Simple Semantic-based Data Augmentation for Named Entity Recognition in Biomedical Texts

Uyen T.P. Phan[1,2], **Nhung T.H. Nguyen**[3]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam
[2]Vietnam National University, Ho Chi Minh city, Vietnam
[3]Department of Computer Science, University of Manchester, UK
ptpuyen@fit.hcmus.edu.vn, nhung.nguyen@manchester.ac.uk

Abstract

Data augmentation is important in addressing data sparsity and low resources in NLP. Unlike data augmentation for other tasks such as sentence-level and sentence-pair ones, data augmentation for named entity recognition (NER) requires preserving the semantic of entities. To that end, in this paper we propose a simple semantic-based data augmentation method for biomedical NER. Our method leverages semantic information from pre-trained language models for both entity-level and sentence-level. Experimental results on two datasets: i2b2-2010 (English) and VietBioNER (Vietnamese) showed that the proposed method could improve NER performance.

1 Introduction

In machine learning and especially deep learning approaches, performance of the trained models is often proportional to the size of the training data. Consequently, for a model to achieve acceptable performance, we need a certain amount of labelled data. This would be an issue for low-resource domain and low-resource languages since annotating labelled data is time-consuming and expensive. To address the issue, data augmentation has been proposed to increase the variety of training data without directly collecting or annotating additional data (Feng et al., 2021).

Intuitively, data augmentation for named entity recognition (NER) task is more difficult to perform than for other sentence-level and sentence-pair tasks. Simple operations used to augment a sentence such as token swap, token deletion, and token insertion (Wei and Zou, 2019) may not work well in the case of NER, especially in the biomedical domain. One of the reasons is that a named entity can be composed by multiple tokens and we have to preserve the semantic of entities after applying those operations. For example, consider the following sentence from the i2b2-2010 corpus (Uzuner et al., 2011) with its entities:

She can be given prn [lasix]$_{Treatment}$ for [weight gain]$_{Problem}$ or [shortness of breath]$_{Problem}$.

If we randomly swap the 'lasix' token with 'weight', the sentence is not semantically correct. Similarly, when the 'weight' token is deleted, the remaining 'gain' token is no longer suitable for an entity of Problem. For the insertion operation, if we randomly insert a token into the sentence, the semantic of the sentence will be changed and we will not be able to assign a suitable entity label for it. As a result, it is necessary to have different augmentation methods specified for NER.

There are several model-based data augmentation methods for NER. Chen et al. (2020) proposed Local Additivity-based Data Augmentation (LADA) that can create virtual samples using interpolation technique. Their exeperimental results showed that LADA could help to produce state-of-the-art (SOTA) on two NER benchmarks including CoNLL 2013 (Tjong Kim Sang and De Meulder, 2003) and GermEval 2014 (Benikova et al., 2014). Meanwhile, Nie et al. (2020) took advantages of the rich semantic information in pre-trained word embeddings to create a semantic augmentation module for NER models. They also reported SOTA performance on some social media corpora.

Obviously, model-based methods can help to improve NER performance, but they are often complicated and difficult to implement. In contrast, rule-based methods are simpler and more intepretable than model-based ones, but still effective. Dai and Adel (2020) adjusted simple operations such as replacement and shuffle to preserve the semantic of both entities and sentences. Specifically, they proposed Synonym Replacement (SR) and Mention Replacement (MR). SR replaces a word in a sentence with a word of the same semantics taken from WordNet. MR replaces the whole entity with another random entity in the same entity type based on the training data; the replacement action for each entity is decided based on the binomial distribution.

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 123–129
May 26, 2022. ©2022 Association for Computational Linguistics

As a result, they could improve the NER performance on both MaSciP and i2b2-2010 corpora.

