Data Augmentation for Layperson's Medical Entity Linking Task

Annisa Maulida Ningtyas
annisa.ningtyas@student.tuwien.ac.at
Technische Universität Wien
Vienna, Austria
Universitas Gadjah Mada
Yogyakarta, Indonesia

Florina Piroi
Research Studios Austria *and*
Artificial Researcher IT GmbH
Vienna, Austria
florina.piroi@researchstudio.at

Linda Andersson
Artificial Researcher IT GmbH
Vienna, Austria
linda.andersson@artificialresearcher.com

Allan Hanbury
Technische Universität Wien
Vienna, Austria
allan.hanbury@tuwien.ac.at

ABSTRACT

Due to the vast amount of health-related data on social media, it is beneficial to monitor health-related issues experienced by the users, such as monitoring adverse drug effects. This problem is known as the Medical Entity Linking (MEL) task, which identifies informal medical terms and normalizes them to corresponding formal medical concepts (Medical Concept Normalization, or MCN). One of the challenges of MEL when working with layperson language is the scarcity of training data and the low coverage of laypersons' health vocabulary. We apply data augmentation methods to existing data sets to increase the size of training corpora and expand the coverage of colloquial health vocabulary. We treat the informal medical phrase identification task as a sequence labelling task for Named Entity Recognition (NER). An extensive evaluation of NER shows that data augmentation may introduce noise that affects the performance of identifying informal medical phrases. Our further investigation on false positive in NER task finds that most of the false positive terms that occurred as a result of augmentation could be identified as medical entities and correctly normalized to SNOMED-CT medical concepts. In contrast, the augmentation techniques in MCN task can significantly improve performance on both the CADEC and PsyTAR data sets, particularly for paraphrase-based augmentation and combinations of augmentation techniques.

CCS CONCEPTS

• **Information systems** → **Information extraction**.

KEYWORDS

medical entity linking, data augmentation, medical concept normalization, named entity recognition

FIRE 2021, December 13–17, 2021, Virtual Conference, India

ACM ISBN 978-1-4503-9596-0/21/12…$15.00
https://doi.org/10.1145/3503162.3503172

ACM Reference Format:
Annisa Maulida Ningtyas, Florina Piroi, Linda Andersson, and Allan Hanbury. 2021. Data Augmentation for Layperson's Medical Entity Linking Task. In *Forum for Information Retrieval Evaluation (FIRE 2021), December 13–17, 2021, Virtual Conference, India.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3503162.3503172

1 INTRODUCTION

In recent years, internet users have increasingly relied on social media platforms to share medical information, discuss medical issues, and obtain medical-related knowledge. Recent studies demonstrate that social media can be beneficial in identifying adverse drug effects [21, 28], monitoring mental health during the COVID-19 pandemic [27, 36], and tracking illicit product sales, such as herbal medicines, during pandemics [17]. Identifying a formal medical concept from unstructured text is useful for various medical applications. For instance, it may assist the pharmaceutical sector in identifying adverse drug effects that are not yet reported via official channels. This is known as Medical Entity Linking (MEL) which identifies and normalizes informal medical words to the formal medical concept. The concise MEL workflow is shown in Figure 1. One of the limitations in MEL research is the scarcity of data sets for training [3].

Automatic detection of medical phrases in social media postings is challenging not least because of the lexical and grammatical diversity of terminology used by individuals that have varied backgrounds and expertise levels [18]. Additional difficulties include the widespread use of informal language, typographic and abbreviation errors, and non-standard syntax [18, 22]. New colloquial medical terms are continually emerging, which is also a consequence of the complexity of the language, in general, and of the language in online communication in particular [13]. As a result, the laypersons' health vocabularies, that is the terminologies and discourse used by laypersons, has little overlap with the vocabulary used by medical experts and medical documents.

Current publicly available data sets for MEL [3, 12, 37] contain mappings between layperson phrases and medical expert phrases. However, less than 10% of the concepts listed in the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) occur in these data sets. Furthermore, in these data sets, 60% of the expert medical phrases have less than 4 informal medical phrases mapped to them. Together these data sets contain a little bit over 8,500

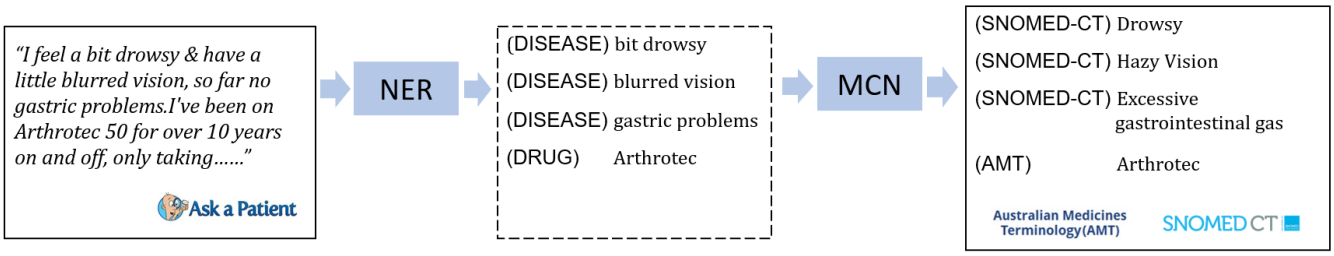

Figure 1: Generic Medical Entity Linking (MEL) Workflow

SNOMED-CT concepts. As deep learning models are data hungry, their use in the biomedical natural language processing is limited as these publicly available data sets are small. Extending such data sets to be able to train deep learning models is expensive as annotations must be done with expert help.

