CASSI: Contextual and Semantic Structure-based Interpolation Augmentation for Low-Resource NER

Tanmay Surana, Ho Thi Nga, Kyaw Zin Tun, Eng Siong Chng
Nanyang Technological University, Singapore
tanmay003@e.ntu.edu.sg, {ngaht, ztkyaw, aseschng}@ntu.edu.sg

Abstract

While text augmentation methods have been successful in improving performance in the low-resource setting, they suffer from annotation corruption for a token-level task like NER. Moreover, existing methods cannot reliably add context diversity to the dataset, which has been shown to be crucial for low-resource NER. In this work, we propose Contextual and Semantic Structure-based Interpolation (CASSI), a novel augmentation scheme that generates high-quality contextually diverse augmentations while avoiding annotation corruption by structurally combining a pair of semantically similar sentences to generate a new sentence while maintaining semantic correctness and fluency. To accomplish this, we generate candidate augmentations by performing multiple dependency parsing-based exchanges in a pair of semantically similar sentences that are filtered via scoring with a pretrained Masked Language Model and a metric to promote specificity. Experiments show that CASSI consistently outperforms existing methods at multiple low-resource levels, in multiple languages, and for noisy and clean text.\footnote{Our code is available at https://github.com/tanmaysurana/CASSI}

1 Introduction

Named Entity Recognition (NER) is the token-level task of classifying terms in text that belong to predefined entities like ‘location’, ‘person’, ‘organization’, etc. It is an important task in NLP that is used in recommendation systems, search and content classification systems for healthcare, academia, news organizations, etc., and as an integral part of other tasks like text summarization (Nallapati et al., 2016), question answering (Zhou et al., 2018; Fabbri et al., 2020), topic modeling (Krasnashchok and Jouili, 2018), etc. Annotating data for a token-level task like NER can be expensive, often requiring domain expertise. Therefore, NER datasets tend to be small, especially for narrow domains and low-resource languages.

Text Augmentation has been a successful method for improving performance in the low-resource setting for various tasks like Natural Language Inference, Sentiment Analysis, etc. It involves creating additional text using the existing training text while generally preserving the labels\footnote{Not all text augmentation methods necessarily preserve labels. For example, Yoon et al. 2021 and Harel-Canada et al. 2022.}. Typical methods for text augmentation include random distortions such as random deletions, insertions, substitutions, (Wei and Zou, 2019), word replacement (Wu et al., 2019; Kobayashi, 2018; Kim et al., 2022), backtranslation (Sennrich et al., 2016; Yu et al., 2018; Coulombe, 2018), paraphrasing (Madnani and Dorr, 2010; Kumar et al., 2019), and label-conditioned text generation through Language Models (Anaby-Tavor et al., 2020; Yoo et al., 2021), GANs (Sun and He, 2020), VAEs (Qiu et al., 2020), etc.

Preserving the token-level entity tags in NER is a challenging task and text augmentation methods often cause annotation corruption, i.e., assignment of incorrect token-level tags. Dai and Adel 2020 avoid the annotation corruption problem by performing random replacements of the same entity type. However, this creates limited diversity in the dataset and generates noisy sentences since the entity replacements don’t take sentence context into account. Methods like backtranslation and paraphrasing show limited improvements and require an additional retagging step, which tends to be erroneous (Sharma et al., 2022; Kyaw, 2022). More effective works attempt to solve the label-preservation problem for low-resource NER by performing augmentations using a label-conditioned language model. The LM is conditioned on the tags by explicitly including the tags in the training
sentences, in a process called label linearization. DAGA (Ding et al., 2020) uses this process to generate new sentences including entity tags. However, these sentences tend to be ungrammatical and disfluent in nature when trained on a small number of examples. MELM (Zhou et al., 2022) uses this process to perform entity replacements through masked generation using a pretrained Masked Language Model, such that the replacements take sentence context into account, but does not introduce new context patterns into the dataset. Both of these methods are also able to add new entities to the training set. These methods outperform simpler methods like random substitution, random deletion, etc. However, these methods involve training a language model on limited data and are therefore still prone to annotation corruption, generating noisy training examples. This also severely limits the number of augmentations that can be added.

Moreover, Lin et al. 2020 show that, in the low-resource setting, decreasing contextual diversity in the training set more significantly impacts NER performance than reducing entity diversity. Therefore, contextual diversity is more important than entity regularity knowledge to generalize to unseen examples in the low-resource setting. While transformer language models are pretrained on large corpora, they are not able to sufficiently capture context patterns for NER when fine-tuned on a small number of examples. Contextual augmentation has been applied to the token-level task of Aspect Term Extraction by Li et al. 2020, who finetune MASS (Song et al., 2019) to generate new partial contexts for English text by replacing ‘O’ tags. However, this requires additional in-domain data. Moreover, MASS is only pretrained on 4 languages: English, German, French, and Romanian. Additionally, Shi et al. 2021 alter the content while maintaining sentence structure by performing random substitutions of same-label spans for few-shot structured prediction tasks where annotations contain structural information like part-of-speech tagging, dependency parsing, and constituency parsing.

