Data Augmentation for Low-Resource Named Entity Recognition Using Backtranslation

Usama Yaseen¹,², Stefan Langer¹,²
¹Technology, Siemens AG Munich, Germany
²CIS, University of Munich (LMU) Munich, Germany
{usama.yaseen,langer.stefan}@siemens.com

Abstract
The state of art natural language processing systems relies on sizable training datasets to achieve high performance. Lack of such datasets in the specialized low resource domains lead to suboptimal performance. In this work, we adapt backtranslation to generate high quality and linguistically diverse synthetic data for low-resource named entity recognition. We perform experiments on two datasets from the materials science (MaSciP) and biomedical domains (S800). The empirical results demonstrate the effectiveness of our proposed augmentation strategy, particularly in the low-resource scenario.

1 Introduction
Most recently, various deep learning methods have demonstrated state of the art performance for many natural language processing tasks such as text classification, sentiment analysis and named entity recognition. The availability of large training datasets is crucial to achieve this improved performance and avoid overfitting. However, in many real-world applications collecting such large training data is not possible. This is especially true for specialized domains, such as the material science or biomedical domain, where annotating data requires expert knowledge and is usually time-consuming and expensive.

Data augmentation (DA) (Simard et al., 1996) has been investigated to overcome this low resource problem. Label preserving synthetic data generation is widely used in computer vision (Krizhevsky et al., 2012; Ciresan et al., 2012; Fawzi et al., 2016) and speech domains (Schlüter and Grill, 2015; Ko et al., 2017). The discrete nature of language makes it difficult to adapt data augmentation strategies from computer vision and speech to natural language processing. Unlike computer vision, where hardcoded transformations (such as rotation, masking, cropping etc.) can be easily applied without changing the label semantics, the manipulation of a single word in a sentence could change its meaning.

Recently, there is an increased interest in applying data augmentation to natural language processing tasks. Most augmentation methods explore sentence-level tasks such as sentiment analysis (Listing et al., 2021), text classification (Wei and Zou, 2019; Xie et al., 2019) and sentence-pair tasks such as natural language inference (Min et al., 2020) and machine translation (Wang et al., 2018). The augmentation methods either employ heuristics such as word replacement (Zhang et al., 2015; Wang et al., 2018; Cai et al., 2020), word swap (Sahin and Steedman, 2018; Min et al., 2020) or random deletion (Wei and Zou, 2019) to generate augmented instances by manipulating a few words in the original sentence; or generate completely artificial instances via sampling from generative models such as variational autoencoders (Yoo et al., 2019; Mesbah et al., 2019) or backtranslation models (Yu et al., 2018; Iyyer et al., 2018).

The sequence labelling tasks such as named entity recognition (NER) and part-of-speech tagging (POS) involves prediction at the token level. This makes applying token-level transformation difficult as such manipulations may change the corresponding token level label. The existing DA methods for sequence labelling uses dependency tree morphing (Sahin and Steedman, 2018), MIXUP (Zhang et al., 2018) to generate queried samples in the active learning scenario (Zhang et al., 2020), sample novel sequences from a trained language model (Ding et al., 2020) and apply pre-defined heuristics such as label-wise token and synonym replacement (Dai and Adel, 2020). The existing sequence labelling DA methods are limiting as they: a). rely on linguistics resources like dependency parser or WordNet b). involves training a language model c). generate grammatically incoherent sequences d). cannot generate linguistically diverse sequences.

Motivated by the advancements in machine trans-
lalion and the availability of high-quality machine translation systems (He, 2015; Wu et al., 2016; Junczys-Dowmunt, 2019), in this work we adapt backtranslation to the task of NER. Backtranslation (BT) can automatically generate diverse paraphrases of a sentence or a phrase by naturally injecting linguistic variations. The injected linguistic variations can be further diversified by introducing layers of intermediate language translations. In this work, we generate paraphrases of one or several phrases in a sentence. We empirically demonstrate the effectiveness of our proposed method on two domain-specific NER datasets.

2 Related Work

There is an abundance of recent work on DA methods for NLP tasks, we refer the readers to Feng et al. for an extensive survey. In this section we narrow our focus to existing DA methods for sequence labelling tasks like NER and POS. We categorize existing DA methods for sequence labelling into two categories:

- **Rule-based:** DA primitives, which use predefined easy-to-compute transformations. We briefly describe six of such transformations proposed in the existing work:

  (a) **NER::Label-wise token replacement (LwTR):** Replace a token with another token of the same entity type at random (Dai and Adel, 2020).