We find two limitations in Dai and Adel (2020)'s approach. Firstly, although the SR operation takes into account the semantic aspect of tokens, it does not consider the semantic at the entity level. Secondly, the MR operation is performed on the entity level randomly, which may cause semantically incorrect sentences. We hypothesise that if we somehow control the semantics in entity and sentence levels in augmentation operations, we could create a meaningful augmented data, hence improving the NER performance. To that end, we propose Semantic Neighbour Replacement (SNR), a simple data augmentation method for biomedical NER that considers the semantic aspects of both entity and sentence levels.

Specifically, at the entity level, unlike MR (Dai and Adel, 2020), we only replace a source entity with a target one if the target entity is in the same entity type and semantically related to the source one. At the sentence level, we only retain sentences that are semantically related to the original sentence. The semantically related entities and sentences are calculated by using pre-trained language models.

We conducted experiments on two biomedical datasets: i2b2-2010 (Uzuner et al., 2011)—an English corpus of clinical records and VietBioNER (Phan et al., 2022)—a Vietnamese corpus of biomedical texts. Experimental results indicate that using SNR, we can improve NER performance on low-resource settings as well as on full training data. In particular, the F1-scores were increased by 0.52% for i2b2-2010 and 1.3% for VietBioNER.

2 Methodology

The core idea of SNR is to replace entities and to control augmented sentences based on semantic similarity. The method can be divided into three consecutive phases: semantic neighbour extraction, entity replacement, and sentence evaluation.

Semantic Neighbour Extraction: Initially, we perform feature extraction for entities using pre-trained language models. An entity embedding is calculated by taking an average of word embeddings in it. Next, we generate sets of semantic neighbors based on cosine similarity. An entity is a semantic neighbor to another entity if both of them belong to the same entity type and have a cosine similarity greater than or equal to a threshold α.

Entity Replacement: During this phase, we generate new sentences by replacing an entity with another random entity in its semantic neighbor set. For each entity type, we just randomly replace one entity of that type in a sentence. As a result, we obtain a set of augmented sentences from original ones.

Sentence Evaluation: Augmented sentences generated in the previous phase are probably semantically incorrect, which may affect the training process. To alleviate the issue, we perform an automatic evaluation to remove augmented sentences that are semantically different from their original sentences. To that end, we firstly represent both original and augmented sentences as vectors by using a pre-trained sentence-level language model. We then use cosine similarity to estimate the semantic similarity between two sentences. If the cosine similarity of an augmented sentence and its original sentence is less than a threshold θ, the augmented sentence will be discarded.

In this paper, the two parameters α and θ will be in ranges of $[0, 1]$. The larger the α, the greater the semantic similarity between entities, but the smaller the number of neighbours. The θ parameter represents the degree of rigour in the automatic evaluation phase. When θ approximates to 1, only sentences that are very close to the meaning of the original sentence are retained. We therefore can keep only a few of the augmented sentences. In contrast, we can keep more sentences as θ approximates to 0. When θ is set to 0, the sentence evaluation phase will be disabled. At this point, we do not discard any augmented sentences from the second phase. We can fine-tune both α and θ to generate suitable augmented data.

3 Experiments

3.1 Datasets

We conduct experiments on the two datasets including i2b2-2010 (English) (Uzuner et al., 2011) and VietBioNER (Phan et al., 2022) (Vietnamese). The i2b2 corpus includes patient records annotated with three named entity categories of Medical Problem, Test, and Treatment. Meanwhile, VietBioNER is constituted by biomedical grey literature specified for tuberculosis. The corpus was annotated with five named entity categories of Organisation, Location, Date and Time, Symptom and Disease, and Diagnostic Procedure. Some statistics of both corpora are reported in Table 1.

	i2b2-2010	VietBioNER
#Sentence	32894	1706
#Sentence in		
Training set	9558	706
Development set	2389	300
Test set	20947	700
Avg. len. of sent.	13	31
#Entity type	3	5
Vocab size	24321	3548

Table 1: The summary statistic of the two datasets.

		MR	ER	SNR
i2b2-2010	S	17	19	12
	M	67	90	61
	L	242	347	239
	F	4462	7308	4626
VietBioNER	S	21	9	7
	M	76	13	13
	L	256	86	84
	F	347	550	459

Table 2: Number of augmented sentences in each training set. **S**mall, **M**edium, **L**arge, and **F**ull sets contain 50 sentences, 150 sentences, 500 sentences, and the complete training set, respectively.

