Could KeyWord Masking Strategy Improve Language Model?

Mariya Borovikova, Arnaud Ferré, Robert Bossy, Mathieu Roche, Claire Nédellec

Could KeyWord Masking strategy improve language model?

Mariya Borovikova[1, 2], Arnaud Ferré[1], Robert Bossy[1], Mathieu Roche[2], and Claire Nédellec[1]

[1]MaIAGE, INRAE, Université Paris-Saclay
[2]TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE

Abstract

This paper presents an enhanced approach for adapting a Language Model (LM) to a specific domain, with a focus on Named Entity Recognition (NER) and Named Entity Linking (NEL) tasks. Traditional NER/NEL methods require a large amounts of labeled data, which is time and resource intensive to produce. Unsupervised and semi-supervised approaches overcome this limitation but suffer from a lower quality. Our approach, called KeyWord Masking (KWM), fine-tunes a Language Model (LM) for the Masked Language Modeling (MLM) task in a special way. Our experiments demonstrate that KWM outperforms traditional methods in restoring domain-specific entities. This work is a preliminary step towards developing a more sophisticated NER/NEL system for domain-specific data.

1 Introduction

Named Entity Recognition (NER) and Named Entity Linking (NEL), also known as Named Entity Disambiguation and Named Entity Normalisation, are important tasks of Natural Language Processing that aim to detect named entities from unstructured text, categorize them (NER) and then link to a knowledge base (NEL) (see figure 2). Traditional approaches to the task [6, 29, 34] require a huge amount of manually labeled data which is resource consuming to produce. Unsupervised [16] and semi-supervised [27] approaches exceed the limit of traditional approaches, but the quality of the results obtained from domain-specific texts tends to diminish [41]. Labelled data scarcity is a major obstacle in achieving high-quality NER and NEL in the biological and biomedical domains. However, lists of relevant terms (lexicons) are usually available. Our work is an original contribution that aims to improve a few-shot technique by fine-tuning a BERT-based model [10] for the Masked LM (MLM) task on biomedical and

epidemiological data. The Masked LM (MLM) task, as introduced in [10], involves predicting or restoring missing tokens in a text given its context. To accomplish this goal, a LM takes a text as input and replaces a random subset of its tokens by a special mask token ([MASK]). The model's performance is then evaluated using accuracy and perplexity metrics. Usually, while adapting an LM to more specific data, the model is fine-tuned on a smaller amount of the relevant texts in the same way. However, the masking procedure can vary depending on a particular purpose of the system. Traditional approaches, such as random masking, may not adequately account for the specific characteristics of these texts. This motivates our proposed approach, which focuses on leveraging semantic information to guide the masking process. Specifically, we pre-fine-tune a BERT-based model to mask only the domain-relevant tokens taken from the lexicons, guided by the assumption that this approach will better account for the linguistic nuances present in these texts. More precisely, we conduct our experiments in the biology and biomedical domains, leveraging publicly available datasets containing biomedical and human disease epidemiology news texts, as well as an in-house dataset focused specifically on plant disease epidemiology. This paper presents a preliminary step towards a more advanced NER/NEL system that could be applied to any domain-specific data. Our work is a part of the BEYOND project [1]. The primary objective of the BEYOND project is to improve epidemiological surveillance strategies. This involves the development of novel risk indicators for plant diseases and the proposal of new surveillance plans to achieve this goal. The rest of the paper is organized as follows: In Section 2, we provide a detailed review of related work. In Section 3, we describe our proposed method. In Section 4, we present our experimental setup and results. Finally, we conclude the paper and discuss future work in Section 5.

Figure 1: This figure provides an example of NER/NEL tasks where short passages are labeled with different entity types (i.e., pests in red, plants in green, microorganisms in emerald, and locations in blue) along with their unique identifiers and the specific resource in the format of Database:id.