In this work, we propose Contextual and Semantic Structure-based Interpolation (CASSI) as a text augmentation scheme for low-resource NER. With the goal of increasing contextual diversity and avoiding annotation corruption, we use dependency parse trees to identify subtrees containing the subject, object, or complement between pairs of semantically similar sentences. We then generate candidate augmentations by performing exchanges between all pairs of subject subtrees and all pairs of object/complement subtrees between the two sentences. To ensure semantic correctness and fluency, we filter the resulting candidates by scoring them with a language model, similar to rescoring methods used in Automatic Speech Recognition (Xu et al., 2018; Chan et al., 2016) and Neural Machine Translation (Gülçehre et al., 2015). We leverage an existing pretrained Masked Language Model to avoid training a language model on limited data. Furthermore, we take steps to reduce the bias from the language model against specific names and towards shorter sentences and high-frequency generic terms. Specifically, we pick the top $k$ scored sen-
Our contributions can be summarized as follows: (1) we present a novel text augmentation scheme that improves context diversity in the dataset while avoiding annotation corruption, (2) we show that our method consistently and significantly outperforms existing methods in multiple languages for monolingual and cross-lingual NER, and (3) we show that our method outperforms existing methods on noisy social media text.

2 Proposed Method

Our proposed method (shown in Figure 1) generates augmentations by combining a pair of semantically similar sentences to generate a new sentence while ensuring a high degree of fluency, contextual and semantic correctness. For every sentence in the training set, we first rank the remaining sentences by sentence similarity (Figure 1a; Section 2.1). Next, we generate multiple candidate augmentations by performing structure-based exchanges between the given sentence and a selected semantically similar sentence using their dependency parse trees (Figure 1b; Section 2.2). We then filter out a subset of the candidate augmentations by scoring them with a pretrained Masked Language Model. We select the final augmentation from this subset using a simple metric to promote specificity and diversity (Figure 1c; Section 2.3).

2.1 Sentence Selection

Since sentence semantics and context clues are important for detecting entities, for a given sentence in the training set we rank the remaining sentences using a sentence similarity metric. Therefore, to generate \( n \) augmentations, we use the top \( n \) sentences according to their similarity to the given sentence.

We use BERTScore (Zhang et al., 2020) as the sentence similarity metric. BERTScore is a token-wise metric used to evaluate text generation for tasks like machine translation and image captioning. It calculates Precision (P) and Recall (R) by matching tokens in the source and target sentences by cosine similarities between their embeddings retrieved from an MLM like BERT (Devlin et al., 2019). For a source sentence \( A \) with normalized embeddings \( \{a_1, ..., a_p\} \) and a target sentence \( B \) with normalized embeddings \( \{b_1, ..., b_q\} \):

\[
P = \frac{1}{p} \sum_{a_i \in A} \max_{b_j \in B} a_i^T b_j; R = \frac{1}{q} \sum_{b_j \in B} \max_{a_i \in A} a_i^T b_j
\]

The final score is the F1 score \( (F1 = 2 \frac{PR}{P+R}) \). We choose BERTScore because it is able to use existing MLMs directly, without additional modifications like calculating sentence-level embeddings (Reimers and Gurevych, 2019). Moreover, we expect the token-wise nature of BERTScore to make it more robust to noisy social media text for a token-level task like NER.

2.2 Generating Candidate Augmentations

Figure 2 shows an example of subtree exchanges to generate candidate augmentations. For a given sentence \( U \) and a selected sentence \( V \), we use dependency parsing to identify subject subtrees and object/complement subtrees. A subject subtree is the subtree of a verb that contains the verb’s subject, and an object/complement subtree is the subtree of a verb or preposition that contains the verb or preposition’s object or complement. To generate candidate augmentations, we perform exchanges within all possible pairs of subject subtrees between \( U \) and \( V \). Similarly, we perform exchanges within
all possible pairs of object/complement subtrees between \( U \) and \( V \). Therefore, each such exchange generates a pair of candidate augmentations.

We choose subjects, objects, and complements for exchanges as they are likely to contain entities and context clues for entities. We use subtrees as the structures to exchange as they contain more information than just the subject, object, or complement or the noun chunks containing them, thus generating more diverse candidate augmentations.