Following Dai and Adel (2020), to simulate a low-resource setting, we create small, medium and large sets with different numbers of sentences: 50, 150 and 500, respectively. These sentences are randomly selected from the training part of each dataset. It is noted that our small, medium and large splits of the i2b2 dataset are different from those by Dai and Adel (2020). Augmentation methods are only applied on the training set, we use the same development and test sets for all experiments.

3.2 Language Models

For semantic neighbour extraction, we use ClinicalBERT (Alsentzer et al., 2019)—a pre-trained language model on clinical text for the i2b2-2010 dataset and PhoBERT (Nguyen and Nguyen, 2020)—a pre-trained language model on Vietnamese Wikipedia and news for VietBioNER.

In sentence evaluation, we employ Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), a sentence-level language model for sentence embeddings, to represent both original and augmented sentences.

We use all the mentioned models with the initialised weights provided by Hugging Face[1].

Regarding the NER task training, we also fine-tune the aforementioned language models on the two corpora.

3.3 Experiment Settings

To show the effectiveness of the proposed method, we conducted the following experiments:

- Baseline: We only trained NER models on the original training data.
- Baseline combined with augmented data: We trained NER models on the original training set and its augmented data created by the following three methods:
  - Mention Replacement (**MR**): We followed the MR method proposed by Dai and Adel (2020).
  - Entity Replacement (**ER**): We only performed the first two phases of our proposed method. The last phase, Sentence Evaluation, was disabled by setting the parameter θ to 0.
  - Semantic Neighbour Replacement (**SNR**): We performed all three phases of our proposed method.

It is noted that since in this paper we focus on biomedical entities, we only created an augmented data for Symptom_and_Disease and DiagnosticProcedure entities in the case of VietBioNER. We however report the NER performance on all five NE categories.

3.4 Experimental Results

Based on the fine-tuning results on the development sets, we selected $\alpha = 0.8$ for all sets of i2b2-2010; for VietBioNER, $\alpha = 0.65$ for the full set, and $\alpha = 0.85$ for the other sets; and $\theta = 0.9$ for all cases across the corpora. The number of augmented sentences generated in each setting are reported in Table 2. Since SNR discards augmented sentences that are not semantically related to the original ones, it is reasonable that the numbers of augmented sentences by SNR is less than or equal to those by MR and ER.

We trained NER models on a combination of augmented and original sentences, and applied them to the corresponding testing sets. The NER performance in terms of F1-scores on those sets

[1] https://huggingface.co/models, https://huggingface.co/sentence-transformers

Method	i2b2-2010				VietBioNER			
	S	M	L	F	S	M	L	F
Baseline	37.13	67.58	75.53	87.21	59.21	70.78	79.48	79.60
+ MR	**39.56**	67.21	76.35	87.54	**60.98**	71.19	79.31	79.00
+ ER (our method)	39.42	68.36	76.33	87.37	59.31	71.94	**79.51**	80.09
+ SNR (our method)	38.75	**69.43**	**76.86**	**87.73**	59.83	**72.14**	79.34	**80.90**

Table 3: NER performance by different augmentation methods in terms of F1-score. Bold numbers indicate the best performance in a specific setting.

