The attack left four wounded on the Ukrainian side*", where recognizing the (non-named) Group is a necessary step before identifying the members of that group – hence the actually wounded persons.

Considering sentence "*Volodymyr Zelensky demands from NATO a clear timeframe for his country's entry into the organization*", not annotating the entities "*his country*" and "*the organization*" (as done in traditional NER) would prevent the later extraction of the relation that is the key information in that statement: the (desired) entry.

In "'*This is nonsense' said U.S. Ambassador to the United Nations Linda Thomas-Greenfield*", the subdivision marker encodes the useful information that this Ambassador role is currently held by Linda Thomas-Greenfield, while the nested entities are necessary to track who that person represents and to whom, hence the practical meaning of that role (and on whose behalf that statement is made).

Finally, in the sentence "*Several EU countries are highly dependent on Russian oil and gas*", marking entity "*gas*" as incomplete allows to retain the information that this dependence is not on gas in general, but specifically the Russian one (e.g. due to existing infrastructure), and this difference has in turn a direct impact on the trade options available to those countries.

B. Appendix: Additional experimental results

As a complement to the baseline results in §6 we report additional results with a nested NER model based on a significantly different approach.

Locate-and-Label (Shen et al., 2021) is a two-stage method that first detects entities then adjusts their exact boundaries. We adapt the authors' implementation[5] to use CamemBERT-base and the French fastText embeddings built from CommonCrawl and Wikipedia (Grave et al., 2018). We run the experiments with the default hyperparameters from Shen et al. (2021)'s ACE05 experiments, then we vary some hyperparameters (maximum epochs increased from 35 to 100, learning rate increased from 3e-5 to 6e-5 and 1e-4).

Micro-averaged results are reported in Table 5. While significantly lower (~5 F1) than the CNN-NER results, the performance on UkraiNER also appears to be highly impacted by hyperparameters. Fine-grained measured as reported in Table 6 reveal that a substantial part of the losses can be attributed to the PERSON class (~15 F1 lower than CNN-NER). However, those results are provided as is for referencing purposes, and we do not apply extensive hyperparameter tuning (or any other model adaptation to better handle a class like PERSON as it appears in UkraiNER) which would be beyond the scope of this work.

Hyperparameters	Precision	Recall	F1
lr = 3e-5, n_{ep} = 35	75.3	72.7	74.0
lr = 3e-5, n_{ep} = 100	78.8	70.5	74.4
lr = 6e-5, n_{ep} = 35	78.7	71.5	74.8
lr = 1e-4, n_{ep} = 35	79.8	71.8	75.6

Table 5: Micro-averaged performance of Locate-and-Label on UkraiNER (full CER annotations).

	Precision	Recall	F1
PERSON	85.2	76.9	80.8
ORGANIZATION	81.3	77.5	79.4
GROUP	67.6	72.4	69.9
LOCATION	80.6	77.1	78.8
DATE	71.0	66.2	68.5
EVENT	63.8	63.1	63.5
PRODUCT	46.1	47.3	46.7
macro	70.8	68.6	69.7

Table 6: Fine-grained and macro-averaged performance for Locate-and-Label on UkraiNER (using the full annotations), with default hyperparameters. Type OTHER is not considered here due to its scarcity in train data and its absence from test data.

[5] https://github.com/tricktreat/locate-and-label