Abstract

Named Entity Recognition (NER) is the task of identifying named entities in texts and classifying them through specific semantic categories, a process which is crucial for a wide range of NLP applications. Current datasets for NER focus mainly on coarse-grained entity types, tend to consider a single textual genre and to cover a narrow set of languages, thus limiting the general applicability of NER systems. In this work, we design a new methodology for automatically producing NER annotations, and address the aforementioned limitations by introducing a novel dataset that covers 10 languages, 15 NER categories and 2 textual genres. We also introduce a manually-annotated test set, and extensively evaluate the quality of our novel dataset on both this new test set and standard benchmarks for NER. In addition, in our dataset, we include: i) disambiguation information to enable the development of multilingual entity linking systems, and ii) image URLs to encourage the creation of multimodal systems. We release our dataset at https://github.com/Babelscape/multinerd.

1 Introduction

Named Entity Recognition (NER) represents a milestone in information extraction, and its aim is to identify and classify key information in unstructured texts, i.e. named entities (Nadeau and Sekine, 2007). It is widely used in a broad spectrum of downstream applications, like machine translation (Babych and Hartley, 2003), question answering (Mollá et al., 2006), automatic text summarization (Aone et al., 1998), and entity linking (Martins et al., 2019), inter alia.

With the advent of pretrained language models like BERT (Devlin et al., 2019) or LUKE (Yamada et al., 2020) – this latter with a particular focus on named entities – the NER field observed astonishing results on conventional benchmarks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). However, such benchmarks are limited in size, cover a single textual genre, and are available only for a narrow set of languages. Moreover, they focus on coarse-grained entity types, and often overlook more complex entities like titles of books, songs and movies. These latter entities are not simple nouns and can be both syntactically and semantically ambiguous. Specifically, they can assume the form of any linguistic constituent (e.g. Singin’ in the Rain) which makes them difficult to extract. Indeed, in the last decade, the OntoNotes 5.0 dataset (Weischedel et al., 2013) has become very popular thanks to its high quality, size and fine-grained categories. Nevertheless, it covers only 3 languages, namely, English, Arabic and Chinese.

Since the manual creation of training data for NER is expensive and time-consuming – especially when many languages have to be covered – several studies have tried to address data scarcity by producing training data automatically (Nothman et al., 2013; Al-Rfou et al., 2015; Tsai et al., 2016; Pan et al., 2017), recently showing that automatically-generated annotations can boast a quality comparable to that of manually-created ones (Tedeschi et al., 2021b). Unfortunately, although these studies have considered a wider range of languages, they have still focused on coarse-grained entities and on a single textual genre, i.e. encyclopedic texts from Wikipedia\(^1\) (Hovy et al., 2013).

In this paper, inspired by the success of the OntoNotes 5.0 dataset and by recent achievements in automatic data creation, we will address the aforementioned gaps and propose the following novel contributions:

1. We design a new language-agnostic methodology for automatically generating high-quality and fine-grained NER annotations by exploiting the texts of Wikipedia\(^2\);

\(^1\)https://en.wikipedia.org/
\(^2\)It is a free-content multilingual wiki containing news articles (https://en.wikinews.org/), as opposed to the encyclopedic articles contained in Wikipedia.
2. We introduce a novel automatically-created benchmark for NER that covers 10 languages, 15 entity types and 2 textual genres, together with a small manually-curated test set for the English language;

3. We extensively evaluate the quality of the data produced on both our manually-annotated test set and standard benchmarks for NER.

Additionally, although in this work we focus on NER, we also contribute to the entity disambiguation (also known as entity linking) task, i.e. the task of linking entities mentioned in texts with their corresponding entry in a knowledge base. Specifically, for a given entity, we provide disambiguation information together with its NER tag in order to enable training, validation and testing of multilingual entity linking models. Finally, we also include image URLs to encourage the creation of multimodal systems. To enable comparability on our benchmark, we release our data and software at https://github.com/Babelscape/multinerd.