  (b) **NER::Synonym replacement (SR):** Replace a token with one of its synonyms retrieved from WordNet at random (Dai and Adel, 2020).

  (c) **NER::Mention replacement (MR):** Replace an entity mention with another entity mention of the same entity type at random (Dai and Adel, 2020).

  (d) **NER::Shuffle within segments (SiS):** Divide the sequence of tokens into segments of the same label and then randomly shuffle the order of segments (Dai and Adel, 2020).

  (e) **POS::Crop Sentences:** Given a dependency tree of the sentence, "crop" a sentence by moving the tree fragments around the root (Sahin and Steedman, 2018).

  (f) **POS::Rotate Sentences:** Given a dependency tree of the sentence, "rotate" a sentence by moving the tree fragments around the root (Sahin and Steedman, 2018).

- **Generative models:** The existing work uses pre-trained language models to generate either part of the sequence or the entire sequence with the corresponding NER tags. Kang et al. proposed Filtered BERT which randomly masks one or several tokens in the original sentence and let BERT (Devlin et al., 2019) predict the masked token. The augmentation is only accepted if the cosine similarity of the word embeddings (computed using fastText embeddings (Bojanowski et al., 2017)) of the original token and the predicted masked token is above a certain threshold. Ding et al. propose a two-step DA process DAGA. First, a shallow language model is trained over linearized sequences of tags and words. Second, sequences are sampled from this language model and delinearized to create new examples.

3 Data Augmentation via Backtranslation

Figure 1 illustrates an example of data augmentation using backtranslation for NER. Note that backtranslation is only applied to the context around the entity mentions. Here the entity mention context is first translated to German and then back to English using an off-the-shelf machine translation system. The backtranslation results in a paraphrase of the original entity mention context. The original entity mention context is replaced with backtranslated context to create the augmented data instance.
is a candidate for the backtranslation. Second, the validity of the segment is determined based on the length of the segment, we only consider segments with three or more tokens as a valid segment for backtranslation. As a final step, the segment tokens are translated to the intermediate language(s) and finally back to the source language; the original segment tokens are replaced with the backtranslated tokens and thus we obtain the augmentation of the original input token sequence. In practice, we use a binomial distribution to randomly decide whether the segment should be backtranslated. Since only the context around the entity mention is backtranslated, it is straightforward to adjust the corresponding BIO-label sequence accordingly for the backtranslated text.

Data augmentation with backtranslation augments the original training set with diverse paraphrases of the entity mention contexts to help the underlying NER model to generalize beyond the standard training set.

4 Experiments and Results

4.1 Datasets

We empirically evaluate backtranslation for NER on two English datasets from the materials science and biomedical domains: MaSciP (Mysore et al., 2019) and S800 (Pafilis et al., 2013). MaSciP contains synthesis procedures annotated with synthesis operations and their typed arguments. S800 consists of PubMed abstracts annotated for organism mentions. We use the original train-dev-test split provided by the authors. The descriptive statistics of the datasets are reported in Appendix (see Table 2).

We simulate low-resource setting as proposed by Dai and Adel; we select 50, 150, 500 sentences from the training set to create the corresponding small, medium and large training sets (denoted as S, M, L in Table 1, whereas the complete training set is denoted as F). Data augmentation is only applied on the training set without altering the development and test set.

4.2 NER Model

We follow the standard approach of modelling the NER task as a sequence labelling task. The mainstream sequence labelling models for NER employ the neural-based encoder and an output tagging component. The typical choice of the encoder is a sequence model such as LSTM (Hochreiter and Schmidhuber, 1997) or more recently a sequence encoder such as Transformer (Vaswani et al., 2017); the output tagging component is usually a conditional random field layer (Lafferty et al., 2001) to model dependencies between neighbouring labels.

We employed the standard BiLSTM-CRF model (Lample et al., 2016) as our backbone model. We experimented with context-independent GloVe embeddings (Pennington et al., 2014) as well as state-of-the-art contextualized BERT embeddings (Devlin et al., 2019). We employed SciBERT (Beltagy et al., 2019), which is based on the BERT model pretrained on scientific publications; our preliminary experiments suggest that SciBERT achieves better performance than BERT. The superiority of domain-specific BERT models on downstream tasks has been observed by existing studies (Gururangan et al., 2020; Dai and Adel, 2020).

