Sartipi-Sedighin at SemEval-2023 Task 2: Fine-grained Named Entity Recognition with Pre-trained Contextual Language Models and Data Augmentation from Wikipedia

Amir Sartipi, Amirreza Sedighin, Afsaneh Fatemi, and **Hamidreza Baradaran Kashani**
Faculty of Computer Engineering, University of Isfahan, Iran
{amirsartipi.msc, amirreza.seddighin1376, a_fatemi, hrb.kashani}@eng.ui.ac.ir

Abstract

This paper presents the system developed by the Sartipi-Sedighin team for SemEval 2023 Task 2, which is a shared task focused on multilingual complex named entity recognition (NER), or MultiCoNER II. The goal of this task is to identify and classify complex named entities (NEs) in text across multiple languages. To tackle the MultiCoNER II task, we leveraged pre-trained language models (PLMs) fine-tuned for each language included in the dataset. In addition, we also applied a data augmentation technique to increase the amount of training data available to our models. Specifically, we searched for relevant NEs that already existed in the training data within Wikipedia, and we added new instances of these entities to our training corpus. Our team achieved an overall F1 score of 61.25% in the English track and 71.79% in the multilingual track across all 13 tracks of the shared task that we submitted to.

1 Introduction

The MultiCoNER 2023 task 2 was initiated with the purpose of developing NER systems that can accurately detect fine-grained NEs across multiple languages. The shared task was organized into 13 tracks, with 12 monolingual tracks and one multilingual track, to facilitate a thorough evaluation of the participating systems (Fetahu et al., 2023b). Despite the inherent complexity and ambiguity of the dataset instances, the task presented two main features that are worth mentioning. The first feature was the identification of fine-grained NEs, which required the systems to detect and classify a wide range of entities with varying levels of specificity. The second feature involved the augmentation of test data for some languages with simulated errors to increase the difficulty and realism of the task (Fetahu et al., 2023a). These features posed significant challenges for the participating systems and necessitated the use of advanced NLP techniques. The work presented in this paper makes two main contributions to the field of NER.

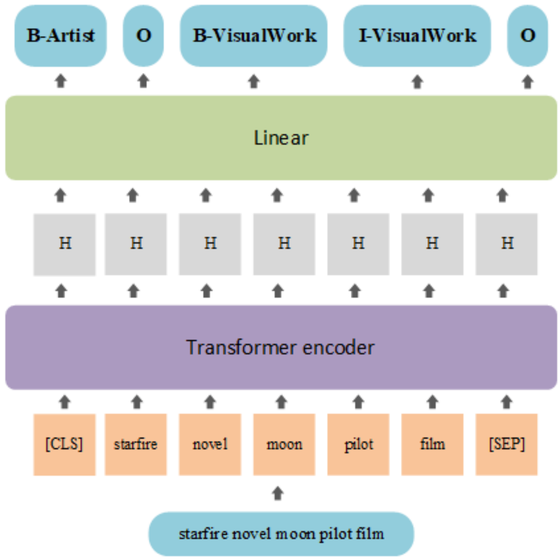

Figure 1: Overall process of fine-tuning NER system

1. We introduce a simple yet effective method for increasing the number of instances in training datasets.
2. We fine-tune (PLMs) for each language in both the multilingual track and monolingual tracks using both the original dataset and the augmented version.

The overall architecture of the model used for fine-tuning can be seen in Figure 1.

2 Related Work

NER is a natural language processing (NLP) task that involves identifying and classifying NEs in text, such as person names, organization names, location names, and others, into predefined categories (Grishman and Sundheim, 1996). NER is widely used in many NLP applications, such as information extraction (Tan, 2022), text summarization (Khademi and Fakhredanesh, 2020), and question answering (McKenna et al., 2021; Mollá

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 565–579
July 13-14, 2023 ©2023 Association for Computational Linguistics

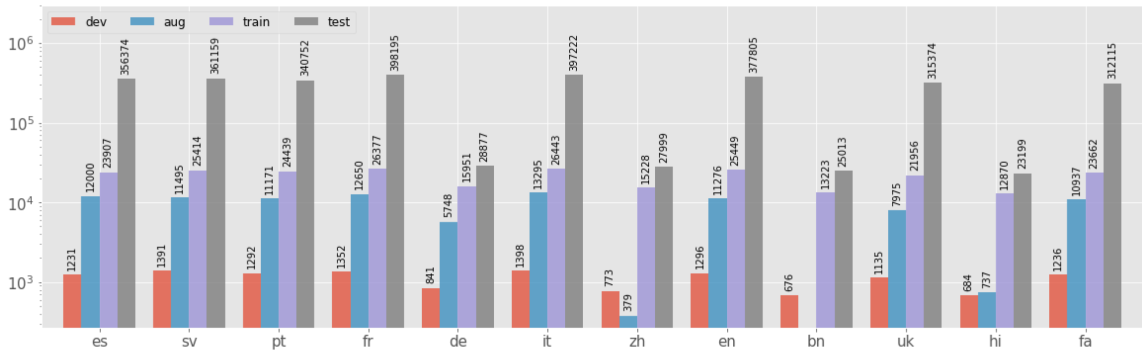

Figure 2: Number of instances for training, development, test, and augmentation set per languages

* The zoomed versions of the pictures are included in the appendix A

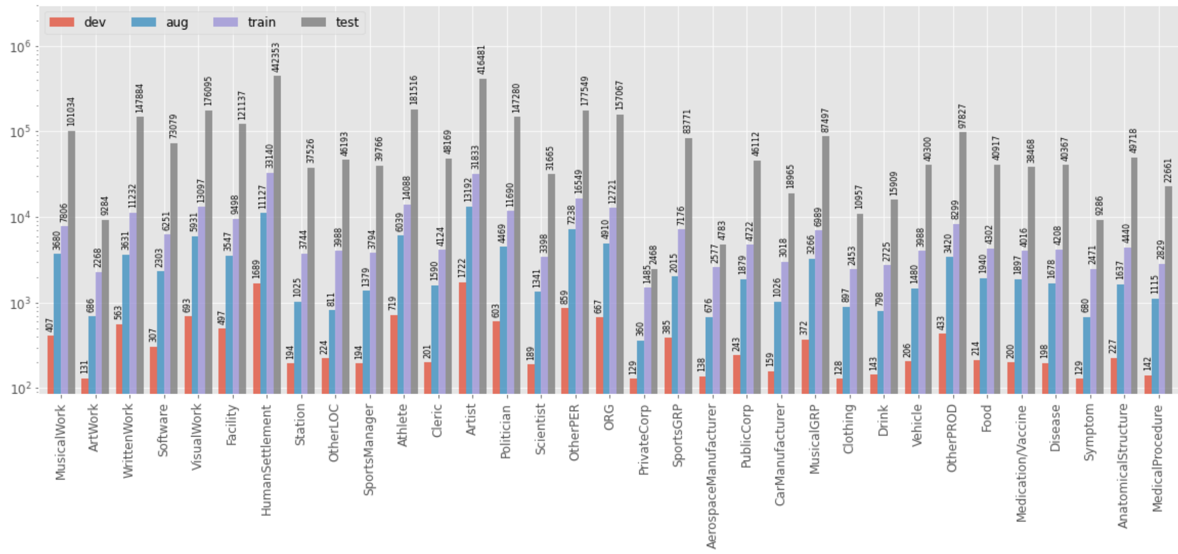

Figure 3: Number of NEs for training, development, test, and augmentation set per fine-grained labels

et al., 2006). Fine-grained NER is a more specific variant of NER that aims to recognize more detailed categories of NEs (Tedeschi and Navigli, 2022). Moreover, various NER datasets have been released in both coarse-grained (Malmasi et al., 2022a; Tjong Kim Sang and De Meulder, 2003; Derczynski et al., 2017) and fine-grained (Fetahu et al., 2023a; Xu et al., 2020; Tedeschi and Navigli, 2022) domains. Additionally, there exists an automatic translation of popular NER benchmarks, for cross-lingual NER evaluation (Sartipi and Fatemi, 2023).

