Performing and Analyzing Named Entity Recognition on Foreign English Contexts

Stanford CS224N Custom Project

Alexander Shan
Department of Computer Science
Stanford University
azshan@stanford.edu

Abstract

The majority of popular English named entity recognition (NER) datasets used to train models source Western contexts of English, despite the existence of many foreign, non-Western contexts of English. In this project, I analyze NER model performance in low-resource, foreign contexts of English. I build a bidirectional LSTM RNN, train it on a popular Western-English dataset and a foreign English context dataset, and compare my results with a state-of-the-art GloVe-based model. Ultimately, I aim to quantify and explain model errors observed when training on a Western-context dataset and testing on foreign contexts (and vice versa). My model performed similarly to the state-of-the-art, indicating that the differences in model architecture did not affect performance in either context. Furthermore, I find that even with RoBERTa pretraining, both models exhibit significant performance losses (over a 10 percent drop in F1) that are salient beyond annotation differences between the two datasets. Additionally, I examine model error, finding that the models were most prone to error when sentence context is ambiguous, especially if the named entities were missing from the pretrain word vectors and training data. I also observe that merging the two models produces significant gains in foreign contexts of English while retaining strong performance in Western contexts.

1 Key Information to include

- Mentor: Rishi Desai
- External Collaborators (if you have any): None
- Sharing project: Extending upon work with the foreign NER dataset from CS195, where I built the dataset

2 Introduction

Named Entity Recognition (NER) is a popular NLP task that has been extensively researched, including NER for low-resource languages (Tsygankova et al., 2021). However, the majority of NER studies use corpora composed of abundant, Western contexts of English. Foreign contexts of English are interesting because they are rich with named entities that do not appear in Western texts, such as cultural names (e.g. "Precious" in South African English) and unique organizations (e.g. the Japanese Diet, analogous to the U.S. Congress). To form a holistic view of the English language, these foreign contexts should also be well-understood. Foreign-context NER can be difficult for models trained on Western text, as words like "Precious" and "Diet" very rarely, if ever, appear in training as a named entity. There have been a few studies on NER within foreign, low-resource contexts of English (Louis et al., 2006), but they use outdated Bayesian-network model approaches and a limited corpus of foreign English, typically only covering one region. However, they demonstrated the limitations of Western-context-trained models tested in a foreign context. This motivates exploring how well modern neural models perform in various foreign contexts of English. To do so, I build a bidirectional LSTM RNN and use StanfordNLP's Stanza GloVe-based NER model to perform NER on two datasets; one composed of Western newswire and one composed of non-Western newswire from around the world. I train both models on these datasets separately and test them on the opposite dataset from which they were trained to observe how well they perform in different English contexts, respectively. Additionally, I include RoBERTa pretraining to see to what degree large models retain the same performance losses seen in the existing literature. Overall, I find that both models experience significant performance losses when trained on one dataset and tested on the other. Moreover, both models experience the greatest error when classifying named entities where textual context is ambiguous, such as when a token could be contextually valid as a location and organization if the word itself is unknown. This effect usually occurs when tokens are not seen in the pretrain word vectors or training. Ultimately, I combine the foreign and Western context models, observing strong performance on both datasets, suggesting that existing models can benefit from training on foreign contexts without compromising their performance on Western texts.

3 Related Work

NER has been explored in non-English languages with great success, with one example being StanfordNLP's open-source Stanza Chinese NER model (Qi et al., 2020; Liu et al., 2022). There has also been research on low-resource foreign languages with notable success using neural based approaches (Cotterell and Duh, 2017). However, low-resource contexts of English have not been recently examined. Louis et al. (2006) showed that Western-English trained models performed poorly in South African contexts of English due to the presence of unknown words such as "Xabanisa", or alternative uses of words like "Peace", a common South African name. However, their model was a simple Bayesian network, so their results are not indicative of how modern neural models might perform, especially with pretraining. Ghaddar et. al (2021) revealed that name-regularity bias still exists in neural models, even when contextual clues are present (Ghaddar et al., 2021). For example, in the sentence "Obama is located in southwestern Fukui Prefecture.", state-of-the-art NER models labeled "Obama" as a person rather than a location. Ghaddar et. al did not examine foreign contexts of English, so while their work could suggest the same phenomenon may occur in foreign contexts, this research remains nontrivial.

4 Approach

4.1: Baselines

Stanza, StanfordNLP's open-source python NLP library, has a BiLSTM NER model in its pipeline which I use as a baseline for the comparison of two datasets. The CoNLL03 dataset (Western English Newswire) has been extensively used, with a state-of-the-art BERT-based model achieving an F1 score of 91.3 on its test set (Lim, 2023). The foreign English dataset is novel and has not been scored on before, so I use Stanza's GloVe-based model with RoBERTa pretraining as a benchmark for my experiments.

4.2: Model Construction

I built a bidirectional LSTM model to observe the differences in performance between English contexts. To build my model, I used Stanza's trainer class script, document and data-loading objects, support for pre-trained vectors, and utility functions (Qi et al., 2020). I also followed Stanza's use of the conditional random field negative log-likelihood function for loss during training (Nowozin, 2011). I wrote my BiLSTM model layer structures, scripts to train, test, score, and analyze the model, and edited some Stanza scripts to accommodate my model's structure when scoring and training.

4.2.1: LSTM Architecture

My RNN architecture consists of an encoder with word-level and character-level representation, and a decoder that uses conditional random fields (CRF) decoding to classify NER tags. This approach is well-documented and is known to perform competitively with other techniques (Qi et al., 2020; Cotterell and Duh, 2017). On the encoder, I use two LSTM layers whose output is then fed to a linear layer. In the decoder, I decode the sequence of labels for the tokens such that the conditional likelihood of the labels is maximized (similar to a Hidden Markov Model). To do so, I use the Viterbi algorithm, which decodes the probability of the most likely sequence of states (named entity labels) for a sequence (text) (Gormley, 2018). I use dropout and optimize for CRF loss during training.