Since we consider all exchanges between \( U \) and \( V \) as candidate augmentations, we avoid selecting duplicate pairs of sentences. Therefore, if \( V \) is selected for \( U \), \( U \) is skipped when selecting sentences for \( V \). For sentences where the dependency parse cannot find at least one subject or one object, we replace entities in the sentence with random entities of the same type to avoid simply oversampling these sentences.

### 2.3 Filtering Candidates

To select the final augmentation, we filter the candidate augmentations in two steps: (1) scoring with a Language Model, and (2) filtering for specificity and diversity. We select just one augmentation for every pair of sentences to increase the diversity of the augmentations.

#### 2.3.1 Language Model Scoring

Language Model scoring is used in Automatic Speech Recognition and Machine Translation to ascertain the correctness of a sentence (transcription or translation) by calculating the probability of occurrence of each token in the sentence. We use LM scoring to score the candidate augmentations for every pair of sentences. We calculate the LM score using a pretrained MLM following Salazar et al. 2020. The score is calculated as the sum of the log-likelihoods of each token in the sentence given the sentence \( \{t_1, \ldots, t_m\} \), the LM score is calculated as:

\[
\text{LMScore}(S) = \frac{1}{M} \sum_{i=1}^{m} \log P(t_i \mid S - \{t_i\}; \Theta)
\]

where \( M \) is the number of words in the sentence, and \( \Theta \) represents the MLM’s parameters.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>LM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I will be staying at the hotel</td>
<td>-0.74</td>
</tr>
<tr>
<td>I will be staying at the Hilton</td>
<td>-1.29</td>
</tr>
<tr>
<td>John will be staying at the hotel</td>
<td>-1.75</td>
</tr>
<tr>
<td>John Doe will be staying at the hotel</td>
<td>-2.19</td>
</tr>
</tbody>
</table>

Table 1: Example of LM bias against specificity. The LM score decreases for more specific sentences. LM scores are pseudo-log-likelihoods calculated using XLM-R\textsubscript{BASE} following Salazar et al. 2020. The scores are normalized by the number of words in the sentence.

#### 2.3.2 Filtering for Specificity and Diversity

As shown in Table 1, LM scoring is inherently biased against specificity. Generic terms are assigned higher probabilities than specific names since they occur more frequently, which is detrimental to a task like NER. To reduce this bias and promote diversity in the augmentations, we pick the top \( k \) LM-scored candidates and further filter them with the geometric mean of the Jaccard similarities between the words in the candidates and the words in the original pair of sentences. Given a sentence \( U \), a sentence \( V \), and a candidate augmentation \( C \) as sets of words:

\[
\text{J-Score}(U, V, C) = \sqrt{\frac{|U \cap C||V \cap C|}{|U \cup C||V \cup C|}}
\]

The selected augmentation is therefore biased towards containing more words from both original sentences, thus increasing the overall specificity of the augmentation. For the same reason, the selected augmentation is likely to be more diverse than the other \( k - 1 \) candidates (since the selected augmentation is biased against being disproportionately more similar to one of the two original sentences).

### 2.4 Post-Processing

In the event that a multi-word entity gets split by dependency parsing (for example, only the last name is picked from a name containing both first and last names), the entity’s tag is fixed according to the IOB scheme to ensure that all entities (1) start with a B- tag, (2) end with an I- tag, and (3) only contain I- tags in between.

### 3 Experiments

We evaluate our augmentation scheme on low-resource monolingual and cross-lingual NER in multiple languages. We also evaluate our method
Table 2: Results of Monolingual NER on subsets of CoNLL 2002/2003. Results are reported as Micro-F1 / Macro-F1. Numbers in bold indicate best performance on noisy social media datasets. We compare our method against MELM (Zhou et al., 2022) and DAGA (Ding et al., 2020).

### 3.1 Datasets

For monolingual NER, we perform experiments on the CoNLL 2002/2003 (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) datasets in four languages: English (En), German (De), Spanish (Es), and Dutch (Nl). For each language, we generate subsets of the training set of size 100, 250, 500, 750, 1k, 2k, and 4k. To concretely study the effects of adding data, we sample the subsets progressively, i.e., each smaller subset is sampled from the next larger subset.

For zero-shot cross-lingual NER, we evaluate models trained using the subsets of the English training set and the standard English dev set on the German, Spanish, and Dutch test sets. Additionally, we evaluate the performance of our method in conditions where dependency parsing is likely to be error-prone. For this, we perform experiments on the following social media datasets: WeiboNER (Peng and Dredze, 2015) - a Chinese social media dataset containing 1.4k sentences in the training set, WNUT-2017 (Derczynski et al., 2017) - an English tweets dataset containing 3.9k sentences in the training set, and ReLDI-NormTagNER-hr 2.0 (Miličević and Ljubešić, 2016) - a Croatian tweets dataset.