MultiCoNER was initially introduced as part of SemEval 2022 Task 11 with the objective of developing multilingual (NER) systems capable of identifying coarse-grained entities. The competition featured a total of 13 tracks, comprising 11 monolingual tracks, one code-mixed track, and one multilingual track (Malmasi et al., 2022b). The MultiCoNER dataset is an extensive multilingual dataset for (NER) that includes three domains: Wiki sentences, questions, and search queries. The dataset is designed to address modern NER challenges, including low-context scenarios, such as short and uncased text, complex entities like movie titles, and long-tail entity distributions (Malmasi et al., 2022a).

In its second iteration, MultiCoNER 2023 aimed to build NER systems capable of identifying NEs across 12 languages, including English (EN), Spanish (ES), Hindi (HI), Bangla (BN), Chinese (ZH), Swedish (SV), Farsi (FA), French (FR), Italian (IT), Portuguese (PT), Ukrainian (UK), and German (DE). The shared task was subdivided into 13 tracks, comprising 12 monolingual tracks and one multilingual track. Two main features of this task are worthy of mention: firstly, the identification of fine-grained NEs, such as Symptom, Politician, and WrittenWork. Secondly, for some languages, namely English, Chinese,

Italian, Spanish, German, French, Portuguese, and Swedish, the test data was augmented with simulated errors to increase the difficulty and realism of the task (Fetahu et al., 2023b).
(Meng et al., 2021) presents several challenges that current datasets and models do not adequately address. These challenges include short-text inputs, long-tail entity distributions, emerging entity types, and complex entities that are linguistically difficult to parse. These challenges pose problems for current NER systems, which are primarily trained on news texts with long sentences that discuss multiple entities. To overcome these challenges, the authors build gazetteers that incorporate external knowledge and contextual information, which is represented using transformers such as BERT. Contextual features from BERT and gazetteers are combined through a fusion process, and the resulting features are then fed into a conditional random field (CRF) layer. This enables the model to incorporate both external knowledge from gazetteers and contextual information from BERT to better handle the challenges. To extend these challenges to multilingual and code-mixed settings, Fetahu et al. (2021) have introduced two datasets: mLOWNER, a multilingual NER training dataset for short texts in six languages, and EMBER, a code-mixed NER dataset covering the same languages as mLOWNER. These datasets can assist in training models to recognize complex NEs and provide a basis for evaluating the models' performance which is included in MultiCoNER.

3 Data

This section will discuss how we increased the number of training instances and some statistics about data.

Augmentation In order to augment the dataset and fine-tune our NER models, we utilized the Wikipedia python library [1] to generate additional instances for some of the shorter instances in the dataset. To accomplish this, we constructed sets of entities from the existing entities in each language, excluding instances labeled as "O". We then used the Wikipedia library to search for these entities, which provided a corresponding paragraph for each entity. In order to segment these paragraphs into sentences, we leveraged Stanza (Qi et al., 2020). For each paragraph, we selected one sentence containing the entity and positioned in the middle or end of the sentence, rather than the beginning. Subsequently, we assigned the "O" tag to the other tokens in the sentence and labeled the corresponding fine-grained category for the entity. We followed this process for all languages, with the exception of BN where no sentence segmenter was available in Stanza. Our aim was to maintain consistency in approach across all languages.
It is important to note that certain entities in Wikipedia had multiple descriptions available, but we opted to utilize only one for the sake of simplicity. Given the time-consuming nature of searching for each entity in Wikipedia, we employed Dask (Rocklin, 2015) to expedite the search process. With the aid of Dask, it took approximately two hours for each language to search all entities. The data presented in Figure 5 include one primary instance for each language, along with an augmented version of that instance. These datasets are publicly available in this[2] Git repository.
The primary motivation behind this work is to increase the diversity of instances that are used for training. Additionally, when a NE appears in a short sentence or with limited context, generating

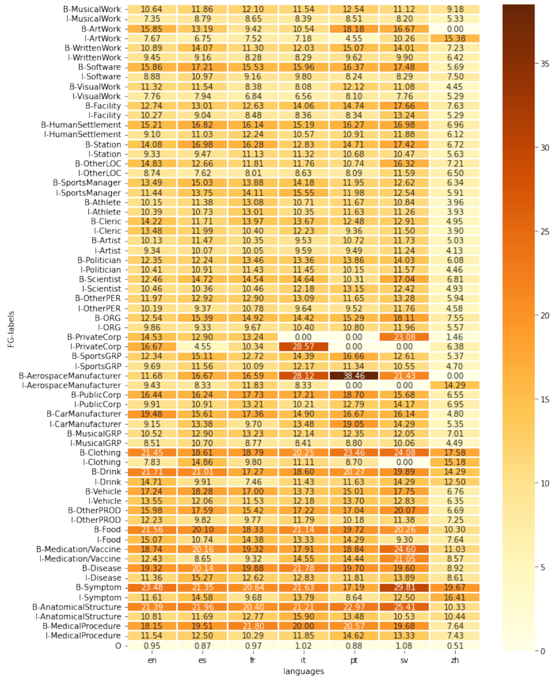

Figure 4: Heat map percentage of corrupted instances in test data for each fine-grained class

[1] https://github.com/goldsmith/Wikipedia

[2] https://github.com/amirsartipi13/MultiCoNER-aug.git

Figure 5: Examples of an instance in training data (Main) and corresponding augmented instance (Aug) separated for each language.

more instances of that NE with different contexts can help to provide a more comprehensive understanding of its meaning and usage. This can improve the model's ability to correctly identify and classify NEs in a variety of different contexts.

Data statistics Table 4 presents an overview of the different sets, while Table 3 provides detailed information about NEs for each category. It is observed that certain categories, such as HumanSettlement and Artist, have a greater number of NEs compared to other classes. Conversely, some classes, such as ArtWork, PrivateCorp, and Clothing, have a notably lower number of NEs. This leads to the conclusion that the classes are not balanced in the training data. The imbalance of data may potentially result in biased predictions during the training process.
Out of all the test sets, corrupted data was found in six languages, namely EN, ES, Fr, IT, PT, SV,

Model Name	Lang
bert-base-spanish-wwm-uncased (Cañete et al., 2020)	ES
bert-base-german-uncased [3]	DE
roberta-hindi [4]	HI
chinese-roberta-wwm-ext (Cui et al., 2020)	ZH
bert-base-swedish-cased (Malmsten et al., 2020)	SV
bert-base-italian-xxl-uncased (Schweter, 2020b)	IT
bert-large-portuguese-cased (Souza et al., 2020)	PT
bert-base-french-europeana-cased (Schweter, 2020a)	FR
banglabert (Bhattacharjee et al., 2022)	BN
roberta-large-wechsel-ukrainian [5]	UK
deberta-v3-large (He et al., 2021)	EN
bert-base-parsbert-uncased (Farahani et al., 2021)	FA
xlm-roberta-large (Conneau et al., 2019)	MULTI

Table 1: Pre-trained models that are used

and ZH. Figure 4 displays the percentage of corrupted data for each fine-grained named entity in each language. For example, no corrupted data was found in certain classes such as PrivateCorp in IT and PT. On the other hand, the highest corruption rate was observed in the B-AerospaceManufacturer class for PT.

