# SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2) 

Besnik Fetahu Sudipta Kar Zhiyu Chen Oleg Rokhlenko Shervin Malmasi<br>Amazon.com, Inc. Seattle, WA, USA<br>\{besnikf, sudipkar, zhiyuche,olegro, malmasi\}@amazon.com


#### Abstract

We present the findings of SemEval-2023 Task 2 on Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). ${ }^{1}$ Divided into 13 tracks, the task focused on methods to identify complex fine-grained named entities (like WrittenWork, Vehicle, MUSICALGRP) across 12 languages, in both monolingual and multilingual scenarios, as well as noisy settings. The task used the MultiCoNER v2 dataset, composed of 2.2 million instances in Bangla, Chinese, English, Farsi, French, German, Hindi, Italian., Portuguese, Spanish, Swedish, and Ukrainian.


MULTICoNER 2 was one of the most popular tasks of SemEval-2023. It attracted 842 submissions from 47 teams, and 34 teams submitted system papers. Results showed that complex entity types such as media titles and product names were the most challenging. Methods fusing external knowledge into transformer models achieved the best performance, and the largest gains were on the Creative Work and Group classes, which are still challenging even with external knowledge. Some fine-grained classes proved to be more challenging than others, such as Scientist, ArtWork, and Privatecorp. We also observed that noisy data has a significant impact on model performance, with an average drop of $10 \%$ on the noisy subset. The task highlights the need for future research on improving NER robustness on noisy data containing complex entities.

## 1 Introduction

Complex Named Entities (NE), like the titles of creative works, are not simple nouns and pose challenges for NER systems (Ashwini and Choi, 2014). They can take the form of any linguistic constituent, like an imperative clause ("Dial M for Murder"), and do not look like traditional NEs (Persons, locations, etc.). This syntactic ambiguity makes it challenging to recognize them based on context.

[^0]We organized the Multilingual Complex NER (MultiCoNER) task (Malmasi et al., 2022b) at SemEval-2022 to address these challenges in 11 languages, receiving a positive community response with 34 system papers. Results confirmed the challenges of processing complex and longtail NEs: even the largest pretrained Transformers did not achieve top performance without external knowledge. The top systems infused transformers with knowledge bases and gazetteers. However, such solutions are brittle against out of knowledgebase entities and noisy scenarios (e.g. spelling mistakes and typos). For entities with fine-grained classes, apart from the entity surface form, the context is critical in determining the correct class.

MULTICONER 2 expanded on these challenges by adding fine-grained NER classes, and the inclusion of noisy input. Fine-grained NER requires models to distinguish between sub-types of entities that differ only at the fine-grained level, e.g. Scientist vs. Athlete. In these cases, it is crucial for models to capture the entity's context. In terms of noise, we assessed how small perturbations in the entity surface form and its context can impact performance. Noisy scenarios are quite common in many applications such as Web search and social media. These challenges are described in Table 1, and our tasks defined below.

1. Monolingual: NER systems are evaluated on monolingual setting, e.g. models are trained and tested on the same language ( 12 tracks in total)
2. Multilingual: NER systems are tested on a multilingual test set, composed from all languages in the monolingual track.

We released the MultiCoNER v2 dataset (Fetahu et al., 2023) to address the aforementioned challenges. MULTICONER v2 includes data from Wikipedia which has been filtered to identify difficult low-context sentences, and further postprocessed. The data covers 12 languages, which

| Challenge | Description |
| :--- | :--- |
| Fine-grained <br> Entities | The entity type can be different based on the context. For example, a creative work entity "Harry <br> Potter and the Sorcerer's Stone" could be s a book or a film, depending on the context. |
| Noisy NER | Gazetteer based models would not work for typos (e.g., "sony xperia" $\rightarrow$ "somy xpria") or <br> spelling errors (e.g., "ford cargo" $\rightarrow$ "f0rd cargo") in entities, degrading significantly their <br> performance. |
| Ambiguous Entities | Some NEs are ambiguous: they are not always entities, e.g. "Inside Out", "Among Us", and <br> and Contexts |
| "Bonanza" may refer to NEs (a movie, video game, and TV show) in some contexts, but not in |  |
| others. Such NEs often resemble regular syntactic constituents. |  |

Table 1: Challenges addressed by MultiConer 2.
are used to define the 12 monolingual subsets of the task. Additionally, the dataset has a multilingual subset which has mixed data from all the languages.

MULTICONER 2 received 842 submissions from 47 teams, and 34 teams submitted system description papers. Results showed that usage of external data and ensemble strategies played a key role in the strong performance. External knowledge brought large improvements on classes containing names of creative works and groups, allowing those systems to achieve the best overall results.

Regarding noisy data, all systems show significant performance drop on the noisy subset, which included simulated typographic errors. Small perturbations to entities had a more negative effect than those to the context tokens surrounding entities. This suggests that current systems may not be robust enough to handle real-world noisy data, and that further research is needed to improve their performance in such scenarios. Finally, NER systems seem to be most robust to noise for PER, while most susceptible to noise for GRP.

In terms of fine-grained named entity types, we observed that performance was lower than the coarse types due to failure to correctly disambiguate sub-classes such as ATHLETE vs. SportsManager. Some of the most challenging fine-grained classes include Privatecorp, Scientist and ArtWork.

## 2 MUlTiCoNER v2 Dataset

The MultiCoNER v2 dataset was designed to address the NER challenges described in §1. The data comes from the wiki domain and includes 12 languages, plus a multilingual subset. Some examples from our data can be seen in Figure 1. For a detailed description of the MULTICONER v2 data, we refer the reader to the dataset paper (Fetahu et al., 2023). The dataset is publicly available. ${ }^{2}$

[^1]

Figure 1: Examples sentences from MultiConer v2.

### 2.1 Languages and Subsets

Multiconer v2 covers 12 languages:

| Bangla (BN) | Chinese (ZH) | English (EN) |
| :--- | :--- | :--- |
| Farsi (FA) | French (FR) | German (DE) |
| Hindi (HI) | Italian (IT) | Portuguese (PT) |
| Spanish (ES) | Swedish (SV) | Ukrainian (UK) |

These languages were chosen to include a diverse typology of languages and writing systems, and range from well-resourced (EN) to lowresourced ones (FA). MuLtiCoNER v2 contains 13 different subsets: 12 monolingual, and a multilingual subset (denoted as MULTI).

Monolingual Subsets Each of the 12 languages has its own subset.

Multilingual Subset This contains randomly sampled data from all the languages mixed into a single subset. This subset is designed for evaluating multilingual models, and should ideally be used under the assumption that the language for each sentence is unknown. From the test set of each language, we randomly selected at most 35,000 samples resulting in a total of 358,668 instances.

### 2.2 Dataset Creation

In this section, we provide a brief overview of the dataset construction process. Additional details are available in Fetahu et al. (2023).

MULTICONER v2 was extracted following the same strategy as Malmasi et al. (2022a), where sen-
tences from the different languages are extracted from localized versions of Wikipedia. We select low-context sentences and the interlinked entities are resolved to the entity types using Wikidata as a reference, according to the NER class taxonomy shown in Table 2. Furthermore, to prevent models from leveraging surface form features, we lowercase the words and remove punctuation. These steps result in more challenging sentences that are more representative of real-world data.

### 2.3 Fine-grained NER Taxonomy

MULTICONER 2 builds on top of the WNUT 2017 (Derczynski et al., 2017) taxonomy entity types, and adds an additional layer of fine-grained types. We also drop the Corporation class, as it overlaps with the Group class. Furthermore, we introduce a new coarse grained class called Medical, which captures entities from the medical domain (e.g. Disease, AnatomicalStructure, etc.). Table 2 shows the 33 fine-grained classes, grouped across 6 coarse types.

The fine-grained taxonomy allows us to capture a wide array of entities, including complex entity structures, such as CW, or entities that are ambiguous without their context, e.g. SCIENTIST vs. ATHLETE as part of the PER coarse grained type.

### 2.4 Noisy Subsets

NER systems are typically trained on carefully curated datasets. However, in real-world scenarios, various errors may arise due to human mistakes. We applied noise only on the test set to simulate environments where NER models are exposed directly to user-generated content.

To evaluate the robustness of NER models, we corrupt $30 \%$ of the test set with various types of simulated errors in 7 languages (EN, ZH, IT, ES, FR, PT, SV). The corruption can impact context tokens and entity tokens. For Chinese, we applied character-level corruption strategies (Wang et al., 2018) which involve replacing characters with visually or phonologically resembled ones. For other languages, we developed token-level corruption strategies based on common typing mistakes made by humans (e.g., randomly substituting a letter with a neighboring letter on the keyboard), utilizing language specific keyboard layouts. ${ }^{3}$

[^2]
### 2.5 Dataset Statistics

Table 3 shows the MuLTICoNER v2 dataset statistics. For most tracks, we released 16k training and 800 development instances (with the exception of $\mathrm{DE}, \mathrm{BN}, \mathrm{HI}, \mathrm{ZH}$ due to data scarcity).

The test splits on the other hand are much larger. This is done for two reasons: (1) to assess the generalizability of NER models in identifying unseen and complex fine-grained entity types, where the entity overlap between train and test sets is small, and and (2) to assess how models handle noise in contextual or entity tokens. For practical reasons, we cap the number of test instances to be less than 250k per subset for most languages (with the exception of DE, BN, HI, ZH which are already small due to data scarcity).

## 3 Task Description and Evaluation

The shared task is composed of 12 monolingual and 1 multilingual track. The multilingual track invited multilingual models capable of identifying entities from monolingual texts from any of the 12 languages. As described in Section 2.4, 30\% of the test sets of the EN, ZH, IT, ES, FR, PT, and SV monolingual tracks are corrupted with simulated noise. We refer the subsets with corruption as noisy subsets and the rest as clean subsets.

For evaluation, we used the macro-averaged F1 scores to evaluate and rank systems. The F1 scores are computed over the fine-grained types using exact matching (i.e. the entity boundary and type must exact match the ground truth), and averaged across all types. We also report the performance on noisy subsets and clean subsets in Appendix A to study the impact on noise in $\S 6$.

## 4 Baseline System

Similar to the 2022 edition (Malmasi et al., 2022b), we train and evaluate a baseline NER system using XLM-RoBERTa (XLM-R) (Conneau et al., 2020), a multilingual Transformer model. The XLM-R model computes a representation for each token, which is then used to predict the token tag using a CRF classification layer (Sutton et al., 2012).

