

SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER)

Shervin Malmasi, Anjie Fang*, Besnik Fetahu*, Sudipta Kar*, Oleg Rokhlenko

Amazon.com, Inc.

Seattle, WA, USA

{malmasi,njfn,besnikf,sudipkar,olegro}@amazon.com

Abstract

We present the findings of SemEval-2022 Task 11 on Multilingual Complex Named Entity Recognition MULTICoNER.¹ Divided into 13 tracks, the task focused on methods to identify complex named entities (like media titles, products, and groups) in 11 languages in both monolingual and multi-lingual scenarios. Eleven tracks were for building monolingual NER models for individual languages, one track focused on multilingual models able to work on all languages, and the last track featured code-mixed texts within any of these languages. The task used the MULTICoNER dataset, composed of 2.3 million instances in Bangla, Chinese, Dutch, English, Farsi, German, Hindi, Korean, Russian, Spanish, and Turkish. Results showed that methods fusing external knowledge into transformer models achieved the best performance. The largest gains were on the Creative Work and Group entity classes, which are still challenging even with external knowledge. MULTICoNER was one of the most popular tasks in SemEval-2022 and it attracted 377 participants during the practice phase. The final test phase had 236 participants, and 55 teams submitted their systems.

1 Introduction

Processing complex and ambiguous Named Entities (NEs) is a challenging NLP task in practical and open-domain settings but has not received sufficient attention from the research community. Complex NEs, like the titles of creative works (movie/book/song/software names) are not simple nouns and are harder to recognize (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 (Person names, locations, organizations). This ambiguity makes it challenging to recognize them

based on their context. Such titles can also be semantically ambiguous, e.g. “On the Beach” can be a preposition or refer to a movie.² Finally, such entities usually grow at a faster rate than traditional categories, and emerging entities pose yet another challenge.

Neural models (e.g. Transformers) have produced high scores on benchmark datasets like CoNLL03/OntoNotes (Devlin et al., 2018). However, as noted by Augenstein et al. (2017), these scores are driven by the use of well-formed news text, the presence of “easy” entities (e.g. person names), and memorization due to entity overlap between train/test sets; these models perform significantly worse on complex/unseen entities (Meng et al., 2021; Fetahu et al., 2021). Researchers using NER on downstream tasks have also noted that a significant proportion of their errors are due to NER systems failing to recognize complex entities (Luken et al., 2018; Hanselowski et al., 2018). Examples of such challenges are highlighted in Table 1.

For this task, we created the MULTICoNER dataset (Malmasi et al., 2022) to address the aforementioned challenges. MULTICoNER provides data from three domains (Wikipedia sentences, questions, and search queries) across 11 different languages, which are used to define 11 monolingual subsets of the shared task. Additionally, the dataset has multilingual and code-mixed subsets.

We received 1,884 submissions from 55 teams during the test phase and 34 system description papers were submitted. Results showed that usage of external data and ensemble strategies played a crucial role in the strong performance on in-domain data and also contributed to domain adaptation. External knowledge brought large improvements on classes containing names of creative works and groups, allowing these systems to achieve the best overall performance.

*These authors contributed equally to this work.

¹<https://multiconer.github.io/>

²<https://www.imdb.com/title/tt0053137>

Challenge	Description
Complex Entities Relevant to all domains	Not all entities are proper names: some types (e.g. creative works) can be linguistically complex. They can be complex noun phrases (Eternal Sunshine of the Spotless Mind), gerunds (Saving Private Ryan), infinitives (To Kill a Mockingbird), or full clauses (Mr. Smith Goes to Washington). Syntactic parsing of such nouns is hard, and most current parsers/NER systems fail to recognize them. The top system from WNUT 2017 achieved 8% recall for creative work entities (Aguilar et al., 2017). Effective evaluation requires corpora with many such entities.
Ambiguous Entities and Contexts Particularly for voice and search domains	Some NEs are ambiguous: they are not always entities, e.g. “Inside Out”, “Among Us”, and “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. News texts have long sentences discussing many entities, but other use cases (search queries, questions) have shorter inputs. Data with minimal context is needed to assess performance of such use cases. Capitalization/punctuation features are large drivers of success in NER (Mayhew et al., 2019), but short inputs (ASR, queries) often lack such surface features. An <i>uncased</i> evaluation is needed to assess model performance.
Emerging Entities For domains with growing entities	All entity types are open classes (new ones are added), but some groups have a faster growth rate, e.g. new books/songs/movies are released weekly resulting in a long-tail distribution. Assessing true generalization requires test sets with many unseen entities, to mimic an open-world setting.

Table 1: Challenges not tackled by current work/datasets, but addressed by the MULTICoNER task and data.

2 MultiCoNER Dataset

The MULTICoNER dataset was designed to address the NER challenges described in §1. It represents three domains (wiki sentences, questions, and search queries) and includes 11 languages, plus multilingual and code-mixed subsets. For a detailed description of the MultiCoNER dataset, we refer the reader to the dataset paper (Malmasi et al., 2022). The dataset is publicly available.³

2.1 NER Taxonomy

MULTICoNER leverages the WNUT 2017 (Derczynski et al., 2017a) taxonomy entity types, which defines the following NER tag-set with six classes:

1. PER: Names of people
2. LOC: Location or physical facilities
3. CORP: Corporations and businesses
4. GRP: All other groups
5. PROD: Consumer products
6. CW: Titles of creative works like movie, song, and book titles

This taxonomy allows us to capture a wide array of entities, including those with more complex entity structures, such as creative works.

2.2 Languages and Subsets

Eleven languages are included in MULTICoNER:

1. Bangla (BN)
2. Chinese (ZH)
3. Dutch (NL)

4. English (EN)
5. Farsi (FA)
6. German (DE)
7. Hindi (HI)
8. Korean (KO)
9. Russian (RU)
10. Spanish (ES)
11. Turkish (TR)

These languages were chosen to include a diverse typology of languages and writing systems, and range from well-resourced (EN) to low-resourced ones (FA).

MULTICoNER contains 13 different subsets: 11 monolingual subsets for the above languages, a multilingual subset (denoted as `MULTI`), and a code-mixed one (`MIX`).

Monolingual Subsets Each of the 11 languages has its own subset, which includes data from all three domains.

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.

Code-mixed Subset This subset contains code-mixed instances, where the entity is from one language and the rest of the text is written in another language. Like the multilingual subset, this subset should also be used under the assumption that the languages present in an instance are unknown.

³<https://registry.opendata.aws/multiconer>

Class	Split	EN	DE	ES	RU	NL	KO	FA	ZH	HI	TR	BN	MULTI	MIX
PER	Train	5,397	5,288	4,706	3,683	4,408	4,536	4,270	2,225	2,418	4,414	2,606	43,951	296
	Dev	290	296	247	192	212	267	201	129	133	231	144	2,342	96
	Test	55,682	55,757	51,497	44,687	49,042	39,237	35,140	26,382	25,351	26,876	24,601	111,346	19,313
LOC	Train	4,799	4,778	4,968	4,219	5,529	6,299	5,683	6,986	2,614	5,804	2,351	54,030	325
	Dev	234	296	274	221	299	323	324	378	131	351	101	2,932	108
	Test	59,082	59,231	58,742	54,945	63,317	52,573	45,043	43,289	31,546	34,609	29,628	141,013	23,111
GRP	Train	3,571	3,509	3,226	2,976	3,306	3,530	3,199	713	2,843	3,568	2,405	32,846	248
	Dev	190	160	168	151	163	183	164	26	148	167	118	1,638	75
	Test	41,156	40,689	38,395	37,621	39,255	31,423	27,487	18,983	22,136	21,951	19,177	77,328	16,357
CORP	Train	3,111	3,083	2,898	2,817	2,813	3,313	2,991	3,805	2,700	2,761	2,598	32,890	294
	Dev	193	165	141	159	163	156	160	192	134	148	127	1,738	112
	Test	37,435	37,686	36,769	35,725	35,998	30,417	27,091	25,758	21,713	21,137	20,066	75,764	18,478
CW	Train	3,752	3,507	3,690	3,224	3,340	3,883	3,693	5,248	2,304	3,574	2,157	38,372	298
	Dev	176	189	192	168	182	196	207	282	113	190	120	2,015	102
	Test	42,781	42,133	43,563	39,947	41,366	33,880	30,822	30,713	21,781	23,408	21,280	89,273	20,313
PROD	Train	2,923	2,961	3,040	2,921	2,935	3,082	2,955	4,854	3,077	3,184	3,188	35,120	316
	Dev	147	133	154	151	138	177	157	274	169	158	190	1,848	117
	Test	36,786	36,483	36,782	36,533	36,964	29,751	26,590	28,058	22,393	21,388	20,878	75,871	20,255
#sentences	Train	15,300	15,300	15,300	15,300	15,300	15,300	15,300	15,300	15,300	15,300	15,300	168,300	1,500
	Dev	800	800	800	800	800	800	800	800	800	800	800	8,800	500
	Test	217,818	217,824	217,887	217,501	217,337	178,249	165,702	151,661	141,565	136,935	133,119	471,911	100,000

Table 2: MULTICONER 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.

2.3 Dataset Creation

The MultiCoNER dataset consists of 11 languages, and three domains (encyclopedia sentences, questions from QA, and Web queries). A detailed overview of the MultiCoNER dataset is provided in the dataset paper (Malmasi et al., 2022).

LOWNER: represents the encyclopedic sentences extracted from the different 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 from (Derczynski et al., 2017b). Manual inspection of 400 sampled English sentences shows that the NER gold labels are 94% accurate.

MSQ-NER: from the MS-MARCO Q&A corpus (Bajaj et al., 2016) *question templates* are extracted by replacing the *entities* with their NER *type* (from the MultiCoNER NER taxonomy). Entities in a question are identified using spaCy.⁴ The templates are translated from English into the rest of the languages.

ORCAS-NER: similar to MSQ-NER, templates from Web user queries are extracted from the ORCAS dataset (Craswell et al., 2020). The templates are translated into the respective languages, and finally, multiple instances are constructed from each template by simply replacing the template slots with actual named entities in the target languages.

2.4 Dataset Statistics

Table 2 shows some statistics of the dataset. For all the tracks, we have released 15,300 training and 800 development instances. In the training splits, the absolute majority of instances are from the Wikipedia domain (i.e. LOWNER), whereas a small number of 100 instances are domain-adaptation data, with 50 instances coming from the Web Questions (i.e. MSQ-NER) and Web Query (i.e. ORCAS-NER) domains, respectively.

The test splits on the other hand are much larger. This is done for mainly two reasons: (1) to be able to assess the generalizability of NER models on unseen and complex entities, and (2) to assess the cross-domain adaptation performance of NER models. For practical reasons, we cap the number of test instances to be at a maximum of 200k per subset, with the exception of the Code-Mixed and Multilingual subsets. The Multilingual test split was generated from the language-specific test splits and was downsampled to contain only 471k instances. On the other hand, for the Code-Mixed subset, we sample test sentences from the language-specific test split, and replace the original entity surface forms with the surface form of the entity in another language, picked at random.

More details on the dataset construction process are available in Malmasi et al. (2022).

⁴<https://spacy.io/>

3 Task Description and Evaluation

The shared task is composed of 11 monolingual and 2 multilingual tracks. The monolingual tracks invited participants to build monolingual models for 11 languages addressed by the shared task. The multilingual track invited multilingual models capable of identifying entities from monolingual texts from any of the 11 languages. The code-mixed track called for models to identify entities in mixed-language texts (any language pair from the 11 languages). That means the multilingual models for multilingual and code mixed tracks should be able to process texts from any language and show competitive performance for all the languages.

We used the macro-averaged F1 scores to evaluate and rank systems. Additionally, we report precision, recall, and per-domain performance.

4 Baseline System

We train and evaluate a baseline NER system using on 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).

The XLM-R baseline is highly suited for multilingual application scenarios, such as our. It supports up to 100 languages and provides a solid baseline upon which the participants can build. The baseline was trained with a learning rate of $2e - 5$ and a maximum number of 50 epochs, with an early stopping criterion of a non-decreasing validation loss for 5 epochs. The code and scripts for the baseline system were provided to the participants to use its functionalities and further extend it with their approaches.⁵

5 Participating Systems and Results

We have received submissions from 55 different teams. Among the monolingual tracks, we have observed the highest participation of 30 teams in the English track. Ordered by the number of participating teams, the other monolingual tracks are Chinese (21), Bangla (18), Spanish (18), Hindi (17), Korean (17), German (16), Dutch (15), Farsi (15), Turkish (15), and Russian (14). The number of participating teams for the Multilingual and Code-mixed tracks are 25

and 21, respectively. Table 3 shows the final rankings for all tracks. Detailed performance breakdown is available in Appendix A.

Most of the top-performing teams aimed at building their system targeting the multilingual track, and then retrained it for the other tracks separately and made submissions to all the 13 tracks. Therefore, in the rest of this section, we will first discuss the approaches by focusing on the multilingual track. Then, we will discuss teams that built their systems for one or more monolingual tracks. Finally, we will summarize the methods (e.g. language models, toolkits) and resources used.

5.1 Top Multilingual Systems

DAMO-NLP (Wang et al., 2022) ranked 1st in the multilingual (MULTI) track and all the monolingual tracks except BN (2nd) and ZH (4th). Given a text, they used a knowledge retrieval module to retrieve K most relevant paragraphs from a knowledge base (i.e. Wikipedia). Paragraphs were concatenated together with the input, and token representations were passed through a CRF to predict the labels. They employed multiple such XLM-RoBERTa models with random seeds and then used a voting strategy to make the final prediction.

USTC-NELSLIP (Chen et al., 2022a) ranked 1st in three tracks (MIX, ZH, BN), and 2nd for all the other tracks. The average performance gap between USTC-NELSLIP and DAMO-NLP is $\approx 3\%$ for all the 13 tracks. USTC-NELSLIP aimed at fine-tuning a Gazetteer enhanced BiLSTM network in such a way that the representation produced for an entity has similarity with the representation produced by a pre-trained language model (LM). They developed a two-step process with two parallel networks, where a Gazetteer-BiLSTM uses a Gazetteer search to produce one-hot labels for each token in a given text and a BiLSTM produces a dense vector representation for each token. Another network uses a frozen XLM-RoBERTa to produce an embedding vector for each token. A KL divergence loss is applied to make the Gazetteer network’s output similar to the LM. These two networks are jointly trained together again and their outputs are fused together for the final prediction.

