kpfriends at SemEval-2022 Task 2: NEAMER - Named Entity Augmented Multi-word Expression Recognizer

Min Sik Oh
Alexa AI*
ohtrent@amazon.com

Abstract

We present NEAMER - Named Entity Augmented Multi-word Expression Recognizer. This system is inspired by non-compositionality characteristics shared between Named Entity and Idiomatic Expressions. We utilize transfer learning and locality features to enhance idiom classification task. This system is our submission for SemEval Task 2: Multilingual Idiomaticity Detection and Sentence Embedding Subtask A OneShot shared task. We achieve SOTA with F1 0.9395 during post-evaluation phase. We also observe improvement in training stability. Lastly, we experiment with non-compositionality knowledge transfer, cross-lingual fine-tuning and locality features, which we also introduce in this paper.

1 Introduction

Multi-Word Expressions (MWEs) are defined as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002). Recent advances in pre-trained language models such as BERT (Devlin et al., 2019) have enhanced performance of Sentence Classification task, however tasks that specifically identify Multi-Word Expressions (MWE) remain unsolved due to its specific idiomatic properties (Garcia et al., 2021; Yu and Ettinger, 2020). This SemEval shared task (Tayyar Madabushi et al., 2022) aims to understand Multi-Word Expressions better by novel classification and sentence similarity tasks.

Named Entity Recognition (NER) is a task to identify Named Entities (People, Organizations etc.) in a sentence. Multiple datasets exist that specifically perform this task, including CoNLL-02/03 Shared Tasks for English, German, Spanish and Dutch (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). Multi-Word Expressions and Named Entities are similar in a

---

Research unrelated to work

MWE	Target	Label
gold mine	This means that search data is a **gold mine** for marketing strategy.	0 (Idiomatic)
gold mine	The hashtag "Qixia **gold mine** incident" has been viewed many million of times on the social media site Weibo.	1 (Non-idiomatic)
gold mine	The **Gold Mine**'s plain frontage & sparse, white-walled dining room suggest that it's a quick-fix refuelling stop rather than a place to linger.	1 (Non-idiomatic)

Table 1: Dataset samples, table from (Tayyar Madabushi et al., 2021). Note that 3rd example is a named entity (The Gold Mine referring to a restaurant).

way that they consist of more than one word but they form a single semantic unit. Thus, Named Entities could be seen as a specific type of Multi-Word Expressions (Jackendoff, 1997; Vincze et al., 2011). However they are different from idiomatic expressions.

We propose **NEAMER - Named Entity Augmented Multi-word Expression Recognizer** that aim to utilize non-compositionality shared between two streams of NLP research. We explore transfer learning between NER and idiom classification tasks. We also experiment with "locality features" to augment representations of text.

We have participated in Subtask A which is a multilingual classification task to determine if a given sentence has correct idiomatic usage or not. We have focused our efforts on the OneShot setting, where the goal is to classify the target sentence utilizing the ZeroShot dataset consisting of idioms not found in test set and the OneShot dataset consisting of 1 idiom-label pair for all idioms in test set.

The dataset has been provided by task organizers (Tayyar Madabushi et al., 2021).

Contributions of this paper are :

- NEAMER system which utilizes transfer learning, NER and other locality features to improve performance and stability of MWE classification task.

- Investigation into transfer learning between NER and idiom classification task.

- Performance and error analysis to understand capabilities of transfer learning, cross-lingual fine-tuning and locality features.

2 Methodology

2.1 Idiom and Named Entity

Idioms and named entities are similar in the way that when they are comprised of multiple words, collocated words encode extra semantics while individual words lose their semantics partially or completely. This property is referred as non-compositionality (Baldwin and Kim, 2010). "In a nutshell" means "very briefly, giving only the main points" (Cambridge, 2022) as an idiom; individual words lose their concrete semantics and only the combination specifies intended meaning. Similarly, "Papa John's" refers to "an American pizza restaurant chain" (Wikipedia, 2022) when used as a named entity; in this case, even grammatical functions of individual words are mostly ignored. This similarity is the basis for the transfer learning experiments we performed.