2 Related Work

2.1 Gold-Standard Data

High-quality annotations are essential for both learning and evaluation of NER systems. Indeed, in the past few decades a large number of NER datasets have been proposed. Initially, the MUC-6 and MUC-7 shared tasks focused on entity names (i.e. persons, locations and organizations), temporal expressions (i.e. dates and times) and number expressions (i.e., currency values and percentages), but only English newswire articles were considered (Grishman and Sundheim, 1996; Chinchor and Robinson, 1997). A few years later, different datasets were derived from Reuters News for the CoNLL-2002 and 2003 shared tasks on language-independent Named Entity Recognition (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), covering four different languages (i.e. Dutch, English, German and Spanish). However, these datasets were limited in size and only four coarse-grained entity types were considered: Person, Location, Organization and Miscellaneous3. Nonetheless, these datasets are still widely used to benchmark NER systems.

Balasuriya et al. (2009) claimed that NER was needed in many domains beyond newswire texts, and introduced WikiGold, a manually-annotated dataset derived from Wikipedia articles. Even so, WikiGold covered coarse-grained entities, was limited in size and considered only the English language. Following the same motivation as Balasuriya et al. (2009), Ritter et al. (2011) introduced a dataset for English tweets using 10 NER classes.

Another considerable step forward was made by Weischedel et al. (2013), who introduced OntoNotes 5.0. This dataset covered 18 fine-grained classes, multiple genres (e.g. newswire and weblogs), and multiple languages (English, Chinese, and Arabic). Thanks to its high quality, it is one of the most widely used datasets for NER.

Finally, another notable dataset was proposed for the WNUT 2017 shared task on emerging and rare entities, covering different textual genres (tweets, YouTube comments, Reddit and StackExchange posts) (Derczynski et al., 2017). However, only the English language and 6 categories were considered.

2.2 Silver-Standard Data

Although the OntoNotes 5.0 dataset constitutes a valuable resource for training and evaluating multilingual and fine-grained NER systems, its applicability is limited to the three languages it covers. Indeed, in the last decade, with the aim of scaling NER to a wider set of languages, more interest has been devoted to automatic data creation.

The first successful attempt in this direction was made by Nothman et al. (2013) who produced the WikiNER dataset. They proposed a strategy for automatically creating multilingual training data for NER by exploiting the texts of Wikipedia and its hypertext organization. In addition, they also used redirect-base heuristics to infer more named-entity mentions. By applying this methodology, they covered 9 languages, but they still focused on the standard coarse-grained entity types.

Adopting a similar strategy, Pan et al. (2017) introduced WikiANN, a language-independent framework for extracting entities from documents. Their procedure was made up of two main steps: i) classify entries in the English Wikipedia into specific entity types, and ii) propagate the annotations to other languages by applying cross-lingual transfer. This procedure yielded massive corpora consisting of 282 languages, but with lower annotation quality and a focus on persons, locations, organizations and geo-political entities.

Finally, Tedeschi et al. (2021b) proposed

3Miscellaneous entities are entities that do not belong to the Person, Location and Organization categories.
WikiNEuRal, an annotation pipeline that effectively combined recent pretrained language models with knowledge-based approaches, and produced high-quality annotations for NER in 9 languages by exploiting Wikipedia texts. Surprisingly, the authors showed that their methodology, in 2 out of 3 settings, produced annotations with a quality even higher than that of manual ones. However, again, only coarse-grained entities were considered.

Despite automatic methods achieved high annotation quality and covered many languages, all of them focused on coarse-grained entities and on a single textual source: Wikipedia. On the other hand, gold-standard datasets focused mainly on the English language. Additionally, none of them included disambiguation information. Evidently, a unified effort to obtain a large-scale multilingual, multi-genre and fine-grained resource for Named Entity Recognition and Disambiguation is still missing.

3 NER Classes

Our 15 NER classes are a subset of the newly introduced 18 classes of Tedeschi et al. (2021a) designed to reduce the intrinsic sparsity of the Entity Linking task. Specifically, they are: Person (PER), Location (LOC), Organization (ORG), Animal (ANIM), Biological entity (BIO), Celestial Body (CEL), Disease (DIS), Event (EVE), Food (FOOD), Instrument (INST), Media (MEDIA), Plant (PLANT), Mythological entity (MYTH), Time (TIME) and Vehicle (VEHI).

We prefer these classes to the OntoNotes ones, because they cover a wider range of macro categories. For instance, the OntoNotes’ PRODUCT class – which groups very heterogeneous entities – is split into FOOD, VEHI and INST. Over and beyond these, our new set contains animals, plants, biological entities, celestial bodies, diseases and mythological entities that are not present in OntoNotes. Table 1 provides textual descriptions and examples of instances concerning our proposed classes. Further details about NER classes are provided in Appendix A.

