A Simple Data Augmentation Method to Improve the Performance of Named Entity Recognition Models in Medical Domain

Abdul Majeed Issifu, Murat Can Ganiz
Department of Computer Engineering, Marmara University
Istanbul/Turkey
abdul.majeed@marun.edu.tr
murat.ganiz@marmara.edu.tr

***Abstract*—Easy Data Augmentation is originally developed for text classification tasks. It consists of four basic methods: Synonym Replacement, Random Insertion, Random Deletion, and Random Swap. They yield accuracy improvements on several deep neural network models. In this study we apply these methods to a new domain. We augment Named Entity Recognition datasets from medical domain. Although the augmentation task is much more difficult due to the nature of named entities which consist of word or word groups in the sentences, we show that we can improve the named entity recognition performance.**

Keywords - Data Augmentation, Medical Data, NER, BERT, BioBERT

I. Introduction

One of the important tasks of Natural Language Processing (NLP) is Named Entity Recognition (NER). NER is a form of information extraction where named entities are extracted from raw text documents. Named entities here represents pre-defined objects, words or word groups that are labeled such as person, location, date, organisation etc. Machine learning and deep learning models performs adequately on various NLP tasks and have so far gotten high accuracy. Sentiment analysis (XLNet [1]), Text classification (ULMFiT [2]), Named entity recognition (LUKE [3]), Question answering (Gated-Attention Reader[4]), Summarizing (RNES [5]) and many more have shown important advancement in the field of NLP. Many of these models are supervised and thus largely depends on labeled data, which is scarce and costly to obtain. The use of semi-supervised learning, transfer learning and data augmentation are the current remedies to this problem. The data augmentation is of course a relatively simpler solution and has its advantages such as allowing to use any kind of supervised algorithm on the top of the augmented data.

Data augmentation is heavily studied in computer vision domain and resulted fairly advanced methods, such as Auto-Augment (AutoAugment [6]: Learning Augmentation Policies from Data), Random Erasing Data Augmentation [7], Albumentations [8]. Data augmentation studies in NLP is relatively new, we have much fewer methods in comparison. Google researchers [9] propose augmentation techniques for text classification where back translation, and TF-IDF were used. Many if not all, the text augmentation in NLP are for text classification and do not apply to NER.

NER on the other hand is a crucial topic that needs much attention especially in the medical domain.

Medical data contains valuable information in the form of narratives or hospital/clinical discharge summaries. There is a large amount of research on how to use these vital information of patience in the hospital to develop artificial intelligent systems and to improve the health care of individuals. Since the data is both crucial and constrained for research, getting them is not easy and labeling them for research need human expertise. Nevertheless access to medical data does not come on a silver plate due to the fact that it contains information that could be used to identify specific individual. So the data is highly protected as Protected Health Information (PHI) In this study, we use 4 basic methods of data augmentation originally proposed by Wie et al [10] to create more realistic and diverse augmented medical data for named entity recognition.

The organization of this paper is as follows. Section II summarizes the background and related work. EDA and NER, the model for data augmentation is provided in Section III. Section IV includes detailed information about the experimental setup, and the results and discussion can be found in Section V. Finally, the paper concludes in Section VI.

II. Background And Related Work

A. Information Extraction of Clinical Data

The importance of information extraction is manifested in the hot research currently going on, especially in the medical domain. Gathering structured data

from medical records, information extraction enables the automation of tasks; as in smart content classification of diseases, integrated research on future infections, management and delivery. Data driven activities like mining of patterns and trends in patient medical history are just but a few to mention. Many Natural Language Processing systems, approaches, models and research techniques have been developed to extract vital information from medical records. These includes the use of ontology-based resources [11], concept mapping [12], grammar structure matching[13], semantic parsing [14] approaches, and rule-based[15] and machine learning[16] systems. The rule-base approach gave less robust due to the fact that for every new corpus, the rules have to be revamped to preserve best performance of the model; this requirement increases the maintenance cost accordingly[17]. A smoking status detection system for patients was developed by Savova et al [18] and later embedded into the clinical Text Analysis and Knowledge Extraction System (cTAKES).