We report the micro-average F1 score as an evaluation metric. We employ early stopping and report the F1 score on the test set using the best performing model on the development set.

4.3 Backtranslation Models

We employed the Huggingface’s Transformers library (Wolf et al., 2020) port of the pretrained English↔German models (Ng et al., 2019) as the underlying backtranslation models for all our experiments.

---

1https://github.com/olivettigroup/annotated-materials-syntheses
2https://github.com/spyysalo/s800
3https://huggingface.co/facebook/wmt19-en-de
4https://huggingface.co/facebook/wmt19-de-en
We report the performance of various augmentation strategies applied to the token-level sequence tagging task of NER. We show that backtranslation can generate high-quality coherent data augmentation techniques especially backtranslation result in better performance when compared to the baseline. However, the average performance improvement due to data augmentation with SciBERT embeddings is lower as compared to the GloVe embeddings.

To quantitatively measure the diversity introduced by various augmentation techniques, we report distinct-1 (Li et al., 2016) in Figure 2. Distinct-1 quantifies the intra-text diversity based on distinct unigrams in each sentence, the value is scaled by the total number of tokens in the sentence to avoid favouring long sentences. Backtranslation yield the highest level of unigram diversity, this is not very surprising as backtranslation is known to generate diverse linguistic variations.

## 5 Conclusion

In this paper, we adapt backtranslation to the token-level sequence tagging task of NER. We show that backtranslation can generate high-quality coherent and linguistically diverse synthetic data for NER. The experiments on two domain-specific datasets demonstrate the effectiveness of backtranslation as a competitive data augmentation strategy for NER.

### 4.4 Hyperparameters

Following existing work (Dai and Adel, 2020), we tune the number of augmentation instances per training instance from a list of numbers: \{1, 3, 6, 10\}. When all data augmentation methods are applied, this tuning list is reduced to: \{1, 2, 3\}. We also tune the probability value \(p\) of the beta distribution which is used to decide if the segment in a sequence should be backtranslated. It is searched over a list of numbers: \{0.1, 0.3, 0.5, 0.7\}. We perform a grid search over these two hyperparameters to find their best combination on the development set.

### 4.5 Results

We report the performance of various augmentation techniques on the test sets in Table 1. For the most part, all data augmentation techniques improve over the baseline; backtranslation results in the biggest average improvement for both context-independent GloVe and contextualized SciBERT embeddings under different data usage percentiles. We attribute the improved performance of backtranslation to the generation of linguistically diverse and meaning-preserving entity mention contexts to enable better generalization of the underlying NER model.

The data augmentation techniques contribute to the biggest improvement in performance when the training sets are small, this effect is reduced as the training sets get larger (see columns S vs F in Table 1). The augmentation on the complete training set even decreases the performances for some augmentation techniques. The performance impact of data augmentation on varying sizes of training sets has also been observed in the existing work (Fadaee et al., 2017; Dai and Adel, 2020; Ding et al., 2020).

We also investigate the effectiveness of data augmentation techniques on the mainstream contextualized (pretrained SciBERT) embeddings. All the augmentation techniques especially backtranslation result in better performance when compared to the baseline. However, the average performance improvement due to data augmentation with SciBERT embeddings is lower as compared to the GloVe embeddings.

### Table 1: F1-score on test sets using different subsets of the training set. Here: S, M, L and F refer to small (50 instances), medium (150 instances), large (500 instances) and full (all instances) set. We repeat all experiments three times with different seeds. Mean values and standard deviations are reported. \(\Delta\) column shows the averaged improvement due to data augmentation for each embedding type across the datasets.

| Embeddings | Method | S | M | L | F | \(\Delta\) | S | M | L | F | \(\Delta\) |
|------------|--------|---|---|---|---|--------|---|---|---|---|--------|
| GloVe      | None   | 61.89±1.3 | 71.76±0.6 | 78.52±0.1 | 79.91±0.1 | 39.78±1.6 | 51.15±1.6 | 64.08±0.8 | 72.73±0.9 | 0.4 |
| LwTR       | 66.88±1.4 | 73.40±1.1 | 77.83±0.1 | 77.51±3.0 | 0.99 | 41.37±0.4 | 51.76±1.0 | 64.97±1.6 | 71.34±0.1 | 0.2 |
| SR         | 67.07±0.8 | 74.56±0.3 | 78.47±0.4 | 79.71±0.3 | 1.9 | 40.24±1.2 | 53.68±0.4 | 62.98±1.4 | 71.77±0.6 | 0.2 |
| MR         | 67.65±1.0 | 74.60±1.3 | 78.04±1.1 | 79.57±0.6 | 1.9 | 41.89±1.4 | 53.24±1.3 | 66.56±1.2 | 70.87±0.5 | 1.2 |
| SiS        | 66.87±2.9 | 73.40±1.5 | 78.95±0.6 | 79.79±0.5 | 1.7 | 41.57±1.8 | 51.83±0.7 | 65.16±1.0 | 71.20±0.6 | 0.5 |
| BT         | 70.11±0.8 | 75.86±0.8 | 78.92±0.2 | 80.30±0.5 | 3.3 | 44.60±1.0 | 53.22±1.3 | 66.76±1.1 | 72.92±0.2 | 2.4 |