4.3: Preparation for analysis
To preprocess my data, I converted the annotations into BIOES format; BIOES marks tokens with their class and position in a named entity span, which can be one of beginning, intermediate, ending, or singular. To maintain consistency between the two datasets, I collapsed the annotations for the foreign English data with 10 classes (none, date, person names, location, facility, organization, miscellaneous, NORP, currency, and product) location into 5 classes (none, person name, location, organization, and miscellaneous). The definitions for each class are included in Appendix A; facility was collapsed into location, and the last 3 were collapsed into miscellaneous). During analysis, the "None" tag is also interchangeably used with "O" to denote a token that is not a named entity. Additionally, I removed the miscellaneous tag from dates in the foreign context dataset to preserve consistency with CoNLL03, making the new tag "none". Finally, I split the data into training, dev, and test sets (75/15/10 split). When scoring the model, I categorized the erroneous behavior into patterns (e.g. true tag "Location", predicted tag "Organization") and examined the particular sentences from the most error-prone patterns to understand model behavior.

5 Experiments

5.1 Data

The first dataset I use is CoNLL03, a popular English NER dataset composed of 1300+ Western-sourced news articles (Sang). I created the second dataset with John Bauer in the Stanford NLP Group, sourcing 1100 news articles from 48 countries, notably excluding Western media to compose the dataset of only foreign English contexts (Shan and Bauer, 2023). The corpus contains many named entities that do not appear in CoNLL03 or GloVe vectors, making it useful for exploring model errors in unfamiliar language contexts. To better understand the difference in language contexts, shown below are two sentences, one from each dataset:

CoNLL03: *Former international goalkeeper Dominique Baratelli is to coach struggling French first division side Nice, the club said on Friday.*

Foreign: *Puerto Iguazú handles twice as many travelers as Ezeiza, Moreno explained Thursday.*

The second sentence may be more difficult to process for Western English speakers, as "Ezeiza" and "Moreno" could be confused for organization, location, or person tags without knowledge of their meanings.

5.2 Evaluation method

I measure the precision, recall, and F1 scores of the models. Precision is a measure of how correct the model's guesses are; recall is a measure of how often the model recognizes each correct tag; F1 is a harmonic mean of the two. My analysis involves explaining model behavior, so I constructed a confusion matrix to hand-analyze and recognize the patterns of errors in foreign contexts.

5.3 Experimental details

I trained and tested Stanza's GloVe-based model with RoBERTa pretraining on each dataset. I trained using stochastic gradient descent (maximum gradient descent steps = 200,000) and then scored the model against the dev and test sets of CoNLL03 and the foreign newswire dataset. To train Stanza's model, I used the following hyperparameters: learning rate=0.1, hidden layer size=256, word and character embeddings size=100, 1 LSTM layer, max gradient norm=5 (for gradient clipping), batch size=32, and dropout p=0.5. Training took around 3 hours for each dataset, with an exception for the combined model, which had to train on a merged dataset (5 hours).

When training my own model, I used almost all of the same hyperparameters as listed above, including using GloVe embeddings for pretraining, along with RoBERTa. The exceptions were that I used two LSTM layers and set my dropout rate to 0.3 since this hyperparameter configuration yielded the highest F1 on the dev sets. Ultimately, the two models' performance was roughly equivalent (see section 5.4). Training took around 5 hours for each dataset, with an exception for the combined model which took 9 hours, likely due to training with a smaller compute cluster.

5.4 Results

Table 1: F1 scores for Stanza GloVe-based model with RoBERTa pretrain

Train Dataset/Eval Set	Foreign Dev	Foreign Test	CoNLL Dev	CoNLL Test
Foreign Eng.	**89**	**90.3**	79.92	74.18
CoNLL03	74.92	77.95	94.71	91.92
Combined	87.37	89.59	**95.74**	**92.11**

Table 2: F1 scores for my model with GloVe, RoBERTa pretrain

Train Set/Eval Set	Foreign Dev	Foreign Test	CoNLL Dev	CoNLL Test
Foreign Eng.	**87.75**	**89.52**	79.61	74.4
CoNLL03	78.13	80.79	**95.48**	91.86
Combined	87.07	89.47	95.15	**92.02**

The results for the Stanza model on CoNLL03 are consistent with the baseline model from section 4.1 Lim (2023). My model and Stanza have similar performance, which is unsurprising because both models have a BiLSTM architecture, optimize the same loss function, and have similar hyperparameters. Both models experience an expected, large drop in performance when trained on the CoNLL03 dataset and tested on the foreign English dataset and vice-versa when training on the foreign English dataset. Even with RoBERTa and GloVe, it appears that the models have difficulty between the datasets. The degree to which the models experience performance losses is surprising since I assumed including pretraining would help the models understand words that they would not see during training. However, considering the existing literature has proven that RoBERTa struggles can struggle with unseen contexts, exhibiting name regularity bias (section 3), so these results are still plausible.

The foreign-trained model has a relatively low F1 on its own dev and test set compared to the CoNLL03-trained model, an unexpected result. The combined model, which merges the foreign English and CoNLL03 models, shows strong performance on all of the evaluation datasets. This is not a very surprising result considering that the combined model had the opportunity to learn the distinct contexts of English and their nuanced differences during finetuning.

Shown below are heat-mapped confusion matrices for the different model performances that are most relevant to subsequent analysis. Red indicates high confusion while green indicates little-to-no confusion, with darker shading indicating a stronger effect. The full, number-based matrices are located in the appendix. My model and Stanza's model had equivalent performance (including types of confusion), so I only included confusion matrices for Stanza's model in this section to avoid redundancy.