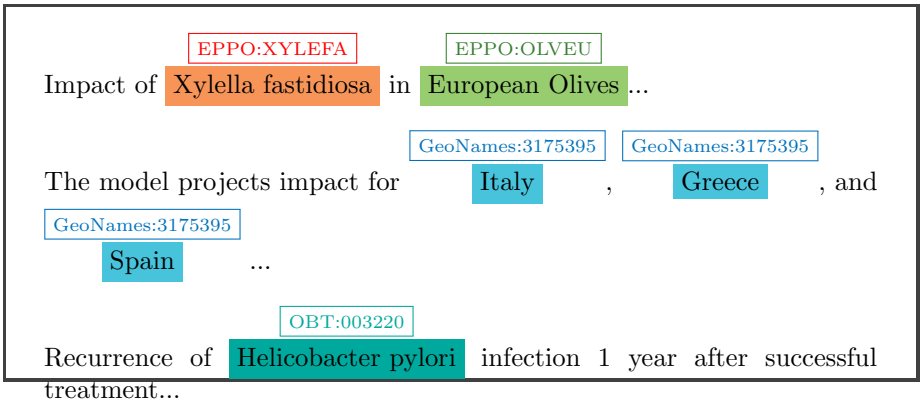
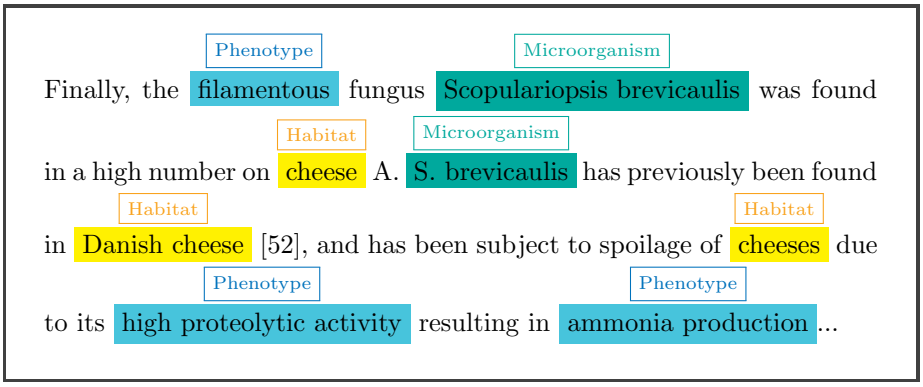


Figure 2: This figure provides an example of NER/NEL tasks where short passages are labeled with different entity types (i.e., pests in red, plants in green, microorganisms in emerald, and locations in blue) along with their unique identifiers and the specific resource in the format of Database:id.



2 Related work

The state-of-the-art NER/NEL models rely on supervised Machine Learning algorithms such as DeepType [29], C-Norm [11], GAN-BERT [18]. Besides, a fine-tuned BERT-based LM with various architectural modifications are mostly used to resolve both NER [33], [39] and NEL [5], [40] tasks.

Traditional unsupervised approaches rely on lexicon-based rules to identify entities within text, often using pre-existing dictionaries or knowledge bases. For example, the system presented in [28] uses syntactic and semantic rules applied to every word in the text, while an approach proposed in [38] involves searching for noun phrases in the text that differ from terms in a given Database by only a few symbols. Modern unsupervised approaches imply the usage of clustering strategies, such as kNN [35]. The state-of-the-art unsupervised NER method, Cyclener [16], trains two functions: one that generates Named entities (NE) from the input text, and another that generates text from a set of NE. The two functions are trained iteratively in a cycle, where the output of one function is used as the input to the other. One of the notable strengths of Cyclener is that it does not require annotated texts, but instead relies on a random set of annotations with the same NE distribution as the texts being analyzed.

Recent research on few-shot named entity recognition (NER) has primarily focused on two approaches: transfer learning and meta-learning. In transfer learning, models pre-trained on large datasets for the same or similar tasks are fine-tuned on smaller, target datasets [8, 13, 17, 22, 23].

On the other hand, proponents of the meta-learning approach train a model on domain-specific data for various tasks and then adapt it to perform the

NER/NEL task with only a few examples [12, 20, 21, 24]. One of the leading meta-learning approaches is proposed in [21], where a generative model rewrites all mentions, and their dense representations are compared with those of the entities in the database in terms of cosine distance. However, we were particularly interested in a transfer learning approach for domain-specific data proposed in [13]. The authors pre-train word embeddings on a large corpus of domain-specific texts, and then fine-tune them on a small labeled NER dataset. What makes this approach interesting is that it relies solely on raw texts written on the same topic during the pre-training stage and significantly improves the NER model's performance.

In addition to these approaches, some researchers have explored using MLM as a preliminary step of training a model for a NER task. For instance, in [26], the authors aim to enhance question-answering systems in the biomedical domain by fine-tuning the LM by masking some of the entities recognized by the SciSpacy system [25]. This approach is similar to the one we propose, but it requires a pre-trained NER system.