4 MultiNERD

In this Section we describe our language-agnostic strategy for automatically generating a fine-grained and multilingual resource to train robust NER and ED systems. Specifically, our methodology widely extends previous state-of-the-art strategies, and it is characterized by the following five steps: i) preprocessing of Wikipedia and Wikinews articles (Section 4.1), ii) identification of entities (Section 4.2), iii) tagging the identified entities with the NER labels (Section 4.3), iv) propagation of the annotations (Section 4.3), v) enhancement of the annotations (Section 4.4).

4.1 Wikitext Preprocessing

Wikipedia and Wikinews articles provide plenty of manually-curated information that can be exploited for the automatic annotation of sentences, i.e. Wikilinks4. However, in addition to Wikilinks, articles may contain elements (e.g. images, tables, formulas and lists) and sections (e.g. see also, references, further readings) that do not correspond to well structured text; therefore, we remove them with the intent of reducing noise. This step converts articles to plain texts containing only Wikilinks.

4.2 Entity or Concept?

Wikilinks provide potential entity mentions. Indeed, some of them correspond to entities (e.g. Elon Musk), and others correspond to concepts (e.g. Table). In order to distinguish between them, we take advantage of the one-to-one linkage between Wikipedia and BabelNet5 (Navigli and Ponzetto, 2012; Navigli et al., 2021) and exploit the concept-vs.-entity categorization provided therein. Although it is evident that for some categories we are only interested in entities in the strictest sense (e.g. PER, ORG and LOC), we need to relax this constraint for other classes (e.g. ANIM, PLANT, FOOD and DIS). Thus, in order to extract animals (e.g. Labrador Retriever), plants (e.g. Pinus), food (e.g. Carbonara) and diseases (e.g. Alzheimer’s disease), among others, we also need to consider elements that are labeled as concepts in BabelNet. This step tells us which Wikilinks have to be annotated with an entity type, and which of them have to be discarded. The full list of design choices is provided in Appendix A.

4.3 Tagging Wikipedia and Wikinews Articles

Semantic Classifier We now aim at providing each (remaining) Wikilink in a Wikipedia (or Wikinews) article with a category $c \in C$, where $C$ 

---

4A Wikilink is a link placed inside an article that points to a Wikipedia page.

5BabelNet (https://babelnet.org) is a wide multilingual semantic network that integrates both lexicographic and encyclopedic knowledge from different sources. We use BabelNet 5.0.
is the set of the NER classes introduced in Section 3. To do this, we introduce a Semantic Classifier that exploits the one-to-one correspondence between Wikipedia pages and BabelNet synsets. We start by manually-annotating 300 synsets to cover as many high-order concepts of the WordNet⁶ (Miller, 1995) nominal taxonomy – which is a subset of the BabelNet taxonomy – as possible. For instance, we label the following high-level synsets as follows:

- animal (bn:00004222n) → ANIM;
- company (bn:00021286n) → ORG;
- town (bn:00077773n) → LOC;

Then, to propagate these annotations to all other synsets in WordNet, we descend through its taxonomy by following hyponymy and has-instance relationships (i.e. parent-to-child relations). For example, all the children of animal (bn:00004222n), e.g. dog (bn:00015267n), inherit the ANIM tag. This step yields 40k high-quality annotated synsets.

At this point, to annotate a Wikilink l in a Wikipedia (or Wikinews) article w, we retrieve its corresponding synset s from BabelNet, and we follow hypernymy relations (child-to-parent relations) until one or more of the 40k synsets in the expanded set is reached⁷. Here, we distinguish between two possible cases:

1. When a single ancestor is reached, or when all the ancestors share the same class, the corresponding annotation is just inherited. For instance, starting from Apple Inc. (bn:03739345n) and climbing the taxonomy, we find only company (bn:00021286n) at distance 1, hence Apple Inc. inherits the ORG annotation.