### 3.2 Experimental Setting

#### 3.2.1 Augmentation

**Baseline Methods** For DAGA and MELM, we follow augmentation settings as described in Ding et al. 2020 and Zhou et al. 2022, respectively.

**CASSI** We use XLM-R BASE (Conneau et al., 2020) for both calculating the pairwise BERTScore and LM scoring. For dependency parsing, we use the following language-specific pipelines from SpaCy (Honnibal et al., 2020): en_core_web_trf for English, de_dep_news_trf...
Table 2 summarizes the results on monolingual NER. Our method consistently outperforms the baseline methods for all languages. There is significant improvement across languages for low-resource levels up to 1k. The improvement becomes marginal for subset sizes larger than 500 sentences, by randomly sampling from the set of augmentations of size \( N \). For each method, we stop adding augmentations if there is no improvement in the best dev set Micro-F1 for three consecutive increments in the number of augmentations. For this purpose, we track the average dev set Micro-F1 across the three runs, so as to report a single augmentation size in each case.

### 3.3 Results

**Monolingual NER** Table 2 summarizes the results on monolingual NER. Our method consistently outperforms the baseline methods for all languages. There is significant improvement across languages for low-resource levels up to 1k. The improvement becomes marginal for subset sizes of 2k and 4k for English, German, and Spanish. This is consistent with observations from Lin et al. 2020 that increasing context diversity in a clean text regular NER dataset has a limited effect on the performance for over 2k sentences when fine-tuning a pretrained transformer. However, the per-

### 3.2.2 NER Training

We perform all experiments using mBERT\textsubscript{BASE-uncased}-BiLSTM-CRF (Vasantharajan et al., 2022a,b). We train with a batch size of 32 for 50 epochs using the AdamW optimizer. All training sets up to size 500 (including augmentations) are trained with a learning rate of \( 2e-4 \) with the number of warm-up steps set to 200. For training sets larger than 500 sentences, we decrease the learning rate and increase the number of warm-up steps proportionately with the increase in training set size. For the CoNLL subsets, we average the performance on three randomly sampled subsets of the same size. For the social media NER datasets, we average the performance on three runs. We report both Micro-F1 and Macro-F1 in each case.

For all augmentation methods, given a gold training set of size \( N \), the number of augmentations we generate ranges from \( N \) up to 10\( N \). For gold training sets of size 1k and larger, we also generate 0.25\( N \), 0.5\( N \), and 0.75\( N \) augmentations, by randomly sampling from the set of augmentations of size \( N \). For each method, we stop adding augmentations if there is no improvement in the best dev set Micro-F1 for three consecutive increments in the number of augmentations. For this purpose, we track the average dev set Micro-F1 across the three runs, so as to report a single augmentation size in each case.
Table 4: Results of NER on Social Media Datasets. Results are reported as Micro-F1 / Macro-F1.

Performance continues to significantly improve for Dutch. We discuss this further in Section 4.1. On average, MELM and DAGA marginally improve performance for 1k sentences but marginally reduce performance for 2k and 4k sentences. Given that performance improvements from augmentations diminish as dataset size increases, we attribute this to the noise in the augmentations outweighing any diversity introduced by them.

Cross-lingual NER Table 3 summarizes the results on zero-shot cross-lingual NER with English as the source language. On average, our method outperforms both MELM and DAGA for all subset sizes. Interestingly, DAGA becomes competitive in the Micro-F1 for resource levels of 2k and 4k sentences. With enough examples, DAGA can produce better context patterns. While these hurt performance in the monolingual setting, they seem to help generalize to unseen entities in other languages. However, it more notably underperforms in the Macro-F1 compared to our method. We posit that this is because the auto-regressive generation of entity tags along with the sentences causes it to disproportionately generate more frequent tags. In the case of MELM, the lack of new context patterns and annotation corruption cause the performance to drop for 2k and 4k sentences.

NER on Noisy Text Table 4 summarizes the results of our experiments on social media datasets. Our method consistently outperforms baseline methods even for noisy social media text where dependency parsing is likely to be erroneous. Moreover, these datasets tend to contain a considerable number of mislabeled examples and a stilted distribution of tags (for example, the ‘person’ tag occurs a lot more frequently than other tags in these datasets), which can make learning a reliable distribution of labels difficult for label-conditioning methods.

