Data Augmentation for Layperson’s Medical Entity Linking Task

@ACL 2021

Problem Description:

Entity linking is the task of identifying mentions in documents and mapping them to their correct concept name in knowledge base.Knowledge Base is a huge collection(4.3 B) words in Biomedical Domain.

There are certain issues that follow up while finding an efficient mapping technique and that is resolved in entity linking tasks.

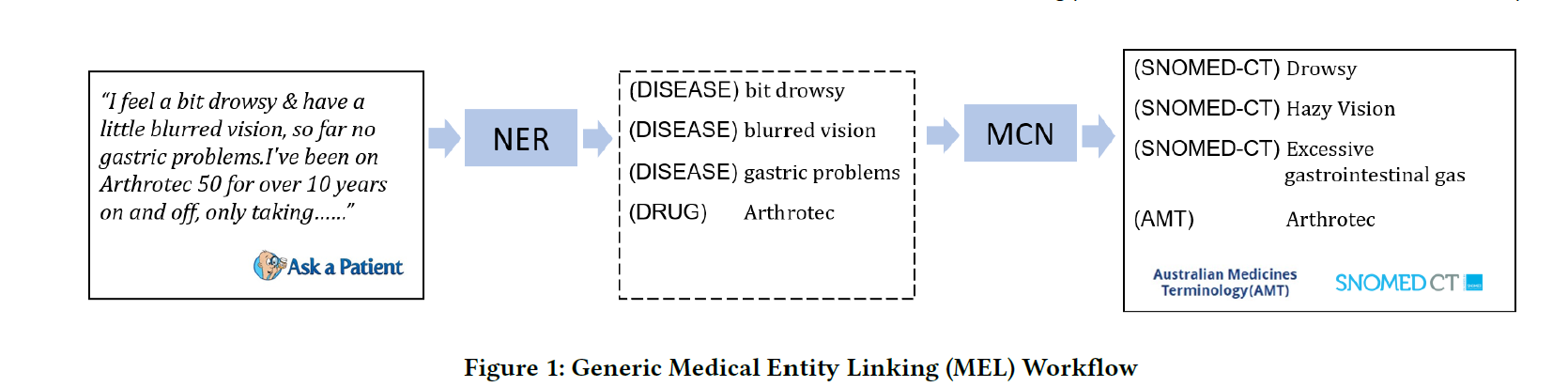
Proposed idea:

We apply data augmentation methods to existing data sets to increase the size of training corpora and expand the coverage of colloquial health vocabulary .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 tasks can significantly improve performance. 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.

One of the limitations in MEL research is the scarcity of data sets for training.

Current publicly available data sets for MEL contain mappings between layperson phrases and medical expert phrases.



Model :

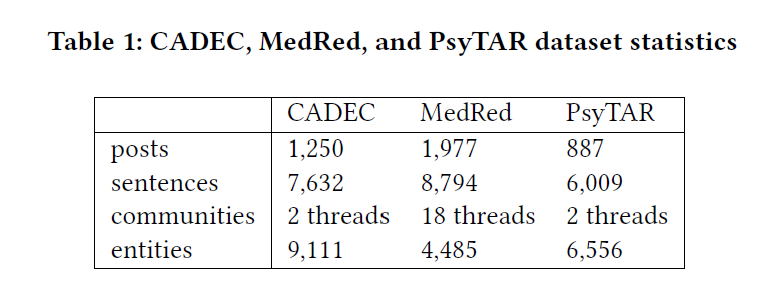
Different Techniques classification are:

Technique 1- correcting typographic & grammatical errors in layperson vocab.

Technique 2 - Semantic information based used for word replacement.by means of distributional word representations (i.e. word embeddings)

Technique 3 : a paraphrasing engine that is based on transformer architecture to generate new data

There are other augmentation techniques available but using layperson’s vocab expands training data significantly.