Our approach to improve the NER/NEL model relies on the intuition of [13] and [26], which is that by masking relevant domain-specific words during MLM pre-training, the model can learn to better represent domain-specific knowledge and improve the performance for other tasks on the same domain. These works were of particular interest to us due to their success in improving algorithm performance without requiring annotated data, but using raw texts from a specific domain instead. However, [26] requires an existing NER system, and both [13] and [26] rely on training and fine-tuning the system on the same corpus, which may lead to unpredictable behavior when applied to new data. To address these limitations, we propose using a KeyWord Masking strategy for a MLM task that we introduce in the section 3.1.

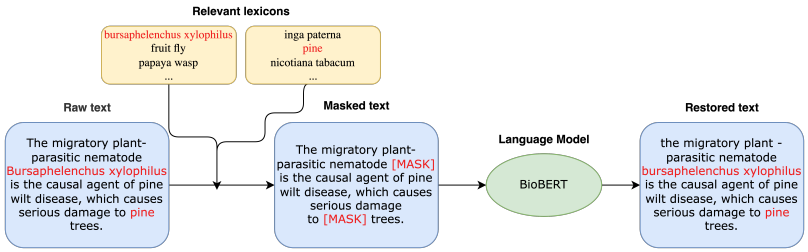
3 Methodology

Building on the insights from [26], we propose that masking some mentions of the relevant entity types can substantially enhance the NER/NEL algorithm performance. Specifically, we compile a comprehensive list of those mentions, based on the entity type semantics, before the fine-tuning. In this section, we describe in detail the KeyWord Masking strategy and the datasets involved. An overview of our approach is shown in Figure 3.

3.1 Masking strategy

The MLM task consists in restoring randomly masked tokens in the input text based on the context provided by the surrounding tokens. Traditional approaches mask randomly chosen tokens (conventionally 15% [36]). Since the final objective of our system is to identify particular entities, we will prioritize masking these mentions. More precisely, we have a lexicon for each entity type which varies for different datasets (see table 1 and section 3.2). These lists were

Figure 3: **Overview of the KeyWord Masking approach.** Relevant entity type mentions are masked in the raw text, based on a comprehensive list compiled from entity type semantics. The Language Model then restores these masked terms.



compiled by gathering relevant terms from domain-specific databases. Then, the list items are masked in the training corpus (see algorithm 1). If the ratio of the masked tokens is below 15%, other tokens are randomly masked while ensuring that if a token is a part of a word, the entire word is masked. The complete process of fine-tuning the model is described in Algorithm 1.

3.2 Domain-specific terms lists

Domain-specific lexicons were created for each dataset separately. For the Plant Health domain, we consulted two sources: a list of pests treated by PESV and a list of pests studied in the BEYOND project. The PESV's list was selected because the Platform monitors plants health at the international level and includes the most dangerous and frequently encountered pests worldwide. The BEYOND project's pests complements the PESV's list with additional pathogens that represent diverse dissemination means. We collected all pests and vulnerable plants names in the EPPO (European and Mediterranean Plant Protection Organization) database [4] and the NCBI taxonomy [32]. The list of plants was compiled using the Encyclopedia of Life resource. By leveraging these sources, the resulting list of pests, pathogens and plants is representative of the types of entities relevant to the specific domain of plant health management.

For Microbiology, we have used a subset of the NCBI taxonomy that contains scientific names of microorganisms.

For geographical entities, we relied on a list of countries and cities from the GeoNames database [2]. The GeoNames database is a comprehensive geographical database for named geographic locations worldwide. Each record in the database contains information, e.g. the name of the location, its coordinates, population size.