QTrade AI (Gan et al., 2022) ranked 3rd in MULTI, 4th in MIX, and 8th in ZH. They used an XLM-RoBERTa encoder and applied sample mixing for data augmentation, along with adversarial training through data noising. For the multilingual

⁵<https://github.com/amzn/multiconer-baseline>

English (EN)			16	Sartipi-Sedighin	0.584	14	CSECU-DSG	0.558	4	RACAI	0.663
Team	F1	Russian (RU)				15	BASELINE	0.522	5	Infrd.ai	0.64
1	DAMO-NLP	0.912	Team	F1	16	B.E.P.	0.513	6	YNUNLP	0.638	
2	USTC-NELSLIP	0.855	1	DAMO-NLP	0.915	German (DE)			7	Sliced	0.63
3	PAI	0.784	2	USTC-NELSLIP	0.838	Team	F1	8	Team Atreides	0.598	
4	ML-HUB	0.781	3	RACAI	0.746	1	DAMO-NLP	0.906	9	brotherhood	0.586
5	RACAI	0.758	4	Sliced	0.737	2	USTC-NELSLIP	0.89	10	MaChAmp	0.565
6	Infrd.ai	0.747	5	YNUNLP	0.73	3	RACAI	0.794	11	MarSan	0.542
7	EURECOM	0.746	6	MaChAmp	0.724	4	Sliced	0.789	12	EURECOM	0.526
8	Sliced	0.745	7	brotherhood	0.703	5	MaChAmp	0.784	13	AaltoNLP	0.518
9	MaChAmp	0.745	8	NetEase.AI	0.698	6	YNUNLP	0.773	14	silpa_nlp	0.514
10	Raccoons	0.742	9	EURECOM	0.682	7	L3i	0.772	15	CSECU-DSG	0.505
11	YNUNLP	0.732	10	MarSan	0.675	8	ML-HUB	0.761	16	B.E.P.	0.451
12	LMN	0.725	11	L3i	0.667	9	brotherhood	0.759	17	L3i	0.448
13	brotherhood	0.724	12	CSECU-DSG	0.631	10	Infrd.ai	0.759	18	Enigma	0.427
14	L3i	0.72	13	B.E.P.	0.6	11	EURECOM	0.744	19	BASELINE	0.394
15	Multilinguals	0.717	14	BASELINE	0.596	12	MarSan	0.731	Multilingual (MULTI)		
16	KDDIE	0.717	15	AutoNER	0.527	13	CSECU-DSG	0.725	Team	F1	
17	MarSan	0.715	Turkish (TR)			14	AaltoNLP	0.714	1	DAMO-NLP	0.853
18	Cardiff NLP	0.709	Team	F1	15	PA Ph&Tech	0.667	2	USTC-NELSLIP	0.853	
19	Lone Wolf	0.698	1	DAMO-NLP	0.887	16	B.E.P.	0.666	3	QTrade AI	0.777
20	MIDAS	0.696	2	USTC-NELSLIP	0.855	17	BASELINE	0.637	4	SeqL	0.755
21	UC3M-PUCPR	0.692	3	SU-NLP	0.72	Chinese (ZH)			5	CMB AI Lab	0.737
22	CSECU-DSG	0.692	4	RACAI	0.704	Team	F1	6	UM6P-CS	0.725	
23	Sartipi-Sedighin	0.675	5	Sliced	0.688	1	USTC-NELSLIP	0.817	7	RACAI	0.721
24	Enigma	0.672	6	MaChAmp	0.676	2	CASIA	0.797	8	Cardiff NLP	0.717
25	DANGNT-SGU	0.669	7	YNUNLP	0.668	3	OPDAI	0.795	9	Sliced	0.711
26	AaltoNLP	0.668	8	ML-HUB	0.658	4	DAMO-NLP	0.781	10	IIE_KDSEC	0.709
27	SPDB I.L.	0.651	9	L3i	0.643	5	NetEase.AI	0.778	11	B.E.P.	0.707
28	silpa_nlp	0.634	10	MarSan	0.611	6	CMB AI Lab	0.764	12	OPDAI	0.695
29	B.E.P.	0.632	11	brotherhood	0.597	7	NCUEE-NLP	0.742	13	brotherhood	0.694
30	BASELINE	0.614	12	EURECOM	0.566	8	QTrade AI	0.74	14	MarSan	0.693
31	AutoNER	0.557	13	CSECU-DSG	0.553	9	CSECU-DSG	0.672	15	Infrd.ai	0.692
Spanish (ES)			14	Sartipi-Sedighin	0.527	10	Multilinguals	0.669	16	HaveNoIdea	0.688
Team	F1	15	BASELINE	0.463	11	L3i	0.669	17	EURECOM	0.681	
1	DAMO-NLP	0.899	16	B.E.P.	0.45	12	Sliced	0.652	18	MaChAmp	0.677
2	USTC-NELSLIP	0.854	Korean (KO)			13	Infrd.ai	0.647	19	YNUNLP	0.668
3	RACAI	0.756	Team	F1	14	MaChAmp	0.638	20	DS4DH	0.652	
4	Infrd.ai	0.753	1	DAMO-NLP	0.886	15	EURECOM	0.634	21	UPB	0.647
5	MaChAmp	0.752	2	USTC-NELSLIP	0.864	16	RACAI	0.627	22	CSECU-DSG	0.644
6	Sliced	0.751	3	RACAI	0.717	17	YNUNLP	0.614	23	NSU-AI	0.642
7	YNUNLP	0.732	4	CMB AI Lab	0.707	18	brotherhood	0.609	24	SPDB I.L.	0.632
8	brotherhood	0.707	5	Sliced	0.707	19	MarSan	0.566	25	L3i	0.612
9	L3i	0.689	6	YNUNLP	0.703	20	SPDB I.L.	0.557	26	BASELINE	0.478
10	PA Ph&Tech	0.689	7	C-3PO	0.675	21	B.E.P.	0.528	Code-Mixed (MIX)		
11	MarSan	0.683	8	UA-KO	0.675	22	BASELINE	0.513	Team	F1	
12	SPDB I.L.	0.673	9	brotherhood	0.674	Hindi (HI)			1	USTC-NELSLIP	0.929
13	CSECU-DSG	0.656	10	Infrd.ai	0.673	Team	F1	2	DAMO-NLP	0.918	
14	EURECOM	0.628	11	MaChAmp	0.654	1	DAMO-NLP	0.862	3	CMB AI Lab	0.846
15	Multilinguals	0.612	12	EURECOM	0.65	2	USTC-NELSLIP	0.846	4	QTrade AI	0.844
16	Sartipi-Sedighin	0.607	13	L3i	0.627	3	RACAI	0.681	5	SeqL	0.803
17	B.E.P.	0.601	14	MarSan	0.623	4	Sliced	0.67	6	IIE_KDSEC	0.796
18	BASELINE	0.578	15	CSECU-DSG	0.621	5	NetEase.AI	0.666	7	RACAI	0.794
19	UC3M-PUCPR	0.568	16	AaltoNLP	0.618	6	Infrd.ai	0.657	8	UM6P-CS	0.792
Dutch (NL)			17	B.E.P.	0.59	7	brotherhood	0.642	9	EURECOM	0.776
Team	F1	18	BASELINE	0.552	8	YNUNLP	0.634	10	OPDAI	0.775	
1	DAMO-NLP	0.905	Farsi (FA)			9	OPDAI	0.629	11	YNUNLP	0.768
2	USTC-NELSLIP	0.877	Team	F1	10	MaChAmp	0.617	12	UC3M-PUCPR	0.764	
3	RACAI	0.784	1	DAMO-NLP	0.897	11	CSECU-DSG	0.577	13	brotherhood	0.759
4	Sliced	0.777	2	USTC-NELSLIP	0.871	12	MarSan	0.563	14	MaChAmp	0.745
5	MaChAmp	0.77	3	RACAI	0.704	13	EURECOM	0.528	15	Sliced	0.727
6	Infrd.ai	0.764	4	Sliced	0.687	14	silpa_nlp	0.515	16	CMNEROne	0.704
7	YNUNLP	0.758	5	YNUNLP	0.672	15	B.E.P.	0.499	17	L3i	0.687
8	brotherhood	0.73	6	brotherhood	0.657	16	L3i	0.497	18	Cardiff NLP	0.681
9	PA Ph&Tech	0.721	7	C-3PO	0.655	17	Enigma	0.486	19	B.E.P.	0.68
10	MarSan	0.711	8	L3i	0.651	18	BASELINE	0.482	20	SPDB I.L.	0.673
11	L3i	0.71	9	MarSan	0.621	Bangla (BN)			21	MarSan	0.67
12	CSECU-DSG	0.679	10	MaChAmp	0.607	Team	F1	22	CSECU-DSG	0.64	
13	EURECOM	0.667	11	AaltoNLP	0.589	1	USTC-NELSLIP	0.842	23	BASELINE	0.581
14	B.E.P.	0.632	12	Sartipi-Sedighin	0.577	2	DAMO-NLP	0.835			
15	BASELINE	0.62	13	EURECOM	0.559	3	NetEase.AI	0.709			

Table 3: Ranking for all of the tracks based on Macro F1. Full forms of the team names “B.E.P.” and “SPDB I.L.” are BaselineExtendinPokemons and SPDB Innovation Lab, respectively.

track, they leveraged an architecture with shared and per-language representations. Finally, they created an ensemble of models trained with different approaches.

SeqL (Hassan et al., 2022) ranked 4th in MULTI 5th in MIX. They train seven XLM-RoBERTa-large and Infoxlm-large models and then used an ensemble approach with voting and score fusion to predict the final labels. They found that the ensemble approach is slightly better than the best single model, and score fusion worked better than simple voting.

CMB AI Lab (PU et al., 2022) ranked 5th in MULTI, 3rd in MIX, 4th in KO, and 6th in ZH. They first utilized a biaffine layer to identify potential entity spans in a sentence, and the extracted spans are then processed with another classifier to obtain their class label. Finally, an ensemble is created by combining different pre-trained encoders and data augmentation techniques based on translations of the original training data. In terms of pre-trained LMs, they used XLM-RoBERTa and mT5.

5.2 Other Noteworthy Systems

RACAI (Pais, 2022) (3rd in ES, NL, RU, KO, FA, DE, HI; 4th in BN, TR; 5th in EN; 7th in MULTI, MIX; 16th in ZH) used XLM-RoBERTa as pre-trained LM and a lateral inhibition layer inspired by the biological mechanism of lateral inhibition. They achieved strong performance in most of the tracks without using any external data.

Sliced (Plank, 2022) (4th in NL, RU, FA, DE, HI; 5th in KO, TR; 6th in ES; 7th in BN; 8th in EN; 9th in MULTI; 12th in ZH; 15th in MIX) used the MaChAmp toolkit (van der Goot et al., 2021) that enables easy exchange of pre-trained LMs for fine-tuning as well as multi-task learning. Within this framework, they have experimented with four different pre-trained LMs and found that XLM-RoBERTa is more efficient for training their system and provides stronger performance.

MaChAmp (van der Goot, 2022) (5th in DE, ES, NL; 6th in RU, TR; 9th in EN; 10th in FA, BN, HI; 11th in KO; 14th in ZH, MIX; 18th in MULTI) first trained a multi-task model on 7 SemEval tasks and then fine-tuned for each task individually. They report that such a multi-tasking and fine-tuning approach is beneficial for a subset of the tasks.

OPDAI (Chen et al., 2022b) (3rd in ZH; 9th in HI; 10th in MIX; 12th in MULTI) used a hybrid technique with multiple stages involving model ensemble using neural model, soft templates, and

Wikipedia lexicons. Their strong performance in ZH is powered by RoBERTa-wwm (Cui et al., 2021) pre-trained on Chinese data and Chinese word embeddings (Song et al., 2018).

CASIA (Fu et al., 2022) only participated in ZH and ranked 2nd. They built a hybrid system based on RoBERTa-wwm and used three training mechanisms (adversarial training, child-Tuning training, and continued pre-training). Additionally, they performed a series of data augmentation steps.

PAI (Ma et al., 2022) (3rd in EN) used string matching to retrieve entities with types from the LUKE entity dictionary (Yamada et al., 2020) for a given text. Then they concatenated the entity information with the input text and fed it to a pre-trained BERT model to build the NER system.

SU-NLP (Çarık et al., 2022) only participated in TR and ranked 3rd. Given an input text, they query an information retrieval (IR) system that indexes Wikipedia articles. Retrieved documents are used as context and a Turkish BERT variant (BERTurk) is used to encode the context and candidate mentions, with classifier heads for NER.

Infrd.ai (He et al., 2022) participated in nine tracks (EN, ES, NL, KO, DE, ZH, HI, BN, MULTI) and their best rank is 4th for ES. They trained a multilingual model with an XLM-RoBERTa base encoder, whose embeddings were passed into a BiLSTM encoder, which finally passed the encoded tokens to a CRF layer for classification. They also used an ensemble strategy with majority voting.

UM6P-CS (Mekki et al., 2022) ranked 6th in MULTI and 8th in MIX. They introduced several self-training and auxiliary tasks that aim to improve NER classification performance on top of XLM-RoBERTa. The auxiliary task of span classification focused on addressing the mention detection performance of the model, which essentially ensures that the model has good coverage of all named entities, regardless of their type. In terms of self-training, the authors predicted weak labels on the unlabelled test set and concatenated both datasets into one. The impact of self-training seems to have a significant impact with 3% improvement in terms of Precision, and 2.24% in terms of F1 score.

Multilinguals (Pandey et al., 2022a,b) participated in EN, ES, and ZH and best rank is 10th in ZH. They applied a BERT encoder with different classification heads: a linear layer, a CRF layer, and a BiLSTM-CRF. BERT and linear approach worked best for EN. For ES and ZH, they pre-trained BERT

using the Whole Word Masking (WWM) learning objective over Wikipedia data and the CRF classification head worked best for these tracks.

L3i (Boros et al., 2022) participated in all tracks and best rank is 7th in DE. They used Sentence-BERT (Reimers and Gurevych, 2019) to retrieve the most similar sentence from the training set and used it as context by adding it to the test text. Their model consists of a BERT encoder with a Transformer layer and a CRF head for classification.

MarSan (Tavan and Najafi, 2022) participated in all tracks and best rank is 9th in FA. They used T5 (monolingual and multilingual) to create feature vector for an input text. Then they performed a subtoken check step to mark the first subword as 1 and others as 0 (Subtoken check increased 4% F1). At the final stage, a Transformer layer is followed by a token prediction layer to perform NER.

TEAM-Atreides (Tasnim et al., 2022) only participated in BN and ranked 8th. They used an ensemble of mono-lingual ELECTRA-based models with majority voting. They also used data augmentation using translation and conducted experiments with non-contextual word embeddings.

UA-KO (Song and Bethard, 2022) ranked 8th in the KO track. They used GeoNames and the Encyclopedia of Korean Culture to incorporate entity names in the training set. Their model uses an ensemble approach with a soft-voting mechanism, combining the monolingual and multilingual models' predictions.

CSECU-DSG (Aziz et al., 2022) participated in all tracks and the best rank is 9th for ZH. The authors propose two approaches: (1) a BiLSTM-CRF that leverages stacked token embeddings from different sources, and (2) a Transformer-based encoder with a feed-forward classification head.

PA Ph&Tech (Lin et al., 2022) participated in ES, DE, and NL and best rank is 9th for NL. They used ensemble embedding from multiple transformers and reinforcement learning was also applied to maximize model accuracy. In an additional setting (Hou et al., 2022), they experimented with an ensemble approach, where they leveraged multiple transformers by assigning different weights in the transformer layers. Meanwhile, data augmentation is also applied to enlarge the training data.

Raccoons (Dogra et al., 2022) ranked 10th in the EN track. They focused on improving word representations for NER through a reinforcement trainer. This was done through a task model and

controller that repeatedly interact to update the embeddings.

AaltoNLP (Pietiläinen and Ji, 2022) participated in five tracks (EN, DE, FA, BN, KO) and the best rank is 11th for FA. Their approach consists of an ensemble strategy where they train two encoders jointly, allowing the models to combine the scores from the different encoders via a linear layer. Different models used different random seeds.

LMN (Lai, 2022) ranked 12th in the EN track. They applied a transfer-based encoder with a feed-forward classification head with a CRF layer. Their best variant used the ALBERT-xxlarge model. They also experimented with entity linking with Wikipedia and augmenting data with entities of the same type.

UC3M-PUCPR (Schneider et al., 2022) participated in EN, ES, MIX, and their best rank is 12th for MIX. They have used an ensemble of language-specific pre-trained LMs with soft-voting to make the final predictions.

NamedEntityRangers (Miftahova et al., 2022) ranked 16th in the MULTI track. They used RemBERT and mT5 to experiment with two approaches, where the first approach is the classical token classification method and the second method uses a template-free paradigm in which an encoder-decoder model translates the input sequence of words to a special output, encoding named entities with the predefined label.

CMNEROne (Dowlagar and Mamidi, 2022) ranked 16th in MIX. Their approach involves fine-tuning multilingual BERT on code-mixed data. To learn language-agnostic features, they pre-trained the model for a downstream task of language identification using the multilingual dataset.

