Natural Language Processing(CS60075)

MULTILINGUAL COMPLEX NAMED ENTITY RECOGNITION (MULTICONER)

FINAL REPORT

Prepared by:
Dnyandeep Chavan (20EX20013)
Nilesh Kumar(20CH30017)
Kinjal Sensharma (20CH3FP31)

Abstract

The work we have done, concentrated on ways to recognize complex named items (such as media titles, products, and groups) in three languages in both monolingual and multilingual contexts. One track concentrated on multilingual models applicable to all languages, the last track featured code-mixed texts within any of these languages, and one track was for developing monolingual NER models for specific languages.

The MULTICONER dataset, which contains 2.3 million occurrences in the following languages English, Hindi, and Bangla, was used for this assignment. The best performance was demonstrated by the results employing techniques that combined external knowledge into transformer models.

The Creative Work and Group Entity classes, which are still difficult even with outside expertise, saw the most gains.

Introduction

In open-domain and practical situations, processing complicated and ambiguous Named Entities (NEs) is a difficult NLP issue that has not gotten enough attention from the research community.

Complex NEs are not simple nouns and are more difficult to recognise, such as the names of creative works (movie, book, music, and software titles). They don't resemble conventional NEs and can take the shape of any linguistic element, such as an imperative phrase ("Dial M for Murder") (Person names, locations, organizations).

It is difficult to identify them based on their context due to this ambiguity. Such titles may also be semantically ambiguous, as in the case of "On the Beach," which may refer to both a movie and a preposition. 2 Finally, because these entities typically expand more quickly than conventional categories, emerging entities present yet another difficulty.

Transformers is one example of a neural network that has achieved great scores on benchmark datasets like CoNLL03/OntoNotes. Although these models perform significantly worse on complex/unseen entities. pointed out that these scores are driven by the use of well-formed news text, the presence of "easy" entities (such as person names), and memorization due to entity overlap between train and test sets. Researchers who employ NER for downstream tasks have also observed that NER systems' failure to recognise complex items accounts for a sizable amount of their errors.

Problem description

There are three monolingual and one multilingual track. The dataset comprised of English, Hindi and Bengali languages in monolingual format. For the multilingual part, we compiled English, Hindi and Bengali datasets together. The task was to prepare a model which did the task of named entity recognition on these datasets. We had six tags namely

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

We evaluated and ranked systems using the macro-averaged F1 scores. We also present precision, recall, and domain-specific performance.

800 development instances and 15,300 training instances were used.
The vast bulk of the cases in the training splits are from the Wikipedia domain (i.e. LOWNER), whereas just 100 instances—50 each from the domains of Web Questions (i.e. MSQ-NER) and Web Query (i.e. ORCAS-NER)—represent domain-adaptation data.

On the other side, the test splits are substantially bigger.

This is done primarily for two reasons:

- To evaluate the generalizability of NER models on
- unknown and complicated thing
- To evaluate the effectiveness of NER models
- for cross-domain adaptation

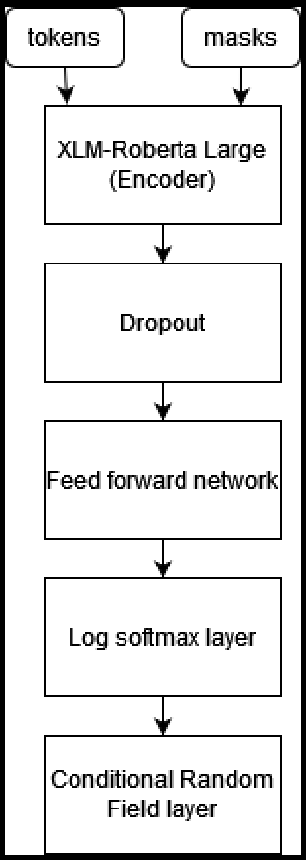
Methodology

Baseline System

Using the multilingual Transformer model XLM-RoBERTa (XLM-R), we train and test a baseline NER system.
Each token's representation is computed using the XLM-R model and utilised in conjunction with a CRF classification layer to forecast the token's tag.

For application scenarios that involve many languages, like ours, the XLM-R baseline is ideally suited.
It offers a strong foundation upon which the participants can build and supports up to 100 different languages.
2 epochs and a learning rate of 0.001 were used to train the baseline model.
The participants were given access to the baseline system's code and scripts so they could use its features and expand it further using their ideas.