Figure 1: Foreign English trained model tested on the CoNLL03 dataset

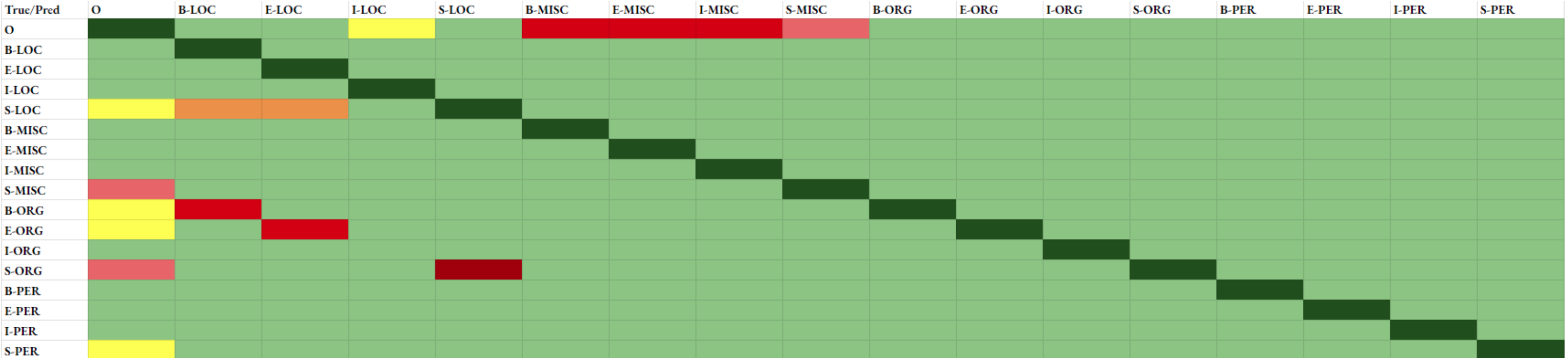
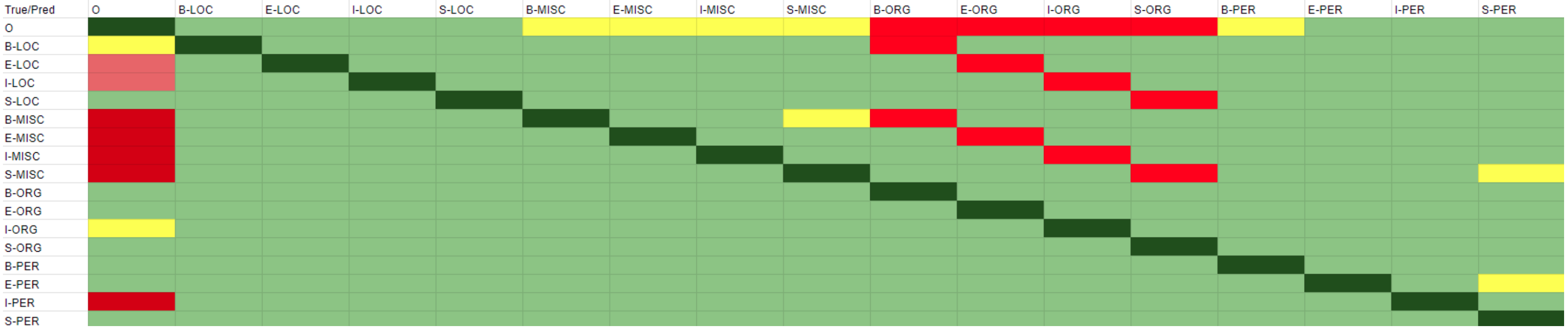


Figure 2: CoNLL03 trained model tested on the foreign English dataset



The matrices show that the foreign English and CoNLL03-trained models demonstrate general confusion between "organization", "location", "none" ("O" in the matrices), and "miscellaneous" when tested on the dataset that they were not trained on. In particular, the foreign English-trained model most often confuses the tag pairs ("location", "organization") and ("none", "miscellaneous") together; the CoNLL03-trained model most often confuses the tag pairs ("none", "miscellaneous"), ("organization", "miscellaneous"), ("organization", "none"), and ("organization", "location").

6 Analysis

6.1 Foreign English trained model performance

Based on the confusion matrix for the foreign English-trained model, I examined the data points in each pattern in which the model demonstrated the greatest error. As seen in figure 1, the model often mistakes "O" (the None class) for miscellaneous, single ("S") miscellaneous/organization tags for "O", and organizations for locations. Further inspection of these errors reveals the reasons for this behavior. With respect to the first pattern, the model often mistakes "O" for miscellaneous when the span of tokens describes an event or currency name. Below are examples to show this pattern, with the named entities bolded:

Example 1: ***Alpine Skiing - Women's World Cup Super G*** *Results*
Example 2: *It said in a statement that it made profits of 4.5 million* ***kroons*** *in November.*
Example 3: *"We will not achieve the full 37 million French* ***franc*** *(net) profit forecast," Frater said.*

This confusion pattern does not reflect a model error, but rather an annotation inconsistency within the CoNLL03 dataset. CoNLL03 labels events/competitions and currencies as miscellaneous, such as the "Euroleague" for competitions and "A$", "US$", and "C$" for currencies (Sang). However, I found hundreds of cases where less common currencies such as the yen (Japan), zlotys (Poland), and lei (Romania) were incorrectly labeled as "O". Similarly, events such as the *"World Cup of Speed Skating"* were erroneously labeled as "O" in CoNLL03. These labeling inconsistencies are likely a result of human error, especially for foreign currency names. CoNLL03 is a widely-used evaluation set for NER research, so these errors should be considered for relabeling by the dataset creators.