B. General Clinical Text Augmentation

Data augmentation in clinical records has no clear-cut method to directly increase size and diversity of labeled data, nevertheless the techniques proposed so far has shown impressive results. Back-translation was proposed by google AI where they take data samples x in a language A to another language B and then translating it back to A to obtain augmented data sample x' [9]. The same research used TF-IDF scores to replaced uninformative words to generate new instances for topic classification tasks. Zhang et al[19], Wei and Zou [10] in their work used *Synonym Replacement* were tokens are replaced by their synonyms from WordNet or a predefined language model [20]. In the work of Wei and Zou, they proposed 4 basic methods: *Synonym Replacement(SR), Random Swap(RS), Random Deletion(RD), and Random Insertion(RI)*. Vary basic methods but very powerful and easy to implement.

C. Text Augmentation for NER Task

Many of the methods of text augmentation are for classification tasks in NLP. To make data augmentation for NER tasks, Xiang Dai et el [21] in their paper "An Analysis of Simple Data Augmentation for Named Entity Recognition" proposed:

- **Label-wise token replacement:** Randomly replace tokens that share same labels with tokens from original data set
- **Mention replacement:** Replace a mentioned entity and its label tokens from original data set that shares same entity label.
- **Shuffle within segments:** Divide sequence into segments and shuffle the tokens in each segment without changing the labels

Their methods improved the performance of both transformer model and Recurrent model after experimenting on two domain-specific data sets MaSciP [22] and i2b2-2010 [23] Tian Kang et al [24] in their work, presents an extension of EDA (Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks) [10] methods, by featuring Unified Medical Language System (UMLS) [11], and adapting the methods for named entity recognition tasks to improve performance of models in both classification and NER in biomedical domain.UMLS a knowledge based system, is not easily accessible,and setting it up is both time consuming and highly costly. This makes it not suitable for low-resource settings. Nevertheless UMLS-EDA [24] enables substantial improvement for NER tasks and also improves the performance of state-of-the-art classification model. In our approach, we extend the methods used, by modifying them to a low cost and easy to use setting, for medical text augmentation.

III. Approach

Our work presents simple data augmentation methods for named entity recognition tasks in medical data mining. We adapt the four easy but robust methods of text augmentation methods used to generate diverse and quality data to train and enhance the performance of medical domain models for named entity recognition tasks. UMLS-EDA adapted EDA to suit NER tasks by adding UMLS, with the notion that it's good performance in text classification can also be realized in NER. We provide low cost and simple way of utilising small amount of domain specific data to enhance models performance with augmentation and transfer learning.

Transfer learning so far have proven to be one of the best ways to improve the performance of deep learning and neural network models. This prompted the idea of fine-tuning the pre-trained model BERT (Bidirectional Encoder Representations)[25] in the medical domain. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) being the first domain-specific BERT-based model is an example of transfer learning which improved and also gain the state-of -the-art model with 0.62% F1 score improvement in biomedical named entity recognition [26]. With this notion at the back of our mind and the fact that data augmentation improves performance of NLP models, we combined the two in our approach to boost the performance of biomedical models.

A. CONTRIBUTION OF PAPER

This section summarises the general contribution of our paper. In this paper we showed how data augmentation can be useful in the uplifting and enhancement of medical data and models. The contributions are :