Augmentation technique can be divided into 3 categories namely character , word and paraphrase augmentation.Paraphrase augmentation has been deployed in these paper. 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.

we use a BiLSTM-CRF based architecture which takes as input the concatenated GloVe and RoBERTa word embeddings, which translates to combining the classical word embeddings (i.e. GloVe) with the contextual embeddings(i.e. RoBERTa). Effective in handling bidirectional sequence data and long-tail sequence tasks.

Training →

We handle this task as a multi-class classification MCN Task. task. The idea is to normalize the informal medical phrases into formal medical concepts, represented by SNOMED-CT codes as class labels.

The data sets we use in this work are CADEC , MedRed, and PsyTAR

Cadec is an annotated data set on patient-reported Adverse Drug Events (ADE), which contains 1,250 annotated texts

MedRed is a medical Named Entity Recognition dataset, which contains 1,977 Reddit posts [24]. For the

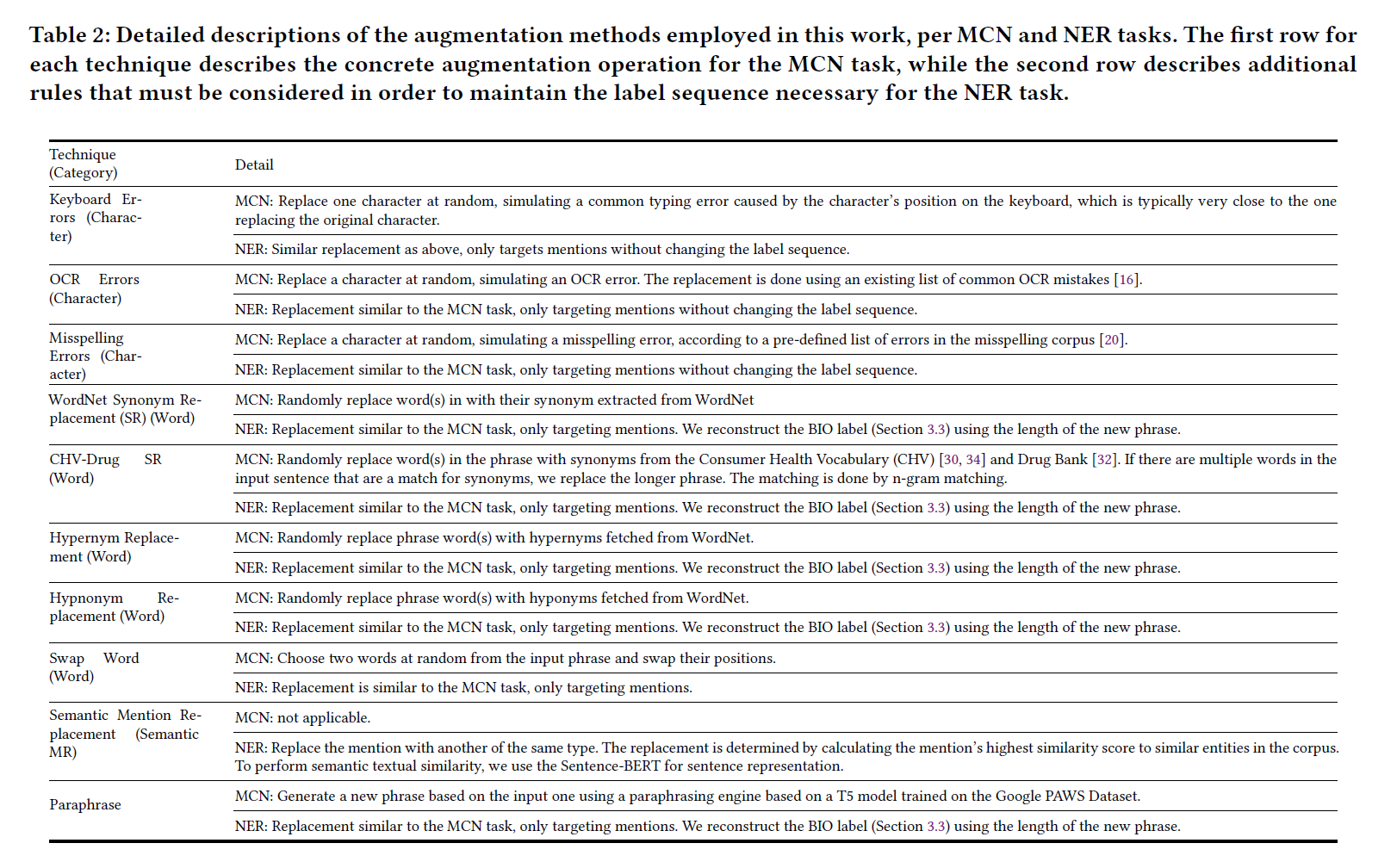
For MCN task, we use two data sets: (1) PsyTAR, a corpus related to ADE and