We have discussed similarities, but what about differences? Idioms and named entities refer to completely different usage of MWEs. Idioms are utilized to improve fluency and understandability, or make language more colloquial (Baldwin and Kim, 2010). Named entities are utilized to specify name of persons, organizations and locations (Tjong Kim Sang and De Meulder, 2003) and do not have such social purpose. Correspondingly we can expect certain knowledge to be easily transferable between two tasks, while it may take more epochs to obtain best final performance due to fundamental difference between tasks leading to necessity for "unlearning" the previous fine-tuned task. We explore the ideas in the experiments.

2.2 Transfer Learning and Stability

As discussed in Section 2.1, idioms and named entities show similar non-compositionality. Thus this is the basis for our transfer-learning experiments, where large language models finetuned on NER task are further trained on idiomatic expression classification task. We investigate following ideas in the experiments:

1. We hypothesize that disparity between task types can bring instability. Large language models are known to be unstable during training (McCoy et al., 2019; Zhou et al., 2020). Language models are trained using Masked LM pre-training task. The aim of the Masked LM task is to classify every masked word to original word, which results in classification of each tokens to 30,000 possible labels. In contrast, the task at hand is much simpler, with the aim being to classify whole sentence into 2 labels according to usage of relevant MWE. NER task can bridge this task complexity gap since the aim is to classify each tokens to 9 labels.

2. We hypothesize that non-compositionality understanding of the model can be shared between tasks. NER systems need to understand non-compositionality to correctly predict B-XXX tags. It also predicts multiple named entities per sentence. Thus we assert that enough non-compositionality understanding is learnt during the NER fine-tuning process compared to Masked LM task where each token is predicted independently.

We additionally hypothesize that language-specific knowledge could be improved for the model through fine-tuning with similar language data, which we perform experiments on.

2.3 Locality Features

We design 5 features that are closely related to MWE usage types. Those are the following:

1. Entity - Whether an MWE contains an NER output span, or an NER output span contains an MWE.

2. Capitalization - Whether any word in the MWE is intentionally capitalized (excluding the first word in a sentence and the case where MWE itself is explicitly capitalized in the dataset).

3. "Be a *" - Whether the MWE starts with a be-verb and the article 'a/an'. Same for Portuguese.

4. "The *" - Whether the MWE starts with "the".

5. Quotation - Whether the MWE is surrounded by quotation marks (" or ').

We name them "locality features" because they expand upon specific position of an MWE by looking at adjacent characters. We encode locality features using a deep neural network to give enough

Feature	Total	0 (Id-iomatic)	1 (Not-idiomatic)
All	4491	2535	1956
"The *"	720	**366**	**354**
Entity	650	94	**556**
Capitalized	634	50	**584**
Quotation	165	**124**	41
"Be a *"	80	**68**	12
Parenthesis	6	**5**	1

Table 2: Label statistics in ZeroShot data

Model	ENG F1
mBERT-base (baseline)	70.7
xlm-roberta-base	75.5
xlm-roberta-large	**79.0**

Table 3: English ZeroShot F1 on validation data

Named Entity Recognition Training

Idiom Classification Training

Figure 1: NER augmented model, see Section 3.2 for details.

significance to the features during training / inference while enabling them to learn complex relationships between the text. This is further informed by label imbalance (excluding "The *" label, which is balanced) shown in Table 2. We perform experiments on whether or not locality features improve the performance on the idiom classification task.

3 Experiment Setup

3.1 Model Selection

Experimental results on English ZeroShot (shown in Table 3) were used to determine pre-trained checkpoints with best performance. We thus selected XLM-Roberta-Large (Conneau et al., 2020) as a starting point for training OneShot models.

The list of checkpoints is: xlm-roberta-base, xlm-roberta-large, xlm-roberta-large-finetuned-conll03-english, xlm-roberta-large-finetuned-conll02-spanish, xlm-roberta-large-finetune-conll03-german, Davlan/xlm-roberta-base-ner-hrl, Davlan/xlm-roberta-large-ner-hrl.