XLM-R is suited for multilingual scenarios, supporting up to 100 languages. It provides a solid baseline upon which the participants can build. It was trained with a learning rate of $2 e-5$ and for 50 epochs, with an early stopping criterion of a non-decreasing validation loss for 5 epochs. The

| PER (Person) | LOC (Location) | GRP (Group) | PROD (Product) | CW (Creative Work) | MED (Medical) |
| :--- | :--- | :--- | :--- | :--- | :--- |
| ARTIST | FACILITY | AEROSPACEMANUFACTURER | CLOTHING | ARTWORK | ANATOMICALSTRUCTURE |
| ATHLETE | HUMANSETTLEMENT | CARMANUFACTURER | DRINK | MUSICALWORK | DISEASE |
| CLERIC | STATION | MUSICALGRP | FOOD | SoFtWARE | MEDICALPROCEDURE |
| POLITICIAN | OTHERLOC | ORG | VEHICLE | VISUALWORK | MEDICATION/VACCINE |
| SCIENTIST |  | PRIVATECORP | OTHERPROD | WRITTENWORK | SYMPTOM |
| SPORTSMANAGER |  | PUBLICCORP |  |  |  |
| OTHERPER |  | SPORTSGRP |  |  |  |

Table 2: MULTICoNER v2 NER taxonomy, consisting of 33 fine-grained classes, grouped across 6 coarse grained types.

| Class | Split | EN | DE | FA | FR | ES | UK | SV | HI | BN | ZH | IT | PT | Multi |
| :--- | :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  | train | 9,294 | 5,508 | 8,006 | 9,295 | 8,360 | 6,441 | 7,695 | 3,609 | 3,778 | 4,862 | 10,387 | 8,241 | 85,476 |
| PER | dev | 481 | 280 | 413 | 483 | 442 | 341 | 445 | 174 | 194 | 239 | 548 | 447 | 4,487 |
|  | test | 137,681 | 11,299 | 115,868 | 141,401 | 125,379 | 96,864 | 111,157 | 5,736 | 6,935 | 9,095 | 160,598 | 120,413 | 180,080 |
|  | train | 4,084 | 2,466 | 3,661 | 5,438 | 3,606 | 2,907 | 3,714 | 1,646 | 1,981 | 2,264 | 5,048 | 3,839 | 40,654 |
| CW | dev | 215 | 127 | 184 | 268 | 183 | 146 | 200 | 90 | 103 | 112 | 267 | 206 | 2,101 |
|  | test | 62,126 | 4,777 | 53,034 | 84,952 | 55,459 | 43,291 | 54,806 | 2,804 | 3,640 | 4,369 | 79,873 | 58,245 | 87,030 |
|  | train | 4,224 | 2,815 | 3,209 | 3,745 | 3,632 | 3,204 | 3,459 | 2,273 | 2,227 | 2,696 | 3,416 | 3,788 | 38,688 |
| GRP | dev | 218 | 177 | 180 | 195 | 195 | 151 | 194 | 143 | 122 | 145 | 173 | 200 | 2,093 |
|  | test | 60,026 | 4,418 | 38,807 | 52,987 | 50,259 | 39,709 | 46,929 | 3,897 | 3,651 | 4,715 | 46,271 | 48,994 | 73,226 |
|  | train | 4,353 | 2,269 | 5,086 | 4,723 | 4,651 | 5,458 | 7,176 | 2,487 | 2,457 | 2,470 | 4,446 | 4,794 | 50,370 |
| LOC | dev | 197 | 117 | 267 | 242 | 230 | 294 | 370 | 133 | 127 | 129 | 248 | 250 | 2,604 |
|  | test | 67,901 | 5,306 | 70,907 | 73,373 | 72,996 | 84,643 | 111,879 | 7,172 | 7,375 | 6,170 | 68,564 | 70,923 | 117,257 |
|  | train | 1,935 | 1,571 | 2,049 | 1,946 | 1,989 | 2,258 | 1,989 | 1,420 | 1,384 | 1,529 | 1,770 | 1,927 | 21,767 |
| PROD | dev | 109 | 78 | 107 | 100 | 100 | 117 | 112 | 74 | 67 | 73 | 86 | 101 | 1,124 |
|  | test | 27,580 | 1,643 | 18,212 | 28,274 | 28,469 | 30,071 | 22,686 | 1,611 | 1,493 | 1,869 | 22,887 | 21,115 | 35,545 |
|  | train | 1,559 | 1,322 | 1,651 | 1,230 | 1,669 | 1,688 | 1,381 | 1,435 | 1,396 | 1,407 | 1,376 | 1,850 | 17,964 |
| MED | dev | 76 | 62 | 85 | 64 | 81 | 86 | 70 | 70 | 63 | 75 | 76 | 88 | 896 |
|  | test | 22,491 | 1,434 | 15,287 | 17,208 | 23,812 | 20,796 | 13,702 | 1,979 | 1,919 | 1,781 | 19,029 | 21,062 | 29,553 |
|  | train | 16,778 | 9,785 | 16,321 | 16,548 | 16,453 | 16,429 | 16,363 | 9,632 | 9,708 | 9,759 | 16,579 | 16,469 | 170,824 |
| Total | dev | 871 | 512 | 855 | 857 | 854 | 851 | 856 | 514 | 507 | 506 | 858 | 854 | 8,895 |
|  | test | 249,980 | 20,145 | 219,168 | 249,786 | 246,900 | 238,296 | 231,190 | 18,399 | 19,859 | 20,265 | 247,881 | 229,490 | 358,668 |

Table 3: MULTICoNER 2 dataset statistics for the different languages for the Train/Dev/Test splits. For each NER class we show the total number of entity instances per class on the different data splits. The bottom three rows show the total number of sentences for each language.
code and scripts for the baseline system were provided to the participants to use its functionalities and further extend it with their approaches. ${ }^{4}$

## 5 Participating Systems and Results

We have received submissions from 47 different teams. Table 4 shows the final rankings for all tracks (fine-grained Macro F1). Among the monolingual tracks, we have observed the highest participation in the English track with 34 teams. Ordered by the number of participating teams, the rest of the monolingual tracks are Chinese (22), German (17), Bangla (18), Spanish (18), Hindi (17), French (17), Portuguese (17), Swedish (16), Italian (15), Farsi (14), and Ukrainian (14). The number of participating teams for the Multilingual track is 18 . Detailed performance breakdown for noisy and clean subsets of English, Spanish, French, Italian, Portuguese, Swedish, and Chinese is available in Appendix A.

### 5.1 Top Systems

DAMO-NLP (Tan et al., 2023) ranked $1^{\text {st }}$ for most tracks, except being $2^{\text {nd }}$ in $\mathrm{BN}, \mathrm{DE}, \mathrm{ZH}$, and

[^3]$4^{t h}$ in HI. They proposed an unified retrievalaugmented system (U-RaNER) for the task. The system uses two different knowledge sources (Wikipedia paragraphs and the Wikidata knowledge graph) to inject additional relevant knowledge to their NER model. Additionally, they explored an infusion approach to provide more extensive contextual knowledge about entities to the model.

PAI (Ma et al., 2023b) ranked $1^{\text {st }}$ in $\mathrm{BN}, \mathrm{DE}, 2^{\text {nd }}$ in FR, $\mathrm{HI}, \mathrm{IT}, \mathrm{PT}, 3^{r d}$ in EN, $4^{t h}$ in $\mathrm{ZH}, 5^{t h}$ in MULTI, $7^{t h}$ in ES, FA, UK, and $8^{t h}$ in SV. They developed a knowledge base using entities and their associated properties like "instanceof", "subclassof" and "occupation" from Wikidata. For a given sentence, they used a retrieval module to gather different properties of the entities by string matching. They observed benefits on the clean subset through the dictionary fusing approach. The same benefits were not observed on the noisy subset.

USTC-NELSLIP (Ma et al., 2023a) ranked $1^{\text {st }}$ in $\mathrm{HI}, 3^{r d}$ in BN, ES, $4^{\text {th }}$ in DE, UK, $5^{\text {th }}$ in IT, SV, $6^{t h}$ in FA, FR, PT, ZH, MULTI, and $8^{t h}$ in EN. They proposed a two-stage training strategy. In the first stage, the representations of gazetteer network and language model are adapted at sentence and entity