3.3 Evaluation method

After fine-tuning our model, the annotated entities of the evaluation dataset are masked. More specifically, we have two evaluation strategies. We mask (1) all

Algorithm 1 KeyWord Masking

Input: raw texts corpus *Corpus*, relevant lexicons *Lexicons* and Language Model M

$min.loss \leftarrow +\infty$
$patience \leftarrow 5$
while $patience > 0$ **do**
  for each $lexicon \in Lexicons$ **do**
    for each $text \in Corpus$ **do**
      for each $term \in lexicon$ **do** REPLACE$(text, term, mask)$
      end for
      while $\frac{length(mask)}{length(text)} < 0.15$ **do**
        REPLACE.RANDOM.ELEMENT$(text, mask)$
      end while
    end for
    FINE-TUNE(M)
    if $LOSS(M) < min.loss$ **then**
      $min.loss \leftarrow +\infty$
      $patience \leftarrow 5$
    else
      $patience \leftarrow patience - 1$
    end if
  end for
end while

the domain-specific entities, i.e. *Pests* and *Plants* for a Plant Health dataset, or (2) random words ($\approx 15\%$). Then, we use the base (non-fine tuned) model, the model fine-tuned in a usual way and the model fine-tuned in a way described in section 3.1 to restore the masked entities. Further, we measure the performance of models for the MLM task by calculating the standard metrics of accuracy and perplexity. Accuracy is the ratio of the number of correct predictions to the number of all the predictions made by a model. Perplexity is calculated as follows:

$$Perplexity(M) = \exp(CrossEntropyLoss(M)) =$$
$$\exp(-\sum_{t \in v} L(t|context) * \log_2 P(t|context)),$$

where M is a language model, t is a token from the vocabulary v of the language model, $L(t|context)$ is the true probability of occurrence of token t in the given context, and $P(t|context)$ is the probability of the occurrence of token t predicted by the model M in the given context. It is used to quantify the dissimilarity between the predicted and actual probability distributions of the text.

The relationship between perplexity and accuracy is not always straightforward. While accuracy assesses how well the model predicts each token, perplexity measures the overall quality of the model's predictions by considering the

Table 1: Masked entities by datasets

Dataset subject matter	Entity type	Number	Examples
Plant Health	Plant	151	*vitis vinifera subsp. vinifera, red rice, tree*
	Pest	96	*l@f. odoratissimum tr4, leafhopper, triozida*
Microbiology (Bacterias and their habitats)	Microorganism	6758474	*Chainia INMI 1349, JRF 142, P.insecticola*
	Habitat	4522	*Ornithodoros moubata, donkey, wild tree*
	Phenotype	574	*oval-shaped, endophytic, osmophile*
Locations (Geographical data in News)	Location (Countries, Cities)	2132976	*Tianxia, Munich, Australian*

probability of the entire generated text. Therefore, these two metrics are complementary in evaluating the performance of MLM models, providing a more nuanced assessment of the model's quality.

3.4 Training data

We would like to evaluate the effectiveness of our approach across different types of domain-specific data. To achieve this, we have chosen three distinct semantic groups of entities types we intend to mask: Plant Health, Microbiology, and Locations. The first two domains are highly specialized and differ significantly from one another, while the third includes entities types that are more likely to appear in general news articles.

Plant Health. In the Plant Health domain, we focus on adapting a LM to *Plant* and *Pest* species. For this purpose, we used an extended version of the corpus we received from the Plateforme d'Épidémiosurveillance en Santé Végétale (PESV) [3], which includes scientific reports and news about plants diseases, their vectors and corresponding pathogens. Additionally, we collected texts that describe plants and/or pests (encyclopedic articles, other scientific reports, etc.) or texts similar to those which will be further processed by the real-time NER+NEL system (news, official reports, etc.). Namely, our efforts were focused on gathering texts from relevant websites where the license explicitly permitted the utilization of the data (e.g., UK Plant Health Information Portal, Missouri Botanical Garden Website, https://www.hortweek.com/news). As a result, we obtained 1311 texts with an average length ranging from 10000 to 20000 characters.

Microbiology. In the Microbiology domain, we fine-tune a LM with a focus on *Microorganisms*, *Habitats* and *Phenotypes*. We accomplish the fine-tuning on the Bacteria Biotope 2019 corpus [7]. The corpus is a dataset of

scientific publications related to bacteria and their habitats, and it is annotated for NER/NEL and Relation extraction tasks. We use raw texts from a training set of the corpus to fine-tune a LM specifically for this domain.

General-domain news. In the General-domain news domain, we focus on *Location* entities and use raw texts from the English conll2003 corpus [31], which is a subset of the Reuters news stories.

4 Experiments

4.1 Evaluation data

To evaluate the performance of our approach, we created or sourced from a publicly available NER/NEL dataset a separate test set for each training set based on its theme. Named entity statistics are provided in Table 2.