2. When two or more ancestors have discordant annotations, then the highest-scoring class⁸.

---

Table 1: Descriptions and instance examples of our NER classes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
<td>People,</td>
<td>Ray Charles, Jessica Alba, Leonardo DiCaprio, Roger Federer, Anna Massey.</td>
</tr>
<tr>
<td>ORG</td>
<td>Associations, companies, agencies, institutions, nationalities and religious or political groups.</td>
<td>University of Edinburgh, San Francisco Giants, Google, Democratic Party.</td>
</tr>
<tr>
<td>LOC</td>
<td>Physical locations (e.g. mountains, bodies of water), geopolitical entities (e.g. cities, states), and facilities (e.g. bridges, buildings, airports).</td>
<td>Rome, Lake Paiku, Chrysler Building, Mount Rushmore, Mississippi River.</td>
</tr>
<tr>
<td>ANIM</td>
<td>Breeds of dogs, cats and other animals, including their scientific names.</td>
<td>Maine Coon, African Wild Dog, Great White Shark, New Zealand Bellbird.</td>
</tr>
<tr>
<td>BIO</td>
<td>Genus of fungus, bacteria and protocists, families of viruses, and other biological entities.</td>
<td>Herpes Simplex Virus, Escherichia Coli, Salmonella, Bacillus Anthracis.</td>
</tr>
<tr>
<td>DIS</td>
<td>Physical, mental, infectious, non-infectious, deficiency, inherited, degenerative, social and self-inflicted diseases.</td>
<td>Alzheimer’s Disease, Cystic Fibrosis, Dilated Cardiomyopathy, Arthritis.</td>
</tr>
<tr>
<td>EVE</td>
<td>Sport events, battles, wars and other events.</td>
<td>American Civil War, 2003 Wimbledon Championships, Cannes Film Festival.</td>
</tr>
<tr>
<td>INST</td>
<td>Technological instruments, mechanical instruments, musical instruments, and other tools.</td>
<td>Spitzer Space Telescope, Commodore 64, Skype, Apple Watch, Fender Stratocaster.</td>
</tr>
<tr>
<td>MEDIA</td>
<td>Titles of films, books, magazines, songs and albums, fictional characters and languages.</td>
<td>Forbes, American Psycho, Kiss Me Once, Twin Peaks, Disney Adventures.</td>
</tr>
<tr>
<td>PLANT</td>
<td>Types of trees, flowers, and other plants, including their scientific names.</td>
<td>Salix, Quercus Petraea, Douglas Fir, Forsythia, Artemisia Maritima.</td>
</tr>
<tr>
<td>MYTH</td>
<td>Mythological and religious entities.</td>
<td>Apollo, Persephone, Aphrodite, Saint Peter, Pope Gregory I, Hercules.</td>
</tr>
<tr>
<td>TIME</td>
<td>Specific and well-defined time intervals, such as eras, historical periods, centuries, years and important days. No months and days of the week.</td>
<td>Renaissance, Middle Ages, Christmas, Great Depression, 17th Century, 2012.</td>
</tr>
</tbody>
</table>

⁶We start from WordNet synsets because they are manually curated.

⁷We employ a Breadth-First Search with max depth = 2.

⁸If more classes have the same score, the corresponding synset is not annotated.
is assigned. Formally, for each NER class $c \in C$, the score is computed as follows:

$$
score(c) = \sum_{a \in A_c} \frac{1}{d(a)},$$

(1)

where $A_c$ is the set of all the ancestors of synset $s$ with tag $c$, and $d(a)$ is a function that returns the distance of $a$ from $s$ in the BabelNet taxonomy. As an example, consider the synset Bill Gates (bn:00010401n). When climbing the taxonomy, at distance 2, the hominid (bn:00044571n) synset is reached, which is classified as ANIM. However, the Bill Gates’ synset is also child of human (bn:00044576n), computer scientist (bn:00021495n) and magnate (bn:00008639n), resulting in a highest score for the PER class.

This procedure allows us to label each Wikilink in a given article with a NER class.

**Tag Propagation** Wikipedia and Wikinews guidelines specify that only the first mention to a certain article has to be linked. This implies that tagging only Wikilinks leads to sparse annotations. To cope with this issue, we employ a simple yet effective exact-match heuristic in which for each Wikilink $l$, with an associated class $c$, we assign the class $c$ to all the expressions $e_i$ in the same document of $l$ such that $e_i = l \lor e_i \in \text{syn}(l)$, where $\text{syn}(l)$ is a function that returns the synonyms of $l$ from BabelNet. Finally, the annotations are converted to BIO format\(^9\).