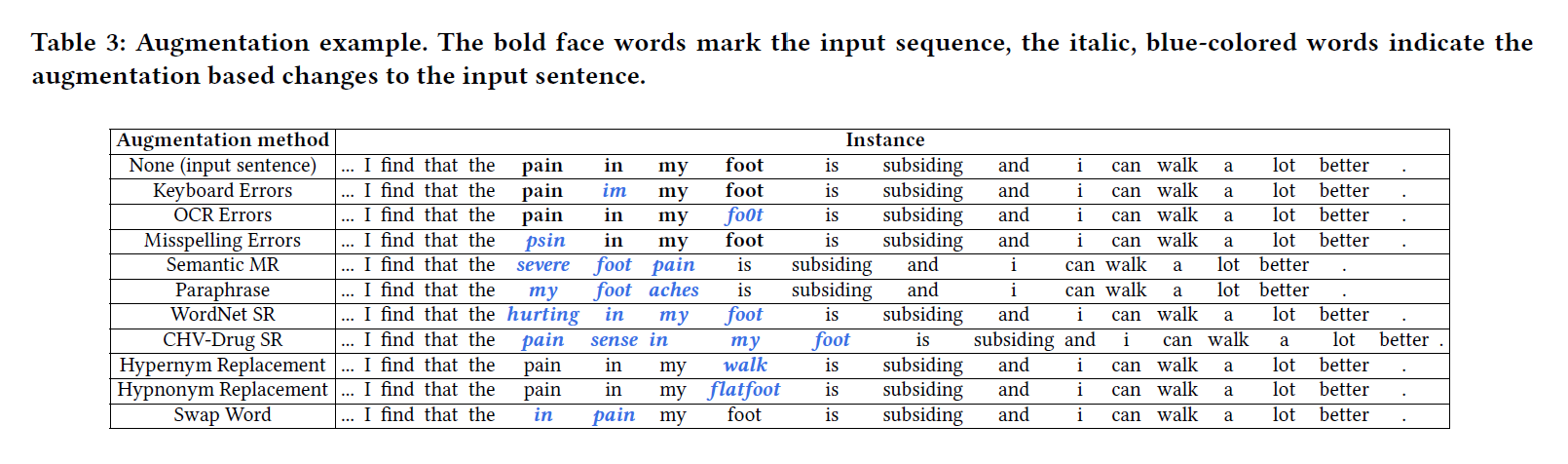
psychiatric medication effectiveness, which contains 887 reviews.