The next pattern I investigated was when the true tag was "organization" and the predicted tag was "O". Every instance of model error occurred when the token span referred to a sports team. In these cases, the sentence context was describing a match score, which never appeared in the training for the foreign English-trained model. As seen in the below examples, without knowledge of the team names, the lack of contextual clues makes it difficult for the model to detect the team names as named entities. This ambiguous context, in combination with the fact that the team names were non-named entity words (e.g. Arsenal, which could have been confused for the traditional meaning of the word), likely caused the model error. Shown below are some illustrative examples from the CoNLL03 dataset to demonstrate this effect, with the named entities bolded:

Example 1: ***Reading*** *50* ***Widnes*** *3*
Example 2: ***Arsenal*** *17 10 5 2 34 16 35*

Similar to the last pattern, the model often predicted "location" when the true tag was "organization" for sentences discussing sports teams. This time, the team names were referring to European cities, which may have elicited name-regularity bias in the model, since the pre-trained word vectors for the city names likely associated them with locations (e.g. *I traveled to Liverpool, England.*). Therefore, when the contextual clues of the city names being organizations were missing (such as in a sports match score), the model predicted "location" since it understood names like "Liverpool" as places from pretraining and finetuning, rather than correctly realizing it refers to Liverpool FC. This confusion was only present when the context was ambiguous, as seen in the below examples; sentences where the teams were described performing an action (e.g. winning, beating, losing, etc.) did not observe this confusion. Therefore, this error is likely attributable to a combination of name-regularity bias and unfamiliar context. If the model was fine-tuned on texts containing sports match scores, it is possible that even without knowing the token (Stefanel Milan, for example), it could predict "organization" correctly by learning this type of sentence context. Indeed, in general, sentences with strong contextual clues (e.g. action words tied to organization names) were correctly labeled by the model, even if the named entities were unknown. Shown below are some illustrative examples from CoNLL03:

Example 1: ***Cibona Zagreb*** *(Croatia) 9 5 4 14.*
Example 2: ***Stefanel Milan*** *(Italy) 9 6 3 15*

6.1.1 Testing on Foreign English Test Set

A surprising result was the relatively low performance of the foreign English-trained model on its own test set. While satisfactory, its performance was noticeably lower than the CoNLL03-trained model's performance on its own test set (see Table 1). Upon inspection of its confusion matrix (see appendix B), the patterns of error were very similar to the CoNLL03-trained model's performance on the foreign English test set, suggesting that they could have the same reason for the error. Indeed, after examining behavior from different subsections of the data, I find that this effect can be attributed to the foreign English model incorrectly classifying named entities that were very different from those in its training data. In particular, the errors frequently occurred in texts sourced from indigenous Australia and New Zealand—regions disproportionately underrepresented in the dataset. Thus, many of the named entities in those texts were still foreign to the model; that is, training on Asian and African named entities did not help the model learn indigenous Australian contexts of English. Therefore, the model's erroneous predictions in this specific case can be explained through the same lens of how the CoNLL03-trained model made mistakes, discussed in the following section.

6.2 CoNLL03 trained model

Similar to section 6.1, I analyzed data points in the most common error patterns from Figure 2, where the CoNLL03-trained model was tested on the foreign English dataset. Similar to how the foreign English-trained model predicts miscellaneous on true "O" tags for currency names and events, the reverse effect is observed; the CoNLL03-trained model expectedly predicts "O" for true miscellaneous tags (which were currencies and events) in the foreign English dataset. As for new patterns, there is an interesting error when the model predicts "O" on true "I-PER" tags, meaning the tag is in the middle of a name. This was most prevalent in Southeast and East-Asian texts, where cultural names are traditionally multi-worded and contain hyphens, such as *Chang Hao-Han*. This is not a popular naming convention in Western English, which appears to cause the model to incorrectly split a single name into multiple names and label the connecting hyphen with "O". Furthermore, even when the context clearly indicates that the name is a single token, this behavior is still observed, such as in the following example with the correct label bolded:

"Giant barrel sponges provide shelter for fish and coral recovering from typhoon damage, she said, citing research conducted by ***Huang Yu-sheng****, a Penghu-based marine biology expert."*

When replacing the Chinese name with a Western name, such as *John Smith*, the model correctly labels the name. This interesting case of model behavior demonstrates a clear limitation of exclusively Western English-trained models, motivating increased diversity of text sources in training.

Another pattern of interest was the model predicting "organization" when the true tag was "O". In contrast to the previous patterns, where unclear contextual clues were a limiting factor for the model, I observed many cases where the sentence's context made an unknown token seem like an organization when it was not. When a model observes a token it does not have a representation for, it is mapped to the UNKNOWN (UNK) representation, which opens room for error because a valuable source of information (word meaning) is missing. Thus, this model limitation is related to an insufficiently complete vocabulary, as the absence of a representation forces the model to only work via context, causing the errors seen in the examples below, where the incorrectly labeled token is bolded.

Example 1: *In 2020, rich natural gas resources offshore Mauritania and Senegal were the subject of the biggest long-term liquefied natural gas (****LNG****) contract signed that year.*
In this example, the context word "contract" likely made the model predict "LNG" as an organization, possibly interpreting "LNG" as a group that formed a contract.
Example 2: *Dr Abbass said it was normal for* ***AI*** *researchers to draw on different forms of communication for their work, including other human languages, body language and even music.*
The context word "researchers" likely caused the model to believe that "AI" was a company employing the aforementioned researchers, thus labeling it as an organization.

The next pattern I analyzed was the model predicting "organization" in place of the true "location" tag. This again turned out to be the result of ambiguous sentence context, so without prior knowledge of the token meanings, it is difficult to ascertain the correct tag. I verified that these errors occurred with words that did not appear in the GloVe word vectors, meaning that the model would have mapped the token to an UNK representation. RoBERTa's sampled texts were not available, but I would not be surprised to find that the words were also missing from its pretrain data. Shown below are some examples of this error pattern with further analysis:

Example 1: ***Puerto Iguazú*** *handles twice as many travelers as* ***Ezeiza*** *, Moreno explained Thursday.*
Puerto Iguazú and Ezeiza are cities in Argentina, but this sentence could also be sensical if they were replaced with two organizations instead, such as United Airlines and Delta Airlines, respectively. Due to these two cities having no known representation, the model likely guessed from the context that they were organizations.

