Unsupervised Neural Machine Translation with Universal Grammar

Zuchao Li[1,2,3,†], **Masao Utiyama**[4,*], **Eiichiro Sumita**[4], **and Hai Zhao**[1,2,3*]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[4]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
charlee@sjtu.edu.cn, {mutiyama, eiichiro.sumita}@nict.go.jp, zhaohai@cs.sjtu.edu.cn

Abstract

Machine translation usually relies on parallel corpora to provide parallel signals for training. The advent of unsupervised machine translation has brought machine translation away from this reliance, though performance still lags behind traditional supervised machine translation. In unsupervised machine translation, the model seeks symmetric language similarities as a source of weak parallel signal to achieve translation. Chomsky's Universal Grammar theory postulates that grammar is an innate form of knowledge to humans and is governed by universal principles and constraints. Therefore, in this paper, we seek to leverage such shared grammar clues to provide more explicit language parallel signals to enhance the training of unsupervised machine translation models. Through experiments on multiple typical language pairs, we demonstrate the effectiveness of our proposed approaches.

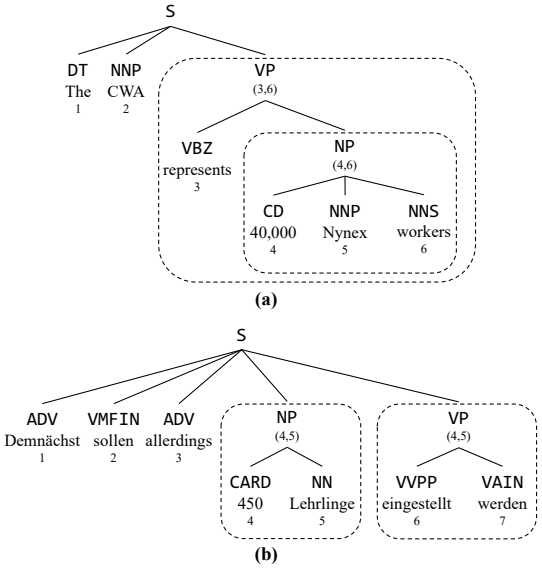

Figure 1: Examples of constituent trees from English Penn Treebank (PTB) and German dataset of SPMRL14 shared task. The dotted box indicates the constituents that can be masked for prediction.

1 Introduction

Recently, Neural Machine Translation (NMT) (Bahdanau et al., 2014; Sutskever et al., 2014) has been greatly developed and become the dominant paradigm in machine translation. On the one hand, the development of deep neural networks such as Transformer (Vaswani et al., 2017; Li et al., 2021a) has played a significant role in NMT's improvements. On the other hand, large-scale parallel corpora like the UN corpus (Ziemski et al., 2016) have also played an important role.

Despite the recent success of NMT in standard benchmarks, the need for large-scale parallel corpora has limited the effectiveness of NMT in many language pairs, especially in low-resource language pairs (Koehn and Knowles, 2017). Unsupervised Neural Machine Translation (UNMT) (Artetxe et al., 2018b) was proposed to alleviate this issue by completely removing the need for parallel data and training an NMT system in a completely unsupervised manner, relying on nothing but monolingual corpora. Unsupervised machine translation does not need the parallel information from parallel sentences; rather, it generally uses embedding alignments, initializes parameters with pre-trained language models, and uses iterative back-translation between two languages to synthesize pseudo parallel corpora for model training (Lample et al., 2018a,c; Yang et al., 2018; Sun et al., 2019; Conneau and Lample, 2019; Li et al., 2020a).

The pseudo parallel data created by iterative back-translation is the key to the success of unsupervised NMT model training (Kim et al., 2020). It takes advantage of the equivalence of translation languages to bring supervision (albeit weak supervision) to model training. Recent results in

*Corresponding author. † This paper was finished when Zuchao Li was a fixed term technical researcher at NICT.
This work was supported by Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3249–3264
November 7–11, 2021.

semi-supervised NMT have demonstrated that further training a UNMT model with true bilingual parallel sentences can lead to better translation performance (He et al., 2016; Kim et al., 2020; Conneau and Lample, 2019; Song et al., 2019a), which suggests that after training, UNMT models are still not optimized because of their lack of explicit supervision.