4 Methodology

In recent years, transformer-based models such as BERT (Devlin et al., 2019) have revolutionized the field of NLP, resulting in significant improvements in NER performance. These models are pre-trained on massive amounts of text data, enabling them to capture complex patterns and relationships between words in the text. They generate highly contextualized embeddings for each token in a sentence, allowing them to understand the meaning of words in context. To leverage the power of these models, we fine-tuned a PLM for each language on the training data.

Hyper-parameters: We used the same Hyper-parameters for all of our experiments. To train our model we used the Hugging Face (Wolf et al., 2020) trainer and all models were trained for 15 epochs and saved the best model according to lower validation loss. We set 32, 2e-5, and 0.01 for batch size, learning rate, and weight decay, respectively.

Fine-tuning During this phase, we utilized transformer-based encoders. The models used for fine-tuning in the evaluation phase are listed in Ta-

[3] https://huggingface.co/dbmdz/bert-base-german-uncased

[4] https://huggingface.co/flax-community/roberta-hindi

[5] https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian

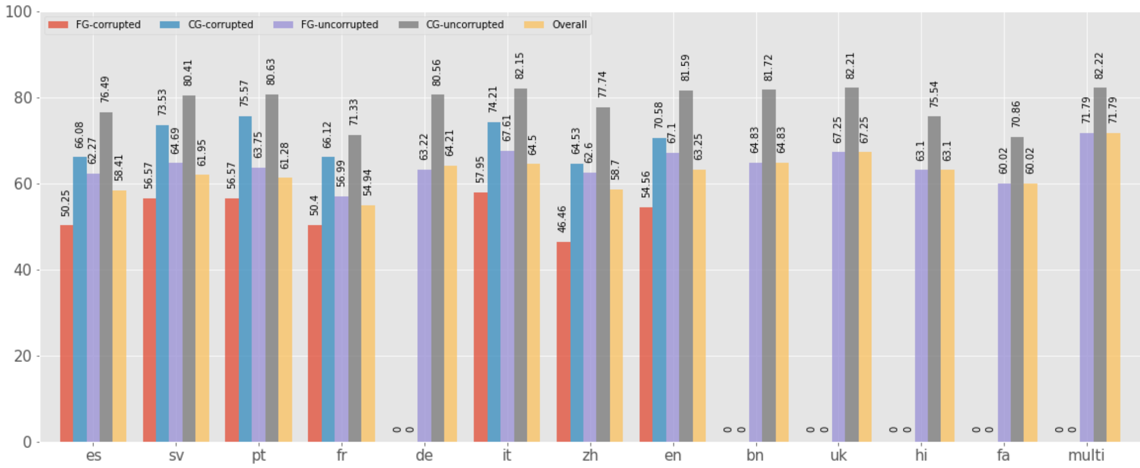

Figure 6: Overall best system results for fine-grained (FG) and coarse-grained (CG)

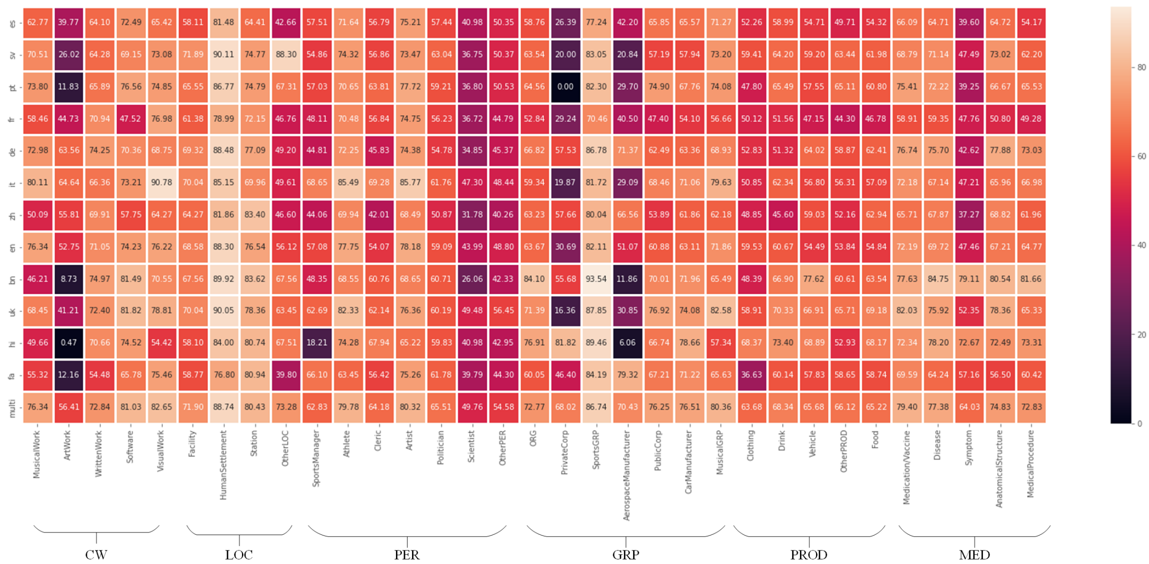

Figure 7: Heat map of the base system which is trained on main training data coarse-grained classes are Creative Works (CW), Location (LOC), Person (PER), Group (GRP), Product (PROD), Medical (MED)

ble 1. Additionally, we also fine-tuned the Roberta-Large (Liu et al., 2019) and Bert-Large-Uncased (Devlin et al., 2018) models during the practice phase, achieving F1 scores of 63.03% and 65.13%, respectively. However, the DeBERTa-v3-Large model yielded a higher F1 score of 65, leading us to choose this model for further analysis.

5 Results and Analysis

In this section, we present the official results as reported by the organizers.

Base Model Figure 7 presents an analysis of the performance of base systems, which were trained on the training data without any augmentation. The results show that recognizing ArtWork and MusicalWork proved to be more challenging within the Creative Work class. Similarly, OtherLoc emerged as the most difficult entity to detect within the Location class. In the Person class, the names of Scientists and OtherPersons were found to be the most challenging entities, while HumanSettlement was more easily recognizable. Moreover, subclasses such as PrivateCorp and AerospaceManufacture within the Group class were particularly demanding, whereas SportGRP had the best f1 values. These findings highlight the categories and languages in which NER systems struggle to accurately detect named entities. Figure 8 illustrates the discrepancies between base systems and augmentation. The data in this table indicates that increasing