3.2 Model Architecture

Our model training scheme and architecture is presented in Figure 1. We fine-tune the model on NER task with selected language. For the experiments, we utilize NER fine-tuned checkpoints as described in Section 3.1 instead of actually performing NER fine-tuning. Then, we train the NER fine-tuned model with text and idiom (MWE) data for the idiom classification task along with selected locality features. We use two layers of fully connected
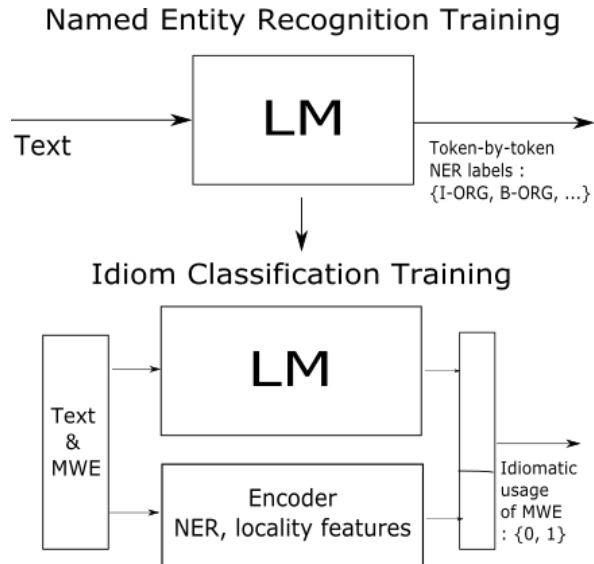
network to encode locality features that are concatenated to the text representation. Locality features used are described in Section 4.5 and implemented in Python to obtain one-hot vectors which are fed into the fully connected network. The feature encoding and hidden layers of FCN are of size 200. In comparision, LM text encoding is 768 as originally used by XLMRobertaForSequenceClassification class in HuggingFace. The size of encoder feature representation is selected to enhance importance of locality features in comparison to LM representation. We use the classification head provided by the same XLMRobertaForSequenceClassification class.

3.3 Training Procedure

We mostly focus on OneShot setting, using both ZeroShot and OneShot data provided. We used a learning rate of 2×10^{-5} and a batch size of 16 for training our models. Models were trained for 24 epochs and the best checkpoints on the evaluation data were selected. Random seeds of 0, 1, 3, 5, 42 are used for initial experiments. If any of the seeds exhibit training failures due to instability (F1 < 0.5), we perform additional experiments with random seeds 49, 81, 100, 121. This resulted in at least 5 checkpoints for our experiments. All provided training data was used for training the models. We picked checkpoints that perform best on respective languages (EN / PT) for evaluation and submission.

Model	Success
XLM-R	55.6%
XLM-R-EngNER	100%
XLM-R-GermanNER	88.9%
XLM-R-EngNER, Augmented	100%

Table 4: Model training success percentage.

Phase	ALL	EN	PT	GL
Baseline	87.7	88.1	87.0	85.4
Evaluation	93.5	**96.1**	89.9	92.1
Post-Evaluation[3]	**94.0**	**96.1**	**91.1**	**92.8**

Table 5: Best submissions.

[1] We implemented our models in HuggingFace (Wolf et al., 2019) and Pytorch (Paszke et al., 2019). We utilize Tesla V100 NVIDIA GPU for training.

4 Results

4.1 Model Stability

We present observed training success rate for each of the models in Table 4. We define training failure as an observance where F1 of the checkpoint is smaller than 0.5. We observe a very high training failure rate for the XLM-R$_{large}$ model (44.4%). We assert that this is due to discrepancy between the pre-training task of MaskedLM and the idiom classification task (more discussion in Section 2.2.)

4.2 Best Submissions

We show our best submissions in Table 5. Our best official submission during evaluation phase is ensemble of 3 checkpoints per language consisting of XLM-R$_{large}$-EngNER & SpaNER, with exception of one XLM-R$_{base}$-EngNER checkpoint[2]. Best post-evaluation submission is ensemble of 5 checkpoints per language consisting of XLM-R$_{large}$-EngNER & SpaNER, selected via process described in Section 3.3. We achieved top 2 during the competition (Section 7). We are currently first place in the post-competition leaderboard (4/15/2022).

4.3 Ensemble Model Performance

We submit our models based on the ensemble model performance shown in Table 6. Checkpoints

---

[1]Galician test data was inferred by Portuguese model for submission.