4 Discussion

4.1 Performance on Dutch

Figure 3 shows the language-wise relative improvement of CASSI and MELM over the Gold-Only baseline for 500 and 1k sentences. For CASSI, across both resource levels, the relative improvements for English, German, and Spanish are very similar whereas the improvement for Dutch is significantly higher. Given that the datasets are from the same domain (news articles), the most salient difference in the experiment settings for these languages is the amount of pretraining they received in mBERT. mBERT is pretrained on the entire Wikipedia dump of the top hundred languages with the largest Wikipedias. The English, German, and Spanish Wikipedias contain 3.3B, 1.1B, and 800M words respectively, however, the Dutch Wikipedia is considerably smaller at 300M words. We suspect that Dutch gains more from contextual diversity in the augmentations due to lower representation during pretraining. This is consistent with observations from Lin et al. 2020 who conclude that pretraining helps capture context patterns when finetuning for NER. This implies that context diversity in the augmentations is more important for low-resource languages with low representation during pretraining. Figure 3 also shows that MELM, an entity replacement method that does not introduce new context patterns, does not follow this pattern.

\[^3\]https://github.com/google-research/bert/blob/master/multilingual.md

4.2 Ablations

We perform ablation studies on the 1k subsets of CoNLL 2002/2003 to evaluate the effect of (1) matching sentences by similarity, and (2) using the J-Score filter for specificity. To assess the quality of the augmentations, we compare the performance on the test set upon adding 10N augmentations (where N = 1k). As mentioned in Section 3.2.2, for each language, we average the results on three sampled subsets of size 1k.

4.2.1 BERTScore

Table 5 compares using BERTScore to select semantically similar sentences with selecting sentences at random. On average, matching sentences by similarity outperforms random selection.

4.2.2 J-Score Filter

We compare the performance of selection from the top k candidates using the J-Score metric for specificity against (1) selecting the candidate with the best LM score, and (2) randomly selecting a candidate from the top k. Table 6 shows these results. On average, filtering using the J-score metric outperforms the other methods. It notably outperforms random selection from top k, showing that using the J-score metric does indeed select sentences with useful properties.

4.3 Quality of Augmentations

Here, we consider the quality of our augmentations by (1) comparing the number of performance-improving augmentations added by each method, and (2) comparing the ‘naturalness’ (via LM scoring) of our augmentations to the gold sentences.

4.3.1 Number of Augmentations

Table 7 summarizes the number of augmentations added for each method on the CoNLL 2002/2003 subsets. We are able to add considerably more performance-improving augmentations than baseline methods. Note that we do not go beyond 10N augmentations in any case.

Additionally, we evaluated all methods on all augmentation increments up to 10N on the 500 sentence subsets of CoNLL 2002/2003 to compare the dev set Micro-F1 for every method. The results are summarized in Figure 4. Our method outperforms the baseline methods on all augmentation increments for all languages.

4.3.2 Naturalness

We performed LM scoring on the 1k subsets of CoNLL 2002/2003, and 1k augmentations generated from them. We scored the sentences using mBERTBASE-uncased (we do not use XLM-RBASE since the augmentations were selected using XLM-RBASE) following Salazar et al. 2020 and normalized them by the number of words (ignoring punctuation). The augmentations are scored separately, i.e., they are not added to the gold-only set. The LM Scores are pseudo-log-likelihoods, with greater values being better. We present the mean and standard deviation of the LM scores over the sets. Table 8 shows that the augmentations score very similarly (albeit marginally lower on average) to the gold sentences.
Figure 4: Dev set Micro-F1 vs Number of Augmentations for different methods on the 500 sentence subsets of CoNLL. Best viewed in color.

Table 8: LM Scores on 1k gold sentences and 1k augmentations. LM scores are pseudo-log-likelihoods calculated using mBERT$_{BASE}$-cased following Salazar et al. 2020. The scores are normalized by the number of words in the sentence.

5 Conclusion

In this work, we propose CASSI as a text augmentation scheme for low-resource NER. CASSI produces contextually diverse sentences while preserving entity tags by structurally combining a pair of semantically similar sentences through the generation of candidate combinations that are filtered for semantic correctness and specificity. It leverages dependency parsers for structural information and a pretrained MLM for scoring. Through extensive experiments, we show that it outperforms baseline methods for multiple languages in monolingual and cross-lingual settings. Furthermore, we show that it is also able to outperform baseline methods under noisy text conditions where dependency parsing is likely to be unreliable.

Limitations

The proposed method requires a dependency parser in the target language which might not be available for some extremely low-resource languages. Currently, Universal Dependencies contains treebanks for 141 languages\(^5\) (Nivre et al., 2020) and SpaCy only supports 24 languages.