The above-described methodology allows us to have at the same time, for a given entity, both the NER annotation and the BabelNet synset. Then, as already mentioned, through BabelNet we can easily access other resources, and retrieve the corresponding Wikipedia and Wikidata pages. Hence, in our dataset we include disambiguation information from the three above-mentioned knowledge bases. Additionally, for a given entity, we include also the corresponding definition and the main image from Wikidata, where this latter can be used to develop multimodal NER and entity linking systems. An instance of our dataset is provided in Table 2.

### Algorithm 1 Self-Improvement Algorithm

**Inputs:** Corpus of raw Wikipedia and Wikinews documents $W$

**Parameters:** Integer $n$, Integer $t$

**Output:** MultiNERD Dataset $D$

1. $A \leftarrow \{w_1, \ldots, w_n\}$, $w_i \in W$
2. $D \leftarrow annotate(A)$
3. $M_D \leftarrow train(D)$
4. for $i \leftarrow 1, \ldots, t$ do
5. $A \leftarrow \{w'_1, \ldots, w'_n\}$, $w'_i \in W$, $A \cap \hat{A} = \emptyset$
6. $A = A \cup \hat{A}$
7. $D = annotate(A, M_D)$
8. $M_D = train(D)$
9. end for
10. return $D$

\(^9\)The BIO tagging scheme (short for Beginning, Intermediate, Out) is a popular format for handling spans of tokens.

<table>
<thead>
<tr>
<th>Token</th>
<th>Label</th>
<th>BabelNet ID</th>
<th>Wikidata ID</th>
<th>Wikipedia ID</th>
<th>Wikipedia Title</th>
<th>Definition</th>
<th>Image URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>O</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>George</td>
<td>B-PER</td>
<td>bn:000040218n</td>
<td>Q2643</td>
<td>12731</td>
<td>George Harrison</td>
<td>English rock star...</td>
<td>george_harrison.jpg</td>
</tr>
<tr>
<td>Harrison</td>
<td>I-PER</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>nor</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ringo</td>
<td>B-PER</td>
<td>bn:00067735n</td>
<td>Q2632</td>
<td>25832</td>
<td>Ringo Starr</td>
<td>Rock star and drummer...</td>
<td>ringo_starr.jpg</td>
</tr>
<tr>
<td>Starr</td>
<td>I-PER</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Example of sentence in our MultiNERD dataset. Together with the NER type (in BIO format), each entity is provided with the following additional information: BabelNet synset, Wikidata ID, Wikipedia ID, Wikipedia title, definition and image url.
In the annotate \((A, M_D)\) function, the neural model \(M_D\) is used to refine the annotations and reduce noise. Specifically, if the NER class of an entity predicted by the neural model is different from the one assigned by the knowledge-based approach (Section 4.3), the corresponding sentence is discarded.

### 5 Experiments

In this Section, we describe our experimental setup (Section 5.1), the datasets\(^{10}\) we use to train (Section 5.2) and evaluate (Section 5.3) our methodology, and finally the results obtained (Section 5.4).

#### 5.1 Experimental Setup

In our experiments, we evaluate the quality of our data in two different settings:

1. In order to compare our dataset against previous state-of-the-art automatically-created datasets, we map our fine-grained annotations to the coarse-grained classes used by these datasets. Then, we train the mBERT + Bi-LSTM + CRF model introduced in Section 4.4 on both our dataset and the other above-mentioned datasets. Finally, we compare the performance of the corresponding systems on gold-standard benchmarks for NER. Appendix A provides class mapping details;

2. To measure the quality of our fine-grained annotations, we manually annotate a random

\(^{10}\)All the datasets used are freely available for research purposes, except for OntoNotes 5.0 (LDC license).
Table 4: Span-based micro F₁ scores obtained by training a reference NER system on different automatically-created datasets (i.e. WikiANN, WikiNER, WikiNEuRal, MultiNERD) and testing on common NER benchmarks.