Example 2: *Two other Cuban exiles, Simon Camacho and Juan Francisco Bermejo, set up the oldest of the* ***Esteli*** *factories, Joya de Nicaragua, in 1968.*
Despite Esteli being a Nicaraguan city, the sentence context also supports Esteli being interpreted as a company (consider replacing Esteli with Nike, for example). Esteli is also missing from the pretrain and finetuning data, supporting the hypothesis that the model incorrectly guessed the tag from context.

Lastly, I examined when the model predicted "organization" in place of the correct "miscellaneous" tag. This was an interesting error to analyze, given that "miscellaneous" encompasses a wide range of named entities, such as national/ethnic identities and works of art. The instances of model error here were also associated with contexts that had multiple interpretations, along with unknown tokens. Shown below are two examples, each one from a different type of miscellaneous tag, accompanied with analysis:

Example 1: *Informs a source associated with* ***Jhalak Dikhhla Jaa****, "We are in discussion with Dheeraj and the deal is likely to be locked soon."*
Jhalak Dikhhla Jaa is a popular Indian reality TV show (true tag is work of art -> miscellaneous). However, given the context of "...a source associated with...", it could be mistaken as an organization. When replacing it with a well-known Western work of art, *War and Peace* (Tolstoy), the model correctly identifies it as miscellaneous. This indicates that the root cause of the error is the unknown token.

Example 2: *The Casanillo indigenous community of the Paraguayan Chaco is composed of about 3,000 natives of various ethnic groups, among which are* ***Toba Maskoy****,* ***Enxet****,* ***Sanapaná****, and* ***Angaite****, which in turn are subdivided into a total of five villages.*
Each of the bolded ethnic minority groups (true tag NORP -> miscellaneous) were mistaken for organizations, a surprising result given that they were explicitly contextualized as ethnic groups. This example demonstrates that model confusion caused by unknown tokens can induce errors, even with contextual clues. Indeed, none of the ethnic NORPs were present in the pretrain or finetuning for the CoNLL03-trained model.

7 Conclusion

Overall, I find limitations in state-of-the-art NER models whose training data are dominated by Western-sourced texts. Similarly, NER models with training data dominated by non-Western sources demonstrate significant performance gaps (observed over 10% loss in F1) when tested on Western texts compared to their own test sets. Erroneous model behavior was most common in sentences with ambiguous contextual clues to signal the meaning of named entities but occasionally existed despite helpful contexts, such as the last example of the analysis section. Despite the inclusion of pre-trained word embeddings from RoBERTa and GloVe, along with sentence contexts that often disambiguate unknown named entities, I learned that it is important to train on diverse data to achieve strong results on a wide variety of sources. When unfamiliar tokens from foreign contexts of English are observed, they often seem to confuse models, forcing them to rely on context clues that are not guaranteed to reveal the correct named entities. Therefore, one solution is to familiarize the models with these previously unseen tokens (to make *Angaite* (NORP) as recognizable as *French* (NORP), for example). That is, a good way to achieve universally strong NER performance is to train on universally sourced data. Indeed, the combined CoNLL03 and foreign English-trained model I built displayed competitive performance on both datasets, suggesting that existing models would also do well to include more foreign-sourced data in their training. In doing so, model performance in lesser-known contexts of English would improve, without compromising model usage in Western contexts. Additionally, I find annotation errors/inconsistencies within CoNLL03, a very popular NER dataset; I recommend that the error patterns I found in the CoNLL03 dataset, discussed in section 6.1, be addressed by the dataset's future users/current maintainers. With this said, I admit some notable limitations to my work. First, the loss in performance appears to be related to incomplete vocabulary, which means that even if models are trained on increasingly diverse and large sets of data, there will still always be more named entities from increasingly foreign language contexts that are missing. Second, the two datasets I compared have annotation differences that I fixed by collapsing the larger class groups of the foreign English dataset into the classes of CoNLL03, but it was not always clear which classes to collapse into which. For example, I collapsed "product" into "miscellaneous", but one could argue that it is better suited to "organization". These design decisions in my experiments caused some confusion themselves, so the F1 scores for the models may be artificially lower when tested on the opposite dataset from which they were trained on. However, I find that there were still many patterns of error that are attributable to model error rather than these annotation discrepancies. In the future, I would like to see if large language models such as GPT-4 also exhibit the same name-recognition bias and errors as seen in this study.

References

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, and Mehdi Rezagholizadeh. 2021. Context-aware adversarial training for name regularity bias in named entity recognition. *Transactions of the Association for Computational Linguistics*, 9:586–604.

Matt Gormley. 2018. Machine learning department school of computer science carnegie mellon ...

David Lim. 2023. Dslim/bert-base-ner · hugging face.

Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li. 2022. Chinese named entity recognition: The state of the art. *Neurocomputing*, 473:37–53.

Anita Louis, Alta De Waal, and Cobus Venter. 2006. Named entity recognition in a south african context. In *Proceedings of the 2006 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries*, SAICSIT '06, page 170–179, ZAF. South African Institute for Computer Scientists and Information Technologists.

Sebastian Nowozin. 2011. Part 4: Conditional random fields - nowozin.net.

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

Sang. Papers with code - conll-2003 dataset.

Alex Shan and John Bauer. 2023. Stanfordnlp/en-foreign-newswire: Ner dataset built from foreign newswire.

Tatiana Tsygankova, Francesca Marini, Stephen Mayhew, and Dan Roth. 2021. Building low-resource NER models using non-speaker annotations. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 62–69, Online. Association for Computational Linguistics.