CG	FG	model	es	sv	pt	fr	de	it	zh	en	bn	uk	hi	fa	multi
CW	MusicalWork	base	62.77	70.51	73.80	58.46	72.98	80.11	50.09	76.34	46.21	68.45	49.66	55.32	76.34
		aug	59.72	69.61	71.39	57.66	72.06	78.01	50.19	75.55	-	66.82	50.68	53.39	74.47
	ArtWork	base	39.77	26.02	11.83	44.73	63.56	64.64	55.81	52.75	8.73	41.21	0.47	12.16	56.41
		aug	36.02	25.87	12.28	44.87	64.00	63.23	53.00	50.85	-	37.25	1.38	9.91	53.90
	WrittenWork	base	64.10	64.28	65.89	70.94	74.25	66.36	69.91	71.05	74.97	72.40	70.66	54.48	72.84
		aug	63.38	65.02	64.22	70.18	73.80	63.53	67.91	70.61	-	72.01	68.77	52.82	71.17
	Software	base	72.49	69.15	76.56	47.52	70.36	73.21	57.75	74.23	81.49	81.82	74.52	65.78	81.03
		aug	72.59	68.21	75.30	48.65	70.25	71.63	56.34	74.30	-	81.23	76.59	64.00	80.70
	VisualWork	base	65.42	73.08	74.85	76.98	68.75	90.78	64.27	76.22	70.55	78.81	54.42	75.46	82.65
		aug	63.35	72.81	73.71	76.54	66.69	89.68	65.10	74.56	-	78.24	56.25	73.24	81.03
LOC	Facility	base	58.11	71.89	65.55	61.38	69.32	70.04	64.27	68.58	67.56	70.04	58.10	58.77	71.90
		aug	57.62	71.95	64.49	61.62	68.14	69.42	63.42	68.62	-	70.42	56.93	56.61	70.82
	HumanSettlement	base	81.48	90.11	86.77	78.99	88.48	85.15	81.86	88.30	89.92	90.05	84.00	76.80	88.74
		aug	80.14	90.85	85.37	77.45	86.26	83.12	81.75	88.05	-	89.75	83.55	74.08	87.68
	Station	base	64.41	74.77	74.79	72.15	77.09	69.96	83.40	76.54	83.62	78.36	80.74	80.94	80.43
		aug	65.47	74.65	73.68	70.34	76.04	68.05	83.16	76.50	-	77.85	78.53	80.28	79.36
	OtherLOC	base	42.66	88.30	67.31	46.76	49.20	49.61	46.60	56.12	67.56	63.45	67.51	39.80	73.28
		aug	43.01	87.58	66.55	45.77	48.42	49.36	48.19	55.25	-	64.64	65.73	39.86	72.30
PER	SportsManager	base	57.51	54.86	57.03	48.11	44.81	68.65	44.06	57.08	48.35	62.69	18.21	66.10	62.83
		aug	56.79	55.85	58.22	49.30	47.25	67.05	44.24	56.09	-	63.45	21.23	66.34	61.51
	Athlete	base	71.64	74.32	70.65	70.48	72.25	85.49	69.94	77.75	68.55	82.33	74.28	63.45	79.78
		aug	72.05	74.46	71.74	70.61	72.30	85.08	69.90	78.85	-	82.27	74.85	63.07	79.09
	Cleric	base	56.79	56.86	63.81	56.84	45.83	69.28	42.01	54.07	60.76	62.14	67.94	56.42	64.18
		aug	57.29	57.41	64.46	57.94	46.15	70.19	40.06	54.44	-	60.96	72.20	57.12	64.47
	Artist	base	75.21	73.47	77.72	74.75	74.38	85.77	68.49	78.18	68.65	76.36	65.22	75.26	80.32
		aug	74.33	74.06	77.39	74.29	73.38	85.13	68.42	78.24	-	76.95	64.59	73.37	79.97
	Politician	base	57.44	63.04	59.21	56.23	54.78	61.76	50.87	59.09	60.71	60.19	59.83	61.78	65.51
		aug	57.34	63.27	60.38	56.41	54.45	62.38	49.08	60.39	-	59.70	60.27	61.11	65.23
	Scientist	base	40.98	36.75	36.80	36.72	34.85	47.30	31.78	43.99	26.06	49.48	40.98	39.79	49.76
		aug	41.96	36.01	38.56	39.26	37.37	49.11	32.00	44.29	-	50.21	46.19	39.07	50.65
	OtherPER	base	50.35	50.37	50.53	44.79	45.37	48.44	40.26	48.80	42.33	56.45	42.95	44.30	54.58
		aug	50.24	50.38	50.87	44.89	46.18	49.48	38.93	47.41	-	56.99	42.66	43.55	54.54
GRP	ORG	base	58.76	63.54	64.56	52.84	66.82	59.34	63.23	63.67	84.10	71.39	76.91	60.05	72.77
		aug	58.82	63.41	64.44	52.45	65.65	58.10	62.10	64.80	-	71.84	76.69	58.68	71.87
	PrivateCorp	base	26.39	20.00	0.00	29.24	57.53	19.87	57.66	30.69	55.68	16.36	81.82	46.40	68.02
		aug	32.83	19.10	0.00	33.52	64.35	21.57	49.33	33.70	-	21.74	83.12	43.58	67.06
	SportsGRP	base	77.24	83.05	82.30	70.46	86.78	81.72	80.04	82.11	93.54	87.85	89.46	84.19	86.74
		aug	76.95	83.62	82.26	70.03	87.12	81.09	79.46	82.61	-	87.53	89.37	83.22	86.62
	AerospaceManufacturer	base	42.20	20.84	29.70	40.50	71.37	29.09	66.56	51.07	11.86	30.85	6.06	79.32	70.43
		aug	40.66	20.03	25.85	43.40	69.98	32.34	67.10	49.70	-	33.39	5.71	79.29	69.13
	PublicCorp	base	65.85	57.19	74.90	47.40	62.49	68.46	53.89	60.88	70.01	76.92	66.74	67.21	76.25
		aug	66.48	57.11	73.86	46.94	62.44	69.17	51.07	61.41	-	75.82	68.68	66.20	76.03
	CarManufacturer	base	65.57	57.94	67.76	54.10	63.36	71.06	61.86	63.11	71.96	74.08	78.66	71.22	76.51
		aug	65.00	57.57	67.37	55.22	62.68	71.26	61.21	62.62	-	73.02	77.98	69.87	75.92
	MusicalGRP	base	71.27	73.20	74.08	56.66	68.93	79.63	62.18	71.86	65.49	82.58	57.34	65.63	80.36
		aug	69.11	72.85	72.65	56.57	66.78	78.17	60.99	71.18	-	82.17	62.10	64.42	78.48
PROD	Clothing	base	52.26	59.41	47.80	50.12	52.83	50.85	48.85	59.53	48.39	58.91	68.37	36.63	63.68
		aug	51.71	59.39	45.57	50.72	48.31	48.68	48.43	59.17	-	59.74	69.11	39.06	60.19
	Drink	base	58.99	64.20	65.49	51.56	51.32	62.34	45.60	60.67	66.90	70.33	73.40	60.14	68.34
		aug	58.56	64.02	63.68	52.40	54.30	59.30	45.00	61.05	-	70.48	72.19	58.80	68.30
	Vehicle	base	54.71	59.20	57.55	47.15	64.02	56.80	59.03	54.49	77.62	66.91	68.89	57.83	65.68
		aug	53.54	59.27	57.03	47.48	62.55	55.78	60.45	54.15	-	65.51	68.89	55.69	64.35
	OtherPROD	base	49.71	63.44	65.11	44.30	58.87	56.31	52.16	53.84	60.61	65.71	52.93	58.65	66.12
		aug	49.16	63.29	64.16	46.07	58.62	55.09	52.87	54.97	-	65.14	52.79	57.56	66.13
	Food	base	54.32	61.98	60.80	46.78	62.41	57.09	62.94	54.84	63.54	69.18	68.17	58.74	65.22
		aug	52.43	61.79	59.31	47.56	60.54	54.32	62.60	55.71	-	68.37	64.91	56.42	64.98
MED	Medication/Vaccine	base	66.09	68.79	75.41	58.91	76.74	72.18	65.71	72.19	77.63	82.03	72.34	69.59	79.40
		aug	65.51	68.62	73.12	57.09	72.59	69.88	67.18	71.60	-	81.61	70.72	66.62	78.24
	Disease	base	64.71	71.14	72.22	59.35	75.70	67.14	67.87	69.72	84.75	75.92	78.20	64.24	77.38
		aug	64.69	71.61	71.09	58.75	75.10	65.05	66.59	67.19	-	75.78	79.13	63.19	76.67
	Symptom	base	39.60	47.49	39.25	47.76	42.62	47.21	37.27	47.46	79.11	52.35	72.67	57.16	64.03
		aug	44.36	47.86	43.03	49.77	43.90	49.48	30.69	48.95	-	53.03	73.86	56.48	65.30
	AnatomicalStructure	base	64.72	73.02	66.67	50.80	77.88	65.96	68.82	67.21	80.54	78.36	72.49	56.50	74.83
		aug	64.24	73.36	65.73	49.25	75.00	64.10	68.83	66.96	-	78.32	71.26	54.90	73.96
	MedicalProcedure	base	54.17	62.20	65.53	49.28	73.03	66.98	61.96	64.77	81.66	65.33	73.31	60.42	72.83
		aug	54.77	62.84	65.04	50.07	73.71	65.13	65.62	65.63	-	66.90	75.41	60.74	72.71
	FG	base	65.73	72.64	71.27	63.48	69.41	75.69	64.30	70.26	73.54	75.17	70.57	66.48	76.03
		aug	65.04	72.64	70.62	63.20	68.60	74.67	63.79	70.31	-	74.96	70.57	64.80	75.32
	CG	base	78.26	83.21	83.34	76.74	84.69	86.02	78.13	83.43	84.86	85.45	78.92	75.12	85.83
		aug	77.24	83.21	82.22	76.03	83.45	84.73	77.80	83.18	-	85.03	78.68	72.98	85.03