Original Sentence		***Significant proportions of the variance*** $_{[Qualifier]}$ *in* ***responsibility*** $_{[Outcome]}$, ***struggle*** $_{[Outcome]}$, *and* ***cooperation*** $_{[Outcome]}$, *however, were not accounted for by* ***therapist process*** $_{[Intervention]}$ *alone.*
Original word sequence		[*"Significant", "proportions", "of", "the", "variance", "in", "responsibility", ",", "struggle", ",", "and", "cooperation", ",", "however", ",", "were", "not", "accounted", "for", "by", "therapist", "process", "alone"*]
Original tag sequence		["B-modifier", "I-modifier", "I-modifier", "I-modifier", "I-modifier", "O", "B-Outcome", "O" , " B-Outcome", "O" , "O" , "B-Outcome", "O", "O" , "O" , "O" , "O" , "O", "O", "O", "B-Intervention", "I-Intervention", "O"]
SR-WordNet (n=2)	**Words**	[*"Significant", "proportions", "of", "the", "discrepancy", "in", "duty", ",", "struggle", ",", "and" , "cooperation", ",", "however" , ",", "were" , "not" , "accounted", "for", "by", "therapist", "process", "alone"*]
	Tags	["B-modifier", "I-modifier", "I-modifier", "I-modifier", "I-modifier", "O", "B-Outcome", "O" , " B-Outcome", "O" , "O" , "B-Outcome", "O", "O" , "O" , "O" , "O" , "O", "O", "O", "B-Intervention", "I-Intervention", "O"]
RI (n=2)	**Words**	[*"Significant", "proportions", "only", "of", "the", "variance", "in", "responsibility", ",", "struggle", ",", "and" , "cooperation", ",", "however" , ",", "were" , "not" , "accounted", "for", "by", "therapist", "process", "alone", "healer"*]
	Tags	["B-modifier", "I-modifier", "I-modifier" ,"I-modifier", "I-modifier", "I-modifier", "O", "B-Outcome", "O" , " B-Outcome", "O" , "O" , "B-Outcome", "O", "O" , "O" , "O" , "O" , "O", "O", "O", "B-Intervention", "I-Intervention", "O", "O"]
RS (n=2)	**Words**	[*"Significant", "responsibility", "of", "the", "variance", "in", "proportions", ",", "struggle", ",",","accounted" , "cooperation", ",", "however" , ",", "were" , "not" , "and", "for", "by", "therapist", "process", "alone"*]
	Tags	["B-modifier", "I-modifier", "I-modifier", "I-modifier", "I-modifier", "O", "B-Outcome", "O" , " B-Outcome", "O" , "O" , "B-Outcome", "O", "O" , "O" , "O" , "O" , "O", "O", "O", "B-Intervention", "I-Intervention", "O"]
RD (n=2)	**Words**	[*"Significant", "proportions", "of", "the", "in", "responsibility", ",", "struggle", ",", "and" , "cooperation", ",", "," , "were" , "not" , "accounted", "for", "by", "therapist", "process", "alone"*]
	Tags	["B-modifier", "I-modifier", "I-modifier", "I-modifier", "O", "B-Outcome", "O" , " B-Outcome", "O" , "O" , "B-Outcome", "O" , "O" , "O" , "O" , "O", "O", "O", "B-Intervention", "I-Intervention", "O"]

Fig. 1: Overview of the affects of augmentation methods on sample data. **SR**: Synonym Replacement, **RI**: Random Insertion, **RS** : Random Swapping, **RD**: Random Deletion. n = number of words transformed. The words marked in red indicate the changes from data augmentation process

TABLE I: Test results of BioBERT with Augmentation EPOCH = 100

Dataset used in fine-tuning	Metric	Original Data	n = 2	n = 16
NCBI disease	Precision (P)	85.21	86.48	86. 48
	Recall (R)	88.23	88.64	88.64
	F1-Score (F1)	86.69	87.55	87.55
Species-800	Precision (P)	70.76	70.76	70.76
	Recall (R)	76.66	76.66	76.66
	F1-Score (F1)	73.59	73.59	73.59

1) We showed that without expensive architecture, data augmentation can boost the performance of biomedical text mining models.
2) We also showed that with little adjustments, the same methods of augmentation used in text classification can be used in biomedical named entity recognition without the need of costly integrated medical software.
3) we also showed that data augmentation in addition to transfer learning is a suitable combination for high performance of biomedical named entity recognition models.