| English (EN) |  |  | 10 | silp_nlp | 65.00 | 4 | NLPeople | 70.76 | 10 | BizNER | 67.71 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | DAMO-NLP | 85.53 | 11 | LSJSP | 64.36 | 5 | IXA/Cogcomp | 69.49 | 11 | LLM-RM | 63.29 |
| 2 | SRC - Beijing | 83.09 | 12 | D2KLab | 62.98 | 6 | USTC-NELSLIP | 68.85 | 12 | D2KLab | 63.29 |
| 3 | PAI | 80.00 | 13 | Sartipi-Sedighin | 61.95 | 7 | PAI | 68.46 | 13 | Sartipi-Sedighin | 63.10 |
| 4 | CAIR-NLP | 79.33 | 14 | SAB | 58.03 | 8 | Sakura | 64.88 | 14 | SAB | 62.30 |
| 5 | KDDIE | 78.06 |  | BASELINE | 57.19 | 9 | garNER | 62.12 | 15 | LSJSP | 53.13 |
| 6 | SRCB | 75.62 | 15 | FII_Better | 52.12 | 10 | Sartipi-Sedighin | 60.02 | 16 | L3i++ | 43.56 |
| 7 | IXA/Cogcomp | 72.82 | 16 | IXA | 25.96 | 11 | D2KLab | 54.20 | 17 | IXA | 26.13 |
| 8 | USTC-NELSLIP | 72.15 | Ukranian (UK) |  |  | 12 | Ertim | 53.77 | Bangla (BN) |  |  |
| 9 | NLPeople | 71.81 | 1 | DAMO-NLP | 89.02 | 13 | SAB | 52.42 | 1 | PAI | 84.39 |
| 10 | BizNER | 70.44 | 2 | CAIR-NLP | 81.29 |  | BASELINE | 51.56 | 2 | DAMO-NLP | 81.60 |
| 11 | Sakura | 70.16 | 3 | IXA/Cogcomp | 75.25 | 14 | IXA | 15.87 | 3 | USTC-NELSLIP | 80.59 |
| 12 | RIGA | 69.30 | 4 | USTC-NELSLIP | 74.37 |  | German (DE) |  | 4 | IXA/Cogcomp | 78.95 |
| 13 | CodeNLP | 63.51 | 5 | NLPeople | 73.41 | 1 | PAI | 88.09 | 5 | NLPeople | 78.24 |
| 14 | Sartipi-Sedighin | 63.25 | 6 | Sakura | 72.31 | 2 | DAMO-NLP | 84.97 | 6 | Sakura | 77.20 |
| 15 | IITD | 63.21 | 7 | PAI | 71.28 | 3 | IXA/Cogcomp | 80.35 | 7 | MLlab4CS | 76.27 |
| 16 | garNER | 62.73 | 8 | Sartipi-Sedighin | 67.25 | 4 | USTC-NELSLIP | 78.71 | 8 | garNER | 73.39 |
| 17 | FII_Better | 61.75 | 9 | garNER | 65.64 | 5 | NLPeople | 77.67 | 9 | silp_nlp | 73.22 |
| 18 | D2KLab | 61.29 | 10 | D2KLab | 64.14 | 6 | Sakura | 76.24 | 10 | CAIR-NLP | 69.46 |
| 19 | silp_nlp | 60.85 | 11 | silp_nlp | 63.18 | 7 | CAIR-NLP | 74.71 |  | BASELINE | 68.24 |
| 20 | Ertim | 59.03 | 12 | SAB | 59.42 | 8 | BizNER | 71.21 | 11 | Sartipi-Sedighin | 64.83 |
| 21 | MEERQAT-IRIT | 58.70 | 13 | LSJSP | 58.07 |  | BASELINE | 67.21 | 12 | VBD_NLP | 64.50 |
| 22 | LSJSP | 57.51 |  | BASELINE | 57.29 | 9 | D2KLab | 67.09 | 13 | BizNER | 64.37 |
| 23 | RGAT | 56.91 | 14 | IXA | 22.81 | 10 | silp_nlp | 64.92 | 14 | D2KLab | 61.43 |
| 24 | CLaC | 55.05 | Portugese (PT) |  |  | 11 | Sartipi-Sedighin | 64.21 | 15 | SAB | 56.01 |
| 25 | L3i++ | 53.00 | 1 | DAMO-NLP | 85.97 | 12 | garNER | 63.88 | 16 | LSJSP | 55.76 |
|  | BASELINE | 52.98 | 2 | PAI | 81.61 | 13 | FII_Better | 55.86 | 17 | L3i++ | 41.33 |
| 26 | VBD_NLP | 52.65 | 3 | CAIR-NLP | 80.16 | 14 | LLM-RM | 55.54 | 18 | IXA | 18.49 |
| 27 | LLM-RM | 52.08 | 4 | BizNER | 72.97 | 15 | SAB | 55.51 | Italian (IT) |  |  |
| 28 | Minanto | 51.47 | 5 | IXA/Cogcomp | 72.28 | 16 | L3i++ | 46.55 | 1 | DAMO-NLP | 89.79 |
| 29 | SAB | 51.41 | 6 | USTC-NELSLIP | 71.26 | 17 | IXA | 16.09 | 2 | PAI | 84.88 |
| 30 | ShathaTaymaaTeam | 50.02 | 7 | Deep Learning Brasil | 70.97 |  | Chinese (ZH) |  | 3 | CAIR-NLP | 83.78 |
| 31 | azaad@BND | 47.42 | 8 | NLPeople | 70.16 | 1 | NetEase.AI | 84.05 | 4 | BizNER | 76.48 |
| 32 | LISAC FSDM-USMBA | 44.00 | 9 | Sakura | 69.98 | 2 | DAMO-NLP | 75.98 | 5 | USTC-NELSLIP | 75.70 |
| 33 | YNU-HPCC | 28.52 | 10 | garNER | 64.51 | 3 | SRCB | 75.86 | 6 | IXA/Cogcomp | 74.67 |
| 34 | IXA | 15.39 | 11 | Sartipi-Sedighin | 61.28 | 4 | PAI | 74.87 | 7 | Sakura | 74.19 |
| Spanish (ES) |  |  | 12 | silp_nlp | 61.05 | 5 | Taiji | 72.52 | 8 | NLPeople | 73.71 |
| 1 | DAMO-NLP | 89.78 | 13 | D2KLab | 60.79 | 6 | USTC-NELSLIP | 66.57 | 9 | garNER | 68.20 |
| 2 | CAIR-NLP | 83.63 | 14 | MEERQAT-IRIT | 59.87 | 7 | NLPeople | 65.96 | 10 | D2KLab | 64.77 |
| 3 | USTC-NELSLIP | 74.44 | 15 | LSJSP | 58.23 | 8 | IXA/Cogcomp | 64.86 | 11 | Sartipi-Sedighin | 64.50 |
| 4 | IXA/Cogcomp | 73.81 | 16 | SAB | 54.12 | 9 | Sakura | 64.61 | 12 | silp_nlp | 63.11 |
| 5 | Sakura | 72.85 |  | BASELINE | 53.52 | 10 | garNER | 63.47 |  | BASELINE | 57.71 |
| 6 | NLPeople | 72.76 | 17 | IXA | 16.97 | 11 | Ertim | 59.45 | 13 | SAB | 57.57 |
| 7 | PAI | 71.67 | French (FR) |  |  | 12 | Sartipi-Sedighin | 58.70 | 14 | FII_Better | 56.36 |
| 8 | BizNER | 71.48 | 1 | DAMO-NLP | 89.59 | 13 | CAIR-NLP | 58.43 | 15 | IXA | 18.41 |
| 9 | garNER | 63.73 | 2 | PAI | 86.17 |  | BASELINE | 58.03 | Multilingual (MULTI) |  |  |
| 10 | D2KLab | 63.17 | 3 | CAIR-NLP | 83.08 | 14 | Janko | 57.90 | 1 | DAMO-NLP | 84.48 |
| 11 | silp_nlp | 62.90 | 4 | BizNER | 78.01 | 15 | YNUNLP | 56.57 | 2 | CAIR-NLP | 79.16 |
| 12 | MEERQAT-IRIT | 60.93 | 5 | IXA/Cogcomp | 74.52 | 16 | D2KLab | 54.92 | 3 | NLPeople | 78.38 |
| 13 | LSJSP | 60.55 | 6 | USTC-NELSLIP | 74.25 | 17 | silp_nlp | 51.65 | 4 | IXA/Cogcomp | 78.17 |
| 14 | Sartipi-Sedighin | 58.41 | 7 | Sakura | 72.86 | 18 | SAB | 44.12 | 5 | PAI | 77.00 |
| 15 | LLM-RM | 54.81 | 8 | NLPeople | 72.85 | 19 | NCUEE-NLP | 44.09 | 6 | USTC-NELSLIP | 75.62 |
| 16 | FII_Better | 54.51 | 9 | Ertim | 66.30 | 20 | L3i++ | 35.34 | 7 | Sakura | 73.82 |
|  | BASELINE | 53.43 | 10 | garNER | 65.68 | 21 | YNU-HPCC | 31.66 | 8 | MaChAmp | 73.74 |
| 17 | SAB | 48.22 | 11 | D2KLab | 64.09 | 22 | IXA | 6.93 | 9 | CodeNLP | 73.22 |
| 18 IXA $\begin{array}{lll}\text { Swedish (SV) } & 16.01\end{array}$ |  |  | 12 | silp_nlp | 62.39 |  | Hindi (HI) |  | 10 | Lumi | 72.15 |
|  |  |  | 13 | MEERQAT-IRIT | 58.90 | 1 | USTC-NELSLIP | 82.14 | 11 | Sartipi-Sedighin | 71.79 |
| 1 | DAMO-NLP | 89.57 | 14 | LSJSP | 56.83 | 2 | PAI | 80.96 | 12 | garNER | 69.16 |
| 2 | CAIR-NLP | 82.88 |  | BASELINE | 55.91 | 3 | IXA/Cogcomp | 79.56 | 13 | LEINLP | 64.63 |
| 3 | IXA/Cogcomp | 76.54 | 15 | SAB | 55.07 | 4 | DAMO-NLP | 78.56 | 14 | D2KLab | 63.83 |
| 4 | BizNER | 76.12 | 16 | Sartipi-Sedighin | 54.94 | 5 | NLPeople | 78.50 |  | BASELINE | 62.86 |
| 5 | USTC-NELSLIP | 75.47 | 17 | IXA | 17.40 | 6 | Sakura | 78.37 | 15 | SAB | 59.55 |
| 6 | NLPeople | 75.08 | Farsi (FA) |  |  | 7 | silp_nlp | 74.32 | 16 | LSJSP | 51.74 |
| 7 | Sakura | 73.79 | 1 | DAMO-NLP | 87.93 | 8 | CAIR-NLP | 72.23 | 17 | SibNN | 50.55 |
| 8 | PAI | 72.38 | 2 | CAIR-NLP | 77.50 | 9 | garNER | 71.23 | 18 | L3i++ | 44.37 |
| 9 | garNER | 67.63 | 3 | BizNER | 73.49 |  | BASELINE | 71.20 |  |  |  |

Table 4: Rankings for all tracks based on Macro F1. The "SRC - Beijing" team is "Samsung Research China - Beijing".
level through minimizing the KL divergence between their representations. In the second stage, two networks are trained together on the NER objective. The final predictions are derived from an ensemble of trained models. The results indicate that the gazetteer played a crucial role in accurately identifying complex entities during the NER process, and the implementation of a two-stage training strategy was effective.

NetEase.AI (Lu et al., 2023) ranked $1^{\text {st }}$ in ZH . Their proposed system consists of multiple modules. First, a BERT model is used to correct any potential errors in the original input sentences. The NER module takes the corrected text as input and consists of a basic NER module and a gazetteer enhanced NER module. This approach boosted the performance on the entity level noise and gave the system a strong advantage over the other teams (Table 11). A retrieval system takes the candidate entity as input and retrieves additional context information, which is subsequently used as input to a text classification model to calculate the probability of the entity's type label. A stacking model is trained to output the final prediction based on the features from multiple modules.

### 5.2 Other Noteworthy Systems

CAIR-NLP (N et al., 2023) ranked $2^{\text {nd }}$ in MULTI, $\mathrm{ES}, \mathrm{FA}, \mathrm{SV}, \mathrm{UK}, 3^{r d}$ in FR, IT, PT, $4^{\text {th }}$ in EN, $7^{\text {th }}$ in DE, $8^{t h}$ in $\mathrm{HI}, 10^{t h}$ in BN , and $13^{\text {th }}$ in ZH . They developed a multi-objective joint learning system (MOJLS) that learns an enhanced representation of low-context and fine-grained entities. In their training procedure they minimize for: 1) representation gaps between fine-grained entity types within a coarse grained type, 2) representation gaps between an input sentence and the input augmented with external information for a given entity, 3) negative log-likelihood loss, 4) biaffine layer label prediction loss. Additionally, external context is retrieved via search engines for an input text, as well as ConceptNet data (Speer et al., 2016) to better represent an entity class with alternative names, aliases, and relation types to other concepts.