In this paper we present a set of data augmentation techniques for textual data that we apply to laypersons' texts in the MEL task. These methods are designed to generate more data by simulating the writing behavior of laypersons.

The first set of techniques is based on typographic and grammatical errors in their writing (e.g. keyboard errors, misspelling, misplaced words). A second set of augmentation techniques is based on semantic information used for word replacement. We replace words with synonyms and with semantically similar words that are found by means of distributional word representations (i.e. word embeddings). A third technique for textual data augmentation uses a paraphrasing engine that is based on transformer architecture to generate new data. We apply the augmentation methods to the data used for the Named Entity Recognition (NER) and the Medical Concept Normalization (MCN) tasks which are part of our MEL model, and investigate their impact on them. Our augmentation methods aim to increase the variations of laypersons' medical phrases in the MEL task, which translates to more data available for training.

2 RELATED WORK

Textual data augmentation addresses the data scarcity problem in supervised learning tasks by adding automatically labelled data to the existing data sets [11]. There are various techniques that can be used in augmenting text data with some more appropriate to the specific task. One such technique is back-translation [6, 8, 26], which augments sentences by performing back translation using transformer-based systems and the Google Translate API. Another technique is word replacement, where the standard method is synonym replacement (replaces words with synonyms extracted from WordNet) [31]. Replacing words can be done by using contextual probabilities extracted from Language Models [14], or by swapping word position in a sentence with randomly chosen words in the original sentence [31].

While these techniques are effective for classification tasks, they are less so for sequence labelling tasks, such as Named Entity Recognition (NER) [11]. Sequence labelling is more prone to data augmentation noise because of the finer granularity (token-level) task. Inspired by augmentation techniques for classification, Kang et al.[11] developed UMLS-EDA, a system that extends Easy Data Augmentation (EDA) approaches [31] by including Unified Medical Language Systems (UMLS), which better suites NER tasks in the Biomedical Domain. Dai and Adel [7], in their work to data augment for the NER task, use mention replacement, which replaces a mention with another from the corpus, of the same type, and synonym replacement. Both studies proposed text augmentation methods for clinical trials documents. Our approach focuses on text data augmentation for laypersons' Medical Entity Linking in user-generated social media postings by employing simple techniques to mimic human writing behavior on medical forums in order to enhance the variations of colloquial medical terms.

3 METHODOLOGY

The data sets we use in this work are CADEC [12], MedRed [24], and PsyTAR [37] (Section 3.1). Each data set contains posts written by laypeople on medical health forums. We explain in this section the augmentation techniques that we use to augment the data from the data sets mentioned above (Section 3.2). We train two separate models for the two tasks (NER and MCN) using training data that has been augmented with these methods. We analyze the models' effectiveness on the test data provided by the original data sets, which has been kept hidden from the model training, by looking at F1 measure scores.

3.1 Data

We demonstrate our data augmentation techniques on the available data sets for the laypersons' MEL task. We use different sets of data for the NER and MCN tasks. There are two available data sets for identifying medical NER for laypersons' language: (1) CADEC, an annotated data set on patient-reported Adverse Drug Events (ADE), which contains 1,250 annotated texts from the "Ask a Patient" forum[1] [12]. The dataset has been previously used to identify adverse drug reactions on social media, see, for example, Tutubalina et al. [29]; (2) MedRed, a medical Named Entity Recognition dataset, which contains 1,977 Reddit posts [24]. For the MCN task, we use two data sets: (1) PsyTAR, a corpus related to ADE and psychiatric medication effectiveness, which contains 887 reviews from "Ask a Patient" forum [37]; and (2) CADEC, which we use as a baseline data set not only for the NER task but also for the MCN task. CADEC contains medical terms associated with medical concepts from standard medical vocabularies, more specifically, from the SNOMED-CT and Australian Medical Terminology (AMT) [12]. Table 1 gives a few details on the datasets' content, describing

[1] https://www.askapatient.com/

the number of social posts per data set, the number of sentences, the number of threads dedicated to specific medical topics of discussions, and the number of medical entities.