		Sentence
i2b2-2010	**Ori**	Her speech was fluent with no [phasic or praxic problems]$_{Problem}$, [dysarthric]$_{Problem}$.
	MR	Her speech was fluent with no [oral lesions]$_{Problem}$, [left coloboma]$_{Problem}$.
	SNR	Her speech was fluent with no [phasic or praxic problems]$_{Problem}$, [slurred speech]$_{Problem}$.
VietBioNER	**Ori**	Tuy nhiên, các xét nghiệm tế bào và vi trùng trong chẩn đoán [lao]$_{Symptom\&Disease}$ có độ nhạy còn thấp. (However, cytology and bacteria tests in the diagnosis of [TB]$_{Symptom\&Disease}$ have low sensitivity.)
	MR	Tuy nhiên, các xét nghiệm tế bào và vi trùng trong chẩn đoán [ho khan]$_{Symptom\&Disease}$ có độ nhạy còn thấp. (However, cytology and bacteria tests in the diagnosis of [dry cough]$_{Symptom\&Disease}$ have low sensitivity.)
	SNR	Tuy nhiên, các xét nghiệm tế bào và vi trùng trong chẩn đoán [bệnh lao]$_{Symptom\&Disease}$ có độ nhạy còn thấp. (However, cytology and bacteria tests in the diagnosis of [TB disease]$_{Symptom\&Disease}$ have low sensitivity.)

Table 4: **Ori**ginal sentences and their augmented sentences with different methods. Blue texts indicates entity replacement.

are reported in Table 3[2]. Generally, we can see that the NER performance was improved when using data augmentation methods on both English and Vietnamese corpora. Detailed results of precision and recall can be found in Appendix A.

Among the four sizes of the data, MR (Dai and Adel, 2020) could obtain the best performance in the small size setting, across the two corpora. This can be explained by the fact that given only 50 sentences in the training, adding more sentences will help the model overcome overfitting. With the medium size sets, MR could improve the performance on VietBioNER but not on i2b2-2010. In contrast, MR could boost F1-scores on the large and full sets on i2b2-2010, but not on VietBioNER.

Regarding SNR, we could have better F1-scores in most settings of medium, large and full sets, on both English and Vietnamese corpora. With the i2b2 English corpus, the proposed methods has an average improvement of 1.23% of F1-scores (SNR) and 0.58% (ER). Meanwhile, that number by MR is 0.26%. For VietBioNER, the average improvement is 0.84%, 0.63%, and -0.12% of F1-scores for SNR, ER, and MR, respectively. It is worth noting that even with a full training set, using SNR to augment the data training could also boost NER performance. In particular, F1-scores were increased by 0.52% for i2b2-2010 and 1.3% for VietBioNER.

Interestingly, while the number of augmented sentences by SNR is lower than those by ER (as shown in Table 2), the NER performance by SNR is better than those by ER in most of the cases across the corpora. This indicates that having augmented sentences semantically related to the original ones in the training data really improves the NER performance, despite the fact that the total number of sentence is not big. For instance, in the case of i2b2-2010, SNR generated about 37% less sentences than ER, but the NER performance by SNR was still better than those by ER.

3.5 Analysis

Although using MR could help improve the NER performance (as illustrated in Table 3), it is inevitable that MR could produce meaningless sentences. We collected such examples and showed them in Table 4. It can be seen that although MR replaced entities in the same type with the original ones, the resulting sentence is meaningless. Meanwhile, SNR controls the semantic at both entity level and sentence level, hence producing a more meaningful sentence close to the original meaning than the one by MR.

Moreover, we observed that most of sentences discarded by the sentence evaluation were semantically incorrect. We report some of discarded sentences in Table 5. It is obvious that the entity re-

[2] We use the IO tagging scheme.

		Sentence
i2b2-2010	Original	He did not sleep at night before and was [extremely fatigued]$_{Problem}$.
	Augmented	He did not sleep at night before and was [some shortness of breath]$_{Problem}$.
VietBioNER	Original	Hình ảnh [X-quang phổi]$_{DiagnosticProcedure}$ chủ yếu là thâm nhiễm 44%... (The [chest X-ray]$_{DiagnosticProcedure}$ image is mainly infiltrative 44%...)
	Augmented	Hình ảnh [chọc dò màng phổi]$_{DiagnosticProcedure}$ chủ yếu là thâm nhiễm 44%... (The [thoracentesis]$_{DiagnosticProcedure}$ image is mainly infiltrative 44%...)