<table>
<thead>
<tr>
<th></th>
<th>CoNLL EN</th>
<th>CoNLL ES</th>
<th>CoNLL NL</th>
<th>CoNLL DE</th>
<th>WikiGold EN</th>
<th>WikiGold EN</th>
<th>WikiGold RU</th>
<th>WikiGold PL</th>
<th>OntoNotes EN</th>
<th>OntoNotes EN</th>
<th>OntoNotes RU</th>
<th>OntoNotes PL</th>
<th>BSNLP EN</th>
<th>BSNLP PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiANN</td>
<td>56.85</td>
<td>53.55</td>
<td>55.76</td>
<td>44.39</td>
<td>57.05</td>
<td>36.43</td>
<td>51.85</td>
<td>53.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WikiNER</td>
<td>73.05</td>
<td>75.07</td>
<td>74.75</td>
<td>64.03</td>
<td>81.98</td>
<td>71.16</td>
<td>65.99</td>
<td>62.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WikiNEuRAL</td>
<td>76.94</td>
<td>77.87</td>
<td>77.40</td>
<td>64.02</td>
<td>82.42</td>
<td>71.98</td>
<td>66.50</td>
<td>62.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MultiNERD w/o SELF-IMPROVEMENT</td>
<td>69.75</td>
<td>71.04</td>
<td>70.58</td>
<td>57.99</td>
<td>76.88</td>
<td>65.12</td>
<td>60.41</td>
<td>58.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MultiNERD (Our Work)</td>
<td><strong>77.11</strong></td>
<td><strong>78.20</strong></td>
<td><strong>77.84</strong></td>
<td><strong>65.22</strong></td>
<td><strong>83.11</strong></td>
<td><strong>72.45</strong></td>
<td><strong>67.39</strong></td>
<td><strong>62.94</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sample of 1K English sentences, and compare the annotations produced by our methodology with the corresponding ground truths.

We implement our model with PyTorch using the Transformers library (Wolf et al., 2019) to load the weights of BERT-base-multilingual-cased (mBERT), and train each model configuration with an early stopping strategy using a patience value of 5. We use Adam optimizer (Kingma and Ba, 2015) with learning rate of $10^{-3}$ and a cross-entropy loss criterion. We repeat each training on 10 different seeds, fixed across experiments, and report the mean of their span F₁ scores computed with the official conlleval script. Further details are provided in Appendix B.

5.2 Training Data

We train our reference model with four different silver-standard datasets:

- **MultiNERD**: the resource created using the steps described in Section 4 from Wikipedia and Wikinews articles with $n = 30K$ and $t = 8$. It covers 10 languages: Chinese, Dutch, English, French, German, Italian, Polish, Portuguese, Russian, and Spanish. Statistics are shown in Table 3.

- **WikiNEuRal** (Tedeschi et al., 2021b): the current best-performing approach for NER silver data creation. It covers 9 languages (i.e. Dutch, English, French, German, Italian, Polish, Portuguese, Russian, and Spanish), and sentences are extracted from Wikipedia.

- **WikiNER** (Nothman et al., 2013): a high-quality automatically-derived dataset for NER from Wikipedia. It covers the same languages as WikiNEuRal.

- **WikiANN** (Pan et al., 2017): a massive dataset for NER consisting of Wikipedia documents annotated in 282 languages.

All datasets are tagged with the four standard entity types (PER, ORG, LOC, MISC), except for WikiANN which does not contain the MISC label. Indeed, when evaluating the WikiANN dataset, only the PER, ORG and LOC classes are evaluated.

5.3 Test Data

Common Benchmarks For our first setting (Section 5.1), we use 5 common gold-standard test sets:


- **WikiGold** (Balasuriya et al., 2009): a set of human-annotated English Wikipedia articles.

- **OntoNotes 5.0** (Pradhan et al., 2012): a popular dataset for NER that includes texts from different textual genres and multiple domains.


All datasets use coarse-grained entity types (i.e. PER, ORG, LOC and MISC), except OntoNotes, which uses fine-grained entities that we convert to coarse-grained classes. Appendix A provides additional details about this mapping. For the WikiGold dataset, which is limited in size, we use the entire dataset as test material. For other datasets, instead, we use the official validation and test splits. Finally, all datasets are converted to BIO format.

---

\(^{12}\)The WikiANN version used in this study corresponds to the dataset of Rahimi et al. (2019), which supports 176 of the 282 languages from the original WikiANN corpus (https://huggingface.co/datasets/wikiann).
Manual Annotation  For our second experimental setting (Section 5.1), due to the absence of NER benchmarks that use our set of categories (Section 3), we conduct a manual evaluation to assess the quality of our dataset. Specifically, we randomly select\(^{13}\) a sample of 1K English sentences, pre-annotated with the NER tags produced using our methodology (Section 4), and confirm or replace the annotations associated with each token in the dataset. The resulting gold-standard dataset is used to analyze the quality of our silver-standard data.