SRCB (Zhang et al., 2023b) ranked $3^{r d}$ in ZH and $6^{t h}$ in EN. The proposed approach, for an input sentence retrieves external evidence coming from Wikidata and Wikipedia, which is concatenated with the original input using special tokens (e.g. "context", "prompt \& description") to allow their models (based on (Li et al., 2020)), to
distinguish the different contexts. To retrieve the external context, the authors first detect entity mentions (Su et al., 2022) from the input sentence, then query the corresponding external sources.

NLPeople (Elkaref et al., 2023) ranked $3^{\text {rd }}$ in MULTI, $4^{\text {th }}$ in FA, $5^{\text {th }}$ in BN, DE, HI, UK, $6{ }^{\text {th }}$ in ES, SV, $7^{\text {th }}$ in $\mathrm{ZH}, 8^{\text {th }}$ in FR, IT, PT, and $8^{t h}$ in EN. They developed a two stage approach. First they extract spans that can be entities, while in the second step they classify spans into the most likely entity type. They augmented the training data with external context by adding relevant paragraphs, infoboxes, and titles from Wikipedia. On languages with smaller test sets, the infoboxes showed to obtain better performance than adding relevant paragraphs.

IXA/Cogcomp (García-Ferrero et al., 2023) ranked $3^{\text {rd }}$ in DE, $\mathrm{HI}, \mathrm{UK}, \mathrm{SV}, 4^{\text {th }}$ in MULTI, BN, $\mathrm{ES}, 5^{\text {th }}$ in PT, FA, FR, $6^{\text {th }}$ in IT, $7^{\text {th }}$ in EN, $8^{\text {th }}$ in ZH , and $8^{t h}$ in EN. They first trained an XLMRoBERTa model for entity boundary detection, by recognizing entities within the dataset and classifying them using the B-ENTITY and I-ENTITY tags. They employed a pre-trained mGENRE entity linking model to predict the corresponding Wikipedia title and Wikidata ID for each entity span based on its context. Then, they retrieved the "part of", "instance of", "occupation" attributes and the article summary from Wikipedia. Finally, they trained a text classification model to categorize each entity boundary into a fine-grained category using the original sentence, entity boundaries and the external knowledge.

Samsung Research China (SRC) - Beijing (Zhang et al., 2023a) ranked $2^{\text {nd }}$ in EN. They fine-tuned a RoBERTa based ensemble system using a variant of dice loss (Li et al., 2019) to enhance the model's robustness on long tail entities. In their case dice loss uses soft probabilities over classes, to avoid the model overfitting on the more frequent classes. Additionally, a Wikipedia knowledge retrieval module was built to augment the sentences with Wikipedia passages.

Sakura (Poncelas et al., 2023) ranked $5^{\text {th }}$ in ES, $6^{\text {th }}$ in BN, DE, HI, UK, $7^{\text {th }}$ in IT, SV, MULTI, $8^{t h}$ in FA, $9^{\text {th }}$ in PT, ZH, and $11^{\text {th }}$ in EN. They used mBART-50 (Tang et al., 2020) to translate data from a source language to other target languages part of the shared task. Then, they aligned the tokens using SimAlign (Jalili Sabet et al., 2020) to annotate the entity tokens in the target language. Using the translated examples they increased the
training data size between 30 K to 102 K sentences depending on the language, providing them with a $1 \%$ increase in terms of macro-F1.

KDDIE (Martin et al., 2023) ranked $5^{t h}$ in EN. Using a retrieval index based on Wikipedia they enrich the original training data with additional sentences from Wikipedia. The data is used to train an ensemble of models, and the final NER scores is based on the vote from the different modules such as BERT-CRF, RoBERTa and DeBERTa.

MLlab4CS (Mukherjee et al., 2023) ranked $7^{\text {th }}$ in BN. MuRIL (Khanuja et al., 2021) was fine-tuned with an additional CRF layer used for decoding. MuRIL is specifically designed to deal with the linguistic characteristics of Indic languages.

CodeNLP (Marcińczuk and Walentynowicz, 2023) ranked $9^{\text {th }}$ in MULTI and $13^{\text {th }}$ in EN. mLUKElarge (Yamada et al., 2020) was fine tuned using different data augmentation strategies, where multiple data instances are concatenated as a single input. Their experiments show that the NER model benefits from the additional context, even when the context was unrelated to the original sentence.
silp_nlp (Singh and Tiwary, 2023) ranked $7^{\text {th }}$ in $\mathrm{HI}, 9^{\text {th }}$ in $\mathrm{BN}, 10^{\text {th }}$ in DE, SV, $11^{\text {th }}$ in ES, UK, $12^{\text {th }}$ in FR, IT, PT, $17^{\text {th }}$ in $\mathrm{ZH}, 19^{\text {th }}$ in EN. Their model is trained in two stages. XLM-RoBERTa is first pre-trained using the multilingual set. Then, the checkpoint is fine-tuned for individual languages.
garNER (Hossain et al., 2023) ranked $8^{t h}$ in BN, $9^{t h}$ in ES, SV, UK, FA, HI, IT, $10^{\text {th }}$ in PT, FR, ZH , $12^{\text {th }}$ in DE, MULTI, and $16^{\text {th }}$ in EN. The authors proposed an approach augmented with external knowledge from Wikipedia. For a given sentence and an entity, the Wikipedia API is called, and the retrieved result is concatenated together with the sentence to provide additional context for token classification. The entities are extracted via spaCy for English, and for other languages XLM-RoBERTa is used to detect entities. The authors performed ablation studies to analyze the model performance and found that the relevance of the augmented context is a significant factor in the model's performance. Useful context can help the model to identify some hard entities correctly, while irrelevant context can negatively affect model's predictions.

Sartipi-Sedighin (Sartipi et al., 2023) ranked $8^{t h}$ in UK, $10^{\text {th }}$ in FA, $11^{\text {th }}$ in BN, DE, IT, PT, MULTI, $12^{t h}$ in $\mathrm{ZH}, 13^{\text {th }}$ in $\mathrm{HI}, \mathrm{SV}, 14^{\text {th }}$ in EN, ES, and $16^{\text {th }}$ in FR. They used a data augmentation approach, where for entities in the training dataset, additional
sentences from Wikipedia are retrieved. The retrieved sentences are used as additional context. Then, they experimented with Transformer based model variations fine-tuned on different languages. Data augmentation helped their model in certain classes, but negatively impacted some other classes by increasing false negatives, e.g. SYMPTOM.

MaChAmp (van der Goot, 2023) ranked $8^{\text {th }}$ in MULTI. mLUKE-large(Yamada et al., 2020) was fine-tuned on data coming from all SemEval2023 text based tasks. For NER a CRF decoding layer used. For hyper-parameters they relied on the MaChAmp toolkit (van der Goot et al., 2021). They also experimented with separate decoders for each language, using intermediate task pre-training with other SemEval tasks, but did not find it useful for further improvements.

D2KLab (Ehrhart et al., 2023) ranked $9^{t h}$ in $\mathrm{DE}, 10^{\text {th }}$ in $\mathrm{ES}, \mathrm{IT}, \mathrm{UK}, 11^{\text {th }}$ in FA, FR, $12^{\text {th }}$ in HI , SV, $13^{\text {th }}$ in PT, $14^{\text {th }}$ in BN, MULTI, $16^{\text {th }}$ in ZH , and $18^{\text {th }}$ in EN. T-NER library (Ushio and CamachoCollados, 2021) was used to fine-tune a Transformer model. They additionally used 10 other publicly available NER datasets, in addition to the data from MultiCoNER 2 and MultiCoNER.

ERTIM (Deturck et al., 2023) ranked $9^{t h}$ in FR, $11^{\text {th }}$ in $\mathrm{ZH}, 12^{\text {th }}$ in FA, and $20^{\text {th }}$ in EN. They finetuned different models for the different languages, e.g. BERT, DistilBERT, CamemBERT, and XLMRoBERTa. Additionally, each input sentence is enriched with relevant Wikipedia articles for additional context. Furthermore, they annotated a set of additional Farsi sentences extracted from news articles, which provides their system with an improvement of $4.2 \%$ in terms of macro-F1 for FA.

LSJSP (Chatterjee et al., 2023) ranked $11^{\text {th }}$ in $\mathrm{SV}, 13^{\text {th }}$ in ES, UK, $14^{\text {th }}$ in FR, $15^{t h}$ in HI, PT, $16^{t h}$ in BN, MULTI, and $22^{\text {nd }}$ in EN. They rely on a nearest neighbor search method, based on FastText's (Bojanowski et al., 2016) implementation, to deal with noisy entities in the dataset. Next, they use pre-trained transformer models, with a CRF layer for NER prediction.

LLM-RM (Mehta and Varma, 2023) ranked $11^{\text {th }}$ in $\mathrm{HI}, 14^{\text {th }}$ in DE, $15^{\text {th }}$ in ES, $27^{\text {th }}$ in EN by fine-tuning XLM-RoBERTa.

MEERQAT-IRIT (Lovon-Melgarejo et al., 2023) ranked $12^{\text {th }}$ in ES, $13^{\text {th }}$ in FR, $14^{\text {th }}$ in PT, $21^{\text {st }}$ in EN. First, they developed hand-crafted tag descriptors for the fine-grained classes, then, an ensemble representation using the original input
and the tag descriptors are used as input to the final CRF layer on top of XLM-RoBERTa.

RIGA (Mukans and Barzdins, 2023) ranked $12^{t h}$ in EN. The original data was augmented using GPT-3 to obtain additional context information, then XLM-RoBERTa (large) was fine-tuned using the adapter fusion approach (Pfeiffer et al., 2021). The additional context extracted through GPT-3 provides them with a performance boost of $4 \%$ in terms of macro-F1. The context is separated from the input sentence using the separator token [SEP].

VBD_NLP (Hoang et al., 2023) ranked $12^{\text {th }}$ in BN and $26^{t h}$ in EN. First, training data was augmented based on BabelNet and Wikipedia redirects to automatically annotate named entities from Wikipedia articles. Then, mDeBERTaV3 with a BiLSTM-CRF layer was fine-tuned for NER. While their model outperformed the baseline in Bangla, it underperformed in English.

SAB (Biales, 2023) ranked $29^{\text {th }}$ in EN, $17^{\text {th }}$ in $\mathrm{ES}, 14^{\text {th }}$ in ES, $\mathrm{HI}, 12^{\text {th }}$ in UK, $16^{\text {th }}$ in $\mathrm{PT}, 15^{\text {th }}$ in FR, DE, BN, MULTI, $13^{t h}$ in FA, IT, $18^{t h}$ in ZH. First, POS tags and dependency relation tags are obtained from open-sourced tools for all languages except $B N$ and MULTI track. XLM-R (base) was fine-tuned under a multi-task setup where POS tags, dependency relations and NER labels are predicted. However, they found that using POS and dependency relation did not improve the results.