Table 1: CADEC, MedRed, and PsyTAR dataset statistics

	CADEC	MedRed	PsyTAR
posts	1,250	1,977	887
sentences	7,632	8,794	6,009
communities	2 threads	18 threads	2 threads
entities	9,111	4,485	6,556

3.2 Data Augmentation for NER and MCN

We propose several data augmentation techniques for NER and MCN tasks. These techniques aim to generate more data to increase the volume of training data for our models. The augmentation techniques we employ mimic laypeople's writing behavior. We divide the techniques into *character augmentation*, *word augmentation*, and *paraphrasing*.

Character augmentation is inspired by the typographical and orthographical errors that occur during written communication [19]. Typographical spelling errors are a form of mistyping (typing a neighbouring letter incorrectly on the keyboard). Orthographical errors are forms of error in guessing or selecting the wrong word, such as *to* instead of *two* or *too*. We also add Optical Character Recognition (OCR) errors in our character augmentation. OCR is used to recognize text from an image source. When images are automatically converted into text, OCR algorithms can produce noise in recognizing characters, such as 'o' and '0'.

Our second augmentation category is word-based. Different from character augmentation, we augment the informal medical phrase(s) at word level and we use several techniques: synonym, hypernym replacement, hyponym replacement, and swap words.

Paraphrase-based augmentation, our last technique employed in this paper, aims to generate new colloquial medical terms by paraphrasing the original informal medical terms.

In our experiments we also apply a fourth augmentation technique, which we call Semantic Mention Replacement, to the NER task only. This method was inspired by the *mention replacement* proposed by Dei and Adel [7]. Table 2 describes each of the augmentation used in this work, by NER and MCN tasks for MEL.

Table 3 illustrates the augmentation approaches used in a NER task. These techniques are equally applicable in MCN tasks that focus solely on the phrase. We notice from the examples in Table 3 that augmentation may lead to different length of the informal medical phrases of interest: (*pain in my foot*) becomes shorter (*my foot aches*) or longer (*pain sense in my foot*). This affects the label sequence of the original sentence for NER task (see the next subsection for a short description of the BIO sequence labelling we use in our work).

3.3 Model Training

We train one model for each of the NER and MCN tasks using deep learning and contextual embedding.

NER Task. We treat entity extraction as a sequence labelling task, where the sequence (one or more words) corresponds to a specific entity type. We use the *BIO* tagged format with types, where *B* marks the beginning word of the sequence, *I* marks words that are not the beginning of the sequence (i.e. are inside the sequence), while *O* marks words not part of the sequence of interest (outside). Each word in the NER dataset which has either a *B* or *I* tag is additionally tagged with either *DIS* or *DRUG* types, where *DIS* marks a disease/symptom and *DRUG* marks a drug. For example, the phrase *bit drowsy* is labeled *B-DIS I-DIS*, and the other words that do not represent the medical terms are labeled as *O*.

For training the NER model we use a BiLSTM-CRF based architecture which takes as input the concatenated GloVe [23] and RoBERTa [15] word embeddings, which translates to combining the classical word embeddings (i.e. GloVe) with the contextual embeddings (i.e. RoBERTa). This combination has been shown to achieve a good performance in the NER task [2]. The BiLSTM-CRF architecture design was proposed by Huang et al. [10], who have shown in their experiments to be effective at handling bidirectional sequence data and long-tail sequence tasks. The output of the NER model is the input sequence labeled with the *BIO* tags.

MCN Task. We handle this task as a multi-class classification task. The idea is to normalize the informal medical phrases into formal medical concepts, represented by SNOMED-CT codes as class labels. Similar to the NER task, we use the concatenated GloVe and RoBERTa word embeddings. We use, then, Gated Recurrent Units (GRU) [5] to learn the SNOMED-CT class labels for medical phrases. We chose GRU over LSTM to better cater to the short input of informal medical entities.

When the MCN module is part of a pipeline, like the one presented in Figure 1, the MCN input can be chosen to be the output of the NER module[2]. Specifically in this work, however, we use the labeled data in CADEC and PsyTAR for the MCN training. The output of the trained model is the input informal medical phrases classified accordingly to the SNOMED-CT.

4 EXPERIMENT SETUP

We designed several experiments to evaluate the effect of the proposed data augmentation methods on the MEL task by comparing the baseline models, trained with the non-augmented data, with models that use the augmented data sets in the training phase, particularly for medical NER and MCN modules. We took the original fold of training, validation, and testing sets provided by the authors of CADEC [12], MedRed [24], and PsyTAR [37]. We only augment the training sets, leaving out the validation sets for evaluating the model training and the testing sets for demonstrating our model's performance. We used CADEC and MedRed datasets to perform medical NER tasks, and the CADEC and PsyTAR datasets for the MCN tasks.

We analyse the impact data augmentation has on the NER task by augmenting different amounts of the baseline training data. For this, we randomly sampled 20%, 50%, and 100% of training data as the source of augmentation, applying all the augmentation techniques described in the previous sections (see also Table 2), except for the Swap Words technique which leads to grammatical errors and injects noise into the sequence labelling task.