[2]The checkpoints were selected according to best performance on validation set.

[3]Experiment performed after end of competition.

Model	ALL	EN	PT	GL
XLM-R	92.7	94.5	89.5	92.3
XLM-R$_{NER_{HRL, 36}}$	92.5	**96.1**	88.4	90.3
XLM-R$_{NER_{ENG, SPA}}$	**94.0**	**96.1**	**91.1**	**92.8**
XLM-R$_{NER}$^{Aug}	92.8	95.6	89.4	90.8

Table 6: Test data F1 performance for ensemble models. All XLM-R models are large variant.
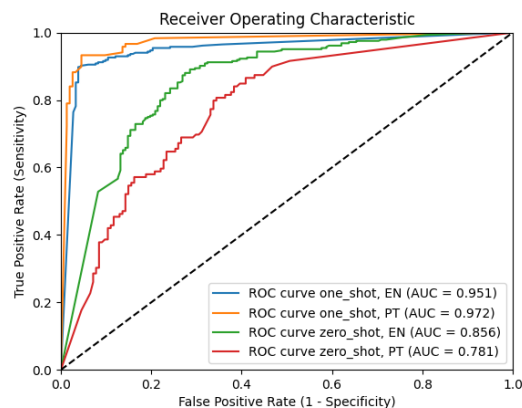

Figure 2: ROC curve of XLM-R$_{NER}$ on validation data for all tasks. We observe very strong prediction ranking capability for both EN and PT (AUC > 0.950) for OneShot task.

for ensemble were selected via the process described in Section 3.3. XLM-R$_{large}$ + NER models (xlm-roberta-large-finetuned-conll03-english, xlm-roberta-large-finetuned-conll02-spanish) that represent transfer learning characteristics perform best, with high F1 score across all languages. Interestingly, locality feature augmentation does not seem to enhance the final output compared to the transfer learning only method. This could be due to model checkpoints not having enough variance between them caused by over-reliance on label imbalance. (More discussion in Section 4.5)

4.4 Average Model Performance

The average F1 scores are presented in Table 7. We observe that additional finetuning on English NER data results in higher performance compared to the baseline XLM-R$_{large}$ model. Augmentation of the model using locality features results in a slight performance increase. Results suggest that NER fine-tuning assists in the idiom classification task, while locality features help relatively less. NER fine-tuning is helpful due to the language model adapting to the non-compositionality expressed in both tasks (more discussion in Section 2.2.)

Model	Average	Ensemble
XLM-R$_{large}$	93.0	94.5
XLM-R$_{large}$-Eng	**94.0**	**96.1**
XLM-R$_{large}$-Eng, Aug	**94.2**	95.6
XLM-R$_{base}$-HRL	91.1	-
XLM-R$_{large}$-HRL	92.9	-
XLM-R$_{large}$-HRL, 36	**94.2**	**96.1**

Table 7: English test data F1

Model	ALL	EN	PT	GL
XLM-R$_{NER}$	62.3	70.8	67.7	44.4
XLM-R$_{NER}$^{Aug}	**64.9**	**72.6**	67.4	**49.2**

Table 8: ZeroShot ensemble test data F1 performance. We note comparatively higher performance for locality feature augmented model on English and Galician data.

4.5 Locality Features

Effect of locality features seem to be marginal, since average F1 (Table 7) only slightly improves in comparison with transfer-learning only model. We also observe lower ensemble performance (Table 6). An enhanced architecture (attention layer in which features explicitly interact with each other) with layer-wise learning rate tuning (to lessen the adverse impact of a cold-start of the feature encoding layers) and dropout (to randomize model training for ensemble enhancement) might be beneficial. We leave it to future work.

We hypothesize that while locality features may be a promising feature to utilize for enhanced architectures, using it by itself may be a relatively too simple indicator. Locality features only require looking at 1~2 specific tokens[4], thus non-compositionality expressed between the tokens themselves is very simple compared to complexity of MWE. An explicit NER feature may also be already encoded in the model via NER fine-tuning step such that no new information is provided during training.