Table 5: Examples of augmented sentences **discarded** by the Sentence Evaluation phase in SNR. Blue texts indicates entity replacement.

placement altered the meaning of those sentences and made them meaningless. As aforementioned, by discarding those sentences, SNR could produce better NER performance, indicating that it is useful to filter augmented sentences based on their semantic relatedness.

4 Conclusion

In this paper, we proposed a semantic-based data augmentation method for the named entity recognition task in the biomedical domain. Our method, namely Semantic Neighbour Replacement (SNR), simply generates more training sentences based on semantics of entity and sentence. Experiments on simulated low-resource settings show that using the proposed method, we can improve F1 score in both English (i2b2-2010) and Vietnamese (VietBioNER) corpora, even on the full training setting. Such results again confirm the importance of semantics in data augmentation. We believe that SNR can be applied to other domains and other languages as long as we have corresponding pre-trained language models.

Similar to previous work, our proposed method only augments in-domain data. Therefore, a followup work would be to study cross-domain augmentation method (Chen et al., 2021), in which we can leverage rich-resource data to enrich low-resource ones.

Acknowledgements

We would like to thank the anonymous reviewers for their useful comments. This research was partially funded by the University of Science, VNU-HCM, Vietnam under grant number CNTT2020-04.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. 2014. GermEval 2014 Named Entity Recognition Shared Task: Companion Paper. In *Proceedings of the KONVENS GermEval workshop*, pages 104–112, Hildesheim, Germany.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised NER. *Proceedings of The 2020 Conference on Empirical Methods in Natural Language Processing*.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. *CoRR*, abs/2010.11683.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Uyen Phan, Phuong Nguyen, and Nhung Nguyen. 2022. A Named Entity Recognition Corpus for Vietnamese

Biomedical Texts to Support Tuberculosis Treatment. In *Proceedings of the 13th Language Resources and Evaluation Conference*. European Language Resources Association.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Ozlem Uzuner, Brett South, Shuying Shen, and Scott DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18:552–6.

Jason Wei and Kai 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)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

A Detailed Results

We report the detailed results of precision, recall and F1-scores on i2b2-2010 in Table 6 and VietBioNER in Table 7.

It is expected that NER performances in terms of recall were mostly improved when using the data augmentation methods. Meanwhile, in terms of precision, the increase or decrease of NER performance was dependent on the data augmentation methods as well as the sizes of the training data. Nevertheless, in the case of full training data, using the SNR method, we could improve the NER performance in both recall and precision across corpora.

Method	Small			Medium			Large			Full		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	43.39	32.45	37.13	66.54	68.65	67.58	74.00	77.13	75.53	86.24	88.20	87.21
+ MR	**44.53**	35.59	**39.56**	63.36	71.55	67.21	73.56	79.35	76.35	86.67	88.42	87.54
+ ER	42.44	**36.79**	39.42	**67.24**	69.52	68.36	72.97	**80.02**	76.33	86.47	88.29	87.37
+ SNR	42.49	35.62	38.75	67.11	**71.90**	**69.43**	**74.37**	79.51	**76.86**	**86.92**	**88.55**	**87.73**

Table 6: NER performance on i2b2-2010 by different augmentation methods in terms of **P**recision, **R**ecall and **F1**-score. Bold numbers indicate the best performance in a specific setting.

Method	Small			Medium			Large			Full		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	56.92	61.69	59.21	67.88	73.93	70.78	**77.12**	81.99	79.48	77.49	81.83	79.60
+ MR	**58.91**	**63.19**	**60.98**	67.79	74.96	71.19	76.60	82.23	79.31	76.85	81.28	79.00
+ ER	57.39	61.37	59.31	**69.70**	74.33	71.94	76.50	**82.78**	**79.51**	77.57	**82.78**	80.09
+ SNR	58.87	60.82	59.83	68.92	**75.67**	**72.14**	76.93	81.91	79.34	**79.09**	**82.78**	**80.90**

Table 7: NER performance on VietBioNER by different augmentation methods in terms of **P**recision, **R**ecall and **F1**-score. Bold numbers indicate the best performance in a specific setting.