For high-resource languages, increasing context diversity provides limited performance improvements for clean text regular NER outside of the low-resource setting.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-GC-2022-005). We would like to acknowledge the High Performance Computing Centre of Nanyang Technological University Singapore, for providing the computing resources, facilities, and services that have contributed significantly to this work.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, N. Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In AAAI Conference on Artificial Intelligence.


Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

\(^5\)https://universaldependencies.org/


Charangan Vasantharajan, Kyaw Zin Tun, Min Zhi Hao, and Chng Eng Siong. 2022a. NERLP. GitHub.


Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with


A Measuring Semantic Diversity of the Augmentations

We measure the semantic diversity of the augmentations when added to the gold sentences using Self-BERTScore, a diversity metric similar to Self-BLEU (Zhu et al., 2018)6. We score the 1k sets of CoNLL 2002/2003 using mBERTBASE-cased in three cases: (i) Gold-Only, (ii) Gold-Only along with 5k augmentations filtered using the J-score filter, and (iii) Gold-Only along with 5k augmentations picked using the Best LM Score. Table 9 shows that augmentations selected using the J-Score metric are more diverse than those selected using the Best LM Score. (The intuition behind this

<table>
<thead>
<tr>
<th>Set</th>
<th>En 1k</th>
<th>De 1k</th>
<th>Es 1k</th>
<th>Ni 1k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold-Only</td>
<td>0.593 ± 0.057</td>
<td>0.622 ± 0.038</td>
<td>0.625 ± 0.055</td>
<td>0.627 ± 0.049</td>
</tr>
<tr>
<td>J-Score</td>
<td>0.599 ± 0.054</td>
<td>0.624 ± 0.034</td>
<td>0.633 ± 0.048</td>
<td>0.623 ± 0.042</td>
</tr>
<tr>
<td>Best LM Score</td>
<td>0.601 ± 0.054</td>
<td>0.629 ± 0.036</td>
<td>0.637 ± 0.048</td>
<td>0.625 ± 0.044</td>
</tr>
</tbody>
</table>

Table 9: Self-BERTScores calculated upon adding 5x augmentations to 1k gold-only sentences. Lower scores imply greater diversity.

may be clear from the following example: given two gold sentences X and Y, 10 words long each with no overlap in words, and two candidate augmentations C1, which contains 5 words from X and 5 words from Y, and C2, which contains 7 words from X and 3 words from Y; intuitively, out of the two, C1 would be more diverse. This is reflected in the J-Scores: C1 has a J-Score of 0.333, whereas C2 has a J-Score of 0.309. Moreover, if the sentences in the original dataset are already well formed - which is the case in CoNLL - LM Scoring is likely to rank C2 over C1.)

Note that the Self-BERTScores of the augmentations are marginally higher than the gold-only sentences. This is expected for two reasons: (1) we pick the most semantically similar sentences for our augmentations, and (2) our augmentations consist of parts of the gold sentences used directly, i.e., the same words in the same order.

B Alternatives to Pairwise BERTScore

While pairwise BERTScore (calculation of BERTScore for every pair of sentences in the dataset) is efficient enough for small datasets, we recognize that using pairwise BERTScore becomes time-consuming for larger datasets since it involves token-wise operations for every pair of tokens in the two sentences. For this reason, we suggest the following alternatives for larger datasets: (1) pairwise cosine similarity using sentence embeddings (Reimers and Gurevych, 2019), and (2) randomly sampling a fixed number of sentences for each source sentence, and selecting the most semantically similar sentences from these random samples using BERTScore or sentence similarity.

C Examples of Generated Augmentations

Table 10 shows some example augmentations taken from a single 1k subset of CoNLL 2003 English. The examples show sentence interpolation, instances where the augmentation process leads to simple short phrase or word replacement, and instances of grammatical incorrectness/incoherence.

---

6Given the nature of our augmentations, the embedding-based BERTScore is a better metric than an n-gram based metric like BLEU. Furthermore, BERTScore outperforms BLEU at measuring diversity, though it still significantly underperforms on human evaluation (Tevet and Berant, 2021). Moreover, our augmentations essentially serve as an edge case for automated metrics since we do not formulate entirely new sentences and instead copy exact phrases from the source sentences.
in the augmentations. (Note that in the examples of simple phrase/word replacements the augmentations can still introduce entities with different tags or introduce an entity in place of a non-entity.)
The European Commission said on Thursday it disagreed with German advice to consumers to shun British lamb until scientists determine whether mad cow disease can be transmitted to sheep.

Bonh has led efforts to protect public health after consumer confidence collapsed in March after a British report suggested humans could contract an illness similar to mad cow disease by eating contaminated beef.