5.4 Results

Coarse-Grained Evaluation  In our first setting (Section 5.1), we measure the effectiveness of our methodology by comparing the quality of the data produced against that of other datasets created using previous state-of-the-art strategies for NER silver-data creation (i.e. datasets listed in Section 5.2). Since past approaches focused on coarse-grained entities, we can compare the quality only for such entity types. The results obtained are shown in Table 4. Although our dataset covers a wider range of categories than its competitors, it nevertheless outperforms all of them on all tested datasets and languages. We attribute this advancement mainly to the self-improvement algorithm introduced in Section 4.4, which iteratively refines the annotations using a better model at each iteration. To demonstrate the impact of our algorithm, we construct baseline versions of MULTI NERD for

\(13\)We ensure that the dataset contains a sufficient number \(n\) of instances for each NER class. We set \(n = 20\). Statistics are provided in the “Support” column of Table 5.

<table>
<thead>
<tr>
<th>Class</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANIM</td>
<td>0.72</td>
<td>0.52</td>
<td>0.60</td>
<td>60</td>
</tr>
<tr>
<td>BIO</td>
<td>0.78</td>
<td>0.66</td>
<td>0.71</td>
<td>32</td>
</tr>
<tr>
<td>CEL</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>25</td>
</tr>
<tr>
<td>DIS</td>
<td>0.99</td>
<td>0.88</td>
<td>0.93</td>
<td>78</td>
</tr>
<tr>
<td>EVE</td>
<td>0.96</td>
<td>0.93</td>
<td>0.94</td>
<td>27</td>
</tr>
<tr>
<td>FOOD</td>
<td>0.83</td>
<td>0.44</td>
<td>0.58</td>
<td>122</td>
</tr>
<tr>
<td>INST</td>
<td>0.90</td>
<td>0.53</td>
<td>0.67</td>
<td>34</td>
</tr>
<tr>
<td>LOC</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>262</td>
</tr>
<tr>
<td>MEDIA</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>41</td>
</tr>
<tr>
<td>O</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>11823</td>
</tr>
<tr>
<td>ORG</td>
<td>0.97</td>
<td>0.95</td>
<td>0.96</td>
<td>59</td>
</tr>
<tr>
<td>PER</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>217</td>
</tr>
<tr>
<td>PLANT</td>
<td>0.73</td>
<td>0.40</td>
<td>0.52</td>
<td>47</td>
</tr>
<tr>
<td>MYTH</td>
<td>0.88</td>
<td>0.96</td>
<td>0.92</td>
<td>23</td>
</tr>
<tr>
<td>TIME</td>
<td>0.99</td>
<td>0.96</td>
<td>0.98</td>
<td>171</td>
</tr>
<tr>
<td>VEH1</td>
<td>0.82</td>
<td>0.85</td>
<td>0.84</td>
<td>27</td>
</tr>
<tr>
<td>ALL</td>
<td>0.90</td>
<td>0.81</td>
<td>0.85</td>
<td>13048</td>
</tr>
</tbody>
</table>

Table 5: Evaluation of the MultiNERD dataset on our manually-annotated English test set.
DE, EN, ES, NL, PL and RU with the same sizes as the corresponding refined versions, but without using our enhancement procedure. As can be observed from Table 4, the refined versions provide an average improvement of almost 7 F1 points. In addition, the wider number of textual genres covered by MultiNERD leads to more robust systems.

Fine-Grained Evaluation Although the coarse-grained evaluation conducted in the previous Section demonstrated that our MultiNERD methodology creates high-quality annotations, independently of the language, it is not sufficient to understand how our annotation pipeline performs on fine-grained classes. Indeed, to measure this, we use a sample of 1K English sentences manually-annotated with fine-grained entities, as explained in Section 5.3, and report the results in Table 5. As expected, the PER, ORG and LOC classes are among the best-performing classes. Similarly, celestial bodies, diseases, events and media also have very high performance, thanks to their occurrences being almost always linked in Wikipedia and Wikinews articles (high recall) and easily distinguishable (high precision). In contrast, animals, biological entities, foods and plants have a high-degree of confusion (lower precision), and are very often not linked (low recall). To better explain this, we report in Figure 1 the confusion matrix of the silver-standard annotations produced by our approach compared to the gold-standard ones. As an example, it can be observed that animals and plants are often confused with each other, mainly because their scientific names are morphologically very close. Similarly, animals and plants are also confused with foods (e.g. Alaskan salmon and Quinoa), and vice versa.