Lastly, we note that we achieve the best ZeroShot setting performance in our experiments with XLM$_{NER}$^{Aug} model which is an ensemble of 3 checkpoints (Table 8). Thus, the locality features could be more promising in the ZeroShot setting where there is less information regarding specific MWE usage. We leave a thorough evaluation to future work.

---

[4]i.e. Capitalization - first letter of words in MWE, Quotation - ' or " before and after MWE. Parenthesis - (or) before and after MWE.

Model	EN	PT	GL
XLM-R$_{large}$-Eng, Spa	94.0	87.5	88.5
XLM-R$_{large}$-German	93.6	87.2	84.2
XLM-R$_{base}$-HRL	91.1	83.6	83.2
XLM-R$_{large}$-HRL	92.9	84.0	83.7
XLM-R$_{large}$-HRL, 36	**94.2**	85.9	87.2

Table 9: Test data average F1 performance for HRL model variants and English, Spanish and German NER fine-tuned model.

4.6 Crosslingual NER Transfer Learning

XLM-R$_{large}$-HRL is an XLM-R$_{large}$ model trained on NER tasks for 10 languages (Arabic, German, English, Spanish, French, Italian, Latvian, Dutch, Portuguese and Chinese). Rationale for fine-tuning this model is to observe the following :

1. Impact of fine-tuning on a model from a pre-trained model trained on NER data from multiple languages. This model has been trained on all CONLL02 / 03 datasets for English, Spanish, Dutch and German, as well as 8 language specific datasets.

2. Impact of fine-tuning on a model which has been pre-trained with capability to perform Portuguese NER task. This model has been trained on Paramopama and Second Harem (Freitas et al., 2010) Portuguese NER datasets.

We show the results in Table 9. We observe that while XLM-R$_{large}$-HRL performs worse on EN F1 than the similarly fine-tuned XLM-R$_{large}$-English and German, training for 36 epochs (50% epoch increase) yields comparable performance. This aligns with our hypothesis that task-to-task training requires "unlearning" partial aspects of the previous task and thus may take longer to train (more discussion in Section 2.1). XLM-R$_{large}$-English was only trained on CoNLL03 English NER task, while HRL models were trained on NER datasets corresponding to 10 languages - this may result in a higher amount of NER task and language specific knowledge that needs to be removed for the model to train properly.

Similarly, we observe worse performance on Portuguese and Galician results for HRL models compared to Spanish fine-tuned model. Portuguese and Galician seem to require more training epochs than English to achieve comparable performance. This may be due to the difference in dataset size per language in both the ZeroShot and OneShot training data for idiom classification task

Feature	LM	NER	Aug
Capitalized (137)	94.2	94.2	91.2
Entity (131)	93.1	93.1	90.8
"The *" (52)	86.5	**92.3**	**92.3**
"Be a *" (13)	100.0	100.0	100.0
Quoted (12)	90.5	90.5	90.5

Table 10: Micro F1 Metrics (validation data) for each locality feature tagged samples corresponding to XLM-R, XLM-R$_{NER}$ and XLM-R$_{NER}$^{Aug}. We observe that transfer learning has improved the performance for "The *" feature. More discussion in Section 5.1.

(English:Portuguese = 2.9:1). We leave training the models on more Portuguese idiom classification datasets and longer epochs to future work.

We also experiment with a model fine-tuned on CoNLL 03 German NER task. We note slightly worse performance for German fine-tuned model compared to models fine-tuned on highly similar languages (English and Spanish NER fine-tuned models). This result seems to suggest that fine-tuning the model on same language for both NER task and Idiom Classification task achieves best performance. More experiments with many languages from other parts of the world could be performed.

5 Error Analysis

5.1 Categorical Performance

We show the F1 metrics for the validation data per each feature in Table 10. We find that the F1 score of "The *" locality feature has increased by 5.8 points after transfer learning is introduced. This locality feature does not directly correspond to NER, and is the only sample-balanced locality feature as shown in Table 2. Thus, we argue that this is further proof of NER transfer learning teaching general non-compositionality to LM that is transferred to MWE classification task.