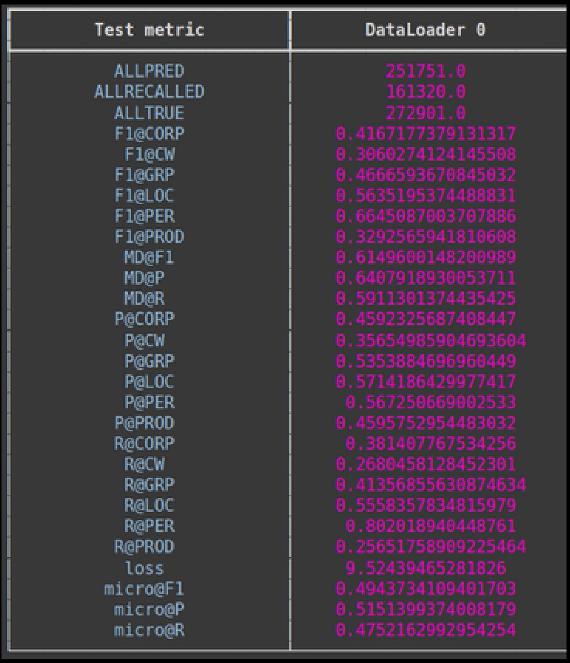
Model Architecture



Results and analysis

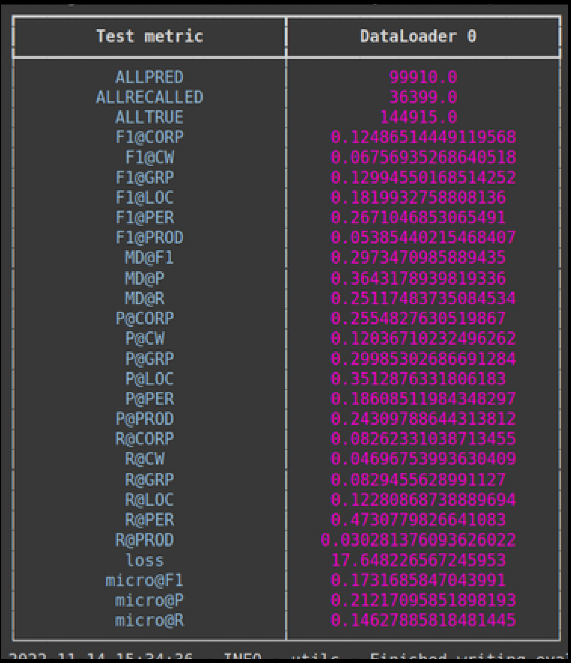
The performance of the model on test data of individual languages is as follows:

English Dataset



Test metric	DataLoader 0
ALLPRED	251751.0
ALLRECALLED	161320.0
ALLTRUE	272901.0
F1@CORP	0.4167177379131317
F1@CW	0.3060274124145508
F1@GRP	0.4666593670845032
F1@LOC	0.5635195374488831
F1@PER	0.6645087003707886
F1@PROD	0.3292565941810608
MD@F1	0.6149600148200989
MD@P	0.6407918930053711
MD@R	0.5911301374435425
P@CORP	0.4592325687408447
P@CW	0.35654985904693604
P@GRP	0.5353884696960449
P@LOC	0.5714186429977417
P@PER	0.567250669002533
P@PROD	0.4595752954483032
R@CORP	0.381407767534256
R@CW	0.2680458128452301
R@GRP	0.41356855630874634
R@LOC	0.5558357834815979
R@PER	0.802018940448761
R@PROD	0.25651758909225464
loss	9.52439465281826
micro@F1	0.4943734109401703
micro@P	0.5151399374008179
micro@R	0.4752162992954254