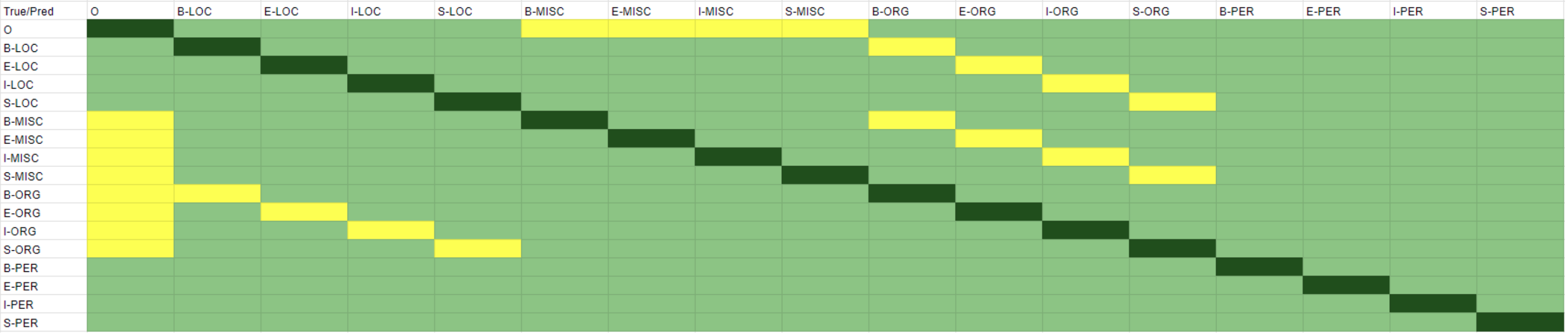
A Appendix: Named entity classes, with definitions and examples

Label Name	Description	Example
Person Name	Names of humans, excluding deities (e.g. "God")	Ryan
Organization	Names of a group or collective body	Supreme Court
Location	A physical place	Paris
Facility (collapsed into location)	A place that is used for a specific utility	Brooklyn Bridge
NORP (collapsed into Misc.)	A national, organizational, religious, or political identity	Chinese
Currency (collapsed into Misc.)	The name of a particular denomination of money	Yen
Product (collapsed into Misc.)	The name of a marketable item or brand	Ford F-150
Miscellaneous	Any other named entity that does not fall into the others (e.g. project initiatives, events, works of art)	Mona Lisa
Date (collapsed into None)	Any time of day, month, or year	2/19/2023
None	Not a named entity	water bottle

B Appendix: Confusion Matrices

B.1 Heat-mapped confusion matrix for foreign English-trained model tested on its own test set

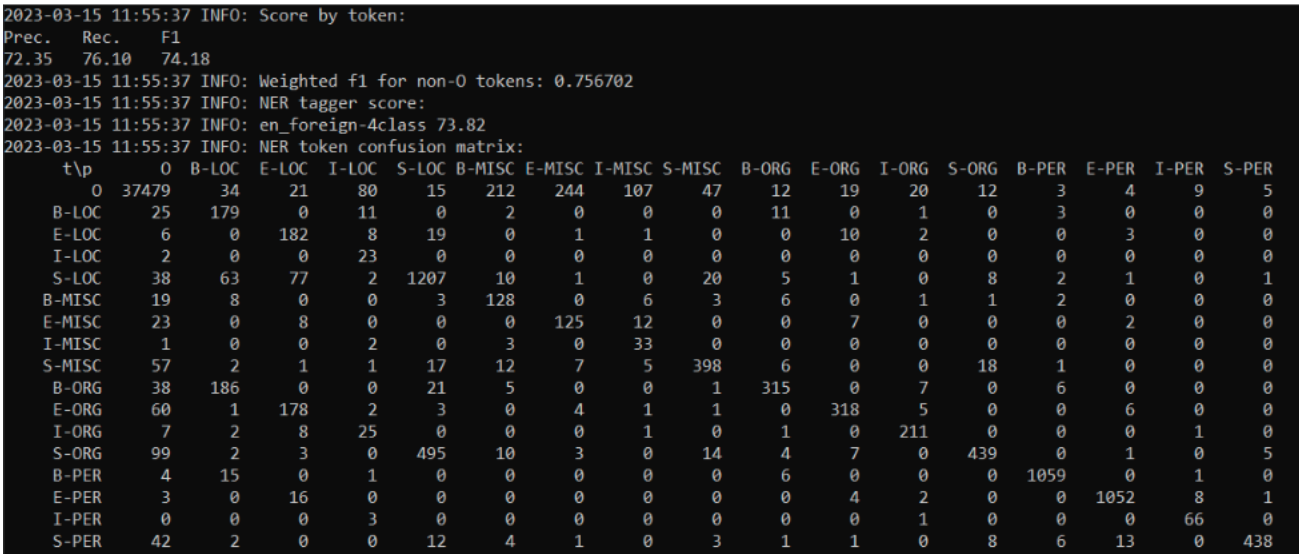
Figure 3: Foreign English trained model tested on the CoNLL03 dataset



True/Pred	O	B-LOC	E-LOC	I-LOC	S-LOC	B-MISC	E-MISC	I-MISC	S-MISC	B-ORG	E-ORG	I-ORG	S-ORG	B-PER	E-PER	I-PER	S-PER
O	⬛	🟩	🟩	🟩	🟩	🟨	🟨	🟨	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
B-LOC	🟩	⬛	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩
E-LOC	🟩	🟩	⬛	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩
I-LOC	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩
S-LOC	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩
B-MISC	🟨	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩
E-MISC	🟨	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩
I-MISC	🟨	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩
S-MISC	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩
B-ORG	🟨	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟩	🟩	🟩	🟩
E-ORG	🟨	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟩	🟩	🟩
I-ORG	🟨	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟩	🟩
S-ORG	🟨	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩	🟩
B-PER	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩	🟩
E-PER	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩	🟩
I-PER	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟩
S-PER	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬛

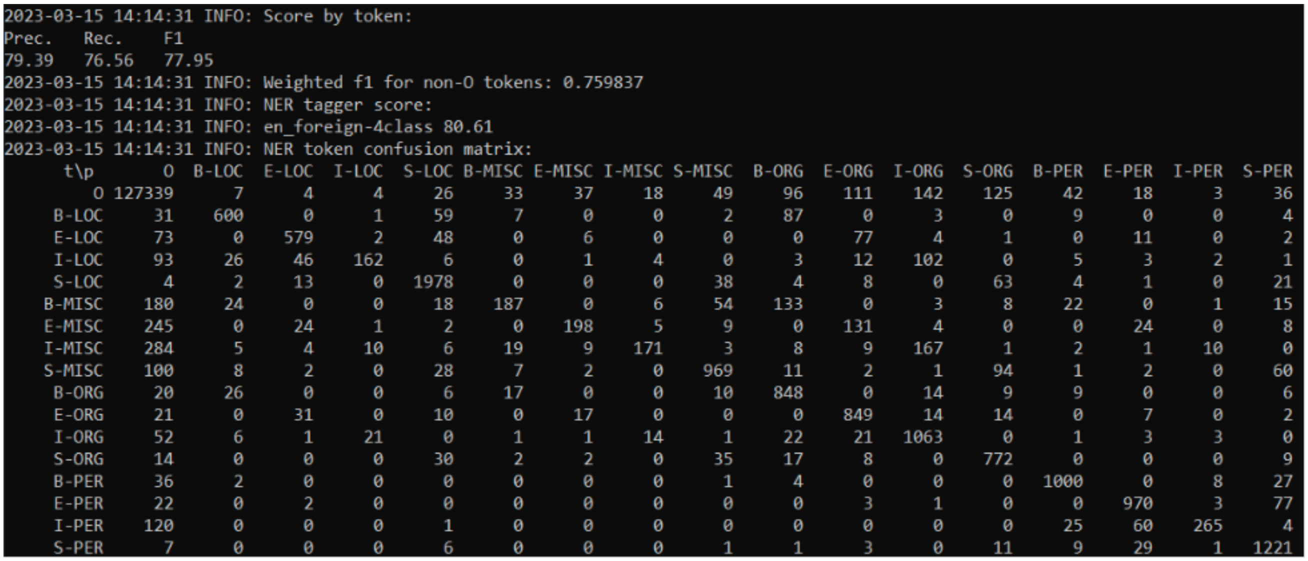
The error patterns with this experiment are noticeably similar to the CoNLL03-trained model tested on the foreign English test set, a finding explained in the analysis section.

B.2 Confusion matrix for foreign English-trained model on CoNLL03 test set



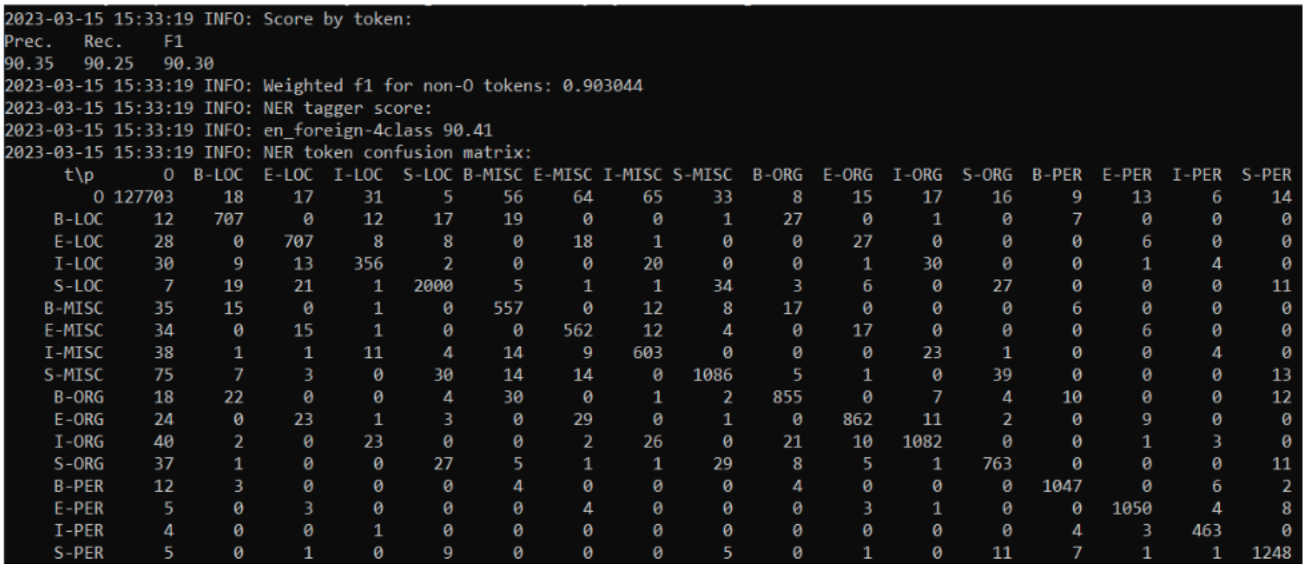
```
2023-03-15 11:55:37 INFO: Score by token:
Prec.   Rec.    F1
72.35   76.10   74.18
2023-03-15 11:55:37 INFO: Weighted f1 for non-O tokens: 0.756702
2023-03-15 11:55:37 INFO: NER tagger score:
2023-03-15 11:55:37 INFO: en_foreign-4class 73.82
2023-03-15 11:55:37 INFO: NER token confusion matrix:
      t\p      O  B-LOC  E-LOC  I-LOC  S-LOC B-MISC E-MISC I-MISC S-MISC  B-ORG  E-ORG  I-ORG  S-ORG  B-PER  E-PER  I-PER  S-PER
        O  37479     34     21     80     15    212    244    107     47     12     19     20     12      3      4      9      5
    B-LOC     25    179      0     11      0      2      0      0      0     11      0      1      0      3      0      0      0
    E-LOC      6      0    182      8     19      0      1      1      0      0     10      2      0      0      3      0      0
    I-LOC      2      0      0     23      0      0      0      0      0      0      0      0      0      0      0      0      0
    S-LOC     38     63     77      2   1207     10      1      0     20      5      1      0      8      2      1      0      1
   B-MISC     19      8      0      0      3    128      0      6      3      6      0      1      1      2      0      0      0
   E-MISC     23      0      8      0      0      0    125     12      0      0      7      0      0      0      2      0      0
   I-MISC      1      0      0      2      0      3      0     33      0      0      0      0      0      0      0      0      0
   S-MISC     57      2      1      1     17     12      7      5    398      6      0      0     18      1      0      0      0
    B-ORG     38    186      0      0     21      5      0      0      1    315      0      7      0      6      0      0      0
    E-ORG     60      1    178      2      3      0      4      1      1      0    318      5      0      0      6      0      0
    I-ORG      7      2      8     25      0      0      0      1      0      1      0    211      0      0      0      1      0
    S-ORG     99      2      3      0    495     10      3      0     14      4      7      0    439      0      1      0      5
    B-PER      4     15      0      1      0      0      0      0      0      6      0      0      0   1059      0      1      0
    E-PER      3      0     16      0      0      0      0      0      0      0      4      2      0      0   1052      8      1
    I-PER      0      0      0      3      0      0      0      0      0      0      0      1      0      0      0     66      0
    S-PER     42      2      0      0     12      4      1      0      3      1      1      0      8      6     13      0    438
```