FII_Better (Lupancu et al., 2023) ranked $13^{\text {th }}$ in DE, $14^{\text {th }}$ in IT, $15^{t h}$ in SV, $16^{\text {th }}$ in ES, and $17^{t h}$ in EN. A BERT model was fine-tuned to label each input token for NER.

IXA (Andres Santamaria, 2023) ranked $14^{\text {th }}$ in FA, UK, $15^{\text {th }}$ in IT, $16^{\text {th }}$ in SV, $17^{\text {th }}$ in DE, HI, FR, PT, $18^{\text {th }}$ in BN, ES, $22^{\text {nd }}$ in ZH , and $34^{\text {th }}$ in EN. XLMRoBERTa was fine-tuned for each track separately.

Janko (Li et al., 2023) ranked $14^{\text {th }}$ in ZH. The authors use the last layer of BERT embeddings to represent input tokens, which is then used in a Bi-LSTM model for NER. Additionally, a dropout layer is added, namely R-DROP.

IITD (Choudhary et al., 2023) ranked $15^{t h}$ in EN . A two-stage pipeline to fine-tune BERT is proposed: the model is first trained with focal loss to avoid class imbalance issues (Lin et al., 2017). Then, each input is augmented with sentences retrieved from MS-MARCO (Nguyen et al., 2016) and KILT (Petroni et al., 2020) datasets.

YNUNLP (Li and Zhou, 2023) ranked $15^{t h}$ in ZH. A BERT based approach with a top CRF layer
for the NER tag prediction was used. Additionally, a R-Drop layer for regularization to increase the model's robustness was used.
$\mathbf{L 3 i}++$ (Gonzalez-Gallardo et al., 2023) ranked $16^{\text {th }}$ in DE, ES, $\mathrm{HI}, 17^{\text {th }}$ in BN, $18^{\text {th }}$ in FA, FR, MULTI, $20^{t h}$ in IT, PT, UK, SV, ZH, and $25^{t h}$ in EN. They submitted three systems. The first model is built with stacked Transformer blocks on top of the BERT encoder with an additional conditional CRF layer. The second one approached the problem with a seq 2 seq framework: sentences and statement templates filled by candidate named entity span are regarded as the source sequence and the target sequence. In the third approach they transformed NER into a QA task, where a prompt is generated for each type of named entity. The third approach showed strong performance in recall but overall performance was better using the stacked approach.

RGAT (Chakraborty, 2023) ranked $23^{r d}$ in EN. They used dependency parse trees from sentences and encode them using a graph attention network. The node representations were computed by taking into account the neighboring nodes and the dependency type. Additionally, they used features from BERT to make the final prediction for a token.

CLaC (Verma and Bergler, 2023) ranked $24^{\text {th }}$ in EN. They fine-tuned XLM-RoBERTa, finding that the span prediction approach is better than the sequence labeling approach.

Minanto (Höfer and Mottahedin, 2023) ranked $28^{t h}$ in EN. XLM-RoBERTa was trained using the training data and a set of translated data from CoNLL 2003 and WNUT 2016 datasets.

## 6 Insights from the Systems

Integrating External Knowledge: To overcome the challenges of complex entities, unseen entities, and low context, the integration of external data was a common theme among the submitted systems, similar to the prior edition. However, this time we observed many new and diverse knowledge sources and novel ways to inject the data into the models for NER prediction. For example, apart from using paragraphs retrieved from Wikipedia using search engine, participating teams used Wikidata, Wikipedia Infoboxes, and ConceptNet. Some of these approaches used knowledge sources to compute better representation of the entity labels.
Multilingual Models: Most participants in the multilingual track opted to use the task's baseline model, XLM-RoBERTa. Additionally, some par-
ticipants used mLUKE, mDEBERTA, and mBERT. In terms of external multilingual resources, participants made mostly use of Wikipedia.

Complex Entities: Our task includes several classes with complex entities such as media titles. The most challenging entities at the coarse level were from PROD class, where the average macroF1 score across all participants was 0.68 . This classes contains challenging entities, with highly complex and ambiguous surface forms, such as Clothing, where the average across all participants was macro- $\mathrm{F} 1=0.58$. There is a high variation among on the challenging coarse types, such as PROD. For instance, for EN the top ranked system, DAMO-NLP, achieves an F1 of 0.88 , while the lowest ranking system IXA achieves a F1 of 0.21 . This is highly related to whether the systems used external knowledge.

Figure 2 shows a confusion matrix of coarsegrained performance. We note that PROD, MED and CW have low recall with more than $25 \%$ of the entities not being identified correctly. GRP is misclassified in $4.2 \%$ of the cases with other types such as LOC or CW, highlighting the surface form ambiguity of this type. On the other hand, PER obtains the highest score with $93.7 \%$, yet at finegrained level often there is confusion among the different PER fine-grained types.


Figure 2: Confusion matrix of baseline performance computed at the coarse type level for the EN test set.

Impact of Fine-grained classes: For coarse types such as PER, participants obtain very high scores, e.g. DAMO-NLP obtains an F1 of 0.97 on the noise-free test set. However, if we inspect the performance at the fine-grained level we notice high variance. For instance, Scientist and OTHERPER obtain significantly lower scores with F1 scores of 0.70 . This gap provides two main insights. First, while the PER class is often very easy to spot, distinguishing the more fine-grained types is much more challenging given their high ambiguity. Second, for fine-grained NER, captur-
ing context is important. In this case we see that for a class like Scientist, where its entities are often in scientific reporting context (e.g. research breakthroughs), pre-trained Transformer models often confuse such entities as either Artist or Politician, for which such models have much more pretrained knowledge. Appendix B provides an in-depth error analysis at the fine-grained entity type level for all coarse grained types.
Impact of Noise: Evaluation on the noisy subsets shows that most of the participants were impacted significantly. Comparing the difference in terms of macro-F1 on the noisy and the clean subsets, we notice that across all participants and languages, there is an average performance drop of $10 \%$. The most impact is observed for ZH , where the gap can be as high as macro-F1 $=\boldsymbol{\nabla} 48 \%$.

Finally, we note that noise is mostly harmful when it affects named entity tokens, while noise on other has a minor impact in terms of NER performance. Across all participants and languages, the average performance dropped $11.1 \%$ when corruption was applied to entity tokens and $4.3 \%$ when it was applied to context tokens.
ChatGPT and LLMs: Our evaluation concluded in Jan 2023, and participants did not use ChatGPT for the submissions. DAMO-NLP (Tan et al., 2023) reported that the performance of ChatGPT on MULTI track is poor and it only achieved $14.78 \%$ F1 score. This matches the results of Lai et al. (2023) where they evaluated ChatGPT on MultiCoNER task from last year (Malmasi et al., 2022b).

## 7 Conclusion

We presented an overview of the SemEval shared task on identifying complex entities in multiple languages. We received system submissions from 47 teams, and 34 system papers. On average, the wining systems for all tracks outperformed the baseline system by a large margin of $35 \% \mathrm{~F} 1$.

All top-performing teams in MULTICoNER 2 utilized external knowledge bases like Wikipedia and gazetteers to provide additional context. We have also observed systems that provided information about the entity classes to help models know the definition of the entity. In terms of modeling, ensemble strategies helped the systems achieve strong performance. Finally, the impact of noise was significant for all submitted systems, with the macro-F1 dropping significantly when compared between the noisy and clean subsets of test data.

## References

Edgar Andres Santamaria. 2023. Ixa at semeval-2023 task 2: Baseline xlm-roberta-base approach. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 379-381, Toronto, Canada. Association for Computational Linguistics.

Sandeep Ashwini and Jinho D. Choi. 2014. Targetable named entity recognition in social media. CoRR, abs/1408.0782.

Siena Biales. 2023. Sab at semeval-2023 task 2: Does linguistic information aid in named entity recognition? In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1131-1137, Toronto, Canada. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

Abir Chakraborty. 2023. Rgat at semeval-2023 task 2: Named entity recognition using graph attention network. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 163-170, Toronto, Canada. Association for Computational Linguistics.

Shilpa Chatterjee, Leo Evenss, Pramit Bhattacharyya, and Joydeep Mondal. 2023. Lsjsp at semeval-2023 task 2: Ftbc: A fasttext based framework with pretrained bert for ner. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1254-1259, Toronto, Canada. Association for Computational Linguistics.

Shivani Choudhary, Niladri Chatterjee, and Subir Kumar Saha. 2023. Iitd at semeval-2023 task 2: A multi-stage information retrieval approach for finegrained named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 800-806, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 84408451, Online. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the wnut2017 shared task on novel and emerging entity recognition. In Proceedings of the 3rd Workshop on Noisy Usergenerated Text, pages 140-147.

Kevin Deturck, Pierre Magistry, Bénédicte DIOTPARVAZ AHMAD, Ilaine Wang, Damien Nouvel, and Hugo Lafayette. 2023. Ertim at semeval-2023
task 2: Fine-tuning of transformer language models and external knowledge leveraging for ner in farsi, english, french and chinese. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 2211-2215, Toronto, Canada. Association for Computational Linguistics.

Thibault Ehrhart, Julien Plu, and Raphael Troncy. 2023. D2klab at semeval-2023 task 2: Leveraging t-ner to develop a fine-tuned multilingual model for complex named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 836-840, Toronto, Canada. Association for Computational Linguistics.

Mohab Elkaref, Nathan Herr, Shinnosuke Tanaka, and Geeth de Mel. 2023. Nlpeople at semeval-2023 task 2: A staged approach for multilingual named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1148-1153, Toronto, Canada. Association for Computational Linguistics.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023. MultiCoNER v2: a Large Multilingual dataset for Finegrained and Noisy Named Entity Recognition.

Iker García-Ferrero, Jon Ander Campos, Oscar Sainz, Ander Salaberria, and Dan Roth. 2023. Ixa/cogcomp at semeval-2023 task 2: Context-enriched multilingual named entity recognition using knowledge bases. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1335-1346, Toronto, Canada. Association for Computational Linguistics.

Carlos-Emiliano Gonzalez-Gallardo, Thi Hong Hanh Tran, Nancy Girdhar, Emanuela Boros, Jose G. Moreno, and Antoine Doucet. 2023. L3i++ at semeval-2023 task 2: Prompting for multilingual complex named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 807-814, Toronto, Canada. Association for Computational Linguistics.

Phu Gia Hoang, Le Thanh, and Hai-Long Trieu. 2023. Vbd_nlp at semeval-2023 task 2: Named entity recognition systems enhanced by babelnet and wikipedia. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1833-1843, Toronto, Canada. Association for Computational Linguistics.