Table 2: Official macro f1 results for both base and augmentation methods in the evaluation phase
* except augmentation version of sv

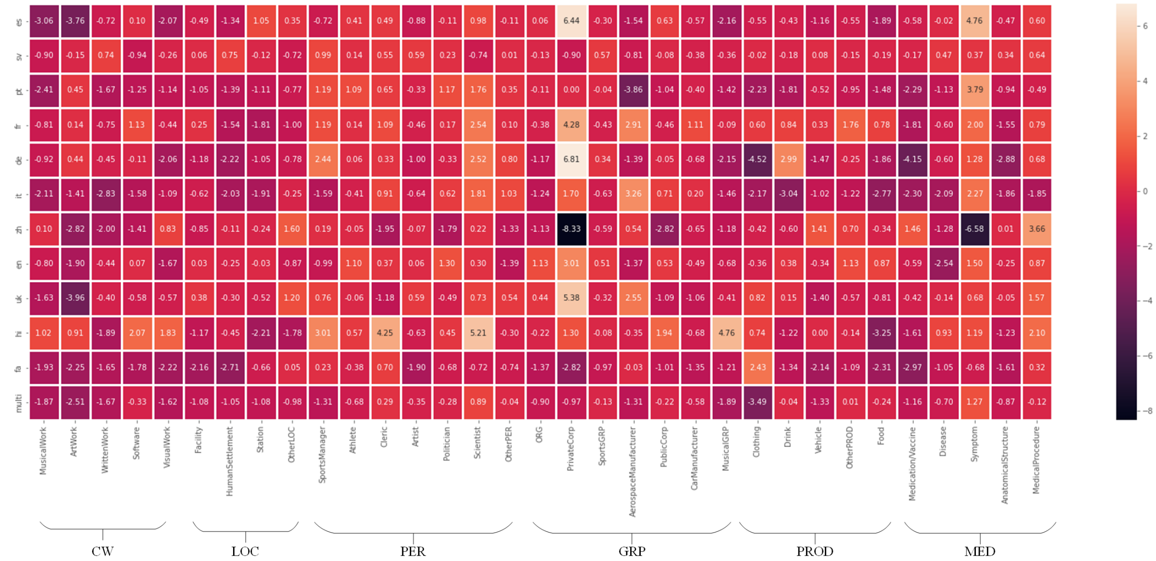

Figure 8: Heat map differences between the base system and the augmentation systems in terms of f1

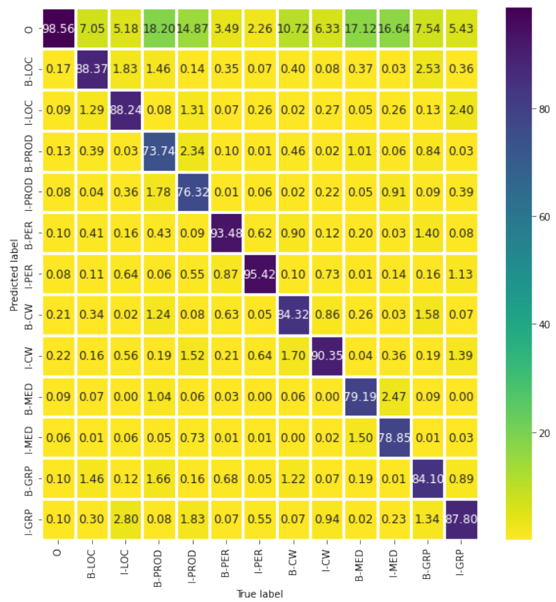

Figure 9: Confusion matrix for multilingual system (base) on coarse-grained labels

the quantity of data in each category can have either a positive effect or a negative one in terms of F1 score, depending on the language and sub-class. The negative impact of augmentation is depicted by the black color. Overall, data augmentation had the most positive impact on sub-classes such as PrivateCorp, Symptoms, and Scientists. Moreover, in terms of languages, Hindi and French exhibited the highest improvements due to data augmentation.

Detailed results Table 2 provides a detailed overview of the two main methods evaluated during the evaluation phase for each language. The last four rows of the table present the overall fine-grained and coarse-grained results of our NER systems.

Overall results As previously noted, certain test sets for specific languages contain corrupted instances. Figure 6 illustrates our best model results, indicating a disparity between the macro F1 value for corrupted and uncorrupted versions of both the fine-grained and coarse-grained datasets. These findings suggest that the models designed for these languages encounter challenges when handling such corrupted data, resulting in a decrease in F1 values.

Error Analysis for Multilingual Track The heat map in Figure 9 illustrates the performance of a multilingual system trained using XLM-Roberta-Large. The map reveals that the model was not always able to accurately assign the "B-" or "I-" tag, and, in some cases, the model wrongly assigned the "O" tag to certain classes. Specifically, 18.20% and 14.87% of B-Prod and I-Prod instances, respectively, were assigned an "O" tag by the model, indicating that further improvements are necessary to enhance the model's ability to recognize and classify NEs more accurately.

6 Conclusion

In this work, we utilized (PLMs) to build a system for recognizing complex NEs. To increase the number of training examples and improve the performance of the system, we applied a simple data augmentation technique. However, we observed that this approach led to mixed results, with improvements in some subclasses but a reverse effect in others.

One possible reason for this outcome is that the augmentation technique involves assigning "O" tags to the rest of the tokens in a sentence, which may lead to some loss of information. Furthermore, the augmented data may be more unbalanced than the original data, with some instances being increased more than others. To address this issue, it may be necessary to use more sophisticated augmentation techniques or balance the data more effectively to ensure that the model can learn from a representative set of examples.

7 Future Work

For instance, a semi-supervised approach could be applied to assign labels to augmented sentences and then add them to the dataset, in order to prevent assigning "O" tags to actual named entity categories.

Acknowledgements

This work has been supported by the Simorgh Supercomputer - Amirkabir University of Technology under Contract No ISI-DCE-DOD-Cloud-900808-1700.

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, 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 User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. ParsBERT: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

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

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer Enhanced Named Entity Recognition for Code-Mixed Web Queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023b. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Mohammad Ebrahim Khademi and Mohammad Fakhredanesh. 2020. Persian automatic text summarization based on named entity recognition. *Iranian*

Journal of Science and Technology, Transactions of Electrical Engineering.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

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.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Nick McKenna, Liane Guillou, Mohammad Javad Hosseini, Sander Bijl de Vroe, and Mark Steedman. 2021. Multivalent entailment graphs for question answering. *CoRR*, abs/2104.07846.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Diego Mollá, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 51–58, Sydney, Australia.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Matthew Rocklin. 2015. Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference*, 130-136. Citeseer.