KDDIE (Martin et al., 2022) only participated in the EN track and ranked 16th. They experimented by fine-tuning BERT and DeBERTa-based models and their best system is a fine-tuned DeBERTa-XLarge model.

DS4DH (Rouhizadeh and Teodoro, 2022) ranked 20th in the MULTI track. Their approach involves fine-tuning different pre-trained LMs (Multilingual-BERT, XLM-RoBERTa-base, XLM-RoBERTa-Large, Distilbert-Multilingual) with different classification heads like CRF and fully-connected layer.

NCUEE-NLP (Lee et al., 2022) ranked 7th in the ZH track. They used external data collected from MSRA, Weibo, People Daily, Boson,

Class	Baseline	DAMO-NLP	USTC-NELSLIP	QTrade AI
PER	63.88	92.07 (+28)	90.76 (+27)	87.20 (+23)
LOC	51.87	86.52 (+35)	86.81 (+35)	80.79 (+29)
CORP	49.61	84.55 (+35)	87.86 (+38)	77.23 (+28)
PROD	44.36	84.32 (+40)	81.05 (+37)	75.23 (+31)
GRP	39.28	79.90 (+41)	81.52 (+42)	71.66 (+32)
CW	37.68	84.49 (+47)	83.81 (+46)	73.85 (+36)

Table 4: F1 scores of the baseline and top three systems in the MULTI track for each class.

CLUNER, and LG, and trained a BiLSTM-CRF model with embeddings from a BERT model pre-trained on Chinese data.

DANGNT-SGU (Nguyen and Huynh, 2022) ranked 25th in the EN track by fine-tuning RoBERTa on the training data.

silpa_nlp (Singh et al., 2022) ranked 14th in HI and BN by fine-tuning XLM-R on the training set.

6 Insights from the Systems

6.1 Advancing the State of the Art

Identifying Complex Entities From the ranking in Table 3, we see that almost all the teams could outperform the official baseline system described in Section 4 in all the tracks. For most of the tracks, the top two teams DAMO-NLP and USTC-NELSLIP’s performance gap is very small compared to third place. To better understand this difference, we look at per-class performance. In Table 4, we show per-class F1 scores for the top three teams in the MULTI track. Although the systems performed better than the official baseline by a large margin, complex entities like creative works, products, and groups are still the most difficult ones to identify. This analysis shows that the largest gains by the top systems leveraging external knowledge came from classes containing complex NEs, e.g. CW and GRP.

Domain Adaptation The official baseline system performed poorly in terms of domain adaptation and achieved much lower F1 in MSQ-NER and ORCAS-NER compared to LOWNER. Intuitively, augmenting the training data with interrogative sentences could be a way to perform better in these domains. However, we observe that the participants could overcome the challenge of domain adaptation without especially including questions and queries

⁶<https://huggingface.co/bert-base-chinese>

⁷<https://github.com/SKTBrain/KoBERT>

⁸<https://github.com/monologg/KoELECTRA>

⁹<https://huggingface.co/kykim/bert-kor-base>

¹⁰<https://huggingface.co/wietsedv/bert-base-dutch-cased>

¹¹<https://huggingface.co/dbmdz/bert-base-german-uncased>

Multilingual
XLM-RoBERTa (XLM-R; Conneau et al. (2020)) : DAMO-NLP, USTC-NELSLIP, QTrade AI, SeqL, CMB AI Lab, RACAI, Sliced, Infrd.ai, UM6P-CS, UA-KO, CSECU-DSG, PA Ph&Tech, Raccoons, AaltoNLP, UC3M-PUCPR, DS4DH, silpa_nlp
mT5 (Xue et al., 2021) : CMB AI Lab, MarSan, NamedEntityRangers
mBERT (Devlin et al., 2019) : Sliced, L3i, PA Ph&Tech, UC3M-PUCPR, CMNEROne, DS4DH
RemBERT (Chung et al., 2021) : Sliced, NamedEntityRangers
English
BERT (Devlin et al., 2019) : PAI, Multilinguals, CSECU-DSG, PA Ph&Tech, Raccoons, UC3M-PUCPR, KDDIE
BigBird RoBERTa (Zaheer et al., 2021) : L3i
T5 (Raffel et al., 2020) : MarSan
XLNet (Yang et al., 2019) : PA Ph&Tech
ALBERT (Lan et al., 2020) : LMN
RoBERTa (Liu et al., 2019) : UC3M-PUCPR, DANGNT-SGU
DistillBERT (Sanh et al., 2019), ELECTRA (Clark et al., 2020) : UC3M-PUCPR
DeBERTa (He et al., 2021) : KDDIE
Spanish
Spanish BERT (Canete et al., 2020) : Multilinguals
BERT-wwm : L3i
Beto (Cañete et al., 2020), Spanish RoBERTa (Gutiérrez-Fandiño et al., 2021) : UC3M-PUCPR
Chinese
RoBERTa-wwm (Cui et al., 2021) : OPDAI, CASIA, Multilinguals
BERT⁶ : L3i, NCUEE-NLP
Korean
KoBERT⁷, Ko-ELECTRA⁸, KR-BERT (Lee et al., 2020), KLUE-RoBERTa (Park et al., 2021) : UA-KO
BERT⁹ : L3i
Bangla
BanglaBERT (Bhattacharjee et al., 2022) : TEAM-Atreides
Dutch
BERT¹⁰ : L3i
Farsi
ParsBERT (Farahani et al., 2020) : L3i
German
BERT¹¹ : L3i
Hindi
IndicBERT (Kakwani et al., 2020) : silpa_nlp
Russian
RuBERT : L3i
Turkish
BERTurk (Schweter, 2020) : SU-NLP, L3i

Table 5: Pre-trained Transformer language models used by the teams for different languages. BERT models for non-English languages are trained on the specific languages’ data with BERT architecture by the community.

in their external data. For example, DAMO-NLP found that their approach of retrieving Wikipedia paragraphs not only provided a strong performance on LOWNER, but also helped with cross-domain transferability.

Adapting to MSQ was easier compared to ORCAS-NER for all the tracks except Bangla. The top systems like DAMO-NLP and USTC-NELSLIP struggled in MSQ-NER for Bangla, while they typically had higher F1 scores for MSQ-NER than ORCAS-NER for the other tracks. This could be an interesting direction to explore in the future.

6.2 Other Insights

External Data In Section 5 we observe that such superior performance by these top systems became possible by exploiting external knowledge during learning and inference. While USTC-NELSLIP used knowledge from pre-trained language models to fine-tune Gazetteer presentations, DAMO-NLP directly used raw texts from Wikipedia to inject context and it gave them an advantage over USTC-NELSLIP in most tracks.

As the availability of external data is higher for English compared to other languages, most of the teams participating in other languages used publicly available pre-trained models for other languages, or translated data from other languages. For example, CASIA augmented data from other languages with translation, and it helped them to secure second place in the Chinese track. In general, a large portion of the participating teams showed that they can do better if they can go beyond the provided training data, and use external data or pre-trained language models for different languages to inject external knowledge in some way.

Modeling Approaches Almost all participating systems relied on publicly available Transformer (Vaswani et al., 2017) based pre-trained language models (Table 5). XLM-RoBERTa (a.k.a. XLM-R) was the most popular choice for building multilingual models. Most of the teams participating in non-English monolingual tracks preferred this particular model to the multilingual variant of BERT.

Other recent language models like T5, ELECTRA, XLNet, and ALBERT were used by some of the teams, but mostly for English. We observed that for non-English languages, many teams used community-developed pre-trained models for other languages like Chinese, Hindi, Spanish, Korean,

Bangla, Turkish, Russian, Farsi, Dutch, and German. Most of such models are trained using the BERT architecture with data for the respective languages. A lot of teams relied on the strength of Conditional Random Field (CRF; Lafferty et al. 2001) for sequence labeling problems and adopted it to gain stronger performance. Very few teams used architectures like LSTMs.

Teams that simply fine-tuned pre-trained language models performed similarly to the baseline system for most of the tracks. Apart from the previously mentioned role of external data, another vital component for strong performance is using ensemble learning strategies. Almost all the strong performing teams trained multiple models and ensembled them for making the final predictions. We have also observed some teams experimenting with adversarial training and reinforcement learning.

7 Conclusion

In this paper, we have presented an overview of the SemEval shared task on identifying complex entities in multiple languages. In this shared task, we have received system submissions from 55 competing teams, and 34 system description papers. On average, the winning systems for all the tracks outperformed the baseline system by a large margin of 35% F1.

Most of the top-performing teams in MULTICONER utilized external knowledge bases like Wikipedia and Gazetteer. They also tend to use XLM-RoBERTa as the pre-trained language model. In terms of modeling approaches, ensemble strategies helped the systems to achieve strong performance. Results from the top teams indicate that identifying complex entities like creative works is still difficult among all the classes even with the usage of external data.

References

- Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153.
- Sandeep Ashwini and Jinho D. Choi. 2014. Targetable named entity recognition in social media. *CoRR*, abs/1408.0782.
- Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity

- recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.
- Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. 2022. CSECU-DSG at SemEval-2022 Task 11: Identifying the Multilingual Complex Named Entity in Text Using Stacked Embeddings and Transformer based Approach. In *The 16th International Workshop on Semantic Evaluation*.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.
- Abhik Bhattacharjee, Tahmid Hasan, Kazi Mubasshir, Md. Saiful Islam, Wasi Ahmad Uddin, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. [Banglabert: Lagnuage model pretraining and benchmarks for low-resource language understanding evaluation in bangla](#). In *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2022*.
- Emanuela Boros, Carlos-Emiliano González-Gallardo, Jose G Moreno, and Antoine Doucet. 2022. L3i at SemEval-2022 Task 11: Straightforward Additional Context for Multilingual Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.
- Buse Çank, Fatih Beyhan, and Reyhan Yeniterzi. 2022. SU-NLP at SemEval-2022 Task 11: Complex Named Entity Recognition with Entity Linking. In *The 16th International Workshop on Semantic Evaluation*.
- 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*.
- Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022a. USTC-NELSLIP at SemEval-2022 Task 11: Gazetteer-Adapted Integration Network for Multilingual Complex Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Ze Chen, Kangxu Wang, Jiewen Zheng, Zijian Cai, Jiarong He, and Jin Gao. 2022b. OPDAI at SemEval-2022 Task 11: A hybrid approach for Chinese NER using outside Wikipedia knowledge. In *The 16th International Workshop on Semantic Evaluation*.
- Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. [Rethinking embedding coupling in pre-trained language models](#). In *International Conference on Learning Representations*.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [Electra: Pre-training text encoders as discriminators rather than generators](#). In *International Conference on Learning Representations*.
- 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 8440–8451, Online. Association for Computational Linguistics.
- Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. Orcas: 18 million clicked query-document pairs for analyzing search. *arXiv preprint arXiv:2006.05324*.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. [Pre-training with whole word masking for chinese bert](#). *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3504–3514.
- Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017a. 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.
- Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017b. [Results of the WNUT2017 shared task on novel and emerging entity recognition](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 140–147. 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. *arXiv preprint arXiv:1810.04805*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics.
- Atharvan Dogra, Prabsimran Kaur, Guneet Singh Kohli, and Jatin Bedi. 2022. Raccoons at SemEval-2022 Task 11: Leveraging Concatenated Word Embeddings for Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Suman Dowlagar and Radhika Mamidi. 2022. CM-NEROne at SemEval-2022 Task 11: Code-Mixed Named Entity Recognition by leveraging multilingual data. In *The 16th International Workshop on Semantic Evaluation*.
- Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *ArXiv*, abs/2005.12515.

- Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Jia Fu, Zhen Gan, Zhucong Li, Sirui Li, Dianbo Sui, Yubo Chen, Kang Liu, and Jun Zhao. 2022. CASIA at SemEval-2022 Task 11: Chinese Named Entity Recognition for Complex and Ambiguous Entities. In *The 16th International Workshop on Semantic Evaluation*.
- Weichao Gan, Yuanping Lin, Guangbo Yu, Guimin Chen, and Qian Ye. 2022. Qtrade AI at SemEval-2022 Task 11: An Unified Framework for Multilingual NER Task. In *The 16th International Workshop on Semantic Evaluation*.
- Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodríguez Penagos, and Marta Villegas. 2021. [Spanish language models](#). *CoRR*, abs/2107.07253.
- Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. [UKP-athene: Multi-sentence textual entailment for claim verification](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.
- Fadi Hassan, Wondimagegnhue Tufa, Guillem Collell, Piek Vossen, Lisa Beinborn, Adrian Flanagan, and Kuan Eeik Tan. 2022. SeqL at SemEval-2022 Task 11: An Ensemble of Transformer Based Models for Complex Named Entity Recognition Task. In *The 16th International Workshop on Semantic Evaluation*.
- JiangLong He, Akshay Uppal, Mamatha N, Shiv Vignesh, Deepak Kumar, and Aditya Kumar Sarda. 2022. Infrd.ai at SemEval-2022 Task 11: A system for named entity recognition using data augmentation, transformer-based sequence labeling model, and EnsembleCRF. In *The 16th International Workshop on Semantic Evaluation*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). In *International Conference on Learning Representations*.
- Changyu Hou, Jun Wang, Yixuan Qiao, Peng Jiang, Peng Gao, Guotong Xie, Qizhi LIN, Xiaopeng Wang, Xiandi Jiang, Benqi Wang, and Qifeng Xiao. 2022. SFE-AI at SemEval-2022 Task 11: Low-Resource Named Entity Recognition using Large Pre-trained Language Models. In *The 16th International Workshop on Semantic Evaluation*.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. [Conditional random fields: Probabilistic models for segmenting and labeling sequence data](#). In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Ngoc Minh Lai. 2022. LMN at SemEval-2022 Task 11: A Transformer-based System for English Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [Albert: A lite bert for self-supervised learning of language representations](#). In *International Conference on Learning Representations*.
- Lung-Hao Lee, Chien-Huan Lu, and Tzu-Mi Lin. 2022. NCUEE-NLP at SemEval-2022 Task 11 Chinese Named Entity Recognition Using the BERT-biLSTM-CRF model. In *The 16th International Workshop on Semantic Evaluation*.
- Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. Kr-bert: A small-scale korean-specific language model. *ArXiv*, abs/2008.03979.
- Qizhi Lin, Changyu Hou, Xiaopeng Wang, Jun Wang, Yixuan Qiao, Peng Jiang, Xiandi Jiang, Benqi Wang, and Qifeng Xiao. 2022. PA Ph&Tech at SemEval-2022 Task 11: NER Task with Ensemble Embedding from Reinforcement Learning. In *The 16th International Workshop on Semantic Evaluation*.
- 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](#). Cite arxiv:1907.11692.
- Jackson Luken, Nanjiang Jiang, and Marie-Catherine de Marneffe. 2018. [QED: A fact verification system for the FEVER shared task](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 156–160, Brussels, Belgium. Association for Computational Linguistics.
- Long Ma, Xiaorong Jian, and Xuan Li. 2022. PAI at SemEval-2022 Task 11: Name Entity Recognition with Contextualized Entity Representations and Robust Loss Functions. In *The 16th International Workshop on Semantic Evaluation*.

- Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.
- Caleb Martin, Huichen Yang, and William Hsu. 2022. Kddie at SemEval-2022 Task 11: Using DeBERTa for Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. In *EMNLP/IJCNLP (1)*, pages 6255–6260. Association for Computational Linguistics.
- Abdellah El Mekki, Abdelkader El Mahdaouy, Mohammed Akallouch, Ismail Berrada, and Ahmed Khoumsi. 2022. UM6P-CS at SemEval-2022 Task 11: Enhancing Multilingual and Code-Mixed Complex Named Entity Recognition via Pseudo Labels using Multilingual Transformer. In *The 16th International Workshop on Semantic Evaluation*.
- 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1499–1512. Association for Computational Linguistics.
- Amina Miftahova, Alexander Pugachev, Artem Skiba, Katya Artemova, Tatiana Batura, Pavel Braslavski, and Vladimir V. Ivanov. 2022. Namedentityrangers at SemEval-2022 Task 11: Transformer-based Approaches for Multilingual Complex Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Dang Tuan Nguyen and Huy Khac Nguyen Huynh. 2022. DANGNT-SGU at SemEval-2022 Task 11: Using Pre-trained Language Model for Complex Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Vasile Pais. 2022. RACAI at SemEval-2022 Task 11: Complex named entity recognition using a lateral inhibition mechanism. In *The 16th International Workshop on Semantic Evaluation*.
- Amit Pandey, Swayatta Daw, and Vikram Pudi. 2022a. Multilinguals at SemEval-2022 Task 11: Transformer Based Architecture for Complex NER. In *The 16th International Workshop on Semantic Evaluation*.
- Amit Pandey, Swayatta Daw, Narendra Babu Unnam, and Vikram Pudi. 2022b. Multilinguals at SemEval-2022 Task 11: Complex NER in Semantically Ambiguous Settings for Low Resource Languages. In *The 16th International Workshop on Semantic Evaluation*.
- Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyeon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. [Klue: Korean language understanding evaluation](#).
- Aapo Pietiläinen and Shaoxiong Ji. 2022. AaltoNLP at SemEval-2022 Task 11: Ensembling Task-adaptive Pretrained Transformers for Multilingual Complex NER. In *The 16th International Workshop on Semantic Evaluation*.
- Barbara Plank. 2022. Sliced at SemEval-2022 Task 11: Bigger, Better? Massively Multilingual Language Models for Multilingual Complex NER. In *The 16th International Workshop on Semantic Evaluation*.
- KEYU PU, Hongyi LIU, Yixiao YANG, Jiangzhou JI, Wenyi LV, and Yaohan He. 2022. CMB AI lab at SemEval-2022 Task 11: A Two-Stage Approach for Complex Named Entity Recognition via Span Boundary Detection and Span Classification. In *The 16th International Workshop on Semantic Evaluation*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Hossein Rouhizadeh and Douglas Teodoro. 2022. DS4DH at SemEval-2022 Task 11: Multilingual Named Entity Recognition Using an Ensemble of Transformer-based Language Models. In *The 16th International Workshop on Semantic Evaluation*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.
- Elisa Terumi Rubel Schneider, Renzo Mauricio Rivera Zavala, Paloma Martinez Fernandez, Claudia Moro, and EMERSON CABRERA PARAISO. 2022. UC3M-PUCPR at SemEval-2022 Task 11: An Ensemble Method of Transformer-based Models for Complex Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Stefan Schweter. 2020. [Berturk - bert models for turkish](#).

- Sumit Singh, Pawankumar Jawale, and Uma Shanker Tiwary. 2022. [silpa_nlp](#) at SemEval-2022 Tasks 11: Transformer based NER models for Hindi and Bangla languages. In *The 16th International Workshop on Semantic Evaluation*.
- Hyunju Song and Steven Bethard. 2022. UA-KO at SemEval-2022 Task 11: Data Augmentation and Ensembles for Korean Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. [Directional skip-gram: Explicitly distinguishing left and right context for word embeddings](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, New Orleans, Louisiana. Association for Computational Linguistics.
- Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.
- Nazia Tasnim, Md. Istiak Hossain Shihab, Asif Shahriyar Sushmit, Steven Bethard, and Farig Sadeque. 2022. TEAM-atreides at SemEval-2022 Task 11: On leveraging data augmentation and ensemble to recognize complex Named Entities in Bangla. In *The 16th International Workshop on Semantic Evaluation*.
- Ehsan Tavan and Maryam Najafi. 2022. Marsan at SemEval-2022 Task 11: Multilingual complex named entity recognition using T5 and transformer encoder. In *The 16th International Workshop on Semantic Evaluation*.
- Rob van der Goot. 2022. Machamp at SemEval-2022 Tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task Multilingual Learning for a Pre-selected Set of Semantic Datasets. In *The 16th International Workshop on Semantic Evaluation*.
- Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. [Massive choice, ample tasks \(MaChAmp\): A toolkit for multi-task 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022. DAMO-NLP at SemEval-2022 Task 11: A Knowledge-based System for Multilingual Named Entity Recognition. In *The 16th International Workshop on Semantic Evaluation*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. [LUKE: Deep contextualized entity representations with entity-aware self-attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. [Big bird: Transformers for longer sequences](#).

Appendix

In this section, we provide the domain specific performance of the teams on each track. For each team, we report precision, recall, and F1 for the three domains, i.e., LOWNER, ORCAS, and MSQ. We also highlight the baseline system’s performance breakdown for each track. Each track’s result is presented in its individual table as listed here:

- Table 6 Bangla (BN)
- Table 7 German (DE)
- Table 8 English (EN)
- Table 9 Spanish (ES)
- Table 10 Farsi (FA)
- Table 11 Hindi (HI)
- Table 12 Korean (KO)
- Table 13 Dutch (NL)
- Table 14 Russian (RU)
- Table 15 Turkish (TR)
- Table 16 Chinese (ZH)
- Table 17 Code-Mixed (MIX)
- Table 18 Multi-lingual (MULTI)

A Detailed Results

A.1 Bangla (BN)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	USTC-NELSLIP	87.32	85.82	86.56	84.2	81.78	82.37	77.21	73.44	75.12	85.84	83.43	84.24
2	DAMO-NLP	86.42	86.21	86.31	83.24	81.25	82.13	73.45	70.9	72.0	84.63	82.53	83.51
3	NetEase.AI	85.83	84.53	85.16	70.32	64.64	67.17	63.36	61.69	62.14	73.78	68.39	70.88
4	RACAI	83.09	83.48	83.28	63.62	60.98	61.6	63.36	58.51	60.69	68.08	65.33	66.28
5	Infrrd.ai	81.52	81.82	81.66	61.59	58.48	59.5	56.06	57.36	56.45	65.68	62.99	63.99
6	YNUNLP	81.46	81.01	81.21	60.64	58.68	59.12	58.58	56.63	57.45	65.11	63.14	63.8
7	Sliced	82.88	83.36	83.1	59.71	57.88	58.06	57.1	57.77	57.12	64.24	62.8	63.05
8	Team Atreides	84.23	82.8	83.48	56.8	52.85	54.23	55.12	58.34	55.44	62.09	58.25	59.75
9	brotherhood	81.56	80.71	81.12	55.18	52.0	53.3	50.91	54.78	51.86	60.33	57.24	58.63
10	MaChAmp	78.44	79.84	79.13	52.18	51.52	51.02	54.05	52.96	52.87	57.25	56.61	56.46
11	MarSan	79.04	79.04	78.98	51.83	48.05	48.83	42.42	50.57	43.92	56.48	53.77	54.22
12	EURECOM	75.36	73.45	74.37	49.33	47.05	48.12	45.52	50.4	45.28	53.78	51.51	52.57
13	AaltoNLP	79.09	78.46	78.74	49.19	43.34	45.78	48.42	47.42	45.84	55.09	49.27	51.79
14	silpa_nlp	76.37	75.97	76.16	47.42	44.68	45.61	44.86	48.77	45.52	53.0	50.34	51.39
15	CSECU-DSG	74.96	74.96	74.95	46.84	43.8	44.85	45.6	48.63	46.14	52.21	49.42	50.55
16	BaselineExtendingPokemons	72.49	75.55	73.96	38.86	40.68	39.2	40.13	44.07	40.97	44.48	46.3	45.07
17	L3i	73.5	72.57	73.01	40.52	38.43	39.08	39.87	42.34	39.82	46.08	43.94	44.81
18	Enigma	73.2	73.34	72.96	41.16	36.48	37.1	39.47	40.82	36.11	46.64	42.03	42.68
19	Baseline	69.27	69.88	69.54	34.12	34.67	34.16	34.03	37.57	34.56	39.29	39.81	39.41

Table 6: Detailed results for Bangla track.

A.2 German (DE)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	94.87	94.92	94.89	84.74	84.4	84.4	84.94	87.8	86.18	90.85	90.5	90.65
2	USTC-NELSLIP	95.8	95.03	95.41	80.5	78.8	79.33	87.06	86.65	86.83	89.88	88.35	89.05
3	RACAI	91.76	91.15	91.44	62.79	61.99	61.89	70.38	73.61	71.61	80.01	78.97	79.39
4	Sliced	90.94	91.04	90.99	61.53	62.81	61.53	71.32	72.85	71.9	78.84	79.18	78.9
5	MaChAmp	89.51	89.87	89.69	61.85	63.71	62.16	69.0	73.04	70.63	78.13	78.83	78.38
6	YNUNLP	90.22	89.6	89.9	59.59	60.56	59.23	70.62	70.7	70.23	77.51	77.42	77.32
7	L3i	90.71	90.72	90.71	60.0	56.23	57.46	64.26	67.57	65.32	78.58	76.18	77.23
8	ML-HUB	88.39	87.7	88.03	59.73	58.93	59.06	60.55	69.53	63.48	76.63	75.8	76.14
9	brotherhood	90.05	89.66	89.85	56.83	55.8	55.78	64.29	67.98	65.53	76.57	75.52	75.94
10	Infrrd.ai	90.65	85.64	88.06	63.26	54.19	57.6	70.52	65.87	67.84	80.05	72.47	75.9
11	EURECOM	88.89	88.68	88.77	56.16	54.01	53.87	62.18	64.02	62.58	75.61	73.84	74.43
12	MarSan	88.4	89.05	88.7	52.14	52.92	51.53	57.75	61.82	58.54	73.1	73.6	73.12
13	CSECU-DSG	86.43	84.81	85.6	56.79	50.75	53.01	61.93	60.99	61.15	74.93	70.47	72.49
14	AaltoNLP	86.49	87.0	86.73	52.31	46.07	48.37	58.33	60.85	59.04	73.16	69.92	71.37
15	PA Ph&Tech	86.08	74.14	79.5	53.42	45.67	48.58	56.6	57.15	55.79	72.65	62.35	66.75
16	BaselineExtendingPokemons	83.81	85.25	84.52	40.75	44.29	42.04	47.58	56.25	50.58	65.44	67.99	66.59
17	Baseline	80.64	81.16	80.83	39.86	41.06	39.96	46.89	55.54	49.6	63.45	64.41	63.74

Table 7: Detailed results for German track.

A.3 English (EN)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	96.68	96.87	96.78	84.27	83.51	83.72	81.76	85.69	83.5	91.54	90.95	91.22
2	USTC-NELSLIP	92.83	91.38	92.09	76.78	75.05	75.59	80.1	83.66	81.74	86.41	84.67	85.47
3	PAI	90.87	91.05	90.96	62.2	60.08	60.58	64.79	68.2	65.91	79.09	77.88	78.37
4	ML-HUB	90.27	87.61	88.9	68.47	56.88	61.6	73.17	67.47	69.46	82.24	74.68	78.14
5	RACAI	88.67	88.08	88.37	59.37	57.87	57.84	67.77	70.11	68.75	76.54	75.35	75.78
6	Infrd.ai	88.29	87.05	87.65	58.0	55.11	56.15	64.7	68.89	65.83	75.97	73.6	74.71
7	EURECOM	88.82	89.06	88.94	54.1	55.31	54.25	62.55	65.93	63.41	74.43	74.9	74.57
8	Sliced	87.47	87.99	87.73	56.99	57.19	56.17	67.39	69.0	68.06	74.53	74.93	74.54
9	MaChAmp	86.21	87.25	86.72	57.3	58.18	57.11	64.97	69.55	66.75	74.16	74.97	74.48
10	Raccoons	87.66	89.39	88.5	53.63	55.09	54.02	63.57	68.39	65.42	73.43	75.05	74.18
11	YNUNLP	86.75	86.92	86.83	53.96	55.4	53.98	64.99	68.33	65.78	72.99	73.64	73.17
12	LMN	87.05	88.71	87.87	50.96	52.17	51.2	58.43	63.84	60.45	71.78	73.33	72.5
13	brotherhood	87.16	86.41	86.78	52.76	51.48	51.67	61.74	65.7	62.73	73.18	71.71	72.35
14	L3i	87.21	87.34	87.26	54.6	47.87	49.71	61.33	64.08	62.57	73.82	70.8	71.96
15	Multilinguals	86.47	87.43	86.94	53.03	48.86	50.16	59.4	59.55	59.11	72.71	71.09	71.74
16	KDDIE	86.65	87.7	87.17	50.36	51.15	50.4	58.29	63.63	60.57	71.34	72.26	71.73
17	MarSan	85.75	86.21	85.96	50.83	52.24	51.16	58.91	64.9	60.64	71.11	71.91	71.45
18	Cardiff NLP	85.93	87.6	86.75	47.63	51.41	49.24	56.23	64.49	58.55	69.72	72.28	70.94
19	Lone Wolf	85.08	85.96	85.51	47.36	48.76	47.75	56.33	62.44	58.47	69.35	70.31	69.77
20	MIDAS	84.68	81.79	83.19	54.75	46.84	49.73	60.63	57.45	58.34	72.95	66.95	69.62
21	UC3M-PUCPR	86.6	87.12	86.84	46.23	46.28	44.25	54.95	57.3	54.07	69.95	69.73	69.24
22	CSECU-DSG	84.76	86.08	85.41	47.0	47.79	47.22	50.45	60.14	54.01	68.72	69.81	69.24
23	Sartipi-Sedighin	82.95	84.69	83.78	44.4	47.16	45.6	46.42	58.19	49.72	66.34	68.79	67.51
24	Enigma	82.55	83.19	82.86	46.14	46.04	45.45	57.07	61.73	58.17	66.97	67.74	67.19
25	DANGNT-SGU	83.6	84.7	84.1	43.14	44.68	43.28	51.75	58.74	53.38	66.51	67.7	66.89
26	AaltoNLP	83.27	84.19	83.73	48.89	33.87	39.24	53.95	54.21	53.51	71.57	63.0	66.85
27	SPDB Innovation Lab	81.35	81.74	81.54	42.16	43.57	42.63	49.45	57.98	52.18	64.52	65.77	65.11
28	silpa_nlp	81.48	80.54	80.99	39.58	38.98	38.65	48.81	54.1	49.95	64.13	63.06	63.42
29	BaselineExtendingPokemons	80.03	82.27	81.11	38.13	42.3	39.96	43.97	57.67	48.84	61.36	65.35	63.24
30	Baseline	78.25	78.0	78.11	38.89	37.47	37.61	46.21	52.2	48.26	62.07	60.97	61.36
31	AutoNER	72.29	74.77	73.35	30.6	33.53	31.02	45.77	49.65	47.21	54.73	57.68	55.72

Table 8: Detailed results for English track.