B. EDA UMLS-EDA

Augmentation operations in computer vision inspired the methods used now in NLP. EDA [10] proposed

TABLE II: Test results of BioBERT with Augmentation. EPOCH = 30

Dataset used in fine-tuning	Matric	Original Data	n = 2	n = 16
NCBI disease	Precision (P)	86.36	86.35	86.35
	Recall (R)	89.06	88.95	88.95
	F1-Score (F1)	87.69	87.63	87.63
Species-800	Precision (P)	70.29	70.29	70.29
	Recall (R)	75.88	75.88	75.88
	F1-Score (F1)	72.98	72.98	72.98

universal data augmentation methods for NLP. Four methods are used to perform augmentation on a randomly selected token in sentence. Their approach was mainly for text classification, therefore needs amendment to suit name entity recognition for token-level prediction.

Most NER models uses the BIO / IOB tag format (**B** : Beginning of entity mentioned, **I**: Inside and part of the entity mentioned except the first **O**: Other words which has no entity) which makes data augmentation in NER more tedious. Kang et al[24] adapted EDA methods to NER by transforming word sequence and the corresponding BIO tag sequences properly, so much that the consistency of both sequence is maintained after augmentation. Each of the following augmentation methods are applied to each given sentence in the training data N times. Stop words are not included.

- **Synonym Replacement** : randomly select n number of words and replace them with their corresponding synonym from WordNet. Tag sequence and word count are maintain as it.
- **Random Insertion** : select a random word, find its random synonym and insert that synonym in the sentence. If the insertion position is at the beginning of an entity (**B**), skip it, else if its at the position inside an entity **I** replace it, with a corresponding *I* tag. Otherwise replace the synonym with a corresponding ***O*** tag
- **Random Swap** : Randomly select two token and swap their positions. Tag sequence and word count are maintained as it is.
- **Random Deletion** : Randomly remove each word in the sentence with a probability p if and only if its not the begining of entity (**B**) tag. If the word is inside and entity (**I**) , delete it and its corresponding *I* tag, else delete an *O* tag.

To illustrate more of the methods used, figure 1 shows an example of how a given sentence and its tag sequences are affected.

C. BioBERT FOR NAMED ENTITY RECOGNITION TASK

BERT [25], unlike other models who do not take into consideration the context of words, is a contextualized word representation model. It is based on a masked language model and pre-trained using bidirectional transformer. BioBERT(Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a great medical model that is based on BERT, pre-trained on large biomedical corpora. Mostly general language models performs less on medical data due the fact that, there are terms, disease names and domain specific used words that are only understood by experts in the medical domain. To make language models like BERT performs well on biomedical data, Lee et al pre-trained their model on PubMed(4.5 bilion number of words), PMC Full-text articles(13.5 bilion number of words), PMC Full-text articles(2.5 bilion number of words) and BooksCorpus(2.5 bilion number of words) and named it as BioBERT. In pre-training of, they first initialized with BERT and for tokenization, they go for WordPiece tokenization[27] . With minimal architectural modification BioBERT is fine-tuned on downstream NLP tasks including NER task.

IV. EXPERIMENTAL SETUP

In order to evaluate our work, we used the standard benchmark datasets usually used to evaluate NER tasks in the medical domain. This includes; NCBI-disease corpus [28], BC5CDR (BioCreative V CDR corpus) [29] and Species-800[30] datasets. Fully annotated medical data sets at the mention and concept level. All data sets are used in their pre-processed version from Lee et al.[26]. We fine-tune BioBERT on this datasets with and without their augmented data. Table III shows the statistics of the datasets used in fine-tuning BioBERT in our work. We try various combinations of the number of augmented data and the number of epochs of training. It shows that small datasets when augmented yields more results than the very large datasets. We use BioBERT version 1.1 (BioBERT-Base v1.1 (+ PubMed 1M)) in all our experiments. This version of BioBERT is pre-trained on BERT-base-Cased.In the fine-tuning we use only 1 NVIDIA GPU 2080ti. The hyper parameters we use are shown in Table IV.