Opel AG together with General Motors came in second place with 49,269 registrations, 16.4 percent of the overall figure.

The BLOM Stock Index which covers both markets rose 0.04 percent to 903.09 and the LISPI index rose 0.02 percent to 81.58.

Defence Minister Volker Ruehe said that German troops would stay on in Bosnia next year as part of an international force to ensure the establishment of peace, a newspaper reported on Saturday.

Iraqi President Saddam Hussein has told visiting Russian ultra-nationalist Vladimir Zhirinovsky that Baghdad wanted to maintain friendship and cooperation with Moscow, official Iraqi newspapers said on Thursday.

They said Zhirinovsky told Saddam before he left Baghdad on Wednesday that his Liberal Democratic party and the Russian Duma (parliament) are calling for an immediate lifting of the embargo imposed on Iraq after its 1990 invasion of Kuwait.

They said German troops told Saddam before he left Baghdad on Wednesday that his Liberal Democratic party and the Russian Duma (parliament) are calling for an immediate lifting of the embargo imposed on Iraq after its 1990 invasion of Kuwait.

Maronite Patriarch Sfeir expressed sorrow over the violations in Sunday’s elections.

Lebed, who has arranged a military truce with separatist rebels in the southern Russia region, was in Moscow this week seeking support for a deal on Chechnya’s political status.

Last October he was invited to attend the referendum held on Iraq’s presidency, which extended Saddam’s term for seven more years.

American world champion Gwen Torrence, the bronze medallist in Atlanta, was second in 11.00.

Hever Golf Rose (11-4), last year’s Prix de l’Abbaye winner at Longchamp, finished third, a further one and a quarter lengths away with the 7-4 favourite Mind Games in fourth.

European champions Juventus will face English league and cup double winners Manchester United in this season’s European Champions’ League.

Ireland midfielder Roy Keane has signed a new four-year contract with English league and F.A. Cup champions Manchester United.

The Netherlands government has ruled out paying ransom money for a Dutch couple kidnapped from their farm, while Costa Rican authorities said on Tuesday they had no leads in the case.

Advanced Medical said it expects to take an unspecified one-time charge to pay for the merger.