[2]This processing is part of our future work.

Table 2: Detailed descriptions of the augmentation methods employed in this work, per MCN and NER tasks. The first row for each technique describes the concrete augmentation operation for the MCN task, while the second row describes additional rules that must be considered in order to maintain the label sequence necessary for the NER task.

Technique (Category)	Detail
Keyboard Errors (Character)	MCN: Replace one character at random, simulating a common typing error caused by the character's position on the keyboard, which is typically very close to the one replacing the original character.
	NER: Similar replacement as above, only targets mentions without changing the label sequence.
OCR Errors (Character)	MCN: Replace a character at random, simulating an OCR error. The replacement is done using an existing list of common OCR mistakes [16].
	NER: Replacement similar to the MCN task, only targeting mentions without changing the label sequence.
Misspelling Errors (Character)	MCN: Replace a character at random, simulating a misspelling error, according to a pre-defined list of errors in the misspelling corpus [20].
	NER: Replacement similar to the MCN task, only targeting mentions without changing the label sequence.
WordNet Synonym Replacement (SR) (Word)	MCN: Randomly replace word(s) in with their synonym extracted from WordNet
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the BIO label (Section 3.3) using the length of the new phrase.
CHV-Drug SR (Word)	MCN: Randomly replace word(s) in the phrase with synonyms from the Consumer Health Vocabulary (CHV) [30, 34] and Drug Bank [32]. If there are multiple words in the input sentence that are a match for synonyms, we replace the longer phrase. The matching is done by n-gram matching.
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the BIO label (Section 3.3) using the length of the new phrase.
Hypernym Replacement (Word)	MCN: Randomly replace phrase word(s) with hypernyms fetched from WordNet.
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the BIO label (Section 3.3) using the length of the new phrase.
Hypnonym Replacement (Word)	MCN: Randomly replace phrase word(s) with hyponyms fetched from WordNet.
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the BIO label (Section 3.3) using the length of the new phrase.
Swap Word (Word)	MCN: Choose two words at random from the input phrase and swap their positions.
	NER: Replacement is similar to the MCN task, only targeting mentions.
Semantic Mention Replacement (Semantic MR)	MCN: not applicable.
	NER: Replace the mention with another of the same type. The replacement is determined by calculating the mention's highest similarity score to similar entities in the corpus. To perform semantic textual similarity, we use the Sentence-BERT for sentence representation.
Paraphrase	MCN: Generate a new phrase based on the input one using a paraphrasing engine based on a T5 model trained on the Google PAWS Dataset.
	NER: Replacement similar to the MCN task, only targeting mentions. We reconstruct the BIO label (Section 3.3) using the length of the new phrase.

Table 3: Augmentation example. The bold face words mark the input sequence, the italic, blue-colored words indicate the augmentation based changes to the input sentence.

Augmentation method	Instance
None (input sentence)	... I find that the **pain** **in** **my** **foot** is subsiding and i can walk a lot better .
Keyboard Errors	... I find that the **pain** ***im*** **my** **foot** is subsiding and i can walk a lot better .
OCR Errors	... I find that the **pain** **in** **my** ***fo0t*** is subsiding and i can walk a lot better .
Misspelling Errors	... I find that the ***psin*** **in** **my** **foot** is subsiding and i can walk a lot better .
Semantic MR	... I find that the ***severe*** ***foot*** ***pain*** is subsiding and i can walk a lot better .
Paraphrase	... I find that the ***my*** ***foot*** ***aches*** is subsiding and i can walk a lot better .
WordNet SR	... I find that the ***hurting*** ***in*** ***my*** ***foot*** is subsiding and i can walk a lot better .
CHV-Drug SR	... I find that the ***pain*** ***sense*** ***in*** ***my*** ***foot*** is subsiding and i can walk a lot better .
Hypernym Replacement	... I find that the pain in my ***walk*** is subsiding and i can walk a lot better .
Hypnonym Replacement	... I find that the pain in my ***flatfoot*** is subsiding and i can walk a lot better .
Swap Word	... I find that the ***in*** ***pain*** my foot is subsiding and i can walk a lot better .

For each applied technique in the character and word augmentation categories, we augment an informal medical term only once. For semantic mention and paraphrase augmentation approaches, we augment a term three times to have a balance of the various annotation categories in the training data (e.g. three character based annotation techniques vs. one Semantic Mention Replacement technique). The combination of original and augmented data is used as training data for the NER model. We repeat the training process five times with different random seeds, in order to see patterns in our model's performance.

While training the models to handle the MCN task, we augmented 100% of the original training data (adding it to the original training data), because we found that the datasets have 60% of the formal medical terminology represented by less than four informal medical phrases. We used similar augmentation techniques as in the NER case to augment the MCN training datasets, except the mention replacement. We also include the Swap Words augmentation method to the experiments. We repeat the training process five times with different random seeds.