A.4 Spanish (ES)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	96.23	96.15	96.19	82.45	80.91	81.33	82.25	84.51	83.1	90.58	89.41	89.94
2	USTC-NELSLIP	90.15	88.08	89.1	80.99	79.17	79.68	84.49	85.95	85.1	86.64	84.39	85.44
3	RACAI	85.29	85.31	85.29	63.37	62.39	61.74	71.98	72.92	72.35	76.21	75.43	75.62
4	Infrrd.ai	85.21	85.55	85.37	62.06	61.71	61.32	66.18	70.93	67.82	75.59	75.11	75.26
5	MaChAmp	84.71	85.34	85.01	61.53	62.99	61.49	67.99	72.11	69.5	74.94	75.66	75.2
6	Sliced	85.68	85.92	85.79	60.74	61.6	60.39	69.21	71.57	70.18	75.15	75.32	75.11
7	YNUNLP	84.18	85.45	84.8	57.41	58.71	56.95	68.33	68.64	68.17	72.93	73.8	73.17
8	brotherhood	85.66	84.73	85.19	51.7	51.52	51.08	59.55	62.29	60.31	71.23	70.35	70.69
9	L3i	83.73	84.32	84.01	48.98	50.03	48.94	52.25	58.52	54.73	68.71	69.34	68.93
10	PA Ph&Tech	82.95	81.48	82.21	51.11	52.8	51.49	51.07	64.25	55.15	68.89	69.23	68.93
11	MarSan	83.12	82.84	82.96	49.26	50.7	48.52	56.64	60.4	57.46	68.65	68.71	68.3
12	SPDB Innovation Lab	83.57	81.69	82.55	49.62	48.7	46.57	60.29	57.49	57.59	68.96	67.24	67.31
13	CSECU-DSG	82.87	79.64	81.2	47.04	41.64	43.17	55.04	53.7	53.72	68.94	63.13	65.62
14	EURECOM	80.25	80.44	80.31	40.24	41.26	40.25	40.59	48.14	43.16	62.49	63.26	62.77
15	Multilinguals	81.13	80.52	80.81	36.73	34.24	34.69	42.77	45.66	43.57	62.27	60.46	61.2
16	Sartipi-Sedighin	77.29	79.74	78.36	37.11	39.41	37.62	42.44	47.33	43.67	59.82	62.03	60.7
17	BaselineExtendingPokemons	77.3	80.34	78.76	34.47	38.82	36.01	40.95	50.19	44.0	58.32	62.22	60.08
18	Baseline	75.66	77.0	76.24	33.32	36.11	33.58	41.34	45.99	43.08	57.07	59.08	57.84
19	UC3M-PUCPR	73.93	72.16	72.89	37.33	36.14	35.42	40.39	41.63	40.88	58.38	56.22	56.79

Table 9: Detailed results for Spanish track.

A.5 Farsi (FA)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	95.96	97.01	96.48	84.99	84.92	84.84	86.79	88.15	87.36	89.81	89.66	89.7
2	USTC-NELSLIP	86.13	84.49	85.29	88.2	86.57	87.2	92.45	91.11	91.75	88.16	86.1	87.05
3	RACAI	80.5	82.15	81.31	63.18	61.96	62.02	70.05	72.19	70.87	70.77	70.45	70.42
4	Sliced	79.17	82.11	80.61	59.39	61.16	59.93	64.35	69.55	66.23	67.83	69.77	68.66
5	YNUNLP	79.54	80.54	80.02	58.25	58.55	57.78	67.51	68.33	67.65	67.29	67.57	67.19
6	brotherhood	81.16	81.69	81.41	56.13	54.41	54.7	61.28	64.6	62.35	66.46	65.53	65.74
7	C-3PO	78.94	81.67	80.27	55.55	55.46	55.08	59.7	65.57	61.84	65.14	66.28	65.51
8	L3i	79.18	80.65	79.89	54.78	55.18	54.63	56.26	63.86	59.08	64.91	65.59	65.11
9	MarSan	76.49	80.84	78.59	51.54	50.99	50.61	55.71	57.7	56.29	61.8	63.06	62.14
10	MaChAmp	75.36	78.68	76.98	48.89	51.69	49.72	50.36	58.05	53.13	59.4	62.43	60.71
11	AaltoNLP	76.46	79.88	78.1	48.19	46.28	46.37	48.18	56.85	51.39	59.19	59.58	58.93
12	Sartipi-Sedighin	75.79	79.59	77.63	44.49	44.77	44.41	45.79	53.09	47.63	57.08	58.64	57.73
13	EURECOM	75.0	77.28	76.06	43.2	42.14	42.28	43.79	48.69	45.16	56.07	56.08	55.91
14	CSECU-DSG	75.96	79.15	77.5	41.98	41.09	41.06	43.58	49.03	45.55	55.8	56.17	55.81
15	Baseline	69.25	74.5	71.67	41.12	39.84	39.03	45.79	48.2	46.59	52.7	53.46	52.24
16	BaselineExtendingPokemons	70.89	77.3	73.91	36.59	39.87	37.62	34.33	46.97	39.02	49.15	54.33	51.26

Table 10: Detailed results for Farsi track.

A.6 Hindi (HI)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	84.85	83.54	84.18	86.82	84.85	85.75	88.54	89.94	89.2	87.27	85.28	86.23
2	USTC-NELSLIP	86.49	83.93	85.18	84.66	82.87	83.16	91.89	90.14	90.94	86.0	83.92	84.64
3	RACAI	82.04	81.99	82.01	64.17	62.46	62.65	75.62	75.96	75.55	69.05	67.77	68.08
4	Sliced	81.48	81.81	81.64	62.82	61.48	61.32	73.49	75.32	73.75	67.93	66.98	67.0
5	NetEase.AI	85.69	82.75	84.15	63.24	57.65	59.88	72.61	73.95	72.38	69.67	64.27	66.63
6	Infrd.ai	79.87	80.01	79.93	61.41	60.08	59.99	71.21	75.78	72.9	66.58	65.6	65.72
7	brotherhood	82.17	81.01	81.58	59.76	56.88	57.68	68.91	74.4	70.59	65.72	63.35	64.23
8	YNUNLP	79.67	79.89	79.77	58.21	57.96	57.35	69.94	72.85	70.7	63.8	63.69	63.39
9	OPDAI	74.82	75.9	75.28	57.86	59.07	57.86	67.39	70.83	68.76	63.03	63.58	62.94
10	MaChAmp	76.31	77.7	76.99	56.15	56.27	55.5	67.53	72.36	69.47	61.9	62.21	61.73
11	CSECU-DSG	75.97	71.45	73.56	55.49	48.73	51.54	66.57	65.47	65.16	61.46	54.77	57.68
12	MarSan	75.09	75.61	75.27	49.77	50.46	49.34	62.45	66.61	63.72	56.39	57.01	56.31
13	EURECOM	68.92	69.69	69.27	48.59	46.71	47.17	53.28	59.51	54.73	53.84	52.36	52.78
14	silpa_nlp	73.9	73.75	73.81	45.53	43.87	44.32	51.24	58.5	51.16	52.44	51.22	51.49
15	BaselineExtendingPokemons	70.78	72.33	71.49	42.34	43.21	42.33	55.07	61.31	55.57	49.68	50.68	49.9
16	L3i	72.38	71.34	71.8	44.0	42.07	42.26	51.0	58.13	52.89	51.01	49.24	49.73
17	Enigma	71.14	71.84	71.03	44.43	39.86	40.47	56.21	58.85	55.9	51.61	48.28	48.62
18	Baseline	65.67	66.44	65.96	41.38	42.59	41.55	54.03	56.67	53.26	48.08	48.98	48.22

Table 11: Detailed results for Hindi track.

A.7 Korean (KO)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	96.58	97.1	96.83	81.1	81.41	81.06	79.44	84.96	81.96	88.55	88.7	88.59
2	USTC-NELSLIP	90.64	90.11	90.37	83.23	81.15	81.82	88.45	87.97	88.19	87.39	85.56	86.36
3	RACAI	85.26	86.81	86.02	58.82	58.58	57.79	69.52	69.63	69.38	72.06	71.93	71.74
4	CMB AI Lab	88.93	88.23	88.57	60.73	47.02	52.7	64.9	58.35	61.09	75.92	66.33	70.7
5	Sliced	84.81	86.93	85.85	55.92	58.41	56.44	65.28	68.6	66.81	69.82	71.94	70.66
6	YNUNLP	84.74	85.42	85.05	57.39	57.11	56.36	66.48	68.39	67.34	70.69	70.51	70.33
7	C-3PO	86.24	87.42	86.8	51.02	49.85	49.69	56.08	57.72	56.27	68.15	67.4	67.49
8	UA-KO	85.91	87.78	86.83	50.59	49.63	49.67	55.92	59.24	56.92	67.72	67.52	67.49
9	brotherhood	85.83	86.67	86.24	50.8	50.69	50.23	57.33	61.72	58.98	67.61	67.5	67.41
10	Infrd.ai	84.15	86.13	85.13	50.75	52.07	50.99	58.9	63.65	60.55	66.69	68.17	67.29
11	MaChAmp	81.48	83.99	82.71	49.31	51.21	49.6	53.61	62.55	57.03	64.68	66.55	65.45
12	EURECOM	86.4	86.63	86.5	46.68	46.14	45.87	50.13	54.66	51.57	65.25	65.14	64.96
13	L3i	83.92	84.93	84.38	42.92	45.9	43.97	48.18	55.21	50.82	61.57	64.09	62.68
14	MarSan	81.58	84.79	83.14	43.31	45.49	43.75	47.76	54.49	49.99	61.13	63.92	62.26
15	CSECU-DSG	82.99	85.12	84.04	43.2	42.04	42.11	48.96	50.41	48.72	62.27	62.14	62.05
16	AaltoNLP	82.06	83.23	82.61	44.96	43.38	42.86	48.62	53.42	50.35	62.72	61.92	61.82
17	BaselineExtendingPokemons	77.42	82.93	80.06	39.65	41.7	40.06	41.75	51.34	44.83	57.46	60.84	58.95
18	Baseline	76.2	76.86	76.46	36.64	38.75	37.0	37.71	46.29	40.03	54.77	56.38	55.25

Table 12: Detailed results for Korean track.

A.8 Dutch (NL)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	97.92	98.0	97.96	81.16	80.39	80.46	83.17	84.26	83.65	90.95	90.14	90.5
2	USTC-NELSLIP	92.14	90.74	91.43	82.63	81.09	81.64	86.3	87.87	86.95	88.56	86.86	87.67
3	RACAI	89.68	89.73	89.7	63.9	63.4	62.83	70.34	72.81	71.2	78.82	78.3	78.41
4	Sliced	89.08	89.42	89.25	61.81	63.1	61.87	69.15	72.56	70.41	77.55	77.95	77.66
5	MaChAmp	88.03	88.65	88.33	61.35	62.83	61.44	67.87	71.78	69.47	76.72	77.43	76.99
6	Infrrd.ai	91.32	85.54	88.31	64.73	56.72	59.74	70.51	68.1	69.08	80.5	73.04	76.4
7	YNUNLP	88.95	88.21	88.56	59.39	59.79	58.78	67.04	70.12	68.25	76.19	75.82	75.82
8	brotherhood	88.86	88.05	88.44	53.74	52.63	52.35	58.5	63.87	60.22	73.96	72.46	73.04
9	PA Ph&Tech	87.49	86.76	87.11	51.28	55.73	52.76	56.39	66.13	59.98	71.06	73.32	72.05
10	MarSan	86.5	87.67	87.06	52.0	52.51	50.26	56.61	61.27	58.27	71.18	71.98	71.13
11	L3i	86.65	87.73	87.15	50.22	50.15	49.39	54.56	60.34	56.74	70.96	71.32	70.96
12	CSECU-DSG	84.82	81.82	83.24	50.84	42.96	45.53	59.8	55.21	57.11	71.74	65.0	67.94
13	EURECOM	82.05	84.25	83.1	45.27	46.43	44.8	49.11	56.91	52.23	66.11	67.75	66.7
14	BaselineExtendingPokemons	81.63	84.91	83.22	37.41	39.6	38.11	40.77	52.49	44.79	61.77	65.07	63.25
15	Baseline	80.66	81.63	81.12	37.16	36.88	36.44	43.4	49.9	45.95	62.04	62.25	62.01
16	Sartipi-Sedighin	80.21	81.07	80.6	29.57	30.59	29.78	34.69	45.4	36.96	57.86	59.07	58.37

Table 13: Detailed results for Dutch track.

A.9 Russian (RU)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	96.37	96.84	96.6	85.89	84.55	85.0	86.89	87.42	87.03	91.93	91.14	91.5
2	USTC-NELSLIP	85.22	83.22	84.2	83.16	81.71	82.23	85.37	86.71	85.91	84.85	82.89	83.82
3	RACAI	82.19	82.07	82.12	66.92	63.51	63.93	76.5	72.78	74.2	75.86	73.83	74.6
4	Sliced	80.65	82.48	81.55	63.54	63.66	62.97	72.14	71.22	71.27	73.59	74.11	73.73
5	YNUNLP	81.41	80.01	80.67	64.38	62.83	62.64	71.95	69.98	70.17	74.09	72.28	72.99
6	MaChAmp	78.65	81.28	79.94	62.64	63.11	62.04	68.82	69.12	68.38	72.0	73.06	72.37
7	brotherhood	80.59	80.92	80.75	57.67	56.22	56.26	64.75	65.26	63.52	71.0	69.84	70.27
8	NetEase.AI	81.07	77.42	79.19	60.42	54.74	56.89	64.69	65.51	63.45	72.61	67.44	69.79
9	EURECOM	80.26	80.11	80.17	54.56	51.04	51.82	63.33	63.7	61.24	69.74	67.19	68.21
10	MarSan	77.99	79.42	78.68	52.33	55.01	53.1	57.5	63.79	58.09	66.83	68.44	67.49
11	L3i	78.64	78.91	78.77	52.76	49.85	50.69	56.76	60.27	57.01	67.89	65.82	66.72
12	CSECU-DSG	75.86	78.26	77.04	44.29	45.97	44.57	53.87	58.07	54.23	62.6	63.9	63.08
13	BaselineExtendingPokemons	71.28	76.78	73.92	41.76	45.55	43.03	44.28	53.55	47.22	58.04	62.4	60.0
14	Baseline	70.58	74.55	72.47	42.57	44.93	43.45	46.36	52.95	48.01	58.42	60.96	59.59
15	AutoNER	62.52	65.16	63.66	37.35	40.19	37.88	51.05	55.08	52.22	51.79	54.33	52.7

Table 14: Detailed results for Russian track.

A.10 Turkish (TR)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DAMO-NLP	96.45	96.42	96.43	86.86	85.39	85.85	89.22	88.4	88.76	89.79	87.82	88.69
2	USTC-NELSLIP	90.32	89.8	90.05	84.22	82.7	83.17	87.64	88.1	87.83	86.62	84.7	85.52
3	SU-NLP	83.53	85.11	84.29	76.0	61.07	67.57	75.13	64.87	68.73	78.86	66.43	72.02
4	RACAI	87.45	88.83	88.13	65.72	64.73	64.04	74.65	73.1	73.59	71.81	70.25	70.42
5	Sliced	86.72	88.51	87.6	62.81	63.47	62.31	70.77	71.38	70.92	69.17	69.22	68.77
6	MaChAmp	84.63	86.93	85.75	61.48	62.81	61.37	65.26	68.63	66.63	67.55	68.3	67.58
7	YNUNLP	86.59	86.99	86.78	61.14	61.04	59.66	73.32	69.45	70.8	68.17	67.05	66.81
8	ML-HUB	84.31	86.92	85.55	61.11	59.47	59.62	56.0	68.23	59.21	66.51	65.98	65.79
9	L3i	85.6	86.99	86.28	57.3	57.15	56.65	63.23	65.47	63.89	64.85	64.26	64.28
10	MarSan	84.73	86.67	85.68	52.39	56.12	53.49	56.27	60.51	57.01	60.22	62.7	61.09
11	brotherhood	85.27	86.86	86.01	50.88	52.78	51.18	57.21	62.29	58.72	59.42	60.68	59.71
12	EURECOM	82.97	85.45	84.18	47.75	50.08	48.41	48.96	55.71	50.97	55.93	57.86	56.57
13	CSECU-DSG	81.86	79.8	80.76	56.5	40.42	46.26	58.08	48.79	52.18	64.57	49.25	55.3
14	Sartipi-Sedighin	79.69	85.74	82.52	42.23	47.34	43.71	45.77	53.34	48.6	50.77	55.81	52.69
15	Baseline	75.87	79.21	77.49	36.2	39.91	37.0	39.19	44.33	40.62	45.31	48.28	46.25
16	BaselineExtendingPokemons	77.39	82.64	79.86	33.63	38.92	35.4	35.41	43.94	38.09	42.74	48.32	44.97

Table 15: Detailed results for Turkish track.