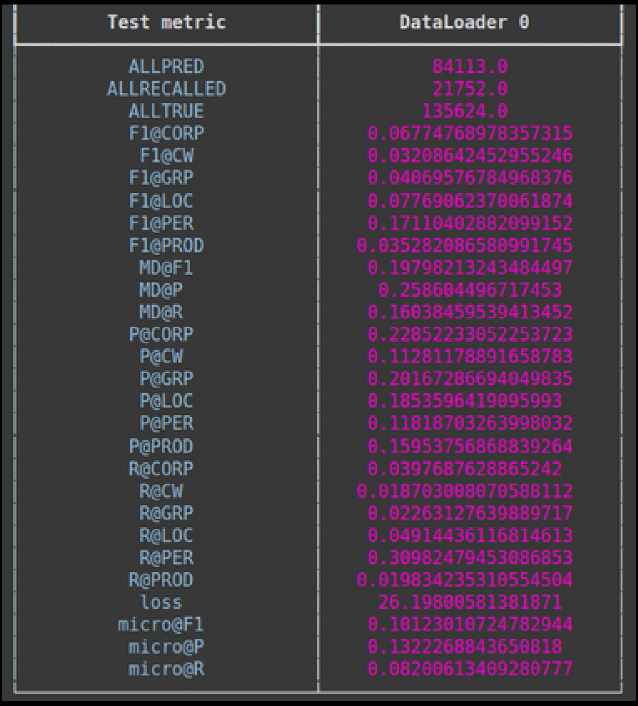
Hindi Dataset



Test metric	DataLoader 0
ALLPRED	99910.0
ALLRECALLED	36399.0
ALLTRUE	144915.0
F1@CORP	0.12486514449119568
F1@CW	0.06756935268640518
F1@GRP	0.12994550168514252
F1@LOC	0.1819932758808136
F1@PER	0.2671046853065491
F1@PROD	0.05385440215468407
MD@F1	0.2973470985889435
MD@P	0.3643178939819336
MD@R	0.25117483735084534
P@CORP	0.2554827630519867
P@CW	0.12036710232496262
P@GRP	0.29985302686691284
P@LOC	0.3512876331806183
P@PER	0.18608511984348297
P@PROD	0.24309788644313812
R@CORP	0.08262331038713455
R@CW	0.046967539936304O9
R@GRP	0.0829455628991127
R@LOC	0.12280868738889694
R@PER	0.4730779826641083
R@PROD	0.030281376093626022
loss	17.648226567245953
micro@F1	0.1731685847043991
micro@P	0.21217095851898193
micro@R	0.14627885818481445

```
2023 11 14 15:34:36   INFO   utils   Finished writing eva
```

Bengali Dataset



Test metric	DataLoader 0
ALLPRED	84113.0
ALLRECALLED	21752.0
ALLTRUE	135624.0
F1@CORP	0.06774768978357315
F1@CW	0.03208642452955246
F1@GRP	0.04069576784968376
F1@LOC	0.07769062370061874
F1@PER	0.17110402882099152
F1@PROD	0.035282086580991745
MD@F1	0.19798213243484497
MD@P	0.258604496717453
MD@R	0.16038459539413452
P@CORP	0.22852233052253723
P@CW	0.11281178891658783
P@GRP	0.20167286694049835
P@LOC	0.1853596419095993
P@PER	0.11818703263998032
P@PROD	0.15953756868839264
R@CORP	0.0397687628865242
R@CW	0.018703008070588112
R@GRP	0.02263127639889717
R@LOC	0.04914436116814613
R@PER	0.30982479453086853
R@PROD	0.019834235310554504
loss	26.19800581381871
micro@F1	0.10123010724782944
micro@P	0.1322268843650818
micro@R	0.08200613409280777

Conclusion

We ran the model in English, Hindi, and Bangla language. It was found that the English dataset performed better than Hindi and Bangla. The model performed well on the English dataset and could recognize the Named Entities with good accuracy. As the other so-called foreign languages, which are said to have particular grammar rules and structural differences that apply solely to them, are not comparable to English or other languages that are similar to it. English has become the standard language for technology due to the consistent support it has received from the normative community. A model is not entirely language-independent just because it lacks explicitly encoded information, according to enough prior research.

Despite the many advantages of working on other languages, most natural language processing (NLP) research focuses on creating techniques that are effective for English. These advantages include a significant social impact, modelling a variety of linguistic variables, avoiding overfitting, and creating engaging machine learning tasks (ML).

In many ways, English and the small group of other high-resource languages are not typical of the rest of the world's tongues. Many resource-rich languages are Indo-European in origin, are predominately spoken in the West, and have poor morphology, meaning that information is primarily expressed syntactically, for example, by using multiple distinct words rather than variation at the word level.

References

1. 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.
2. Gustavo Aguilar, Suraj Maharjan, Adrian Pastor LópezMonroy, 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.
3. 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
4. Sandeep Ashwini and Jinho D. Choi. 2014. Targetable named entity recognition in social media. CoRR, abs/1408.0782.
5. 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