We use the micro F1-score to evaluate the NER and MCN models. We calculate the statistical significance using the *paired t-test* with $p < 0.05$. Means and standard deviations are reported on each evaluation result.

Furthermore, we conduct false positive analysis terms predicted by NER module, to identify whether the terms can potentially be either a DIS (disease/symptom) or DRUG entity. To ensure the

reliability of our analysis, we perform manual assessment by the medical experts. Moreover, to support the analysis, we provide the false positive terms with their formal medical concept predicted by our MCN module. The false positive analysis resembles the end-to-end MEL workflow as described in Figure 1. However, we did not conduct an evaluation for the end-to-end workflow in the whole data sets, as this is our plan for future work.

The implementation of character augmentation uses the `nlpaug` library, [16], and word augmentation uses the `niacin` library [20], except for the CHV-DRUG Synonym Replacement and Semantic mention replacement. We used the Flair framework [1] to train our models.

4.1 Evaluation Metrics

We evaluate our model performance using the F1-Measure.

Computing F1 for the NER Task. When assessing the models trained for the NER task, we use two methods ouf of the four described in the literature [25] to decide on the relevancy of the results[3]:

- Strict evaluation (exact-span and type matching)
- Partial span matching (regardless to the type)

Precision and Recall for strict and partial span matching are computed using the MUC scoring system, created for evaluating Information Extraction systems [4]. The MUC scoring categories, adapted for the SemEVAL tasks [25] and which are the ones we use in our evaluation, can be described as comparing a system's output to the ground truth. We give below an explanation of each of the MUC categories:

- Correct (COR): The output of the system and the ground truth are in agreement;
- Incorrect (INC): the output of a system does not match the ground truth;
- Partial (PAR): the system and the ground truth are "similar" but not identical;
- Missing (MIS): A ground truth is not captured by a system.
- Spurious (SPU): the system generates an output that isn't present in the ground truth;

To compute Precision and Recall we need to define the True Positives (TP), False Negatives (FN), and False Positives (FP) quantities. We define them as sums of MUC categories as describe in the Evaluation of SemEval-2013 Task 9.1 [4]:

$$TP_{strict} = COR \tag{1}$$

$$TP_{partial} = COR + w * PAR \tag{2}$$

$$ACT = COR + INC + PAR + SPU \tag{3}$$

$$POS = COR + INC + PAR + MIS \tag{4}$$

where, TP_{strict} is the TP for strict evaluation, $TP_{partial}$ is the TP for partial evaluation, and w is the weight of the PAR score, which we set to 0.5. Meanwhile, ACT is the set of annotations returned by the model, and POS is the set of the ground truth annotations.

Using these definitions we can compute the Precision and Recall for the strict and partial evaluations:

$$\begin{aligned} Precision_{strict} &= \frac{COR}{ACT} \\ Recall_{strict} &= \frac{COR}{POS} \end{aligned} \tag{5}$$

$$\begin{aligned} Precision_{partial} &= \frac{COR + 0.5 * PAR}{ACT} \\ Recall_{partial} &= \frac{COR + 0.5 * PAR}{POS} \end{aligned} \tag{6}$$

Finally, the F1 scores for the partial and strict evaluations are defined as the harmonic mean for the respective precision and recall quantities.

Computing F1 for the MCN Task. We evaluate the MCN model performance by using standard Precision, Recall and F1-Measure. Since MCN is a multi classification task, we only report the micro-averaged F1 scores. Micro-averaged F1-scores calculated by counting overall TP, FN, and FP from all of the classes.

5 RESULTS AND DISCUSSIONS

Data Augmentation for NER task. The experiments aim to evaluate the effect of data augmentation on the output of the NER model in our MEL workflow (Figure 1). Table 4 shows the evaluation result for the various NER models trained on the different datasets we created with the augmentation techniques given in Section 3.2).

From the experiments results, we conclude that, compared to the baseline, the models for all augmentation approaches underperformed on both the CADEC and MedRed datasets. Moreover, the partial evaluation performs better than the strict evaluation. This is is to be expected as the partial evaluation is more relaxed than the strict evaluation, allowing for more TP cases. Among the proposed techniques, the model that performs best is the one that uses word augmentation to increase the amount of training data by 20%. According to our experiment, adding more data does not guarantee that the model's performance will improve. In this respect, we plan for further analysis on the quality of augmentation results.

As discussed in [7], data augmentation may alter the ground truth label or create erroneous occurrences. This adverse effect may be particularly pronounced in large training sets. Additionally, previous research has shown that data augmentation approaches are beneficial in low-resource settings with a small training set, which is not the goal of our experiment. We presume that the nature of the informal medical phrases may introduce more noise which confuses the model [9]. We further analyze the implication in the failure analysis in Section 5.1.