A.11 Chinese (ZH)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	USTC-NELSLIP	92.76	89.42	91.01	77.96	74.73	75.64	85.94	84.84	85.37	83.94	80.07	81.69
2	CASIA	88.52	81.48	84.61	81.04	72.75	75.88	86.56	80.62	83.37	84.77	76.13	79.7
3	OPDAI	85.14	85.44	85.27	75.46	75.81	74.69	80.9	83.93	82.24	80.54	79.47	79.54
4	DAMO-NLP	89.75	87.87	88.77	74.41	73.09	72.24	80.57	80.97	80.1	80.64	77.45	78.06
5	NetEase.AI	89.73	85.48	87.47	75.69	70.89	72.01	77.71	78.92	77.84	81.52	75.63	77.77
6	CMB AI Lab	90.18	86.67	88.37	74.36	64.08	68.54	76.66	75.66	75.53	81.28	72.29	76.36
7	NCUEE-NLP	85.28	81.26	83.1	70.9	68.38	68.5	75.33	76.8	75.27	77.01	72.99	74.18
8	QTrade AI	88.76	85.43	86.98	69.19	66.37	66.3	76.88	77.45	76.88	76.91	72.82	74.0
9	CSECU-DSG	84.85	83.33	84.04	59.16	59.61	58.05	65.15	71.02	66.86	68.55	67.61	67.22
10	Multilinguals	84.48	83.01	83.72	59.48	59.2	57.91	61.93	71.1	64.78	68.39	67.22	66.95
11	L3i	84.5	82.14	83.25	59.63	59.46	58.07	63.71	70.65	65.08	68.62	67.07	66.91
12	Sliced	85.54	84.25	84.86	58.6	56.1	55.02	64.92	67.65	65.09	67.99	65.07	65.21
13	Infrd.ai	82.98	77.89	80.19	59.28	56.65	56.03	62.74	70.05	64.26	67.83	64.16	64.68
14	MaChAmp	83.38	81.82	82.55	57.16	55.83	54.48	61.16	65.41	61.14	66.45	63.88	63.81
15	EURECOM	83.53	81.39	82.25	57.51	54.4	53.25	63.81	68.65	64.31	66.89	63.43	63.4
16	RACAI	86.53	83.53	84.91	58.07	52.36	51.29	65.45	65.12	63.34	68.17	62.05	62.7
17	YNUNLP	83.01	82.63	82.79	53.44	51.46	50.3	64.27	67.0	64.48	63.99	61.41	61.38
18	brotherhood	85.04	81.55	83.11	53.03	49.84	49.69	58.74	62.61	59.25	64.08	60.05	60.86
19	MarSan	81.95	80.59	81.19	48.76	46.19	44.49	57.56	61.96	58.11	60.15	57.1	56.64
20	SPDB Innovation Lab	80.88	79.4	80.06	45.56	46.47	43.97	53.92	59.42	54.94	57.25	57.09	55.74
21	BaselineExtendingPokemons	74.57	78.13	76.23	43.8	44.51	41.39	52.72	56.93	51.41	54.44	55.06	52.8
22	Baseline	73.7	73.18	73.29	43.39	42.43	40.54	45.66	53.57	46.53	53.51	52.32	51.3

Table 16: Detailed results for Chinese track.

A.12 Code-Mixed (MIX)

Rank	Team	LOWNER			ORCAS-NER			MSQ-NER			Average		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	USTC-NELSLIP	95.5	94.99	95.21	89.1	88.64	88.74	93.32	93.39	93.22	93.21	92.61	92.9
2	DAMO-NLP	93.88	93.36	93.57	88.37	87.25	87.68	91.39	90.24	90.56	92.35	91.24	91.79
3	CMB AI Lab	93.68	91.22	92.35	79.89	68.21	73.2	85.02	76.56	80.0	88.4	81.27	84.62
4	QTrade AI	93.21	90.36	91.65	78.06	71.21	73.93	82.86	77.44	79.42	87.12	82.01	84.35
5	SeqL	91.55	90.92	91.16	67.04	65.81	65.89	76.32	75.28	74.93	81.1	79.72	80.29
6	IIE_KDSEC	87.45	88.35	87.8	67.28	67.25	66.88	75.48	76.15	75.29	79.52	79.74	79.59
7	RACAI	89.43	89.6	89.42	65.78	65.48	65.29	74.53	74.51	73.87	79.58	79.23	79.37
8	UM6P-CS	87.7	88.16	87.82	67.08	66.62	66.5	75.06	74.57	74.11	79.38	79.08	79.21
9	EURECOM	86.63	87.25	86.82	63.58	62.71	62.71	73.99	74.11	73.28	77.84	77.37	77.6
10	OPDAI	87.26	85.98	86.45	61.87	61.12	61.0	72.78	71.82	71.49	77.79	77.23	77.46
11	YNUNLP	85.85	87.03	86.33	62.48	62.85	62.12	72.23	72.62	71.77	76.64	76.98	76.78
12	UC3M-PUCPR	87.8	86.74	87.15	60.94	58.86	59.48	71.92	70.74	70.57	77.15	75.64	76.36
13	brotherhood	87.84	87.51	87.57	60.83	60.13	59.87	71.28	70.51	70.0	76.62	75.45	75.91
14	MaChAmp	85.37	86.66	85.82	58.05	60.8	58.66	69.41	69.87	68.55	73.85	75.4	74.52
15	Sliced	86.96	88.05	87.41	54.91	55.8	54.4	68.51	67.59	66.63	72.67	73.23	72.74
16	CMNEROne	83.05	83.24	82.97	51.68	52.34	51.36	63.81	64.02	63.05	70.41	70.62	70.44
17	L3i	73.32	71.49	71.9	56.08	56.22	55.53	66.74	66.4	65.79	68.93	68.73	68.7
18	Cardiff NLP	72.56	73.5	72.66	57.17	58.84	57.45	67.71	69.29	67.96	67.4	69.03	68.07
19	BaselineExtendingPokemons	72.97	72.47	72.31	56.13	55.7	55.45	67.01	66.59	66.01	67.92	68.27	67.99
20	SPDB Innovation Lab	76.07	76.58	76.08	51.73	52.82	51.78	64.59	65.22	64.05	66.96	67.8	67.32
21	MarSan	73.27	70.17	70.76	54.73	56.43	54.75	65.63	66.03	64.59	67.36	67.41	67.03
22	CSECU-DSG	68.11	68.02	67.62	53.46	53.4	52.53	63.54	63.26	62.32	64.23	64.36	64.03
23	Baseline	74.03	74.23	73.72	38.85	37.13	37.15	51.45	50.92	49.86	59.1	57.66	58.14

Table 17: Detailed results for Code-Mixed track.

A.13 Multilingual (MULTI)

Team		LOWNER			ORCAS-NER			MSQ-NER			Avg.		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1. DAMO-NLP	BN	86.4	85.64	86.01	79.18	76.72	77.77	71.85	69.23	70.28	79.14	77.2	78.02
	DE	94.29	94.51	94.4	81.87	82.11	81.62	84.91	86.01	85.31	87.02	87.54	87.11
	EN	96.52	96.66	96.59	82.26	81.45	81.57	82.4	83.87	83.04	87.06	87.33	87.06
	ES	95.63	95.82	95.72	79.61	79.07	78.85	83.16	86.22	84.48	86.13	87.04	86.35
	FA	95.94	96.97	96.45	82.98	82.85	82.69	86.15	86.4	86.11	88.36	88.74	88.42
	HI	84.64	83.8	84.2	82.42	80.96	81.46	87.74	88.59	88.1	84.93	84.45	84.58
	KO	96.59	97.07	96.83	79.29	79.7	79.18	80.88	84.72	82.6	85.59	87.17	86.2
	NL	97.65	97.78	97.72	79.1	79.03	78.56	84.46	84.16	84.27	87.07	86.99	86.85
	RU	96.0	96.6	96.3	82.19	81.52	81.42	86.19	85.92	85.9	88.12	88.01	87.87
	TR	97.66	97.87	97.77	82.66	82.18	82.0	87.69	84.86	86.17	89.34	88.31	88.65
	ZH	88.37	87.9	88.12	67.74	66.83	65.59	78.02	79.74	78.09	78.05	78.16	77.27
	Avg.	93.61	93.69	93.65	79.94	79.31	79.16	83.04	83.61	83.12	85.53	85.54	85.31
2. USTC-NELSLIP	BN	87.48	85.7	86.57	83.99	81.65	82.29	76.6	73.16	74.71	82.69	80.17	81.19
	DE	95.8	95.1	95.45	80.0	79.0	79.25	87.47	85.36	86.33	87.76	86.49	87.01
	EN	93.02	91.63	92.32	77.19	76.11	76.36	80.01	84.03	81.69	83.41	83.92	83.46
	ES	90.05	88.11	89.06	80.65	79.38	79.62	84.84	87.69	86.09	85.18	85.06	84.92
	FA	86.23	84.9	85.54	87.81	86.34	86.9	93.02	91.14	92.04	89.02	87.46	88.16
	HI	86.74	83.83	85.25	84.65	82.8	83.17	91.51	90.64	91.04	87.63	85.76	86.48
	KO	90.62	90.44	90.53	83.34	81.15	81.87	87.36	87.29	87.27	87.11	86.29	86.55
	NL	91.93	90.93	91.43	82.65	81.4	81.81	86.89	86.71	86.58	87.16	86.35	86.61
	RU	84.8	83.42	84.1	82.57	81.15	81.62	86.88	86.67	86.64	84.75	83.75	84.12
	TR	90.32	89.77	90.04	83.52	82.58	82.82	86.43	87.53	86.94	86.76	86.63	86.6
	ZH	92.99	89.98	91.44	77.42	72.83	74.39	85.02	82.84	83.87	85.14	81.88	83.23
	Avg.	90.0	88.53	89.25	82.16	80.4	80.92	86.0	85.73	85.75	86.06	84.89	85.3
3. QTrade AI	BN	85.09	85.1	85.08	70.37	68.98	69.22	65.93	63.96	64.69	73.8	72.68	73.0
	DE	93.08	92.36	92.72	73.04	72.64	72.36	77.95	76.61	76.69	81.36	80.54	80.59
	EN	89.84	89.02	89.42	67.71	67.3	67.14	72.03	72.45	71.82	76.53	76.26	76.13
	ES	87.98	86.53	87.24	72.32	71.78	71.51	77.0	80.07	78.03	79.1	79.46	78.93
	FA	81.96	81.97	81.94	71.21	70.69	70.59	75.71	76.16	75.42	76.29	76.27	75.99
	HI	84.18	83.48	83.81	72.33	70.98	71.3	79.48	81.06	80.13	78.66	78.5	78.41
	KO	87.57	87.49	87.52	68.01	67.34	67.08	74.83	75.64	75.0	76.81	76.82	76.53
	NL	90.99	90.1	90.54	72.66	72.37	72.06	76.97	77.82	77.05	80.21	80.1	79.88
	RU	83.38	82.4	82.89	74.83	72.91	73.03	80.89	76.97	78.38	79.7	77.43	78.1
	TR	88.81	88.64	88.72	74.18	73.22	73.09	79.34	77.84	78.41	80.78	79.9	80.07
	ZH	88.06	85.55	86.74	68.26	66.02	65.88	77.85	77.3	77.34	78.06	76.29	76.65
	Avg.	87.36	86.6	86.97	71.36	70.38	70.3	76.18	75.99	75.72	78.3	77.66	77.66
4. SeqL	BN	86.08	84.98	85.51	67.19	64.55	65.63	60.09	61.73	60.53	71.12	70.42	70.56
	DE	92.87	91.98	92.42	69.63	68.72	68.75	72.43	76.2	73.71	78.31	78.97	78.29
	EN	90.42	88.95	89.67	64.87	63.63	63.9	69.69	72.79	70.85	74.99	75.12	74.81
	ES	88.39	86.19	87.26	69.13	68.01	68.19	72.47	78.55	74.79	76.66	77.58	76.75
	FA	83.53	81.32	82.38	69.12	67.58	68.15	70.29	73.81	70.87	74.31	74.24	73.8
	HI	84.84	82.7	83.73	69.59	67.18	68.01	76.47	80.6	78.26	76.97	76.82	76.67
	KO	87.59	87.27	87.42	65.93	64.75	64.9	70.51	75.32	72.58	74.68	75.78	74.97
	NL	91.14	89.69	90.41	68.73	67.82	67.92	73.22	76.14	74.11	77.7	77.89	77.48
	RU	84.08	81.8	82.91	72.86	69.85	70.68	78.32	76.7	76.91	78.42	76.12	76.83
	TR	89.37	88.01	88.68	70.95	69.69	69.88	73.83	76.25	74.8	78.05	77.98	77.79
	ZH	88.15	85.89	86.95	63.15	59.56	59.44	71.81	72.57	71.1	74.37	72.67	72.5
	Avg.	87.86	86.25	87.03	68.29	66.49	66.86	71.74	74.61	72.59	75.96	75.78	75.5
5. CMB AILab	BN	89.09	81.88	85.32	72.13	52.88	60.4	66.07	55.27	59.97	75.76	63.34	68.56
	DE	94.33	89.61	91.9	73.18	58.05	63.61	81.46	69.55	74.44	82.99	72.4	76.65
	EN	91.86	87.1	89.4	68.42	51.66	57.75	77.06	64.82	70.07	79.12	67.86	72.4
	ES	89.97	84.68	87.21	71.83	56.53	62.11	78.19	70.48	73.89	80.0	70.56	74.4
	FA	85.07	80.51	82.66	72.96	56.35	62.92	77.93	65.99	70.8	78.65	67.62	72.13
	HI	87.89	79.94	83.66	73.64	57.26	63.76	82.92	74.45	78.35	81.48	70.55	75.26
	KO	89.46	84.83	87.06	71.92	55.66	61.62	79.96	67.73	72.91	80.45	69.41	73.86
	NL	92.65	88.07	90.29	72.96	57.99	63.43	80.4	68.32	73.26	82.0	71.46	75.66
	RU	84.39	80.89	82.57	73.96	59.41	64.69	84.19	69.33	75.59	80.85	69.88	74.28
	TR	90.65	87.57	89.07	74.6	58.82	64.69	81.79	69.13	74.55	82.35	71.84	76.11
	ZH	91.05	83.63	87.11	70.03	50.03	57.0	76.11	64.96	69.57	79.06	66.21	71.23