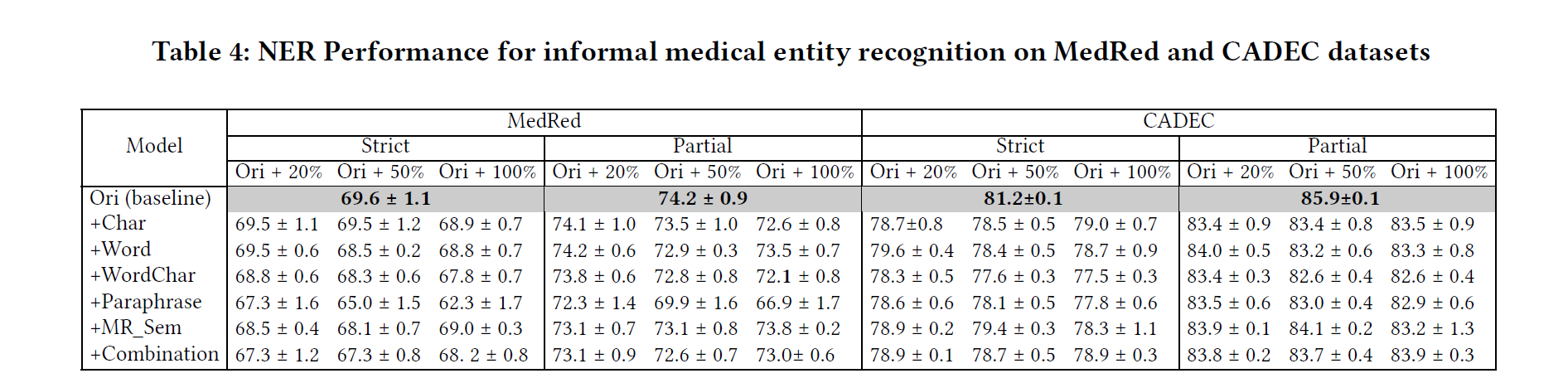
mention replacement

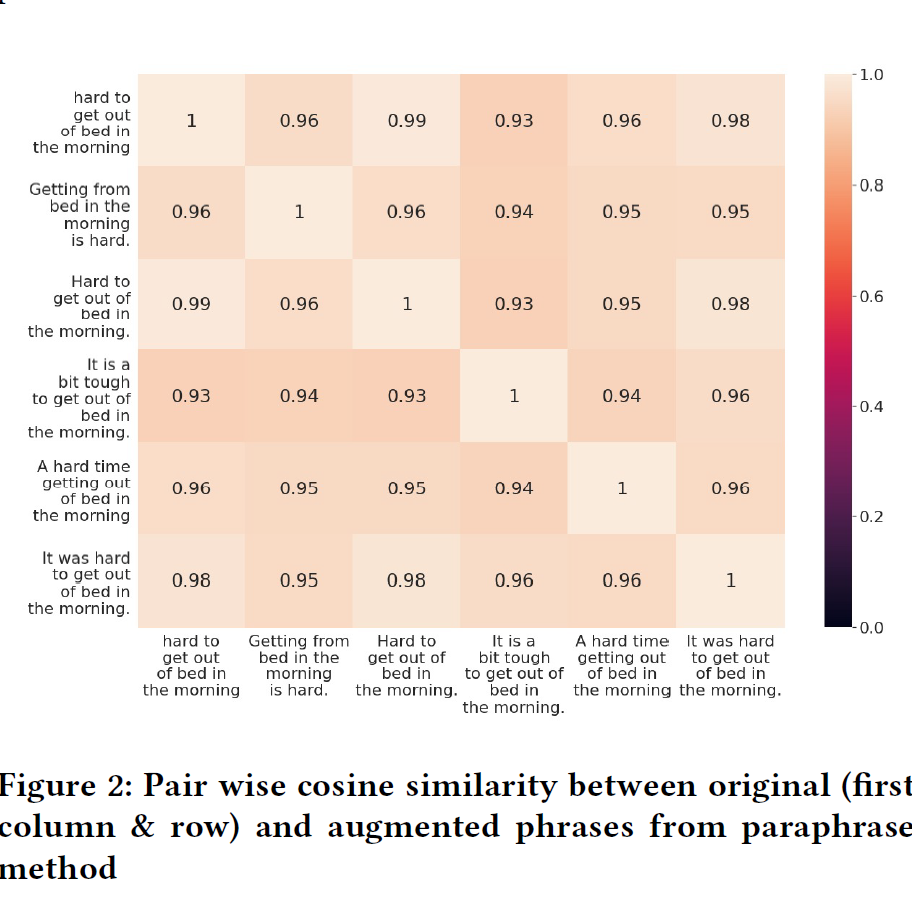
Comparison / Results →

The implementation of character augmentation uses the nlpaug library, and word augmentation uses the library ,niacin except for the CHV-DRUG Synonym Replacement and Semantic.









Bottlenecks:

Methodology for how augmentation actually work is not provided in the paper