Plant Health. To the best of our knowledge, there is currently no publicly available dataset for NER/NEL in the Plant Health domain. Therefore, we have constructed a very small one that we introduce here. "Plant Health Risks Identification from textual data" is a new open-source test dataset for the evaluation of NER/NEL in the Plant Health domain. It is a small dataset of 23 representative manually annotated texts that contain relevant and representative information on Plant Health Monitoring. Specifically, these texts are official reports or news articles and describe the occurrence of a pest on a particular plant in a specific geographical zone. During the text selection process, we consulted with experts in the Plant Health domain from the EPPO (European and Mediterranean Plant Protection Organization) and requested a list of the currently monitoring pests. We then collected texts that cover all the pests from the list, with each pest occurring multiple times under different names to ensure comprehensive coverage. We made sure that there is no document overlap between the training and test sets. During the manual annotation process, we labeled the mentions of four entity types (see table 2). Only two of them, *Pest* and *Plant*, are used for the method described in this article. To normalize the entities, we assigned EPPO [4] and NCBI [32] labels for *Plant* and *Pest* entities respectively. The other two annotated entities types, *Date* and *Location*, will be used in subsequent work for NER/NEL system evaluation. We used *GeoNames* [2] labels for *Location* entities, while temporal entities were normalized with the TIMEX3 [9] format. The dataset with its full annotation guide and a more detailed description are publicly available through open access and can be found at the following link[1].

Microbiology. To evaluate the performance of our approach in restoring entities related to microbiology, we use the Bacteria Biotope 2019 corpus [7] development set which contains 100 documents.

Geographical data. In order to measure the quality of our method in reconstructing geographical entities, we use the GeoVirus corpus [14]. The

[1] https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi%3A10.57745%2FHVPITE

dataset consists of 229 news articles that describe events related to epidemics and/or global disease outbreaks.

Table 2: Test corpora statistics

Dataset	Plant Health					Bacteria Biotope (dev)				GeoVirus
Entity type	Plant	Pest	Date	Location	All	Microorganism	Habitat	Phenotype	All	Location
Total entities	86	188	93	131	400	402	610	161	1073	1981
Unique entities	24	43	61	71	199	137	130	44	311	569

4.2 Experimental protocol

All the data and results processing were conducted using Python Programming language 3.8 [30]. The main libraries used for this project are: PyTorch [15] and transformers [37].

In our experiments, we have fine-tuned BERT [10] and BioBERT [19] models. We selected BERT because it is a widely used model accross various domains, and BioBERT, as it is the current State-of-the-art model in the biomedical domain, which includes Microbiology and Plant Health. Both models were chosen for their ease of use and relatively light computational requirements.

For both models training is done with an Adam optimizer and a learning rate $5e-5$ for 40 epochs.

4.3 Results

Tables 3 and 4, along with Figures 4 and 5 present the results based on accuracy and perplexity indicators, respectively. The evaluation of BERT and BioBERT models was performed on different datasets with and without fine-tuning, using standard masking and a KeyWord Masking approach (KWM). The *Dataset* column indicates the evaluation dataset. The *Masked tokens* column describes whether random or domain-specific tokens were masked during the evaluation process. When masking specific entities, all the entities mentioned in Tables 3 and 4 were masked in the texts. The *Model* column contains the name of the pre-trained model, and the columns *without fine-tuning*, *standard masking*, and *KWM* contain the corresponding accuracy or perplexity indicators. The best results for each dataset and model are shown in bold. To ensure the reliability of our results, we trained our models 10 times and present the mean and standard deviation values.

Our experiments reveal a clear distinction between the two masking strategies. Models fine-tuned with the KWM strategy outperform those fine-tuned using the standard approach for restoring domain-specific entities. Conversely, for

Table 3: Accuracy comparison of BERT and BioBERT models fine-tuned by different masking techniques. The values in bold are the best for each task and model.