Md Zobaer Hossain, Averie Ho Zoen So, Silviya Silwal, H. Andres Gonzalez Gongora, Ahnaf Mozib Samin, Jahedul Alam Junaed, Aritra Mazumder, Sourav Saha, and Sabiha Tahsin Soha. 2023. garner at semeval-2023: Simplified knowledge augmentation for multilingual complex named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 823-835, Toronto, Canada. Association for Computational Linguistics.

Antonia Höfer and Mina Mottahedin. 2023. Minanto at semeval-2023 task 2: Fine-tuning xlm-roberta for named entity recognition on english data. In Proceedings of the 17th International Workshop on Semantic

Evaluation, pages 1127-1130, Toronto, Canada. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1627-1643, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613.

Jiankuo Li, Zhengyi Guan, and Haiyan Ding. 2023. Janko at semeval-2023 task 2: Bidirectional 1stm model based on pre-training for chinese named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 958-962, Toronto, Canada. Association for Computational Linguistics.

Jing Li and Xiaobing Zhou. 2023. Ynunlp at semeval2023 task 2: The pseudo twin tower pre-training model for chinese named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1619-1624, Toronto, Canada. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5849-5859, Online. Association for Computational Linguistics.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced nlp tasks. arXiv preprint arXiv:1911.02855.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. CoRR, abs/1708.02002.

Jesus Lovon-Melgarejo, Jose G. Moreno, Romaric Besançon, Olivier Ferret, and Lynda Lechani. 2023. Meerqat-irit at semeval-2023 task 2: Leveraging contextualized tag descriptors for multilingual named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 878-884, Toronto, Canada. Association for Computational Linguistics.

Ruixuan Lu, Zihang Tang, Guanglong Hu, Dong Liu, and Jiacheng Li. 2023. Netease.ai at semeval-2023 task 2: Enhancing complex named entities recognition in noisy scenarios via text error correction and external knowledge. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 897-904, Toronto, Canada. Association for Computational Linguistics.

Viorica-Camelia Lupancu, Alexandru-Gabriel Platica, Cristian-Mihai Rosu, Daniela Gifu, and Diana Trandabat. 2023. Fii_better at semeval-2023 task 2: Multiconer ii multilingual complex named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1107-1113, Toronto, Canada. Association for Computational Linguistics.

Jun-Yu Ma, Jia-Chen Gu, Jiajun Qi, Zhenhua Ling, Quan Liu, and Xiaoyi Zhao. 2023a. Ustc-nelslip at semeval-2023 task 2: Statistical construction and dual adaptation of gazetteer for multilingual complex ner. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 651-659, Toronto, Canada. Association for Computational Linguistics.

Long Ma, Zeye Sun, Jiawei Jiang, and xuan li. 2023b. Pai at semeval-2023 task 4: A general multi-label classification system with class-balanced loss function and ensemble module. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 256-261, Toronto, Canada. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. Multiconer: A large-scale multilingual dataset for complex named entity recognition. In Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, pages 3798-3809. International Committee on Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Association for Computational Linguistics.

Michał Marcińczuk and Wiktor Walentynowicz. 2023. Codenlp at semeval-2023 task 2: Data augmentation for named entity recognition by combination of sequence generation strategies. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1798-1804, Toronto, Canada. Association for Computational Linguistics.

Caleb Martin, Huichen Yang, and William Hsu. 2023. Kddie at semeval-2023 task 2: External knowledge injection for named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1498-1501, Toronto, Canada. Association for Computational Linguistics.

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 62566261, Hong Kong, China. Association for Computational Linguistics.

Rahul Mehta and Vasudeva Varma. 2023. Llm-rm at semeval-2023 task 2: Multilingual complex ner using xlm-roberta. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 453-456, Toronto, Canada. Association for Computational Linguistics.

Eduards Mukans and Guntis Barzdins. 2023. Riga at semeval-2023 task 2: Ner enhanced with gpt-3. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 331-339, Toronto, Canada. Association for Computational Linguistics.

Shrimon Mukherjee, Madhusudan Ghosh, girish ., and Partha Basuchowdhuri. 2023. Mllab4cs at semeval2023 task 2: Named entity recognition in lowresource language bangla using multilingual language models. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1388-1394, Toronto, Canada. Association for Computational Linguistics.

Sangeeth N, Biswajit Paul, and Chandramani Chaudhary. 2023. Cair-nlp at semeval-2023 task 2: A multi-objective joint learning system for named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1926-1935, Toronto, Canada. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. choice, 2640:660.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. Kilt: a benchmark for knowledge intensive language tasks. arXiv preprint arXiv:2009.02252.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 487-503, Online. Association for Computational Linguistics.

Alberto Poncelas, Maksim Tkachenko, and Ohnmar Htun. 2023. Sakura at semeval-2023 task 2: Data augmentation via translation. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1718-1722, Toronto, Canada. Association for Computational Linguistics.

Amir Sartipi, Amirreza Sedighin, Afsaneh Fatemi, and Hamidreza Baradaran Kashani. 2023. Sartipisedighin at semeval-2023 task 2: Fine-grained named entity recognition with pre-trained contextual language models and data augmentation from wikipedia. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 565-579, Toronto, Canada. Association for Computational Linguistics.

Sumit Singh and Uma Shanker Tiwary. 2023. Silp_nlp at semeval-2023 task 2: Cross-lingual knowledge transfer for mono-lingual learning. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1183-1189, Toronto, Canada. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. CoRR, abs/1612.03975.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. arXiv preprint arXiv:2208.03054.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. Foundations and Trends® in Machine Learning, 4(4):267-373.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Damo-nlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 2014-2028, Toronto, Canada. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Asahi Ushio and Jose Camacho-Collados. 2021. TNER: An all-round python library for transformerbased named entity recognition. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 53-62, Online. Association for Computational Linguistics.

Rob van der Goot. 2023. Machamp at semeval-2023 tasks $2,3,4,5,7,8,9,10,11$, and 12 : On the effectiveness of intermediate training on an uncurated collection of datasets. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 230-245, Toronto, Canada. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multitask learning in NLP. In Proceedings of the 16th

Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 176-197, Online. Association for Computational Linguistics.

Harsh Verma and Sabine Bergler. 2023. Clac at semeval-2023 task 2: Comparing span-prediction and sequence-labeling approaches for ner. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 1558-1561, Toronto, Canada. Association for Computational Linguistics.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2517-2527.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entityaware self-attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6442-6454, Online. Association for Computational Linguistics.

Haojie Zhang, Xiao Li, Renhua Gu, Xiaoyan Qu, Xiangfeng Meng, Shuo Hu, and Song Liu. 2023a. Samsung research china - beijing at semeval-2023 task 2 : An al-r model for multilingual complex named entity recognition. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 114-120, Toronto, Canada. Association for Computational Linguistics.

Yuming Zhang, Hongyu Li, Yongwei Zhang, Shanshan Jiang, and Bin Dong. 2023b. Srcb at semeval-2023 task 2: A system of complex named entity recognition with external knowledge. In Proceedings of the 17th International Workshop on Semantic Evaluation, pages 671-678, Toronto, Canada. Association for Computational Linguistics.

## Appendix

## A Detailed Results for Noisy Test Sets

In this section, we provide the detailed performance for a subset of the monolingual tracks that contain a noisy test subset. For each team, we report the F1 scores for the clean subset and the subset with entity level and context level noise.

- Table 5 English (EN)
- Table 6 Italian (IT)
- Table 7 Spanish (ES)
- Table 8 French (FR)
- Table 9 Portuguese (PT)
- Table 10 Swedish (SV)
- Table 11 Chinese (ZH)

| Rank | Team | Clean Subset F1 | Noisy Subset F1 | Entity Noise F1 | Context Noise F1 | Macro F1 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | DAMO-NLP | 88.13 | 79.76 | 79.07 | 86.30 | 85.53 |
| 2 | Samsung Research China - Beijing | 85.36 | 77.94 | 77.33 | 83.74 | 83.09 |
| 3 | PAI | 86.16 | 65.41 | 63.23 | 84.74 | 80.00 |
| 4 | CAIR-NLP | 81.29 | 74.89 | 74.58 | 77.83 | 79.33 |
| 5 | KDDIE | 80.08 | 73.50 | 73.03 | 78.05 | 78.06 |
| 6 | SRCB | 79.74 | 66.21 | 65.06 | 76.56 | 75.62 |
| 7 | IXA/Cogcomp | 76.64 | 64.36 | 63.81 | 69.59 | 72.82 |
| 8 | USTC-NELSLIP | 74.87 | 65.76 | 65.35 | 69.18 | 72.15 |
| 9 | NLPeople | 76.00 | 62.23 | 60.93 | 74.52 | 71.81 |
| 10 | BizNER | 72.12 | 66.64 | 66.32 | 69.65 | 70.44 |
| 11 | Sakura | 72.86 | 64.06 | 63.79 | 66.39 | 70.16 |
| 12 | RIGA | 70.74 | 66.07 | 65.84 | 68.23 | 69.30 |
| 13 | CodeNLP | 66.04 | 57.84 | 57.58 | 60.17 | 63.51 |
| 14 | Sartipi-Sedighin | 67.10 | 54.56 | 53.68 | 62.86 | 63.25 |
| 15 | IITD | 67.52 | 53.59 | 52.82 | 60.47 | 63.21 |
| 16 | garNER | 65.25 | 56.96 | 56.73 | 58.90 | 62.73 |
| 17 | FII_Better | 65.67 | 52.74 | 51.87 | 60.60 | 61.75 |
| 18 | D2KLab | 64.72 | 53.54 | 53.12 | 57.24 | 61.29 |
| 19 | silp_nlp | 62.59 | 56.96 | 56.91 | 57.22 | 60.85 |
| 20 | Ertim | 61.85 | 52.78 | 52.76 | 52.75 | 59.03 |
| 21 | MEERQAT-IRIT | 60.46 | 54.72 | 54.68 | 55.03 | 58.70 |
| 22 | LSJSP | 60.67 | 50.48 | 50.44 | 50.59 | 57.51 |
| 23 | RGAT | 61.29 | 47.15 | 46.56 | 52.04 | 56.91 |
| 24 | CLaC | 57.68 | 49.06 | 48.91 | 50.26 | 55.05 |
| 25 | L3i++ | 55.87 | 46.70 | 46.47 | 48.68 | 53.00 |
| 26 | VBD_NLP | 57.00 | 42.44 | 41.45 | 51.06 | 52.65 |
| 27 | LLM-RM | 54.73 | 46.30 | 46.17 | 47.45 | 52.08 |
| 28 | Minanto | 53.43 | 47.00 | 47.03 | 46.48 | 51.47 |
| 29 | SAB | 54.28 | 44.96 | 44.82 | 46.16 | 51.41 |
| 30 | ShathaTaymaaTeam | 52.34 | 44.78 | 45.09 | 41.67 | 50.02 |
| 31 | azaad@BND | 50.09 | 41.28 | 41.09 | 42.94 | 47.42 |
| 32 | LISAC FSDM-USMBA | 47.36 | 36.58 | 36.27 | 39.32 | 44.00 |
| 33 | YNU-HPCC | 29.95 | 25.31 | 25.31 | 25.21 | 28.52 |
| 34 | IXA | 16.88 | 11.84 | 11.49 | 15.09 | 15.39 |