We also find that Capitalized and Entity F1 scores have stayed the same after the introduction of NER transfer learning, and it has actually decreased by 2~3 points after locality feature augmentation. We also observe a recall decrease of 0.214 (0.357 -> 0.143) as shown in Table 11. As discussed in Section 4.5, this is due to over-reliance on training data label imbalance.

5.2 Sample Analysis

We list the prediction improvements between base XLM-R$_{large}$ model and NER transfer-learning

	Pred 0	Pred 1
Label 0 (Idiomatic)	5	9
Label 1 (Non-idiomatic)	0	117

	Pred 0	Pred 1
Label 0 (Idiomatic)	2	12
Label 1 (Non-idiomatic)	0	117

Table 11: Confusion matrix for Entity in non-augmented models(XLM-R, XLM-R$_{NER}$) vs augmented model (XLM-R$_{NER}$^{Aug}).

based models in Appendix A. Interestingly, we observe that 6 out of 9 sample prediction improvements for English model are also observed with HRL, German[5] models. This strongly suggests that shared characteristics are present between NER transfer-learning based models. We also observe that the model output changes are not associated with named entities, strengthening our hypothesis of general non-compositionality knowledge transfer between tasks.

6 Conclusion

We present NEAMER - Named Entity Augmented Multi-word Expression Recognizer. This system explores how we can utilize non-compositionality shared between Named Entity and Idiomatic Expressions. We find that the NER transfer learning variant achieves the best MWE classification OneShot performance. We also observe high training stability. We investigate non-compositionality knowledge transfer between tasks and obtain promising results across experiments.

7 Rank Information

During the official evaluation phase, we were top 2 in Subtask A (One-Shot) leaderboard with F1 score of 0.9346 (Table 5). We trained 50 checkpoints and measured F1 on English and Portuguese separately. Checkpoints were generated via process described in 3.3. Best English performing checkpoints inferred on English test submission data, while best Portuguese performing checkpoints inferred on Galician as well as Portuguese test submission data. Finally, we ensembled best performing models on each language using different strategies (including top 3, top 5, top 10) to optimize generalization performance.

---

[5]German model is not trained on CoNLL03 English data, making the result more interesting.

8   Acknowledgements

References

Timothy Baldwin and Su Nam Kim. 2010. *Handbook of Natural Language Processing, Second Edition*. CRC Press, Boca Raton, USA.

Cambridge. 2022. In a nutshell. In *Cambridge Dictionary*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. 2010. Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

R. Jackendoff. 1997. *The Architecture of the Language Faculty*. Linguistic inquiry monographs. MIT Press.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *CoRR*, abs/1911.02969.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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

Veronika Vincze, István Nagy T., and Gábor Berend. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 289–295, Hissar, Bulgaria. Association for Computational Linguistics.

Wikipedia. 2022. Papa John's — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Papa%20John's&oldid=1081001743. [Online; accessed 07-April-2022].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP),
pages 4896–4907, Online. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal.
2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions.
CoRR, abs/2004.13606.

A   Prediction Improvements

We list the classification improvements[6] in validation dataset observed across NER transfer learning models in comparison to base XLM-R$_{large}$ model. The NER transfer learning models we compare are English, German, and HRL (10 languages). We find 6 samples that prediction have improved consistently across all 3 models, which is 66.7% of prediction improvements in English model.

MWE	Sentence	Feature
high life	"This is the story of "Memo Fantasma" or "Will the Ghost," who started life in the Medellín Cartel, funded the bloody rise of a paramilitary army, and today lives the high life in Madrid."	"The *"
home run	He is the only player to hit at least 30 home runs in 15 seasons and is one of only four players to produce at least 17 seasons with 150 or more hits.	-
health check	Big Tech Show · Why your DNA may be your next health check	-
pillow slip	By morning most of it is on the pillow slip, and soap and water will clean up the rest."	"The *"
pillow slip	"Her pillow slip by now was very much askew; one ear pointed northward, the other southeast, and she could only see out of one eye."	-
dry land	And God called the dry land Earth; and the gathering together of the waters called he Seas: and God saw that it was good.	"The *"

Table 12: Improved samples due to NER fine-tuning.

---

[6]Wrong prediction in XLM-R$_{large}$ model, but correct prediction in NER transfer learning models.