Amir Sartipi and Afsaneh Fatemi. 2023. Exploring the potential of machine translation for generating named entity datasets: A case study between persian and english.

Stefan Schweter. 2020a. Europeana bert and electra models.

Stefan Schweter. 2020b. Italian bert and electra models.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Samantha Swee Yun Tan. 2022. Named entity recognition for information extraction.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Liang Xu, Yu Tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. CLUENER2020: fine-grained named entity recognition dataset and benchmark for chinese. *CoRR*, abs/2001.04351.

A Appendix

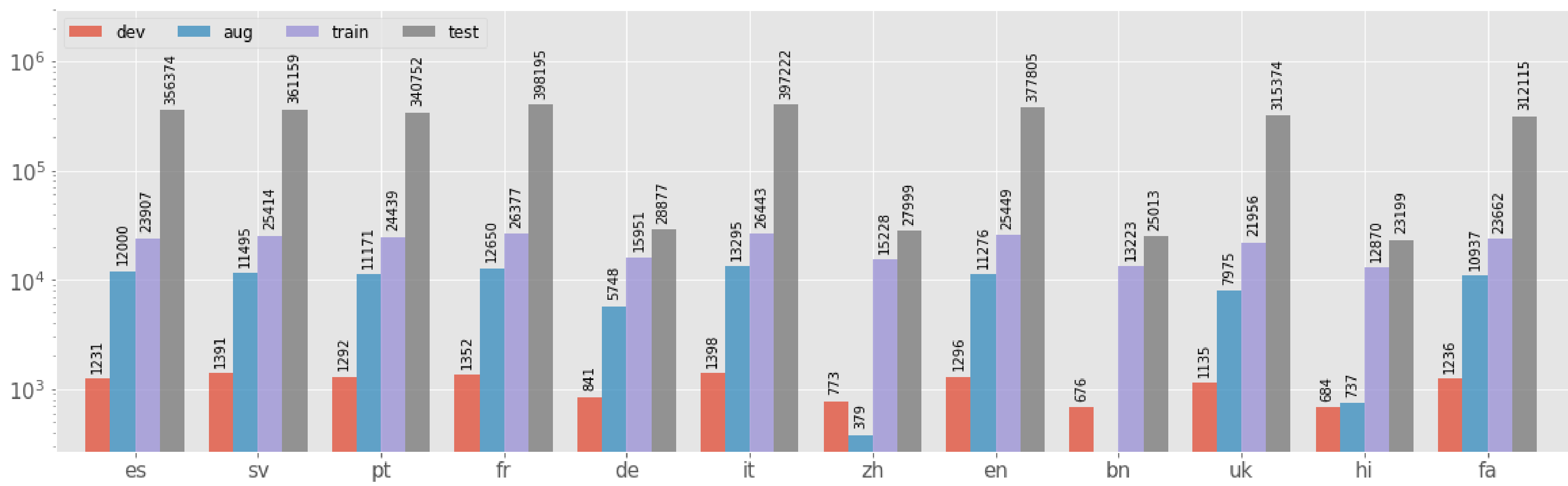

Figure 10: This is the zoomed version of Figure 2

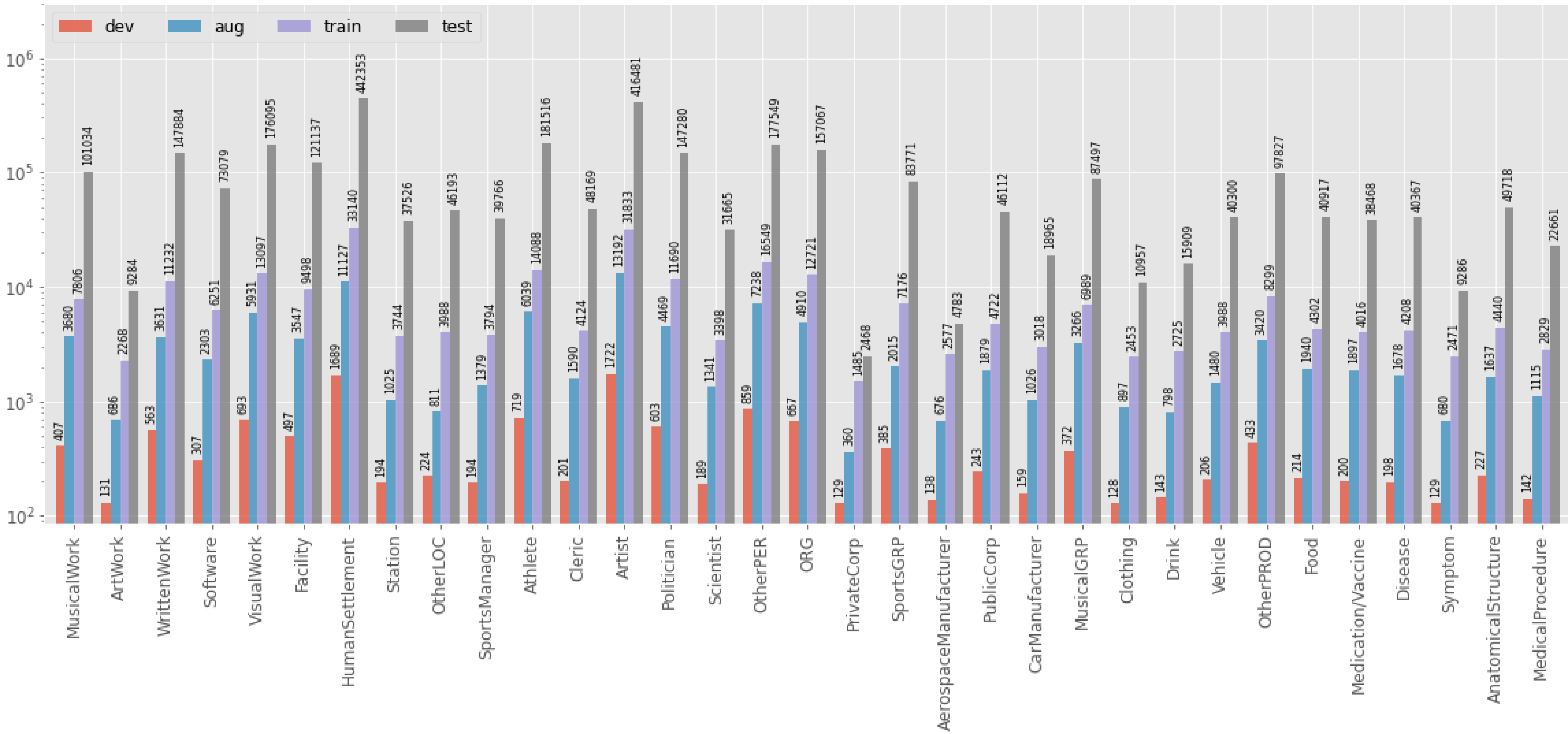