Data Augmentation for MCN tasks. Table 5 summarizes each trained model's performance on the baseline and enhanced corpora in terms of micro F1-score, for the MCN task. Each data augmentation techniques was applied once to the entire training corpus. The experiments were repeated five times, and the statistical significance was determined using the paired student t-test ($p < 0.05$).

In comparison to NER tasks, augmentation techniques applied to data for training MCN models outperform our baseline model trained on original data. The paraphrasing technique outperforms other augmentation on PsyTar dataset. Our paraphrasing engine

[3]The relevancy is used, then, to compute Precision and Recall metrics, which are part of the F1 score.
[4]https://www.cs.york.ac.uk/semeval-2013/task9/

Table 4: NER Performance for informal medical entity recognition on MedRed and CADEC datasets

Model	MedRed						CADEC					
	Strict			Partial			Strict			Partial		
	Ori + 20%	Ori + 50%	Ori + 100%	Ori + 20%	Ori + 50%	Ori + 100%	Ori + 20%	Ori + 50%	Ori + 100%	Ori + 20%	Ori + 50%	Ori + 100%
Ori (baseline)	**69.6 ± 1.1**			**74.2 ± 0.9**			**81.2±0.1**			**85.9±0.1**		
+Char	69.5 ± 1.1	69.5 ± 1.2	68.9 ± 0.7	74.1 ± 1.0	73.5 ± 1.0	72.6 ± 0.8	78.7±0.8	78.5 ± 0.5	79.0 ± 0.7	83.4 ± 0.9	83.4 ± 0.8	83.5 ± 0.9
+Word	69.5 ± 0.6	68.5 ± 0.2	68.8 ± 0.7	74.2 ± 0.6	72.9 ± 0.3	73.5 ± 0.7	79.6 ± 0.4	78.4 ± 0.5	78.7 ± 0.9	84.0 ± 0.5	83.2 ± 0.6	83.3 ± 0.8
+WordChar	68.8 ± 0.6	68.3 ± 0.6	67.8 ± 0.7	73.8 ± 0.6	72.8 ± 0.8	72.1 ± 0.8	78.3 ± 0.5	77.6 ± 0.3	77.5 ± 0.3	83.4 ± 0.3	82.6 ± 0.4	82.6 ± 0.4
+Paraphrase	67.3 ± 1.6	65.0 ± 1.5	62.3 ± 1.7	72.3 ± 1.4	69.9 ± 1.6	66.9 ± 1.7	78.6 ± 0.6	78.1 ± 0.5	77.8 ± 0.6	83.5 ± 0.6	83.0 ± 0.4	82.9 ± 0.6
+MR_Sem	68.5 ± 0.4	68.1 ± 0.7	69.0 ± 0.3	73.1 ± 0.7	73.1 ± 0.8	73.8 ± 0.2	78.9 ± 0.2	79.4 ± 0.3	78.3 ± 1.1	83.9 ± 0.1	84.1 ± 0.2	83.2 ± 1.3
+Combination	67.3 ± 1.2	67.3 ± 0.8	68. 2 ± 0.8	73.1 ± 0.9	72.6 ± 0.7	73.0± 0.6	78.9 ± 0.1	78.7 ± 0.5	78.9 ± 0.3	83.8 ± 0.2	83.7 ± 0.4	83.9 ± 0.3

Table 5: MCN performance on CADEC and PsyTAR.

Method	CADEC	PsyTAR
None	72.5 ± 4.9	79.6 ± 0.2
Character	74.5 ± 5.4	79.0 ± 0.04
Word	76.8 ± 4.6	80.2 ± 0.5
Paraphrase	77.1 ± 4.2	**81.4 ± 0.5**
WordChar	76.9 ± 5.5	80.3 ± 0.1
Combination	**78.8 ± 3.3**	81.2 ± 0.4

is based on a *Text-to-Text Transfer Transformer* fine-tuned on the Google PAWS data set [33, 35]. The engine is used to paraphrase the informal medical terms, for example, *"hard to get out of bed in the morning"* augmented to be *"getting from bed in the morning is hard"*. According to our findings, the paraphrasing engine could preserve the semantics of informal medical terms while producing new semantically similar phrases to the original terms. Figure 2 shows the cosine similarity between the input and augmented phrases.

Figure 2: Pair wise cosine similarity between original (first column & row) and augmented phrases from paraphrase method

We find that word replacement introduces new vocabulary to the corpus, which allows the model to handle the out-of-vocabulary problems on the test set. This finding is aligned with Wei and Zou's work [31]. Additionally, in CADEC experiments, combining all augmentation techniques (Combination) outperforms other techniques. This is a result of the increased variety of informal medical terms introduced in the training data. This diverse training data may help prevent over-fitting of the models. Meanwhile, character augmentation performed poorly in comparison to other techniques. On PsyTAR data, character-based augmentation methods cannot be used to improve the baseline model.