Table 5: Detailed results for the English track.

| Rank | Team | Clean Subset F1 | Noisy Subset F1 | Entity Noise F1 | Context Noise F1 | Macro F1 |
| :---: | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | DAMO-NLP | 91.85 | 85.89 | 85.30 | 90.99 | 89.79 |
| 2 | PAI | 88.94 | 76.53 | 75.14 | 89.56 | 84.88 |
| 3 | CAIR-NLP | 85.08 | 81.00 | 80.58 | 84.63 | 83.78 |
| 4 | BizNER | 77.24 | 74.81 | 74.34 | 79.33 | 76.48 |
| 5 | USTC-NELSLIP | 78.06 | 70.65 | 70.05 | 75.91 | 75.70 |
| 6 | IXA/Cogcomp | 78.16 | 67.66 | 66.77 | 75.95 | 74.67 |
| 7 | Sakura | 76.67 | 69.03 | 68.53 | 73.18 | 74.19 |
| 8 | NLPeople | 77.45 | 65.88 | 64.58 | 78.87 | 73.71 |
| 9 | garNER | 70.16 | 63.99 | 63.53 | 67.81 | 68.20 |
| 10 | D2KLab | 68.17 | 57.68 | 57.07 | 63.15 | 64.77 |
| 11 | Sartipi-Sedighin | 67.61 | 57.95 | 57.16 | 65.31 | 64.50 |
| 12 | silp_nlp | 64.53 | 60.13 | 60.00 | 61.00 | 63.11 |
| 13 | SAB | 60.36 | 51.60 | 51.15 | 55.56 | 57.57 |
| 14 | FII_Better | 60.32 | 47.85 | 46.76 | 58.36 | 56.36 |
| 15 | IXA | 20.05 | 14.82 | 14.38 | 18.84 | 18.41 |

Table 6: Detailed results for the Italian track.

| Rank | Team | Clean Subset F1 | Noisy Subset F1 | Entity Noise F1 | Context Noise F1 | Macro F1 |
| :---: | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | DAMO-NLP | 91.74 | 85.81 | 85.29 | 91.06 | 89.78 |
| 2 | CAIR-NLP | 85.03 | 80.66 | 80.44 | 82.92 | 83.63 |
| 3 | USTC-NELSLIP | 77.25 | 68.52 | 68.00 | 73.73 | 74.44 |
| 4 | IXA/Cogcomp | 77.65 | 66.09 | 65.48 | 72.30 | 73.81 |
| 5 | Sakura | 75.42 | 67.39 | 66.93 | 72.01 | 72.85 |
| 6 | NLPeople | 77.22 | 63.53 | 62.43 | 74.76 | 72.76 |
| 7 | PAI | 79.35 | 55.25 | 53.16 | 75.10 | 71.67 |
| 8 | BizNER | 72.60 | 69.11 | 69.08 | 69.53 | 71.48 |
| 9 | garNER | 66.19 | 58.43 | 58.21 | 60.35 | 63.73 |
| 10 | D2KLab | 66.69 | 55.75 | 55.26 | 60.47 | 63.17 |
| 11 | silp_nlp | 64.88 | 58.77 | 58.66 | 59.71 | 62.90 |
| 12 | MEERQAT-IRIT | 63.04 | 56.42 | 56.16 | 58.78 | 60.93 |
| 13 | LSJSP | 63.39 | 54.46 | 54.26 | 56.21 | 60.55 |
| 14 | Sartipi-Sedighin | 62.27 | 50.25 | 49.70 | 55.57 | 58.41 |
| 15 | LLM-RM | 57.42 | 49.32 | 49.19 | 50.47 | 54.81 |
| 16 | FII_Better | 58.96 | 44.77 | 43.57 | 56.07 | 54.51 |
| 17 | SAB | 50.83 | 42.62 | 42.58 | 42.87 | 48.22 |
| 18 | IXA | 17.65 | 12.16 | 11.83 | 14.59 | 16.01 |

Table 7: Detailed results for the Spanish track.

| Rank | Team | Clean Subset F1 | Noisy Subset F1 | Entity Noise F1 | Context Noise F1 | Macro F1 |
| :---: | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | DAMO-NLP | 91.62 | 85.14 | 84.58 | 90.49 | 89.59 |
| 2 | PAI | 89.50 | 78.71 | 77.64 | 88.96 | 86.17 |
| 3 | CAIR-NLP | 84.67 | 79.54 | 79.22 | 82.55 | 83.08 |
| 4 | BizNER | 79.09 | 75.63 | 75.29 | 78.92 | 78.01 |
| 5 | IXA/Cogcomp | 78.60 | 65.81 | 64.93 | 74.14 | 74.52 |
| 6 | USTC-NELSLIP | 76.81 | 68.49 | 67.93 | 73.55 | 74.25 |
| 7 | Sakura | 75.58 | 66.86 | 66.38 | 71.26 | 72.86 |
| 8 | NLPeople | 77.12 | 63.40 | 62.02 | 76.51 | 72.85 |
| 9 | Ertim | 69.73 | 58.60 | 57.77 | 66.09 | 66.30 |
| 10 | garNER | 68.09 | 60.22 | 59.77 | 64.27 | 65.68 |
| 11 | D2KLab | 67.70 | 56.05 | 55.30 | 63.12 | 64.09 |
| 12 | silp_nlp | 64.40 | 58.04 | 57.81 | 60.02 | 62.39 |
| 13 | MEERQAT-IRIT | 61.29 | 53.65 | 53.23 | 57.32 | 58.90 |
| 14 | LSJSP | 58.74 | 52.60 | 52.34 | 54.93 | 56.83 |
| 15 | SAB | 57.98 | 48.61 | 48.19 | 52.41 | 55.07 |
| 16 | Sartipi-Sedighin | 56.99 | 50.40 | 50.22 | 52.05 | 54.94 |
| 17 | IXA | 18.90 | 13.89 | 13.49 | 17.29 | 17.40 |

Table 8: Detailed results for the French track.

| Rank | Team | Clean Subset F1 | Noisy Subset F1 | Entity Noise F1 | Context Noise F1 | Macro F1 |
| :---: | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | DAMO-NLP | 87.33 | 83.38 | 83.04 | 88.00 | 85.97 |
| 2 | PAI | 84.56 | 76.12 | 75.50 | 82.87 | 81.61 |
| 3 | CAIR-NLP | 81.73 | 77.10 | 76.94 | 78.61 | 80.16 |
| 4 | BizNER | 74.36 | 70.35 | 70.12 | 72.81 | 72.97 |
| 5 | IXA/Cogcomp | 76.00 | 65.54 | 64.91 | 72.32 | 72.28 |
| 6 | USTC-NELSLIP | 74.04 | 65.91 | 65.49 | 70.37 | 71.26 |
| 7 | Deep Learning Brasil | 72.07 | 68.91 | 68.75 | 70.11 | 70.97 |
| 8 | NLPeople | 74.50 | 62.22 | 61.27 | 73.48 | 70.16 |
| 9 | Sakura | 72.74 | 64.76 | 64.29 | 69.57 | 69.98 |
| 10 | garNER | 66.81 | 60.04 | 59.82 | 61.52 | 64.51 |
| 11 | Sartipi-Sedighin | 63.75 | 56.57 | 56.30 | 59.32 | 61.28 |
| 12 | silp_nlp | 63.07 | 57.23 | 56.99 | 59.93 | 61.05 |
| 13 | D2KLab | 64.44 | 53.98 | 53.47 | 59.01 | 60.79 |
| 14 | MEERQAT-IRIT | 61.82 | 56.17 | 56.01 | 58.00 | 59.87 |
| 15 | LSJSP | 60.63 | 53.60 | 53.32 | 56.23 | 58.23 |
| 16 | SAB | 57.55 | 47.56 | 47.21 | 51.08 | 54.12 |
| 17 | IXA | 18.40 | 13.91 | 13.61 | 17.95 | 16.97 |

Table 9: Detailed results for the Portuguese track.

| Rank | Team | Clean Subset F1 | Noisy Subset F1 | Entity Noise F1 | Context Noise F1 | Macro F1 |
| :---: | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | DAMO-NLP | 91.08 | 86.76 | 86.34 | 91.41 | 89.57 |
| 2 | CAIR-NLP | 84.54 | 79.75 | 79.49 | 80.75 | 82.88 |
| 3 | IXA/Cogcomp | 80.75 | 68.69 | 67.99 | 74.81 | 76.54 |
| 4 | BizNER | 77.23 | 73.87 | 73.54 | 77.78 | 76.12 |
| 5 | USTC-NELSLIP | 78.51 | 69.64 | 69.22 | 72.87 | 75.47 |
| 6 | NLPeople | 79.31 | 67.15 | 66.30 | 75.22 | 75.08 |
| 7 | Sakura | 76.74 | 68.12 | 67.66 | 71.74 | 73.79 |
| 8 | PAI | 81.53 | 55.22 | 53.04 | 77.04 | 72.38 |
| 9 | garNER | 70.40 | 62.19 | 61.86 | 66.01 | 67.63 |
| 10 | silp_nlp | 67.15 | 60.87 | 60.53 | 63.74 | 65.00 |
| 11 | LSJSP | 67.23 | 58.63 | 58.18 | 64.13 | 64.36 |
| 12 | D2KLab | 66.78 | 55.80 | 55.29 | 61.14 | 62.98 |
| 13 | Sartipi-Sedighin | 64.69 | 56.57 | 56.15 | 60.38 | 61.95 |
| 14 | SAB | 61.58 | 51.14 | 50.90 | 52.90 | 58.03 |
| 15 | FII_Better | 56.66 | 43.11 | 42.20 | 50.67 | 52.12 |
| 16 | IXA | 27.96 | 21.72 | 21.30 | 26.72 | 25.96 |

Table 10: Detailed results for the Swedish track.