Dataset	Masked tokens	Model	without fine-tuning	standard masking	KWM
Plant Health	*Plants* and *Pests* entities	BERT	0.02	0.09±0.02	**0.21±0.02**
		BioBERT	0.04	0.14±0.02	**0.56±0.03**
	Random	BERT	0.37	**0.52±0.02**	0.3±0.02
		BioBERT	0.41	**0.56±0.02**	0.46±0.02
BB	*Microorganisms*, *Habitats* and *Phenotypes* entities	BERT	0.02	0.02±0.01	**0.06±0.01**
		BioBERT	0.03	0.03±0.01	**0.08±0.03**
	Random	BERT	0.37	**0.38±0.01**	0.12±0.00
		BioBERT	**0.46**	0.43±0.01	0.15±0.00
GeoVirus	*Locations* entities	BERT	0.08	0.14±0.03	**0.21±0.1**
	Random	BERT	0.30	**0.41±0.03**	0.08±0.00

random word masking, models fine-tuned using the standard approach perform better than those fine-tuned with KWM. Moreover, we observe that BioBERT outperforms BERT on the biomedical (Bacteria Biotope (BB)) and epidemiological (Plant Health) texts, whereas BERT performs better on general domain texts (GeoVirus).

Another observation concerns the perplexity. A lower perplexity value indicates that the model is more confident in predicting the next token. Our results show that the perplexity is consistently lower on the test set when using the standard fine-tuning approach. This implies that the model fine-tuned with the standard approach has a higher level of confidence in predicting a masked token. This finding underscores the importance of fine-tuning methodology in achieving optimal performance in the MLM task.

Table 4: Perplexity of BERT and BioBERT models fine-tuned by different masking techniques. The values in bold are the best for each task.

Dataset	Masked tokens	Model	without fine-tuning	standard masking	KWM
Plant Health	*Plants* and *Pests* entities	BERT	12253.5	**2.7±0.2**	5.1±2.1
		BioBERT	968.5	2.2±0.2	**1.3±0.3**
	Random	BERT	11350.6	**1.9±0.2**	2.7±1.1
		BioBERT	1014.3	**1.7±0.2**	2.2±0.1
BB	*Microorganisms*, *Habitats* and *Phenotypes* entities	BERT	26170240	**6.2±3.2**	6.7±1.4
		BioBERT	313426	6.3±2.1	**5.9±2.3**
	Random	BERT	2605567.2	**1.9±0.5**	3.4±0.4
		BioBERT	61243.3	**1.7±0.9**	3.1±1.9
GeoVirus	*Locations* entities	BERT	2432350	**4.8±3.6**	15±2.7
	Random	BERT	2875939	**7.3±5.2**	20±6.1

Figure 4: Accuracy comparison.

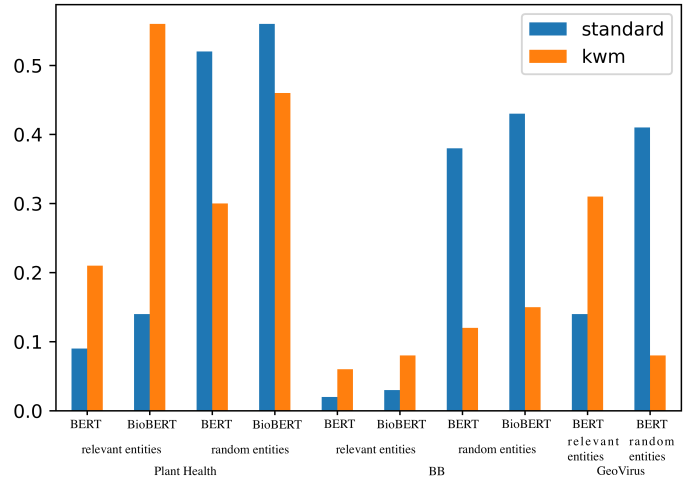
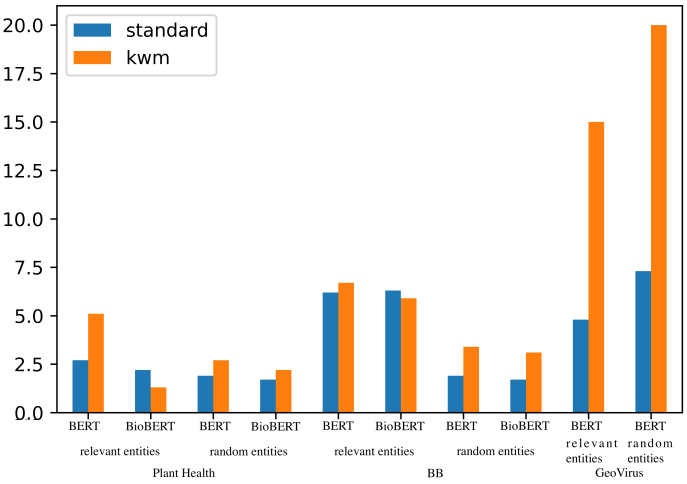


Figure 5: Perplexity comparison.