	Avg.	89.67	84.43	86.93	72.33	55.88	62.0	78.73	67.28	72.13	80.25	69.19	73.69
6. UM6P-CS	BN	81.82	81.7	81.75	62.2	59.46	60.15	61.75	58.43	59.8	68.59	66.53	67.24
	DE	90.69	90.55	90.62	63.99	64.33	63.51	73.59	72.19	72.64	76.09	75.69	75.59
	EN	87.81	87.46	87.63	59.2	58.28	58.05	69.68	68.15	68.81	72.23	71.3	71.5
	ES	86.06	85.57	85.81	63.52	63.03	62.49	73.01	73.74	73.13	74.2	74.11	73.81
	FA	80.46	80.72	80.58	63.04	62.53	62.37	69.88	71.14	70.07	71.13	71.46	71.01
	HI	81.21	80.49	80.83	64.59	61.92	62.33	79.6	76.31	77.83	75.13	72.91	73.66
	KO	84.56	86.02	85.28	59.03	58.9	58.33	66.54	70.03	67.97	70.04	71.65	70.53
	NL	89.13	88.99	89.05	63.86	63.77	63.09	72.86	71.28	71.91	75.28	74.68	74.69
	RU	81.19	81.65	81.41	66.65	64.66	64.71	76.58	72.15	73.96	74.81	72.82	73.36
	TR	87.33	87.73	87.53	65.65	65.2	64.58	76.34	74.47	75.28	76.44	75.8	75.8
	ZH	84.55	83.4	83.97	60.37	57.44	56.97	69.06	71.16	69.74	71.32	70.67	70.22
	Avg.	84.98	84.93	84.95	62.92	61.77	61.51	71.72	70.82	71.01	73.21	72.51	72.49
7. RACAI	BN	82.81	82.79	82.79	62.85	60.41	61.11	59.47	60.58	59.86	68.38	67.93	67.92
	DE	91.27	91.07	91.17	63.67	63.11	62.63	70.67	72.1	71.04	75.2	75.43	74.95
	EN	87.83	88.18	88.01	59.72	58.49	58.29	67.49	69.47	68.27	71.68	72.04	71.52
	ES	83.38	83.45	83.4	61.91	61.09	60.66	67.61	73.13	69.95	70.96	72.56	71.34
	FA	79.7	82.46	81.03	63.68	62.93	62.85	68.12	70.83	68.97	70.5	72.07	70.95
	HI	82.33	81.99	82.15	65.02	63.16	63.46	75.29	77.86	76.27	74.21	74.34	73.96
	KO	85.51	87.23	86.34	59.83	59.88	58.95	67.11	70.08	68.36	70.82	72.4	71.22
	NL	89.08	89.5	89.29	63.63	63.22	62.66	69.13	70.72	69.75	73.95	74.48	73.9
	RU	80.36	82.19	81.26	66.4	63.83	63.98	75.57	73.34	74.1	74.11	73.12	73.11
	TR	86.53	88.54	87.52	65.51	64.64	64.21	72.34	73.61	72.8	74.79	75.6	74.84
	ZH	84.72	84.42	84.57	59.53	57.11	56.4	66.33	70.17	67.17	70.19	70.56	69.38
	Avg.	84.87	85.62	85.23	62.89	61.62	61.38	69.01	71.08	69.69	72.25	72.78	72.1
8. Cardiff NLP	BN	82.43	83.67	83.02	60.67	60.32	60.1	59.19	59.87	59.05	67.43	67.95	67.39
	DE	91.24	91.47	91.35	62.09	63.47	62.06	70.23	72.8	71.19	74.52	75.91	74.87
	EN	87.7	88.63	88.16	57.69	58.55	57.49	66.69	68.7	67.5	70.69	71.96	71.05
	ES	85.4	86.71	86.04	61.36	62.47	61.2	68.96	74.82	71.23	71.91	74.67	72.82
	FA	79.43	82.98	81.16	60.08	61.88	60.55	66.17	71.89	68.33	68.56	72.25	70.01
	HI	82.11	82.29	82.18	63.18	63.07	62.6	73.93	78.68	75.94	73.07	74.68	73.57
	KO	84.79	87.72	86.22	55.74	58.53	56.38	66.13	70.77	68.25	68.88	72.34	70.29
	NL	89.03	90.18	89.59	61.04	62.53	61.12	68.23	72.6	70.03	72.77	75.1	73.58
	RU	79.89	82.87	81.35	62.96	62.97	62.03	73.59	72.84	72.74	72.14	72.89	72.04
	TR	86.6	89.3	87.92	62.78	63.69	62.28	72.14	73.33	72.59	73.84	75.44	74.26
	ZH	85.06	85.05	85.05	55.35	55.0	53.08	64.74	69.65	66.51	68.38	69.9	68.21
	Avg.	84.88	86.44	85.64	60.27	61.13	59.9	68.18	71.45	69.4	71.11	73.01	71.64
9. Sliced	BN	82.89	83.11	82.97	59.22	57.37	57.58	56.15	57.69	56.51	66.09	66.06	65.68
	DE	90.83	90.88	90.85	61.76	62.8	61.72	70.31	71.86	70.87	74.3	75.18	74.48
	EN	87.58	87.91	87.74	57.15	57.24	56.36	67.62	68.84	67.97	70.78	71.33	70.69
	ES	85.59	85.81	85.69	60.51	61.49	60.28	69.32	73.36	70.77	71.81	73.55	72.25
	FA	79.75	82.35	81.03	59.67	61.46	60.2	63.45	68.85	65.38	67.62	70.89	68.87
	HI	81.16	81.29	81.21	62.7	61.39	61.23	73.22	76.61	74.29	72.36	73.1	72.24
	KO	84.49	86.53	85.5	55.3	57.96	55.91	64.05	68.25	65.92	67.95	70.91	69.11
	NL	88.7	89.21	88.95	61.48	62.92	61.64	69.04	72.58	70.42	73.07	74.9	73.67
	RU	80.34	82.38	81.34	64.09	64.0	63.41	74.2	73.51	73.43	72.88	73.29	72.73
	TR	86.6	88.45	87.5	62.71	63.5	62.27	70.09	70.85	70.33	73.13	74.27	73.36
	ZH	85.86	84.44	85.1	58.35	55.95	54.95	65.57	68.28	65.85	69.93	69.56	68.63
	Avg.	84.89	85.67	85.26	60.27	60.55	59.6	67.55	70.06	68.34	70.9	72.09	71.06
10. IIE_KDSEC	BN	82.17	81.64	81.86	59.9	56.97	57.87	58.43	57.74	57.53	66.84	65.45	65.75
	DE	90.17	90.43	90.3	62.06	62.48	61.57	69.19	72.44	70.21	73.8	75.12	74.02
	EN	87.3	87.48	87.39	57.62	57.44	56.75	66.92	69.35	67.66	70.61	71.42	70.6
	ES	85.51	85.07	85.28	61.75	62.09	61.25	67.08	73.81	69.43	71.45	73.66	71.98
	FA	79.36	81.72	80.5	60.97	60.08	59.99	63.57	68.11	64.71	67.96	69.97	68.4
	HI	81.06	80.14	80.59	63.67	61.61	62.21	73.87	77.78	75.2	72.86	73.18	72.66
	KO	84.54	86.11	85.3	57.64	58.01	57.34	65.12	70.12	67.09	69.1	71.41	69.91
	NL	88.41	89.01	88.7	61.86	62.45	61.45	68.59	71.4	69.5	72.95	74.28	73.22
	RU	79.81	81.62	80.7	65.4	64.32	64.15	72.79	72.06	71.48	72.67	72.66	72.11
	TR	86.37	88.14	87.24	63.1	62.88	62.25	68.35	70.43	68.9	72.61	73.82	72.8
	ZH	85.15	82.83	83.91	58.91	54.57	54.43	66.85	68.59	66.58	70.31	68.66	68.31
	Avg.	84.53	84.93	84.71	61.17	60.26	59.93	67.34	70.17	68.03	71.01	71.78	70.89

11. B.E.P	BN	82.24	82.26	82.23	60.92	57.93	58.85	59.92	56.44	57.4	67.7	65.54	66.16
	DE	90.68	90.47	90.57	61.17	61.23	60.51	70.51	71.47	70.36	74.12	74.39	73.82
	EN	87.72	87.39	87.55	57.43	55.95	55.79	66.24	66.78	65.91	70.46	70.04	69.75
	ES	86.07	85.34	85.7	61.03	60.65	60.09	68.97	73.12	70.36	72.02	73.04	72.05
	FA	80.28	80.42	80.34	59.67	59.43	58.96	65.17	66.57	65.2	68.37	68.81	68.17
	HI	80.81	80.26	80.5	63.6	60.79	61.29	76.33	75.39	75.51	73.58	72.15	72.43
	KO	84.97	85.84	85.4	57.15	57.65	56.65	65.8	68.49	66.7	69.31	70.66	69.58
	NL	88.97	88.79	88.87	61.39	61.5	60.85	67.05	70.13	68.06	72.47	73.47	72.59
	RU	81.67	80.97	81.32	64.11	63.15	62.83	73.87	72.41	72.28	73.22	72.18	72.14
	TR	86.98	87.56	87.26	62.79	61.74	61.23	72.42	71.26	71.45	74.06	73.52	73.31
	ZH	84.83	83.46	84.12	58.83	53.95	53.3	67.71	67.5	65.38	70.46	68.3	67.6
	Avg.	85.02	84.8	84.9	60.74	59.45	59.12	68.54	69.05	68.06	71.43	71.1	70.69
12. OPDAI	BN	82.79	82.06	82.42	59.68	50.37	54.03	58.27	53.76	55.55	66.92	62.06	64.0
	DE	90.97	90.33	90.65	59.6	56.78	57.75	67.82	68.07	67.45	72.8	71.73	71.95
	EN	87.98	87.95	87.97	51.85	49.03	49.94	63.56	65.24	64.06	67.8	67.4	67.32
	ES	84.12	83.18	83.64	53.54	50.86	51.5	64.2	66.7	64.98	67.29	66.91	66.71
	FA	80.9	81.79	81.33	57.93	52.49	54.77	62.88	63.56	62.62	67.24	65.95	66.24
	HI	81.59	80.68	81.13	61.16	53.24	56.5	72.25	72.98	72.45	71.67	68.97	70.02
	KO	85.44	86.11	85.77	58.73	51.67	54.56	65.73	65.7	65.58	69.96	67.83	68.63
	NL	89.39	89.03	89.21	58.69	56.36	57.07	66.34	68.08	66.93	71.47	71.16	71.07
	RU	81.44	81.61	81.53	63.88	55.46	58.98	70.79	66.93	68.21	72.04	68.0	69.57
	TR	86.58	87.47	87.02	60.69	56.24	57.92	66.58	66.02	66.1	71.28	69.91	70.35
	ZH	86.47	85.8	86.08	68.15	66.7	66.1	82.63	84.11	83.11	79.08	78.87	78.43
	Avg.	85.24	85.09	85.16	59.45	54.47	56.28	67.37	67.38	67.0	70.69	68.98	69.48
13. brotherhood	BN	82.62	82.8	82.7	60.63	57.71	58.47	57.14	57.33	56.89	66.79	65.95	66.02
	DE	90.46	89.99	90.22	59.96	59.81	59.29	68.52	70.3	68.86	72.98	73.37	72.79
	EN	87.53	87.65	87.59	56.32	55.8	55.43	63.92	66.72	64.83	69.26	70.06	69.28
	ES	83.11	82.3	82.7	57.19	56.52	56.11	63.23	69.44	65.8	67.85	69.42	68.2
	FA	80.22	82.03	81.11	57.32	56.51	56.39	61.97	66.71	63.37	66.5	68.42	66.95
	HI	81.19	80.59	80.88	61.79	59.02	59.66	72.26	76.23	73.69	71.74	71.95	71.41
	KO	85.03	86.45	85.74	56.45	56.91	55.95	63.5	68.34	65.33	68.33	70.57	69.01
	NL	88.79	88.94	88.86	59.28	59.62	58.89	64.91	68.78	66.29	70.99	72.44	71.35
	RU	80.39	81.61	80.99	60.8	59.46	59.18	65.77	67.56	66.18	68.99	69.54	68.79
	TR	86.41	88.05	87.22	60.66	59.96	59.59	67.52	70.86	68.66	71.53	72.95	71.82
	ZH	85.77	83.95	84.8	58.11	53.34	53.1	64.98	68.96	66.02	69.62	68.75	67.97
	Avg.	84.68	84.94	84.8	58.96	57.7	57.46	64.88	68.29	65.99	69.51	70.31	69.42
14. MarSan	BN	81.56	81.36	81.42	59.37	56.93	57.76	55.92	56.57	54.9	65.62	64.95	64.69
	DE	90.43	90.55	90.46	62.24	63.31	62.27	69.96	73.92	70.82	74.21	75.93	74.52
	EN	86.88	87.4	87.1	57.16	56.77	56.27	63.0	67.82	64.38	69.02	70.66	69.25
	ES	85.62	85.01	85.27	59.28	60.47	59.45	64.83	73.28	67.66	69.91	72.92	70.79
	FA	78.55	79.51	78.97	55.59	54.98	54.96	58.27	63.37	59.0	64.14	65.95	64.31
	HI	79.63	78.53	79.01	58.95	57.04	57.49	71.59	74.39	71.75	70.06	69.99	69.42
	KO	83.58	84.05	83.78	55.49	55.08	54.6	61.29	65.69	62.36	66.79	68.28	66.91
	NL	88.09	88.54	88.29	60.07	61.31	60.12	67.07	72.15	68.55	71.74	74.0	72.32
	RU	80.59	79.57	80.04	61.8	59.8	60.18	69.08	70.0	67.99	70.49	69.79	69.4
	TR	85.94	86.8	86.34	62.51	62.63	62.0	65.92	69.66	67.05	71.46	73.03	71.8
	ZH	84.14	83.3	83.68	58.8	55.43	55.24	66.48	70.16	67.09	69.81	69.63	68.67
	Avg.	84.09	84.06	84.03	59.21	58.52	58.21	64.86	68.82	65.6	69.39	70.47	69.28
15. Infrd.ai	BN	79.97	80.51	80.18	59.53	55.61	56.49	58.78	55.59	56.88	66.09	63.9	64.52
	DE	88.97	88.84	88.9	60.59	59.98	59.48	67.99	68.85	67.77	72.52	72.56	72.05
	EN	86.12	86.47	86.29	57.76	55.14	55.44	65.26	67.91	66.29	69.72	69.84	69.34
	ES	84.4	84.64	84.51	60.32	59.01	58.59	64.46	69.94	66.42	69.73	71.2	69.84
	FA	78.94	81.21	80.04	58.66	56.92	57.06	63.36	65.37	63.88	66.99	67.83	66.99
	HI	78.69	78.11	78.37	61.63	58.28	59.05	72.43	76.04	73.63	70.92	70.81	70.35
	KO	82.69	84.76	83.69	57.2	56.44	55.88	64.15	67.99	65.73	68.01	69.73	68.43
	NL	87.77	87.82	87.79	61.02	60.09	59.71	67.85	70.08	68.79	72.21	72.66	72.1
	RU	78.57	81.61	80.04	63.47	61.07	61.3	72.86	71.51	71.29	71.63	71.4	70.88
	TR	84.87	87.51	86.16	61.93	60.12	59.97	67.84	68.31	67.87	71.54	71.98	71.33
	ZH	81.81	81.9	81.78	56.59	53.22	52.84	62.12	65.99	62.94	66.84	67.04	65.86
	Avg.	82.98	83.94	83.43	59.88	57.81	57.8	66.1	67.96	66.5	69.65	69.9	69.24