5.1 Failure Analysis

This section aims to discuss and analyze the false positives that only occurred on our NER augmentation model and not on the NER base model. We conducted a manual assessment done by two medical experts focused on inspecting the false-positive terms. The assessment aimed to identify whether the detected text spans can be either a disease/symptom (DIS) or a drug (DRUG) entity.

We focused on the CADEC dataset and captured 452 distinct spans of false-positive informal medical terms extracted from the combination of all augmentation techniques. Among these 452 terms, both experts agree that 407 are medical terms, and 10 are not medical terms. The Cohens kappa coefficient for this task is 0.3 (fair agreement). Based on the expert assessment, The terms that are not disease/symptom are actually related to medical procedures, body parts, or medical tests, whereas the terms that are not drug are nutritional or dietary supplements.

Furthermore, in order to validate the 407 terms identified as medical terms, we used the MCN model to cross-check them with the formal medical concept. Among the 407 terms, 343 of the formal medical concepts are correctly identified. For instance, medical terms *muscle ached* and *unable to lose weight* can be normalized to SNOMED-CT medical concept of *myalgia* and *weight gain*. This observation demonstrates that the false positive for informal medical phrases could lead to correct formal medical phrases. Based on this assessment, we argue that 343 of 452 (75%) false-positive terms that only occurred on our augmentation model could be considered a medical entities.

6 CONCLUSION

We described in this paper the data augmentation applied by us to the Medical Entity Linking in informal medical phrases, which consists of NER and MCN tasks. The evaluation results show that for NER tasks, augmentation approaches could not outperform the baseline model on CADEC and MedRed datasets. Word augmentation technique outperforms compare to other techniques. Based on the NER false-positive analysis, we discovered that 343 of the

452 false-positive terms caused by augmentation could be classified as a medical entity and correctly normalized to the SNOMED-CT medical concept.

Meanwhile, the augmentation techniques in MCN can significantly improve performance on both the CADEC and PsyTAR data sets, particularly for paraphrases and combinations of techniques. Compared to the baseline model performance, the models trained on data augmented by paraphrasing by a combination of augmentation techniques achieve the best performance. Models trained on data augmented by character-based methods performed poorly when compared to our models trained on data augmented by other techniques.

Data augmentation methods for the training of models for the NER task remains a challenge which we will address in future work. In this respect We aim to explore paraphrasing techniques in more detail, since it achieve a best performance for the (augmented) PsyTar data set and second best for the (augmented) CADEC data set in the MCN task. From our observations to this date, the paraphrasing-based technique produces a promising variations of laypersons medical phrases. Additionally, we plan to expand the variations of laypersons' medical phrases with distant supervision techniques.

ACKNOWLEDGMENTS

The present work is part of PhD research funded by Indonesia-Austria Scholarship Programme (IASP), a joint scholarship between the Indonesian Ministry of Education and Culture (KEMDIKBUD) and Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH). Reference number: B/1863/D3/KD.02.00/2019. We also acknowledge the contributions of Ryza Yasha and Ranny Fitriani Herwati for their expertise on analysing false positive data.

REFERENCES

[1] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 54–59. https://doi.org/10.18653/v1/N19-4010

[2] A. Akbik, D. Blythe, and R. Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*. 1638–1649.

[3] M. Basaldella, F. Liu, E. Shareghi, and N. Collier. 2020. COMETA: A Corpus for Medical Entity Linking in the Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3122–3137. https://doi.org/10.18653/v1/2020.emnlp-main.253

[4] N. Chinchor and B. Sundheim. 1993. MUC-5 Evaluation Metrics. In *Proceedings of the 5th Conference on Message Understanding* (Baltimore, Maryland) *(MUC5 '93)*. Association for Computational Linguistics, USA, 69–78. https://doi.org/10.3115/1072017.1072026

[5] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[6] J. Corbeil and H. A. Ghadivel. 2020. BET: A Backtranslation Approach for Easy Data Augmentation in Transformer-based Paraphrase Identification Context. *CoRR* abs/2009.12452 (2020). arXiv:2009.12452 https://arxiv.org/abs/2009.12452

[7] X. Dai and H. Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 3861–3867. https://doi.org/10.18653/v1/2020.coling-main.343

[8] M. Fadaee, A. Bisazza, and C. Monz. 2017. Data Augmentation for Low-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 567–573. https://doi.org/10.18653/v1/P17-2090

[9] J. Gu and Z. Yu. 2020. Data Annealing for Informal Language Understanding Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3153–3159. https://doi.org/10.18653/v1/2020.findings-emnlp.282

[10] Z. Huang, W. Xu, and K. Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991 [cs.CL]

[11] T. Kang, A. Perotte, Y. Tang, C. Ta, and C. Weng. 2020. UMLS-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association* 28, 4 (12 2020), 812–823. https://doi.org/10.1093/jamia/ocaa309 arXiv:https://academic.oup.com/jamia/article-pdf/28/4/812/36642183/ocaa309_supplementary_data.pdf

[12] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics* 55 (2015), 73 – 81. https://doi.org/10.1016/j.jbi.2015.03.010

[13] D. Kershaw, M. Rowe, and P. Stacey. 2016. Towards Modelling Language Innovation Acceptance in Online Social Networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (San Francisco, California, USA) *(WSDM '16)*. Association for Computing Machinery, New York, NY, USA, 553–562. https://doi.org/10.1145/2835776.2835784

[14] S. Kobayashi. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 452–457. https://doi.org/10.18653/v1/N18-2072

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[16] E. Ma. 2019. NLP Augmentation. https://github.com/makcedward/nlpaug.