4.4 Discussion

Based on our findings, the strategy of masking the keywords is advantageous for restoring domain-specific mentions. Therefore, we assume that models fine-tuned in this way captures better the semantics of the masked words lexical

group and that it will improve the quality while being used for the NER/NEL task where the entities belong to the same lexical group. However, this masking approach seems to worsen the overall understanding of the language and seems to have no advantage unless it is specifically targeted towards a related lexical group of masked tokens.

This is reasonable, as the standard fine-tuning approach involves training the language model on the entire text data with randomly chosen tokens to mask, while the KWM strategy trains the model to focus on implicit information that is helpful in predicting specific entities. We believe our method can enhance the performance of MLMs in NER/NEL tasks, but it may reduce its ability to capture the overall structure and patterns of the input data, making it less suitable for general text processing. Further research is needed to determine whether KWM fine-tuned LMs will perform well for NER/NEL tasks, and we plan to conduct additional testing in future work.

5 Conclusion

In this work, we aimed to improve the LM understanding of domain-specific entities in the fields of biomedicine and epidemiology for it further usage. The results of our experiments show that our approach outperforms the traditional methods on restoring domain-specific entities. Our code is available on https://github.com/project178/KeyWord-Masking-strategy.

This work is a preliminary step towards developing an effective NER/NEL system for domain-specific data. Future research will focus on exploring the impact of various lexicons and corpora (diversity, coverage, etc.) on the performance of our method. We will also explore combining our approach with unsupervised or semi-supervised NER/NEL algorithms. These experiments will provide a deeper understanding of the factors that influence the effectiveness of our approach and will guide the development of a more advanced system.

5.0.1 Acknowledgements

The authors would like to express their sincere gratitude to the ANR-20-PCPA-0002, BEYOND [1] for providing the funding that made this research possible.

References

[1] BEYOND: Building epidemiological surveillance & prophylaxis with observations near & distant, homepage at https://www6.inrae.fr/beyond/. Last accessed 06 Feb 2023

[2] GeoNames, https://gd.eppo.int/. Last accessed 06 Feb 2023

[3] PESV Homepage, https://gd.eppo.int/. Last accessed 06 Feb 2023

[4] EPPO (2023), EPPO Global Database (available online), https://plateforme-esv.fr. Last accessed 06 Feb 2023

[5] Ayoola, T., Fisher, J., Pierleoni, A.: Improving entity disambiguation by reasoning over a knowledge base. arXiv preprint arXiv:2207.04106 (2022)

[6] Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., Auli, M.: Cloze-driven pretraining of self-attention networks. arXiv preprint arXiv:1903.07785 (2019)

[7] Bossy, R., Deléger, L., Chaix, E., Ba, M., Nédellec, C.: Bacteria Biotope 2019 (2022). https://doi.org/10.57745/PCQFC2, https://doi.org/10.57745/PCQFC2

[8] Chen, X., Li, L., Fei, Q., Zhang, N., Tan, C., Jiang, Y., Huang, F., Chen, H.: One model for all domains: Collaborative domain-prefix tuning for cross-domain ner. arXiv preprint arXiv:2301.10410 (2023)

[9] Derczynski, L., Llorens, H., Saquete, E.: Massively increasing TIMEX3 resources: A transduction approach. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 3754–3761. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/451_Paper.pdf

[10] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423

[11] Ferré, A., Deléger, L., Bossy, R., Zweigenbaum, P., Nédellec, C.: C-norm: a neural approach to few-shot entity normalization. BMC bioinformatics **21**(23), 1–19 (2020)

[12] Fritzler, A., Logacheva, V., Kretov, M.: Few-shot classification in named entity recognition task. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. pp. 993–1000 (2019)

[13] Gligic, L., Kormilitzin, A., Goldberg, P., Nevado-Holgado, A.: Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. Neural Networks **121**, 132–139 (2020)

[14] Gritta, M., Pilehvar, M.T., Collier, N.: Which melbourne? augmenting geocoding with maps. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1285–1296 (2018)

[15] Imambi, S., Prakash, K.B., Kanagachidambaresan, G.: Pytorch. Programming with TensorFlow: Solution for Edge Computing Applications pp. 87–104 (2021)