Even though the quality of the annotations produced by our approach for any particular one of the 10 languages covered herein is strongly dependent on the quality of the corresponding Wikipedia and Wikinews dumps, we expect comparable performance on all other languages, as suggested by the statistics in Table 3 which show strong consistency across languages.

6 Conclusions

In this work we introduced MultiNERD, a novel resource for training robust multilingual and fine-grained Named Entity Recognition (and Disambiguation) systems. To create it, we presented a new language-agnostic strategy for generating high-quality silver-standard NER and ED annotations. This strategy uses a knowledge-based semantic classifier to automatically annotate Wikipedia and Wikinews articles, and then iteratively enhances the annotations produced by means of a self-improvement algorithm which builds upon neural models. Our experiments showed that MultiNERD outperformed previous state-of-the-art data-production methods across all tested languages and domains, while covering a much wider set of NER categories. Additionally, we also included image URLs in our dataset to encourage the development of multimodal NER and ED systems. This visual information could also be exploited to further improve the quality of the annotations by ensembling the predictions of NLP and Computer Vision models. We release our MultiNERD dataset and software at https://github.com/Babelscape/multinerd.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union’s Horizon 2020 research and innovation programme.

References


A NER Classes

Additional Details The OntoNotes 5.0 dataset distinguishes between locations, facilities and geopolitical entities – which are closely related – while the LOC class in our set clusters them. Moreover, we ignore the cardinals, ordinals, quantities and monetary values because their instances are not linked in Wikipedia, and their identification requires language-specific solutions. We also remove the PHY class (i.e. physical phenomena) due to its extreme rarity (i.e. 0.4 ‰ of entities belong to the PHY class according to Tedeschi et al., 2021a).

Mappings In order to measure the quality of our data against that of previous state-of-the-art automatically-generated datasets, which cover only the four coarse-grained entity types, we need to map our 15 classes to those entities. To do so, we analyze how elements belonging to our classes are labeled in the CoNLL dataset. The resulting class mapping is provided in Table 6. Similarly, for evaluating system performance on the OntoNotes dataset, we map its 18 classes to the 4 CoNLL classes. In this case, we use the mapping provided by Tedeschi et al. (2021b). Once mapped, as in Tedeschi et al. (2021a), the datasets are refined through a NER classifier trained on WikiNER.

Entity or Concept? Some of the entities that we are interested in (e.g. animals and diseases) are labeled as concepts in BabelNet, hence we need to accurately specify for which categories concepts are allowed. Our choices are reported in Table 6.

B Training Details

All model training was carried out on an NVIDIA GeForce RTX 3090 architecture. It required ∼45 s/epoch on the CoNLL and WikiANN datasets,
<table>
<thead>
<tr>
<th>Hyperparameter name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of Bi-LSTM layers</td>
<td>2</td>
</tr>
<tr>
<td>LSTM hidden size</td>
<td>512</td>
</tr>
<tr>
<td>batch size</td>
<td>128</td>
</tr>
<tr>
<td>learning rate</td>
<td>0.001</td>
</tr>
<tr>
<td>dropout</td>
<td>0.5</td>
</tr>
<tr>
<td>gradient clipping</td>
<td>1.0</td>
</tr>
<tr>
<td>adam $\beta_1$</td>
<td>0.9</td>
</tr>
<tr>
<td>adam $\beta_2$</td>
<td>0.999</td>
</tr>
<tr>
<td>adam $\epsilon$</td>
<td>1e-8</td>
</tr>
</tbody>
</table>

Table 7: Hyperparameter values of the reference model.

whereas it required $\sim$6 min/epoch on the WikiNER, WikiNEuRal and MultiNERD datasets. The full list of hyperparameter values of the best performing model – the one used to compare the quality of the different datasets – is shown in Table 7.