[17] T. K. Mackey, J. Li, V. Purushothaman, M. Nali, N. Shah, C. Bardier, M. Cai, and B. Liang. 2020. Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales: Infoveillance Study on Twitter and Instagram. *JMIR Public Health Surveill* 6, 3 (25 Aug 2020), e20794. https://doi.org/10.2196/20794

[18] S. Mesbah, J. Yang, R.-J. Sips, M. Valle Torre, C. Lofi, A. Bozzon, and G.-J. Houben. 2019. Training Data Augmentation for Detecting Adverse Drug Reactions in User-Generated Content. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2349–2359. https://doi.org/10.18653/v1/D19-1239

[19] K. Min, W. H. Wilson, and Y.-J. Moon. 2000. Typographical and Orthographical Spelling Error Correction. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA), Athens, Greece. http://www.lrec-conf.org/proceedings/lrec2000/pdf/221.pdf

[20] D. Niederhut. 2020. niacin: A Python package for text data enrichment. *The Journal of Open Source Software* 5, 50, Article 2136 (June 2020), 2136 pages. https://doi.org/10.21105/joss.02136

[21] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. 2015. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association* 22, 3 (2015), 671–681. https://doi.org/10.1093/jamia/ocu041

[22] N. Pattisapu, V. Anand, S. Patil, G. Palshikar, and V. Varma. 2020. Distant supervision for medical concept normalization. *Journal of Biomedical Informatics* 109 (2020), 103522. https://doi.org/10.1016/j.jbi.2020.103522

[23] J. Pennington, R. Socher, and C. D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[24] S. Scepanovic, E. Martin-Lopez, D. Quercia, and K. Baykaner. 2020. Extracting Medical Entities from Social Media. In *Proceedings of the ACM Conference on Health, Inference, and Learning* (Toronto, Ontario, Canada) *(CHIL '20)*. Association for Computing Machinery, New York, NY, USA, 170–181. https://doi.org/10.1145/3368555.3384467

[25] I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA, 341–350. https://aclanthology.org/S13-2056

[26] R. Sennrich, B. Haddow, and A. Birch. 2015. Improving Neural Machine Translation Models with Monolingual Data. *CoRR* abs/1511.06709 (2015). arXiv:1511.06709 http://arxiv.org/abs/1511.06709

[27] T. Tabak and M. Purver. 2020. Temporal Mental Health Dynamics on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.7

[28] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, and V. Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics* 84, June (2018), 93–102. https://doi.org/10.1016/j.jbi.2018.06.006

[29] E. Tutubalina and S. Nikolenko. 2017. Combination of Deep Recurrent Neural Networks and Conditional Random Fields for Extracting Adverse Drug Reactions from User Reviews. *Journal of healthcare engineering* 2017 (2017), 9451342. https://doi.org/10.1155/2017/9451342

[30] V. Vydiswaran, Q. Mei, D. Hanauer, and K. Zheng. 2014. Mining Consumer Health Vocabulary from Community-Generated Text. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2014 (2014), 1150–9.

[31] J. Wei and K. Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6382–6388. https://doi.org/10.18653/v1/D19-1670

[32] D. Wishart, Y. D. Feunang, A. Guo, E. J. Lo, A. Marcu, J. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* 46 (2018), D1074 – D1082.

[33] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

[34] Q. T. Zeng and T. Tse. 2006. Exploring and Developing Consumer Health Vocabularies. *Journal of the American Medical Informatics Association* 13, 1 (2006), 24–29. https://doi.org/10.1197/jamia.M1761

[35] Y. Zhang, J. Baldridge, and L. He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

[36] J. Zhou, H. Zogan, S. Yang, S. Jameel, G. Xu, and F. Chen. 2021. Detecting Community Depression Dynamics Due to COVID-19 Pandemic in Australia. *IEEE Transactions on Computational Social Systems* (2021), 1–10. https://doi.org/10.1109/TCSS.2020.3047604

[37] M. Zolnoori, K. W. Fung, T. B. Patrick, P. Fontelo, H. Kharrazi, A. Faiola, N. D. Shah, Y. S. Shirley Wu, C. E. Eldredge, J. Luo, M. Conway, J. Zhu, S. K. Park, K. Xu, and H. Moayyed. 2019. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief* 24 (2019), 103838. https://doi.org/10.1016/j.dib.2019.103838