[16] Iovine, A., Fang, A., Fetahu, B., Rokhlenko, O., Malmasi, S.: Cyclener: an unsupervised training approach for named entity recognition. In: Proceedings of the ACM Web Conference 2022. pp. 2916–2924 (2022)

[17] Jia, C., Liang, X., Zhang, Y.: Cross-domain ner using cross-domain language modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2464–2474 (2019)

[18] Jiang, S., Cormier, S., Angarita, R., Rousseaux, F.: Improving text mining in plant health domain with gan and/or pre-trained language model. Frontiers in Artificial Intelligence **6** (2023)

[19] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[20] Li, J., Chiu, B., Feng, S., Wang, H.: Few-shot named entity recognition via meta-learning. IEEE Transactions on Knowledge and Data Engineering **34**(9), 4245–4256 (2022). https://doi.org/10.1109/TKDE.2020.3038670

[21] Li, X., Li, Z., Zhang, Z., Liu, N., Yuan, H., Zhang, W., Liu, Z., Wang, J.: Effective few-shot named entity linking by meta-learning. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE). pp. 178–191. IEEE (2022)

[22] Liu, Z., Jiang, F., Hu, Y., Shi, C., Fung, P.: Ner-bert: A pre-trained model for low-resource entity tagging. arXiv preprint arXiv:2112.00405 (2021)

[23] Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., Fung, P.: Crossner: Evaluating cross-domain named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 13452–13460 (2021)

[24] Ming, H., Yang, J., Jiang, L., Pan, Y., An, N.: Few-shot nested named entity recognition. arXiv e-prints pp. arXiv–2212 (2022)

[25] Neumann, M., King, D., Beltagy, I., Ammar, W.: Scispacy: Fast and robust models for biomedical natural language processing. BioNLP 2019 p. 319 (2019)

[26] Pergola, G., Kochkina, E., Gui, L., Liakata, M., He, Y.: Boosting low-resource biomedical qa via entity-aware masking strategies. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1977–1985 (2021)

[27] Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. arXiv preprint arXiv:1705.00108 (2017)

[28] Popovski, G., Kochev, S., Korousic-Seljak, B., Eftimov, T.: Foodie: A rule-based named-entity recognition method for food information extraction. ICPRAM **12**, pp–915 (2019)

[29] Raiman, J., Raiman, O.: Deeptype: multilingual entity linking by neural type system evolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

[30] van Rossum, G.: Python programming language, v. 3.8.15 (2022), available at https://www.python.org/downloads/release/python-3815/. Last accessed 06 Feb 2023

[31] Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050 (2003)

[32] Schoch, C., Ciufo, S., Hotton, C., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J., Sun, L., Turner, S., Karsch-Mizrachi, I.: Ncbi taxonomy: A comprehensive update on curation, resources and tools. Database **2020** (08 2020). https://doi.org/10.1093/database/baaa062, https://www.ncbi.nlm.nih.gov/taxonomy. Last accessed 06 Feb 2023

[33] Ushio, A., Camacho-Collados, J.: T-ner: an all-round python library for transformer-based named entity recognition. arXiv preprint arXiv:2209.12616 (2022)

[34] Wang, C., Sun, X., Yu, H., Zhang, W.: Entity disambiguation leveraging multi-perspective attention. IEEE Access **7**, 113963–113974 (2019)

[35] Wang, S., Li, X., Meng, Y., Zhang, T., Ouyang, R., Li, J., Wang, G.: k nn-ner: Named entity recognition with nearest neighbor search. arXiv preprint arXiv:2203.17103 (2022)

[36] Wettig, A., Gao, T., Zhong, Z., Chen, D.: Should you mask 15% in masked language modeling? arXiv preprint arXiv:2202.08005 (2022)

[37] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6

[38] Xu, J., Gan, L., Cheng, M., Wu, Q.: Unsupervised medical entity recognition and linking in chinese online medical text. Journal of healthcare engineering **2018** (2018)

[39] Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: Deep contextualized entity representations with entity-aware self-attention. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6442–6454. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.523, https://aclanthology.org/2020.emnlp-main.523

[40] Yamada, I., Washio, K., Shindo, H., Matsumoto, Y.: Global entity disambiguation with bert. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3264–3271 (2022)

[41] Zhang, S., Elhadad, N.: Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of biomedical informatics **46**(6), 1088–1098 (2013)