B.3 Confusion matrix for CoNLL03-trained model on foreign English test set



```
2023-03-15 14:14:31 INFO: Score by token:
Prec.   Rec.    F1
79.39   76.56   77.95
2023-03-15 14:14:31 INFO: Weighted f1 for non-O tokens: 0.759837
2023-03-15 14:14:31 INFO: NER tagger score:
2023-03-15 14:14:31 INFO: en_foreign-4class 80.61
2023-03-15 14:14:31 INFO: NER token confusion matrix:
      t\p       O  B-LOC  E-LOC  I-LOC  S-LOC B-MISC E-MISC I-MISC S-MISC  B-ORG  E-ORG  I-ORG  S-ORG  B-PER  E-PER  I-PER  S-PER
        O  127339      7      4      4     26     33     37     18     49     96    111    142    125     42     18      3     36
    B-LOC      31    600      0      1     59      7      0      0      2     87      0      3      0      9      0      0      4
    E-LOC      73      0    579      2     48      0      6      0      0      0     77      4      1      0     11      0      2
    I-LOC      93     26     46    162      6      0      1      4      0      3     12    102      0      5      3      2      1
    S-LOC       4      2     13      0   1978      0      0      0     38      4      8      0     63      4      1      0     21
   B-MISC     180     24      0      0     18    187      0      6     54    133      0      3      8     22      0      1     15
   E-MISC     245      0     24      1      2      0    198      5      9      0    131      4      0      0     24      0      8
   I-MISC     284      5      4     10      6     19      9    171      3      8      9    167      1      2      1     10      0
   S-MISC     100      8      2      0     28      7      2      0    969     11      2      1     94      1      2      0     60
    B-ORG      20     26      0      0      6     17      0      0     10    848      0     14      9      9      0      0      6
    E-ORG      21      0     31      0     10      0     17      0      0      0    849     14     14      0      7      0      2
    I-ORG      52      6      1     21      0      1      1     14      1     22     21   1063      0      1      3      3      0
    S-ORG      14      0      0      0     30      2      2      0     35     17      8      0    772      0      0      0      9
    B-PER      36      2      0      0      0      0      0      0      1      4      0      0      0   1000      0      8     27
    E-PER      22      0      2      0      0      0      0      0      0      0      3      1      0      0    970      3     77
    I-PER     120      0      0      0      1      0      0      0      0      0      0      0      0     25     60    265      4
    S-PER       7      0      0      0      6      0      0      0      1      1      3      0     11      9     29      1   1221
```

B.4 Confusion matrix for foreign English-trained model on foreign English test set



```
2023-03-15 15:33:19 INFO: Score by token:
Prec.   Rec.    F1
90.35   90.25   90.30
2023-03-15 15:33:19 INFO: Weighted f1 for non-O tokens: 0.903044
2023-03-15 15:33:19 INFO: NER tagger score:
2023-03-15 15:33:19 INFO: en_foreign-4class 90.41
2023-03-15 15:33:19 INFO: NER token confusion matrix:
      t\p       O  B-LOC  E-LOC  I-LOC  S-LOC B-MISC E-MISC I-MISC S-MISC  B-ORG  E-ORG  I-ORG  S-ORG  B-PER  E-PER  I-PER  S-PER
        O  127703     18     17     31      5     56     64     65     33      8     15     17     16      9     13      6     14
    B-LOC      12    707      0     12     17     19      0      0      1     27      0      1      0      7      0      0      0
    E-LOC      28      0    707      8      8      0     18      1      0      0     27      0      0      0      6      0      0
    I-LOC      30      9     13    356      2      0      0     20      0      0      1     30      0      0      1      4      0
    S-LOC       7     19     21      1   2000      5      1      1     34      3      6      0     27      0      0      0     11
   B-MISC      35     15      0      1      0    557      0     12      8     17      0      0      0      6      0      0      0
   E-MISC      34      0     15      1      0      0    562     12      4      0     17      0      0      0      6      0      0
   I-MISC      38      1      1     11      4     14      9    603      0      0      0     23      1      0      0      4      0
   S-MISC      75      7      3      0     30     14     14      0   1086      5      1      0     39      0      0      0     13
    B-ORG      18     22      0      0      4     30      0      1      2    855      0      7      4     10      0      0     12
    E-ORG      24      0     23      1      3      0     29      0      1      0    862     11      2      0      9      0      0
    I-ORG      40      2      0     23      0      0      2     26      0     21     10   1082      0      0      1      3      0
    S-ORG      37      1      0      0     27      5      1      1     29      8      5      1    763      0      0      0     11
    B-PER      12      3      0      0      0      4      0      0      0      4      0      0      0   1047      0      6      2
    E-PER       5      0      3      0      0      0      4      0      0      0      3      1      0      0   1050      4      8
    I-PER       4      0      0      1      0      0      0      0      0      0      0      0      0      4      3    463      0
    S-PER       5      0      1      0      9      0      0      0      5      0      1      0     11      7      1      1   1248
```