Gold Sentence Pair Augmentation

<table>
<thead>
<tr>
<th>Gold Sentence Pair</th>
<th>Augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>The European Commission said on Thursday it disagreed with German advice to</td>
<td>The European Commission said on Thursday humans could contract an illness</td>
</tr>
<tr>
<td>consumers to shun British lamb until scientists determine whether mad cow disease</td>
<td>similar to mad cow disease by eating contaminated beef.</td>
</tr>
<tr>
<td>can be transmitted to sheep.</td>
<td></td>
</tr>
<tr>
<td>Bonh has led efforts to protect public health after consumer confidence collapsed</td>
<td>Bonh has led efforts to protect public health after consumer confidence</td>
</tr>
<tr>
<td>in March after a British report suggested humans could contract an illness</td>
<td>collapsed in March after a British report suggested humans could contract</td>
</tr>
<tr>
<td>similar to mad cow disease by eating contaminated beef.</td>
<td>an illness similar to mad cow disease by eating contaminated beef.</td>
</tr>
<tr>
<td>Opel AG together with General Motors came in second place with 49,269 registrations,</td>
<td>Opel AG together with General Motors rose 0.04 percent to 903.09 and the</td>
</tr>
<tr>
<td>16.4 percent of the overall figure.</td>
<td>LISPI index rose 0.02 percent to 81.58.</td>
</tr>
<tr>
<td>Defence Minister Volker Ruehe said that German troops would stay on in Bosnia</td>
<td>Defence Minister Volker Ruehe said that German troops would stay on in Bosnia</td>
</tr>
<tr>
<td>next year as part of an international force to ensure the establishment of peace,</td>
<td>next year as part of to maintain friendship and cooperation with Moscow a</td>
</tr>
<tr>
<td>a newspaper reported on Saturday.</td>
<td>newspaper reported on Saturday.</td>
</tr>
<tr>
<td>Iraqi President Saddam Hussein has told visiting Russian ultra-nationalist</td>
<td>They said Saddam before he left Baghdad on Wednesday that his Liberal</td>
</tr>
<tr>
<td>Vladimir Zhirinovsky that Baghdad wanted to maintain friendship and cooperation</td>
<td>Democratic party and the Russian Duma (parliament) are calling for an</td>
</tr>
<tr>
<td>with Moscow, official Iraqi newspapers said on Thursday.</td>
<td>immediate lifting of the embargo imposed on Iraq after its 1990 invasion of</td>
</tr>
<tr>
<td>They said Zhirinovsky told Saddam before he left Baghdad on Wednesday that his</td>
<td>Kuwait.</td>
</tr>
<tr>
<td>Liberal Democratic party and the Russian Duma (parliament) are calling for an</td>
<td>They said German troops told Saddam before he left Baghdad on Wednesday</td>
</tr>
<tr>
<td>immediate lifting of the embargo imposed on Iraq after its 1990 invasion of Kuwait.</td>
<td>that his Liberal Democratic party and the Russian Duma (parliament) are</td>
</tr>
<tr>
<td>Defence Minister Volker Ruehe said that German troops would stay on in Bosnia</td>
<td>calling for an immediate lifting of the embargo imposed on Iraq after its</td>
</tr>
<tr>
<td>next year as part of an international force to ensure the establishment of peace,</td>
<td>1990 invasion of Kuwait.</td>
</tr>
<tr>
<td>a newspaper reported on Saturday.</td>
<td></td>
</tr>
<tr>
<td>Lebed, who has arranged a military truce with separatist rebels in the southern</td>
<td>Lebed, who has arranged a military truce with separatist rebels in the</td>
</tr>
<tr>
<td>Russia region, was in Moscow this week seeking support for a deal on Chechnya’s</td>
<td>southern Russia region, was in Moscow this week seeking to attend the</td>
</tr>
<tr>
<td>political status.</td>
<td>referendum held on Iraq’s presidency which extended Saddam’s term for</td>
</tr>
<tr>
<td>Last October he was invited to attend the referendum held on Iraq’s presidency,</td>
<td>seven more years.</td>
</tr>
<tr>
<td>which extended Saddam’s term for seven more years.</td>
<td></td>
</tr>
<tr>
<td>Africans seeking to renew or obtain work and residence rights say Prime Minister</td>
<td>Africans seeking to renew or obtain work and residence rights was asked at a</td>
</tr>
<tr>
<td>Alain Juppe’s proposals are insufficient as hunger strike enters 49th day in</td>
<td>news conference if India’s decision to block adoption of the accord in Geneva</td>
</tr>
<tr>
<td>Paris church and Wednesday rally attracts 8,000 sympathisers.</td>
<td>would lead to an arms race with Pakistan and China.</td>
</tr>
<tr>
<td>Foreign Minister I.K. Gujral was asked at a news conference if India’s decision</td>
<td></td>
</tr>
<tr>
<td>to block adoption of the accord in Geneva would lead to an arms race with Pakistan</td>
<td></td>
</tr>
<tr>
<td>and China.</td>
<td></td>
</tr>
<tr>
<td>American world champion Gwen Torrence, the bronze medallist in Atlanta, was</td>
<td>American world champion Gwen Torrence, the bronze medallist in Atlanta, was</td>
</tr>
<tr>
<td>second in 11.00.</td>
<td>second in Longchamp.</td>
</tr>
<tr>
<td>Hever Golf Rose (11-4), last year’s Prix de l’Abbaye winner at Longchamp,</td>
<td></td>
</tr>
<tr>
<td>finished third, a further one and a quarter lengths away with the 7-4 favourite</td>
<td></td>
</tr>
<tr>
<td>Mind Games in fourth.</td>
<td></td>
</tr>
<tr>
<td>European champions Juventus will face English league and cup double winners</td>
<td>European champions Juventus will face a new four-year contract with English</td>
</tr>
<tr>
<td>Manchester United in this season’s European Champions’ League.</td>
<td>league and F.A. Cup champions Manchester United in this season’s European</td>
</tr>
<tr>
<td>Irish midfielder Roy Keane has signed a new four-year contract with English league</td>
<td>Champions’ League.</td>
</tr>
<tr>
<td>and F.A. Cup champions Manchester United.</td>
<td></td>
</tr>
<tr>
<td>The Netherlands government has ruled out paying ransom money for a Dutch couple</td>
<td>The Netherlands government has ruled out if expects to take an unspecified</td>
</tr>
<tr>
<td>kidnapped from their farm, while Costa Rican authorities said on Tuesday they had</td>
<td>one-time charge to pay for the merger, while Costa Rican authorities said on</td>
</tr>
<tr>
<td>no leads in the case.</td>
<td>Tuesday they had no leads in the case.</td>
</tr>
<tr>
<td>Advanced Medical said it expects to take an unspecified one-time charge to pay for</td>
<td></td>
</tr>
<tr>
<td>the merger.</td>
<td></td>
</tr>
</tbody>
</table>

Table 10: Example augmentations generated from a 1k subset of CoNLL 2003 English. Instances of grammatical incorrectness/incoherence are underlined in red.