Figure 11: This is the zoomed version of Figure 3

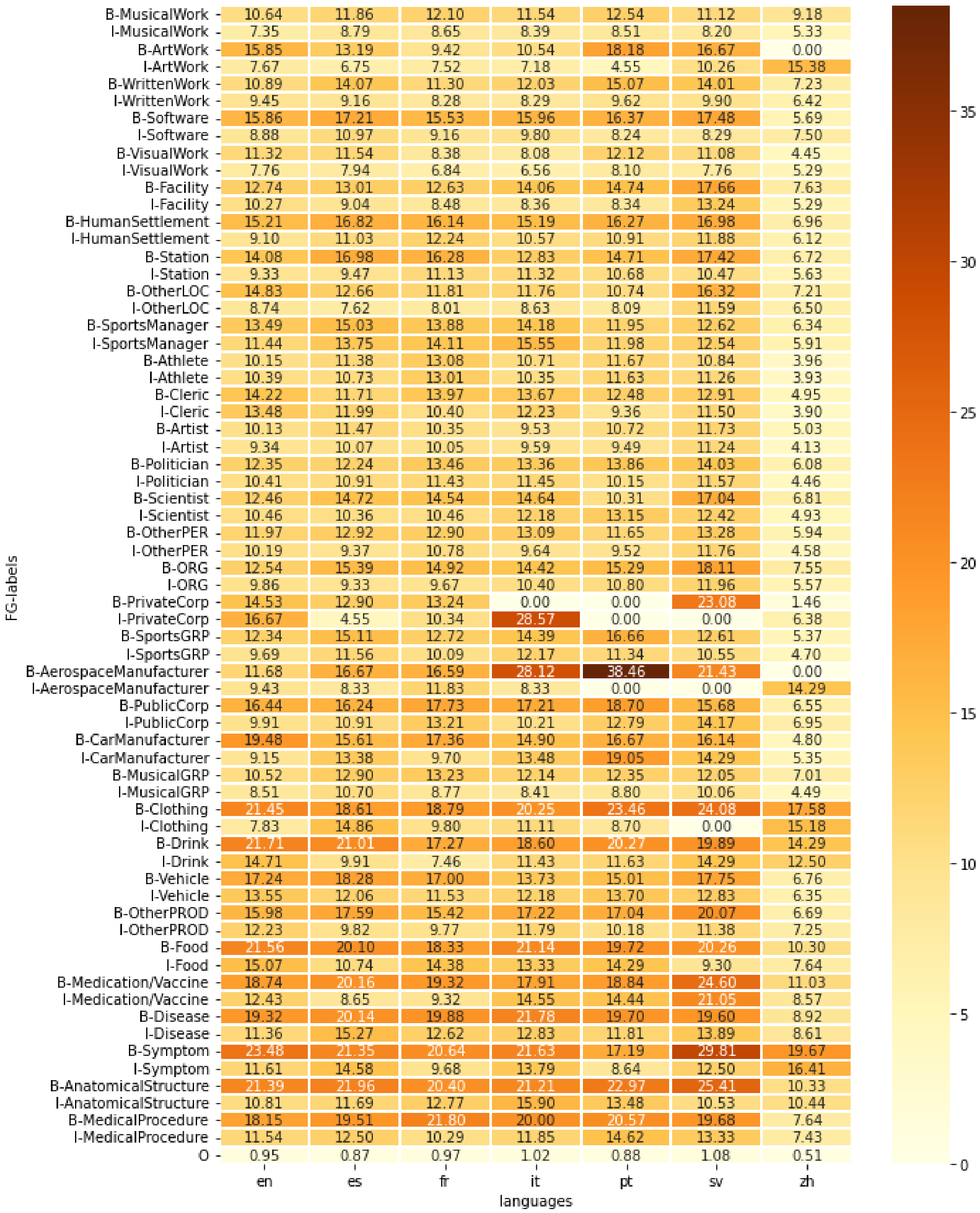

Figure 12: This is the zoomed version of Figure 4

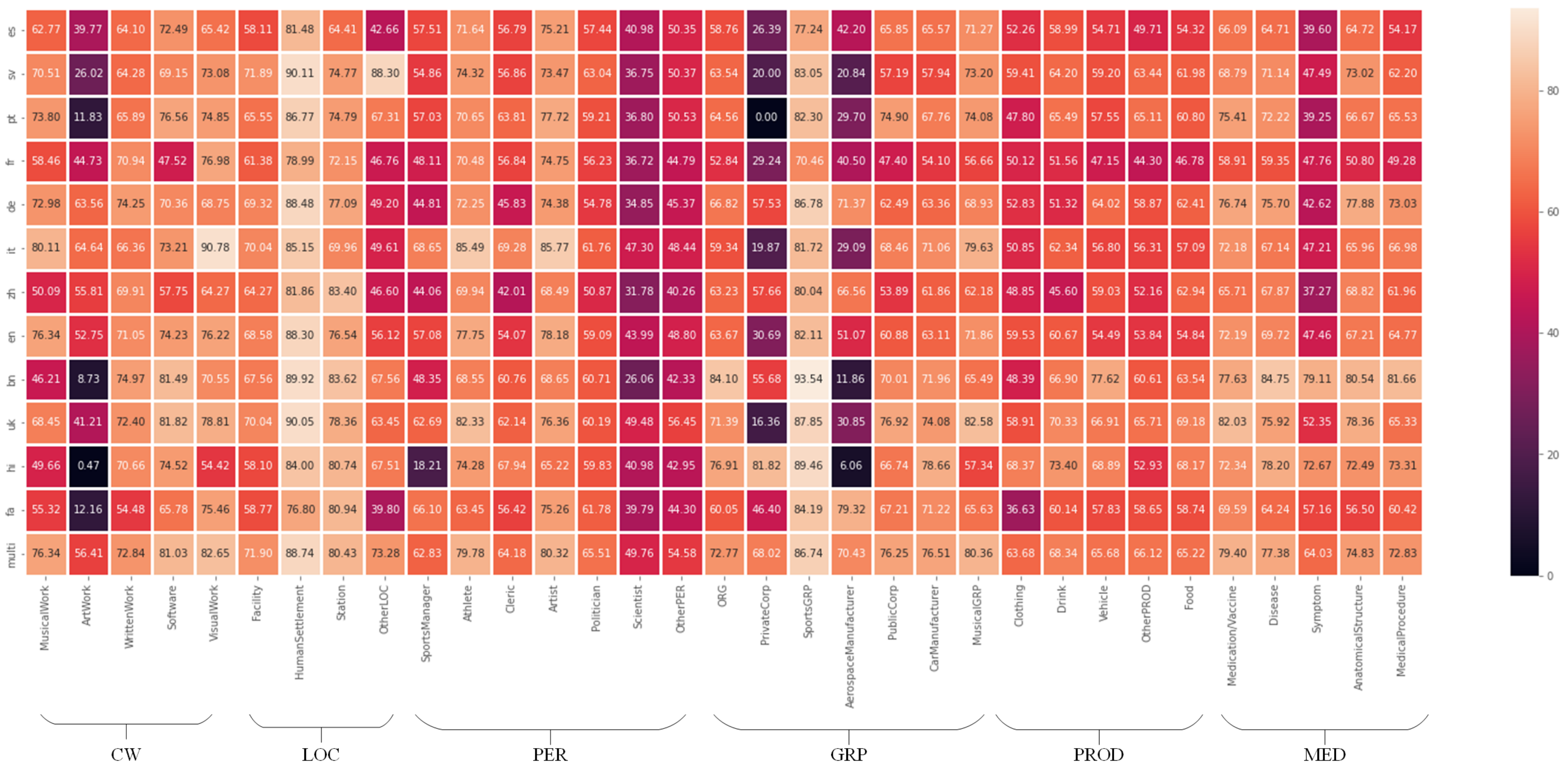

Figure 13: This is the zoomed version of Figure 7

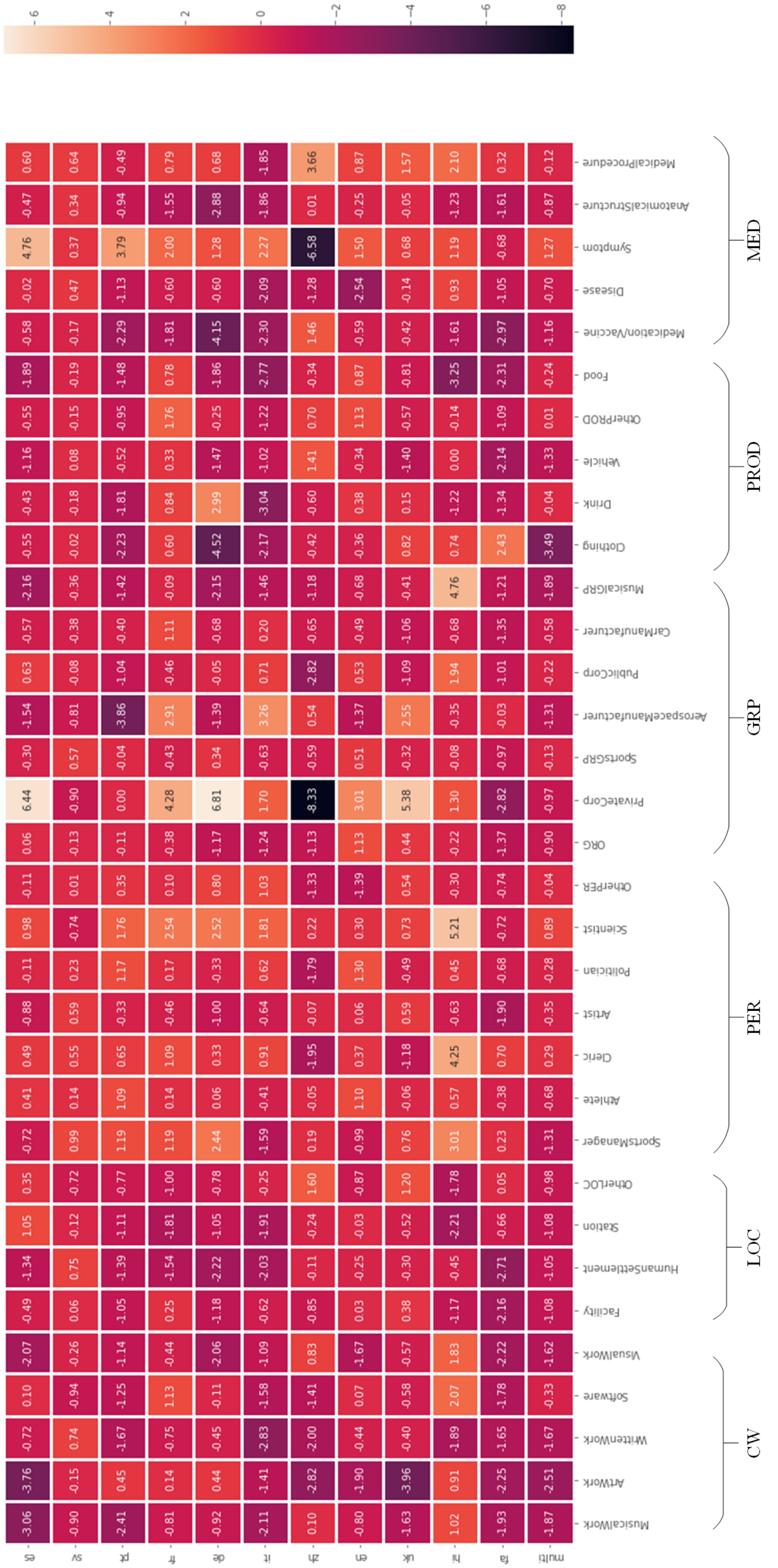

Figure 14: This is the zoomed version of Figure 8

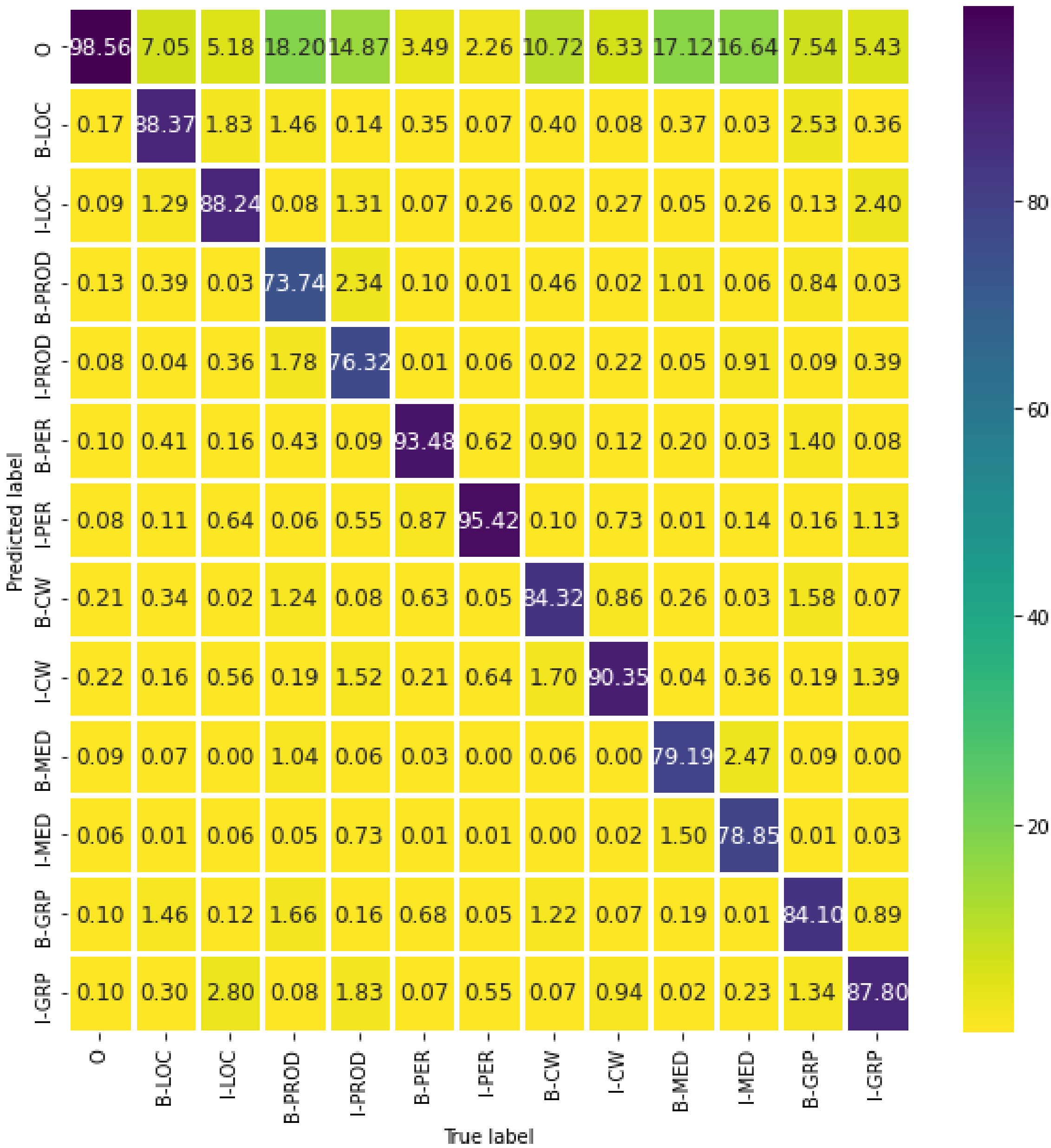

Figure 15: This is the zoomed version of Figure 9