| Rank | Team | Clean Subset F1 | Noisy Subset F1 | Entity Noise F1 | Context Noise F1 | Macro F1 |
| :---: | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | NetEase.AI | 88.47 | 69.05 | 67.80 | 86.43 | 84.05 |
| 2 | DAMO-NLP | 82.91 | 54.32 | 52.26 | 81.45 | 75.98 |
| 3 | SRCB | 87.19 | 39.39 | 35.37 | 88.37 | 75.86 |
| 4 | PAI | 86.23 | 41.90 | 38.57 | 85.31 | 74.87 |
| 5 | Taiji | 75.70 | 61.39 | 60.53 | 72.28 | 72.52 |
| 6 | USTC-NELSLIP | 70.10 | 55.06 | 54.16 | 64.75 | 66.57 |
| 7 | NLPeople | 71.43 | 48.95 | 47.91 | 61.57 | 65.96 |
| 8 | IXA/Cogcomp | 70.35 | 48.37 | 47.36 | 59.88 | 64.86 |
| 9 | Sakura | 68.79 | 51.20 | 50.43 | 59.73 | 64.61 |
| 10 | garNER | 67.50 | 50.17 | 49.57 | 58.23 | 63.47 |
| 11 | Ertim | 64.26 | 44.38 | 43.26 | 59.44 | 59.45 |
| 12 | Sartipi-Sedighin | 62.60 | 46.46 | 46.10 | 49.10 | 58.70 |
| 13 | CAIR-NLP | 62.89 | 44.74 | 43.84 | 56.16 | 58.43 |
| 14 | Janko | 62.45 | 44.70 | 44.14 | 52.15 | 57.90 |
| 15 | YNUNLP | 61.45 | 42.69 | 42.17 | 50.29 | 56.57 |
| 16 | D2KLab | 58.75 | 43.34 | 42.67 | 48.12 | 54.92 |
| 17 | silp_nlp | 54.65 | 42.11 | 41.57 | 48.95 | 51.65 |
| 18 | SAB | 47.71 | 33.37 | 32.46 | 42.82 | 44.12 |
| 19 | NCUEE-NLP | 51.36 | 18.24 | 15.39 | 43.74 | 44.09 |
| 20 | L3i++ | 38.02 | 27.13 | 26.63 | 32.99 | 35.34 |
| 21 | YNU-HPCC | 34.24 | 24.07 | 23.52 | 32.50 | 31.66 |
| 22 | IXA | 8.06 | 4.49 | 4.35 | 5.20 | 6.93 |

Table 11: Detailed results for the Chinese track.

## B Fine-Grained Results Analysis

Figure 3 shows the misclassification across the different fine-grained types for the baseline approach on the EN test set. An ideal classifier would have a $100 \%$ performance on the diagonal.

CW. For this class, the baseline has low recall, with many of the entities being missed ( O tag). In terms of misclassifying the fine-grained types, we note that the highest confusion is between MUSICALWORK and VisualWork, with $7.4 \%$ of false positives.

GRP. In the case of GRP, we notice a high confusion between ORG, PUBLICCORP and Privatecorp, with error rates going up to $26.3 \%$. This highlights the difficulty of the different fine-grained classes, where context capture is important. Even more importantly in this particular problem of fine-grained NER, external knowledge or world knowledge of entities is crucial to distinguish between such fine-grained differences. In this case, external knowledge about different corporations may be necessary to correctly distinguish between different named entity types.

LOC. For this class, most of the errors are between FACILITY and OTHERLOC.
PER. In the case of PER, SportsMAnAGER is confused as Athlete in $41.2 \%$ of the cases (this is because many sports managers are former athletes). The PER coarse type is highly challenging in some of the fine-grained types, given that the surface forms can be highly ambiguous, and only the context can differentiate between the different types (Athlete, Scientist, Artist, etc.)

MED. In this case, we notice a high confusion between DISEASE and SYmptom, with $21.6 \%$. This is an interesting insights, given that often, names for diseases and symptoms are used interchangeably (i.e., a symptom may cause a disease that is referred using the same name).

PROD. Finally, here we notice that DRINK and FoOD are often confused with each other with $10.7 \%$. This highlights some of the ambiguous cases where a drink may be considered both, e.g. milk. Finally, the most misclassification happen between VEHICLE and OTHERPROD. A potential cause for this is the lack of detailed type assignment of entities in Wikidata, which may lead to such misclassifications, i.e. OTHERPROD entities may actually belong to VEHICLE, however they are not explicitly associated with this type in Wikidata.
$\begin{array}{llllllllllllllllllllllllllllllllllllllllllllll} \\ \text { MusicalWork -58.6 } & 0.3 & 2.1 & 0.5 & 7.4 & 0.2 & 0.5 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 1.5 & 0.0 & 0.0 & 0.2 & 0.2 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 1.5 & 0.1 & 0.1 & 0.0 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 26.0\end{array}$







 $\begin{array}{lllllllllllllllllllllllllllllllllllllllllll}\text { SportsManager - } \begin{array}{llllllllllllll} & 0.0 & 0.0 & 0.0 & 0.1 & 0.2 & 0.0 & 0.1 & 0.0 & 0.0 & 40.6 & 41.2 & 0.1 & 3.8 \\ 3.1 & 0.2 & 7.2 & 0.3 & 0.0 & 1.1 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0\end{array} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.8\end{array}$ $\begin{array}{lllllllllllllllllllllllllllllllllllllllllll}\text { Athlete - } 0.0 & 0.0 & 0.1 & 0.0 & 0.2 & 0.1 & 0.1 & 0.0 & 0.0 & 3.3 & 74.7 & 0.1 & 8.0 & 1.3 & 0.3 & 8.5 & 0.2 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.7\end{array}$
 $\begin{array}{llllllllllllllllllllllllllllllllllllllllllll}\text { Artist }-0.2 & 0.0 & 0.3 & 0.0 & 0.6 & 0.1 & 0.1 & 0.0 & 0.0 & 0.1 & 2.0 & 0.8 & 78.1 & 3.0 & 0.8 & 9.1 & 0.4 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 1.2 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 2.7\end{array}$

 $\begin{array}{lllllllllllllllllllllllllllllllllllllll}\text { OtherPER- } 0.1 & 0.1 & 0.7 & 0.1 & 0.6 & 0.2 & 0.6 & 0.0 & 0.0 & 0.4 & 7.6 & 3.2 & 17.8 & 8.9 & 4.7 & 48.6 & 0.9 & 0.0 & 0.2 & 0.1 & 0.0 & 0.1 & 0.2 & 0.0 & 0.0 & 0.1 & 0.1 & 0.1 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 4.5\end{array}$

 $\begin{array}{lllllllllllllllllllllllllllllllllllllllllllll}\text { SportsGRP }-0.0 & 0.0 & 0.1 & 0.2 & 0.5 & 0.4 & 3.5 & 0.1 & 0.3 & 0.3 & 1.6 & 0.0 & 0.2 & 0.2 & 0.0 & 0.3 & 3.7 & 0.1 & 78.4 & 0.0 & 0.3 & 1.0 & 0.5 & 0.0 & 0.0 & 0.3 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 8.1\end{array}$ AerospaceManu. - $\begin{array}{llllllllllllllllllllllllllllllllllllllll} & 0.0 & 0.0 & 0.2 & 0.2 & 0.1 & 0.5 & 1.6 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.7 & 0.5 & 10.1 & 0.3 & 0.7 & 62.2 & 7.1 & 2.4 & 0.1 & 0.0 & 0.0 & 2.6 & 2.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 7.6\end{array}$

 | CarManufacturer - 0.0 | 0.0 | 0.1 | 0.3 | 0.2 | 0.5 | 1.2 | 0.0 | 0.0 | 0.0 | 0.8 | 0.0 | 0.2 | 0.2 | 0.1 | 0.3 | 6.5 | 0.3 | 3.5 | 3.0 | 6.1 | 59.1 | 0.1 | 0.1 | 0.2 | 5.2 | 3.4 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 8.4 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | $\begin{array}{lllllllllllllllllllllllllllllllllllllllllll}\text { MusicalGRP - } 3.6 & 0.0 & 0.5 & 0.5 & 4.2 & 1.3 & 0.5 & 0.0 & 0.1 & 0.0 & 0.4 & 0.0 & 8.7 & 0.1 & 0.0 & 0.3 & 3.9 & 0.2 & 0.7 & 0.1 & 0.5 & 0.1 & 56.1 & 0.0 & 0.1 & 0.1 & 0.3 & 0.2 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 17.4\end{array}$



 Vehicle - $\begin{array}{llllllllllllllllllllllllllllllllllllllll} & 0.1 & 0.2 & 0.2 & 0.4 & 0.7 & 0.6 & 1.0 & 0.4 & 0.6 & 0.0 & 0.4 & 0.1 & 0.2 & 0.3 & 0.1 & 1.0 & 2.2 & 0.0 & 1.0 & 1.0 & 0.3 & 2.7 & 0.1 & 0.1 & 0.1 & 46.9 & 12.5 & 0.1 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 26.7\end{array}$ | OtherPROD -0.2 | 0.2 | 0.4 | 2.3 | 0.4 | 1.2 | 0.4 | 0.2 | 0.5 | 0.0 | 0.1 | 0.0 | 0.2 | 0.1 | 0.1 | 0.3 | 1.6 | 0.0 | 0.3 | 0.5 | 0.7 | 1.4 | 0.2 | 0.8 | 0.2 | 8.3 | 43.3 | 0.5 | 0.3 | 0.1 | 0.0 | 0.5 | 0.5 | 34.5 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |

 $\begin{array}{lllllllllllllllllllllllllllllllllllllllllll}\text { Medication- } 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.1 & 0.4 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.2 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.1 & 0.6 & 0.0 & 0.7 & 1.2 & 71.1 & 0.5 & 0.1 & 0.4 & 1.5 & 22.6\end{array}$ $\begin{array}{lllllllllllllllllllllllllllllllllllllllllll}\text { Disease - } 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.5 & 61.3 & 6.0 & 1.1 & 0.8 & 29.6\end{array}$ $\begin{array}{llllllllllllllllllllllllllllllllllllllllllll}\text { Symptom }-0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3 & 21.6 & 43.1 & 0.5 & 0.8 & 33.4\end{array}$
 MedicalProcedure - $\left.\begin{array}{lllllllllllllllllll} & 0.0 & 0.0 & 0.3 & 0.3 & 0.0 & 0.3 & 0.2 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.3 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.2 & 0.0 & 0.0 & 1.6 & 0.1 & 0.8 & 0.8 & 0.5 & 0.7 & 50.9 & 42.1\end{array}\right]$


Figure 3: Confusion matrix of baseline performance computed at the fine-grained level for the EN test set.


[^0]:    ${ }^{1}$ https://multiconer.github.io

[^1]:    ${ }^{2}$ https://registry.opendata.aws/multiconer

[^2]:    ${ }^{3}$ We extended the keyboard layouts in this library to include 7 languages: https://github.com/ranvijaykumar/typo

[^3]:    ${ }^{4}$ https://github.com/amzn/multiconer-baseline