16. HaveNoldea	BN	82.01	80.5	81.17	57.25	51.15	52.6	56.97	52.24	53.64	65.41	61.3	62.47
	DE	91.22	90.04	90.61	63.33	61.86	61.18	75.42	72.88	73.66	76.66	74.93	75.15
	EN	88.34	87.01	87.65	58.52	55.63	55.5	67.91	66.85	67.23	71.59	69.83	70.13
	ES	87.7	84.1	85.82	61.96	59.84	59.27	70.31	70.84	70.51	73.33	71.59	71.87
	FA	79.44	76.16	77.65	53.27	50.53	50.87	60.15	60.02	58.97	64.28	62.24	62.49
	HI	80.84	78.05	79.34	57.6	53.9	54.46	74.25	71.63	72.43	70.89	67.86	68.74
	KO	83.01	81.49	82.16	51.96	47.59	48.03	63.46	63.07	62.57	66.14	64.05	64.26
	NL	89.99	87.36	88.63	63.26	61.13	60.44	72.94	70.77	71.56	75.39	73.09	73.55
	RU	82.66	80.28	81.41	65.65	61.35	61.82	77.69	70.21	73.32	75.33	70.61	72.19
	TR	87.22	86.57	86.87	63.86	61.5	61.22	72.33	68.81	70.2	74.47	72.29	72.76
	ZH	83.1	71.09	76.34	56.85	48.15	49.48	65.75	63.87	63.44	68.57	61.03	63.09
	Avg.	85.05	82.06	83.42	59.41	55.69	55.9	68.83	66.47	67.05	71.1	68.07	68.79
17. EURECOM	BN	79.9	78.85	79.33	54.31	53.41	53.54	54.12	55.24	53.31	62.78	62.5	62.06
	DE	89.29	89.02	89.15	59.53	60.47	59.52	67.38	71.35	68.7	72.06	73.61	72.46
	EN	86.11	86.01	86.06	55.55	55.49	55.0	62.94	66.83	64.13	68.2	69.44	68.4
	ES	85.28	84.32	84.8	59.56	60.08	59.23	65.83	69.89	67.37	70.22	71.43	70.47
	FA	77.82	78.34	78.06	55.33	54.92	54.87	57.17	62.99	58.75	63.44	65.42	63.89
	HI	78.35	76.35	77.29	55.74	54.91	55.04	69.25	72.53	69.84	67.78	67.93	67.39
	KO	82.15	83.8	82.96	53.99	53.35	53.26	58.34	65.45	61.01	64.83	67.53	65.74
	NL	87.59	87.71	87.65	59.47	60.39	59.45	65.7	69.59	67.04	70.92	72.56	71.38
	RU	79.1	79.66	79.37	60.59	59.66	59.48	69.16	69.25	68.08	69.62	69.52	68.97
	TR	84.83	86.71	85.76	60.28	60.43	59.82	65.44	69.03	66.69	70.18	72.05	70.75
	ZH	83.34	82.02	82.66	57.82	54.94	54.48	64.36	68.51	65.12	68.51	68.49	67.42
	Avg.	83.07	82.98	83.01	57.47	57.1	56.7	63.61	67.33	64.55	68.05	69.13	68.08
18. MaChAmp	BN	78.09	79.68	78.87	52.07	51.38	50.89	53.49	53.03	52.46	61.22	61.36	60.74
	DE	89.24	89.56	89.4	62.45	64.2	62.79	70.25	72.33	71.08	73.98	75.36	74.42
	EN	85.93	86.97	86.44	58.08	58.48	57.7	64.65	69.39	66.2	69.56	71.62	70.11
	ES	84.81	85.53	85.15	61.77	63.48	61.94	67.16	72.73	69.23	71.25	73.91	72.1
	FA	75.24	78.49	76.82	49.14	51.84	49.9	49.79	58.31	52.85	58.06	62.88	59.86
	HI	76.69	78.04	77.35	56.29	56.45	55.72	67.68	73.68	70.04	66.89	69.39	67.7
	KO	81.05	83.76	82.38	49.07	51.05	49.34	53.97	62.91	57.36	61.36	65.91	63.03
	NL	87.56	88.34	87.94	61.45	63.0	61.62	68.38	72.58	70.23	72.46	74.64	73.26
	RU	78.31	81.24	79.74	62.04	62.51	61.46	68.92	68.97	68.14	69.75	70.91	69.78
	TR	84.48	87.23	85.81	61.76	63.15	61.68	64.52	67.95	65.9	70.25	72.78	71.13
	ZH	79.96	80.37	80.13	49.66	48.44	46.96	58.89	63.62	59.88	62.83	64.15	62.32
	Avg.	81.94	83.56	82.73	56.71	57.63	56.36	62.52	66.86	63.94	67.06	69.36	67.68
19. YNUNLP	BN	79.19	77.07	78.05	57.95	51.59	53.7	55.19	55.49	53.56	64.11	61.38	61.77
	DE	87.61	86.41	86.99	59.12	58.75	57.98	69.29	69.04	67.19	72.01	71.4	70.72
	EN	85.67	84.19	84.9	54.77	52.48	52.38	64.73	63.75	63.04	68.39	66.81	66.77
	ES	85.22	81.31	83.2	57.7	56.3	55.91	66.15	68.41	66.0	69.69	68.67	68.37
	FA	79.42	75.53	77.39	56.48	54.03	54.38	63.67	64.01	62.28	66.52	64.52	64.68
	HI	77.69	74.36	75.93	59.47	56.21	57.07	71.12	73.9	71.03	69.43	68.16	68.01
	KO	81.56	80.87	81.18	56.52	54.11	54.28	59.42	65.49	60.53	65.83	66.82	65.33
	NL	87.43	84.81	86.08	58.73	57.78	57.11	67.24	67.15	66.17	71.14	69.91	69.78
	RU	81.74	75.93	78.72	62.54	58.78	59.31	70.27	68.38	67.57	71.52	67.69	68.53
	TR	84.4	83.52	83.95	60.08	57.53	57.37	66.84	67.32	65.46	70.44	69.46	68.93
	ZH	80.19	77.99	79.0	51.79	47.17	47.22	61.37	63.07	61.05	64.45	62.75	62.42
	Avg.	82.74	80.18	81.4	57.74	54.98	55.16	65.03	66.0	63.99	68.5	67.05	66.85
20. DS4DH	BN	79.42	75.85	77.57	53.09	45.66	47.87	54.34	48.87	50.71	62.29	56.79	58.71
	DE	88.21	85.58	86.87	55.3	52.95	52.98	66.38	63.79	64.55	69.96	67.44	68.13
	EN	86.21	82.8	84.44	52.18	48.88	49.18	61.27	57.99	59.06	66.55	63.22	64.23
	ES	85.31	80.72	82.92	55.12	51.71	52.0	63.9	62.0	62.66	68.11	64.81	65.86
	FA	81.89	74.94	78.22	56.3	51.66	53.26	63.25	61.66	61.71	67.15	62.75	64.4
	HI	79.81	74.44	76.95	57.56	51.0	53.0	71.66	69.79	70.06	69.68	65.08	66.67
	KO	84.45	81.23	82.8	52.55	49.93	49.99	62.8	60.73	61.42	66.6	63.96	64.74
	NL	88.16	84.7	86.38	55.24	52.71	52.7	65.16	62.9	63.74	69.52	66.77	67.61
	RU	82.52	75.96	79.09	59.17	54.36	55.35	68.95	62.22	64.87	70.22	64.18	66.44
	TR	86.83	83.17	84.96	56.56	52.84	53.15	67.34	62.2	64.38	70.25	66.07	67.5
	ZH	83.24	78.17	80.52	53.77	47.15	48.15	61.73	61.21	60.78	66.25	62.18	63.15
	Avg.	84.19	79.78	81.88	55.17	50.8	51.6	64.25	61.21	62.18	67.87	63.93	65.22

21. UPB	BN	79.84	80.64	80.2	52.51	51.07	50.88	52.09	50.36	49.95	61.48	60.69	60.34
	DE	87.67	88.32	87.99	52.24	53.77	52.18	60.43	65.32	61.79	66.78	69.14	67.32
	EN	85.14	86.16	85.64	50.6	51.15	50.14	57.73	62.56	59.02	64.49	66.62	64.93
	ES	83.81	84.02	83.91	52.05	52.73	51.33	57.96	64.4	60.21	64.61	67.05	65.15
	FA	77.06	80.14	78.57	53.26	53.7	52.46	57.45	62.04	58.86	62.59	65.29	63.3
	HI	77.53	77.68	77.58	55.0	54.26	53.98	65.64	69.92	66.47	66.06	67.29	66.01
	KO	81.16	84.35	82.71	50.74	51.9	50.41	56.73	62.38	58.56	62.88	66.21	63.89
	NL	86.73	87.61	87.14	53.31	53.99	52.73	59.68	64.38	61.06	66.57	68.66	66.98
	RU	77.93	80.93	79.39	57.27	57.01	56.06	63.2	65.12	62.73	66.13	67.69	66.06
	TR	83.19	85.37	84.26	53.39	52.94	52.04	58.02	59.78	57.85	64.87	66.03	64.72
	ZH	82.89	82.21	82.49	52.98	50.47	48.96	58.15	63.05	58.42	64.67	65.24	63.29
Avg.		82.09	83.4	82.72	53.03	53.0	51.92	58.83	62.66	59.54	64.65	66.36	64.73
22. CSECU-DSG	BN	77.5	76.82	77.12	51.42	46.24	47.93	52.23	52.55	51.63	60.38	58.54	58.89
	DE	87.12	86.34	86.72	52.5	53.92	52.52	62.89	65.57	63.82	67.5	68.61	67.69
	EN	84.06	83.34	83.69	49.84	49.8	48.91	58.88	60.99	59.47	64.26	64.71	64.02
	ES	82.93	81.62	82.26	51.62	52.18	51.02	60.08	64.29	61.72	64.88	66.03	65.0
	FA	77.28	77.87	77.55	52.12	51.76	51.45	58.85	63.0	59.66	62.75	64.21	62.88
	HI	76.99	75.54	76.23	56.3	54.03	54.51	67.76	70.11	67.08	67.01	66.56	65.94
	KO	80.51	82.81	81.62	49.44	50.17	49.04	58.58	61.56	59.6	62.85	64.85	63.42
	NL	86.53	85.57	86.05	53.53	54.14	53.05	62.14	65.41	63.34	67.4	68.37	67.48
	RU	77.98	78.23	78.09	54.22	53.26	52.63	66.57	63.92	64.18	66.26	65.14	64.97
	TR	83.99	84.09	84.03	54.47	54.08	53.26	61.25	63.19	61.46	66.57	67.12	66.25
	ZH	81.91	80.29	81.03	50.09	48.27	47.08	56.59	60.81	57.6	62.86	63.12	61.9
Avg.		81.53	81.14	81.31	52.32	51.62	51.04	60.53	62.85	60.87	64.79	65.21	64.4
23. NSU-AI	BN	77.78	77.27	77.49	52.46	49.81	50.37	49.13	48.34	48.07	59.79	58.47	58.64
	DE	87.9	87.79	87.84	53.35	53.64	52.74	58.78	63.51	60.52	66.68	68.31	67.03
	EN	84.65	85.02	84.83	50.6	50.22	49.68	54.44	59.15	56.25	63.23	64.8	63.59
	ES	83.59	82.96	83.27	52.78	52.49	51.58	55.14	61.31	57.59	63.84	65.59	64.15
	FA	75.33	77.73	76.47	50.9	51.12	50.45	54.1	60.42	56.47	60.11	63.09	61.13
	HI	77.1	76.02	76.53	55.3	52.65	53.21	65.03	67.59	65.28	65.81	65.42	65.01
	KO	81.62	83.14	82.35	53.53	52.63	52.3	57.73	61.64	59.15	64.29	65.8	64.6
	NL	87.03	86.79	86.91	54.37	54.41	53.51	59.23	64.16	61.11	66.88	68.45	67.17
	RU	78.68	79.38	79.01	58.83	56.94	57.01	64.38	66.72	64.85	67.29	67.68	66.96
	TR	83.23	84.38	83.79	54.85	53.59	53.2	57.77	59.41	58.08	65.28	65.79	65.02
	ZH	81.64	80.89	81.19	52.25	49.82	48.45	60.37	62.88	59.94	64.76	64.53	63.19
Avg.		81.69	81.94	81.79	53.57	52.48	52.05	57.83	61.38	58.85	64.36	65.27	64.23
24. SPDB Innovation Lab	BN	75.59	75.54	75.54	45.87	41.92	42.27	53.4	47.64	49.55	58.29	55.03	55.79
	DE	85.92	84.87	85.38	50.94	51.48	50.04	65.47	64.83	64.63	67.45	67.06	66.68
	EN	83.15	82.06	82.59	47.89	46.77	46.15	59.29	58.03	58.41	63.44	62.29	62.38
	ES	82.49	80.59	81.5	49.9	49.15	48.04	61.47	61.4	61.13	64.62	63.71	63.56
	FA	76.55	76.23	76.35	50.3	48.04	48.11	58.45	57.38	56.84	61.76	60.55	60.43
	HI	75.35	73.81	74.51	55.39	51.67	52.38	72.27	68.99	70.21	67.67	64.83	65.7
	KO	79.77	80.56	80.14	47.68	47.32	46.18	58.55	59.16	58.25	62.0	62.34	61.52
	NL	85.65	84.98	85.31	51.67	52.05	50.66	66.56	63.74	64.68	67.96	66.92	66.88
	RU	78.46	76.82	77.63	56.74	54.1	53.82	68.18	60.85	63.58	67.8	63.92	65.01
	TR	83.54	83.65	83.58	53.9	53.08	52.11	64.62	61.86	62.75	67.35	66.2	66.15
	ZH	81.39	78.83	80.02	51.15	46.6	46.01	60.06	59.38	57.95	64.2	61.61	61.33
Avg.		80.71	79.81	80.23	51.04	49.29	48.71	62.57	60.3	60.73	64.78	63.13	63.22
25. L3i	BN	73.57	72.55	72.84	47.0	42.43	43.79	38.16	39.75	38.04	52.91	51.58	51.56
	DE	89.55	88.49	89.01	57.48	56.04	55.68	62.63	65.6	63.51	69.89	70.04	69.4
	EN	85.81	85.55	85.67	53.37	50.62	50.86	54.89	55.68	54.55	64.69	63.95	63.69
	ES	84.27	83.74	83.98	55.13	53.47	53.37	57.22	62.82	59.23	65.54	66.68	65.53
	FA	75.71	75.4	75.51	47.17	43.21	44.43	48.19	53.31	49.09	57.02	57.31	56.34
	HI	70.72	68.68	69.54	46.78	42.45	43.33	49.58	56.2	50.6	55.69	55.78	54.49
	KO	80.02	78.75	79.37	45.4	41.19	42.21	47.43	50.83	47.26	57.62	56.92	56.28
	NL	87.12	87.18	87.13	57.05	55.68	55.37	60.83	64.2	61.96	68.33	69.02	68.15
	RU	77.55	77.69	77.59	54.56	50.71	51.51	58.71	60.39	57.88	63.6	62.93	62.33
	TR	83.21	85.44	84.3	54.03	50.87	51.34	53.43	56.44	53.88	63.56	64.25	63.18
	ZH	82.39	80.96	81.59	54.62	50.17	49.54	57.08	61.1	56.76	64.7	64.08	62.63
Avg.		80.9	80.4	80.59	52.05	48.8	49.22	53.47	56.94	53.89	62.14	62.05	61.23

26. Baseline	BN	73.32	74.14	73.54	39.28	36.45	37.49	39.87	41.36	40.03	50.83	50.65	50.35
	DE	83.69	82.35	82.93	41.91	40.47	40.63	49.96	53.5	50.94	58.52	58.77	58.17
	EN	79.78	78.97	79.29	41.19	38.93	39.36	44.36	49.88	45.8	55.11	55.93	54.82
	ES	67.52	64.25	65.53	36.1	33.13	33.33	36.75	39.54	37.08	46.79	45.64	45.32
	FA	47.01	45.37	45.46	32.2	26.58	28.23	34.42	35.25	34.46	37.88	35.74	36.05
	HI	72.06	71.1	71.38	44.95	41.59	42.63	52.53	55.79	53.4	56.51	56.16	55.8
	KO	42.84	37.41	39.8	39.07	34.38	35.76	35.81	41.62	36.13	39.24	37.8	37.23
	NL	68.94	65.02	66.73	40.6	38.48	38.59	45.62	49.21	46.64	51.72	50.9	50.65
	RU	51.22	53.0	51.17	35.76	30.7	32.65	32.34	34.37	32.27	39.77	39.36	38.7
	TR	57.51	53.81	55.17	36.67	33.55	34.09	37.01	39.82	37.62	43.73	42.39	42.29
	ZH	77.29	77.5	77.36	42.83	42.68	41.32	46.75	54.27	49.89	55.62	58.15	56.19
	Avg.	65.56	63.9	64.4	39.14	36.09	36.73	41.4	44.96	42.21	48.7	48.32	47.78

Table 18: Detailed results for the Multi-lingual track. Full form of B.E.P. is BaselineExtendingPokemons.