V. RESULTS AND DISCUSSION

In this section, we discuss the results of the experimental setup. We observe that the number of epochs

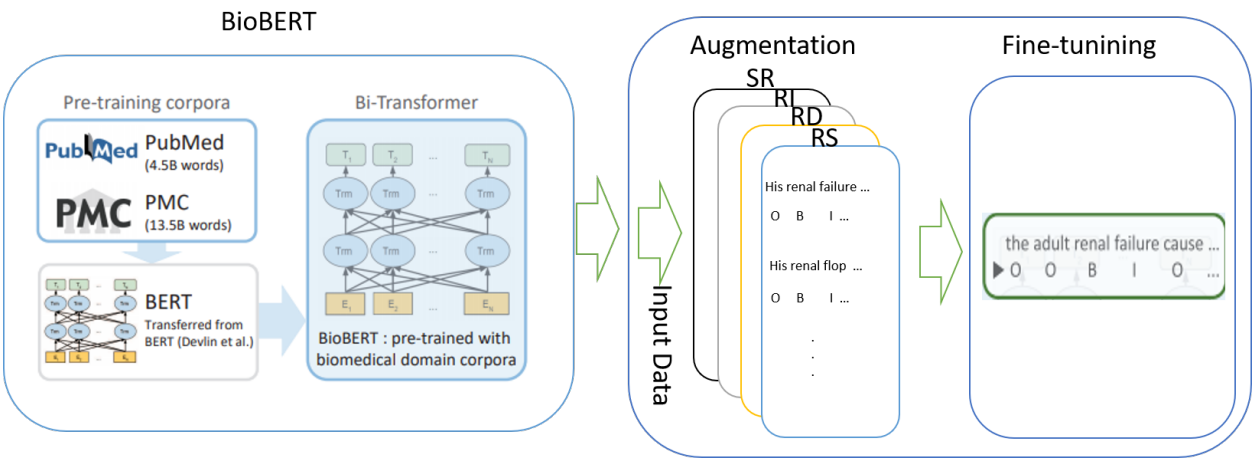

Fig. 2: Overview of our general approach, pre-training of BioBERT, data augmentaion and fine-tuning for NER task.

Dataset	Entity	# Annotation
NCBI disease corpus	disease NE	6881
BC5CDR	disease NE	12694
Species-800	Species NE	3708

TABLE III: Datasets used in fine-tuning BioBERT. '#': number of NE: Named Entity

The number of annotation is provided as shown in BioBERT paper.

Hyper-parameter	Value
mini-batch size	32
epochs	3 , 30 & 100
max. sequence length	192
weight decay rate	0.01, 0.0
optimizer	Adam

TABLE IV: Hyperparameters for BioBERT fine-tuning

affects the performance BioBERT trained on augmented data. When the number of augmentation is n = 2, that is increasing the labeled data two folds, F1 score yields as good as the original model. Increasing the number of n, improves the performance of the model as can be seen in Table 3. NCBI-disease dataset with seven thousand annotations shows clear improvement on the model when augmented with n=16. It yields F1 score of 87.55% better than original 86.69% before augmentation. Similarly even after augmentation, the models still gets the same precision, recall and f1 score values as original using species-800 dataset. This tells us that the with small amount of data instances, augmentation increases the performance of the model vividly and with relatively large data,there is still improvement which is minor. Overall our results show that EDA can successfully be adapted for NER in medical domain. As the future work we would like to explore additional methods that are more specific to the NER and medical domain.

A. Accuracy Metrics

In general the standard metrics used as accuracy metrics are Precision, Recall, F1-Score, and Accuracy Score. Precision shows the percentage of true labels among all the labels; Recall measures the percentage of true labels in the dataset being recalled; F1-Score is the harmonic mean of the precision and recall. For the evaluation metrics in our work, we used entity level precision, recall and F1 score.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 * precision * recall}{precisiom + recall}$$

where:

TP - True Positive;
TN - True Negative
FP - False Positive;
FN - False Negative

VI. Conclusion

In this study we use EDA, which is originally proposed for augmenting text classification datasets and

applied it on augmenting NER datasets which is very different and much more difficult to augment due to the risk of distorting sentences syntactically and semantically. There is also an additional complexity due the complex terminology and structure of sentences in the medical domain. Our approach when applied to medical NER datasets shows promising results. Our experiments show that the performance is sensitive to the augmentation factor n, which shows how much each labeled instance is augmented, and the epoch of the deep learning algorithms used in NER. High number of epoch increase the performance of NER models. Increasing the number of n, improves the performance of the model as can be seen in Table 3. NCBI-disease dataset with seven thousand annotations shows clear improvement on the model when augmented with n=16. It yields F1 score of 87.55% better than original 86.69% before augmentation. Similarly even after augmentation, the models still gets the same precision, recall and f1 score values as original using species-800 dataset. This tells us that the with small amount of data instances, augmentation increases the performance of the model vividly and with relatively large data,there is still improvement which is minor. Overall our results show that EDA can successfully be adapted for NER in medical domain. As the future work we would like to explore additional methods that are more specific to the NER and medical domain.

REFERENCES

[1] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[2] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[3] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "Luke: deep contextualized entity representations with entity-aware self-attention," *arXiv preprint arXiv:2010.01057*, 2020.

[4] B. Dhingra, H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Gated-attention readers for text comprehension," *arXiv preprint arXiv:1606.01549*, 2016.

[5] Y. Wu and B. Hu, "Learning to extract coherent summary via deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[6] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[7] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13001–13008, 2020.

[8] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

[9] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," *arXiv preprint arXiv:1904.12848*, 2019.

[10] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[11] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[12] F. Bakhshi-Raiez, L. Ahmadian, R. Cornet, E. de Jonge, and N. F. de Keizer, "Construction of an interface terminology on snomed ct," *Methods of Information in Medicine*, vol. 49, no. 04, pp. 349–359, 2010.

[13] J.-w. Fan, E. W. Yang, M. Jiang, R. Prasad, R. M. Loomis, D. S. Zisook, J. C. Denny, H. Xu, and Y. Huang, "Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1168–1177, 2013.

[14] S. Jonnalagadda, T. Cohen, S. Wu, and G. Gonzalez, "Enhancing clinical concept extraction with distributional semantics," *Journal of biomedical informatics*, vol. 45, no. 1, pp. 129–140, 2012.

[15] S. Sohn, Z. Ye, H. Liu, C. G. Chute, and I. J. Kullo, "Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports," *AMIA summits on translational science proceedings*, vol. 2013, p. 249, 2013.

[16] M. Torii, K. Wagholikar, and H. Liu, "Using machine learning for concept extraction on clinical documents from multiple data sources," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 580–587, 2011.

[17] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 473–480, 2002.

[18] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[19] H. Cai, H. Chen, Y. Song, C. Zhang, X. Zhao, and D. Yin, "Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight," *arXiv preprint arXiv:2004.02594*, 2020.

[20] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," *arXiv preprint arXiv:1805.06201*, 2018.

[21] X. Dai and H. Adel, "An analysis of simple data augmentation for named entity recognition," *arXiv preprint arXiv:2010.11683*, 2020.

[22] S. Mysore, Z. Jensen, E. Kim, K. Huang, H.-S. Chang, E. Strubell, J. Flanigan, A. McCallum, and E. Olivetti, "The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures," *arXiv preprint arXiv:1905.06939*, 2019.

[23] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/va challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

[24] T. Kang, A. Perotte, Y. Tang, C. Ta, and C. Weng, "Umls-based data augmentation for natural language processing of clinical research literature," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 812–823, 2021.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[27] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[28] R. I. Doğan, R. Leaman, and Z. Lu, "Ncbi disease corpus: a resource for disease name recognition and concept normalization," *Journal of biomedical informatics*, vol. 47, pp. 1–10, 2014.

[29] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "Biocreative v cdr task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, 2016.

[30] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen, "The species and organisms resources for fast and accurate identification of taxonomic names in text," *PloS one*, vol. 8, no. 6, p. e65390, 2013.