| SciBERT    | None   | 53.22 | 53.68 | 51.76 | 58.97 | 0.4 | 0.99 | 1.9 | 1.9 | 1.7 | 3.3 |
| LwTR       | 51.15 | 53.68 | 51.76 | 58.97 | 0.4 | 0.99 | 1.9 | 1.9 | 1.7 | 3.3 |
| SR         | 53.68 | 51.76 | 58.97 | 0.4 | 0.99 | 1.9 | 1.9 | 1.7 | 3.3 |
| MR         | 53.24 | 51.83 | 65.16 | 1.0 | 41.57±1.8 | 51.83±0.7 | 65.16±1.0 | 71.20±0.6 | 0.5 |
| SiS        | 53.22 | 53.24 | 51.83 | 1.0 | 41.57±1.8 | 51.83±0.7 | 65.16±1.0 | 71.20±0.6 | 0.5 |
| BT         | 58.97 | 58.97 | 58.97 | 58.97 | 0.4 | 0.99 | 1.9 | 1.9 | 1.7 | 3.3 |

4.3.2 Impact of different training set sizes.

The impact of different training set sizes on the performance is shown in Table 1. The augmentation on the complete training set even decreases the performances for some augmentation techniques. The performance impact of data augmentation on varying sizes of training sets has also been observed in the existing work (Fadaee et al., 2017; Dai and Adel, 2020; Ding et al., 2020).
References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Sci bert: A pretrained language model for scientific text. 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 2019, Hong Kong, China, November 3-7, 2019, pages 3613–3618. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers, pages 567–573. Association for Computational Linguistics.

Ahluwais Fawzi, Horst Samulowitz, Deepak S. Turaga, and Pascal Frossard. 2016. Adaptive data augmentation for image classification. In 2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016, pages 3688–3692. IEEE.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. CoRR, abs/2105.03075.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 8342–8360. Association for Computational Linguistics.

Zhongjun He. 2015. Baidu translate: Research and products. In Proceedings of the Fourth Workshop on Hybrid Approaches to Translation, HyTra@ACL 2015, July 31, 2015, Beijing, China, pages 61–62. The Association for Computer Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Comput., 9(8):1735–1780.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 1875–1885. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1, pages 225–233. Association for Computational Linguistics.

Min Kang, Kye Lee, and Youngho Lee. 2021. Filtered bert: Similarity filter-based augmentation with bidirectional transfer learning for protected health information prediction in clinical documents. Applied Sciences, 11:3668.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In 2017 IEEE International Conference on Acoustics, Speech and
Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. **Switchout: an efficient data augmentation algorithm for neural machine translation.** In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 856–861. Association for Computational Linguistics.

Jason W. 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 2019, Hong Kong, China, November 3-7, 2019, pages 6381–6387. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. **Transformers: State-of-the-art natural language processing.** In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020, pages 38–45. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. **Unsupervised data augmentation.** CoRR, abs/1904.12848.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2019. **Data augmentation for spoken language understanding via joint variational generation.** In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 7402–7409. AAAI Press.
## A Datasets

|                  | MaSciP   |     | S800   |     |
|------------------|----------|-----|--------|-----|
|                  | Train    | Dev | Test   | Train| Dev | Test |
| Number of sentences | 1,899 | 112 | 162 | 5,733 | 830 | 1,630 |
| Number of mentions | 18,896 | 1,190 | 1,259 | 2,557 | 384 | 767 |
| Number of unique mentions | 4,707 | 590 | 605 | 1,070 | 194 | 3781 |
| Number of entity types | 21    | 20  | 21   | 1    | 1   | 1    |

Table 2: The descriptive statistics of the datasets.