Universal grammar (UG) is a notion in linguistics and philosophy that goes back at least to Roger Bacon's observation, "*in its substance, grammar is one and the same in all languages, even if it accidentally varies*" (Bacon, 1902). Chomsky (1965a,b) developed a universal grammar theory. The idea of a universal grammar states that all human languages are species of a common genus because they have all been shaped by a factor that is common to all human beings (Lappin and Shieber, 2007; Nivre, 2015). Therefore, in this paper, we leverage this grammar commonality to derive additional supervision to enhance UNMT training. In other words, our proposed method is built on the existence of universal grammar. If there is no cross-lingual commonality and definitional similarity in the syntactic structure, then we will not be able to obtain weakly supervised signals for UNMT.

Specifically, we choose the grammar representation framework of constituent syntax as the research object. Unlike typical approaches to leveraging syntax information, rather than adopting a syntactic encoder to enhance representations, we focus on acquiring more supervision by finding commonalities between two languages' syntaxes and demonstrate this supervision by training UNMT models. Since different languages often share some of the same constituent types (syntax categories), predicting these matching constituents in model training can be used for a weak alignment. As shown in Figure 1, although the two sentences are not parallel, during the training, the model is exposed to both `NP` and `VP` constituents, and a weak alignment between these constituents can be used to enhance the UNMT training, i.e., the `NP` constituents in English and the `NP` constituents in German (the same to `VP`, `PP`, ..., etc.) are more likely to be parallel. Notably, our method is only an application of Universal Grammar in UNMT, but far from all applications since we only leverage a very small part of Universal Grammar (universal constituent and syntactic label definition).

Masked Language Modeling (MLM) is a commonly used training approach for language modeling. In MLM, some of the tokens in the sentence are masked, and then the model is required to predict these masked tokens at their placeholders. Based on MLM, we propose a CONSTMLM approach that also draws from constituent syntax. In our CONSTMLM, constituents are masked, and the model is tasked with predicting both the tokens in a constituent and the constituent's syntactic category. Masking large constituents will present too difficult a problem for the model, as there will be insufficient context, so we also propose BTLM, a method of leveraging back-translation to provide more context and alleviate this issue. We then implement CONSTBTLM based on the CONSTMLM, which leverage our proposed BTLM. To accommodate both UNMT and language modeling training, we have prepared both encoder-decoder models and encoder-only models for our CONSTBTLM, BTLM, and CONSTMLM approaches.

In our experiments, we demonstrated the effectiveness of leveraging universal grammar and of our proposed approaches on multiple unsupervised translation tasks. Our proposed approaches show consistent improvements compared to the baselines in these tasks. We also present a significantly boosted performance on several low-resource semi-supervised tasks. These results verify that universal grammar commonalities can bring additional supervision information to bolster the training of unsupervised and low-resource translation models.

2 The Proposed Approaches

2.1 Background

We formally present the background of our baseline UNMT system in terms of unsupervised machine translation between languages L_1 and L_2. Our UNMT model follows an encoder-decoder architecture as in standard NMT. We use a joint subword (Sennrich et al., 2016b) vocabulary shared between languages and share parameters between source→target and target→source models to take advantage of multilingualism (Edwards, 2002). In this framework, three training methods are indispensable for the feasibility of unsupervised machine translation: initialization, denoising generation, and iterative back-translation. UNMT models typically use denoising generation and iterative back translation simultaneously by alternating between the two methods in a single phase rather than separately in multiple phases. The model is

given monolingual data $\{X_i\}$ in language L_1 and $\{Y_j\}$ in language L_2. $|X|$ and $|Y|$ are the number of sentences in monolingual data $\{X_i\}$ and $\{X_j\}$, respectively.

Initialization Initialization is a crucial step for bootstrapping UNMT models. The initialization process injects non-randomized cross- or multilingual knowledge into a UNMT model. In general, two types of initialization are usually adopted (Lample et al., 2018c). The first entails initializing the embedding layer of a UNMT model with pre-trained embeddings, while the second uses a pre-trained language model with the same structure as the UNMT encoder to initialize the embedding layer and most of the neural network parameters in the encoder and decoder (Conneau and Lample, 2019). The experimental performance in (Conneau and Lample, 2019) shows that using a pre-trained language model to initialize a UNMT model can produce better performance, so we choose this as our method of initialization.

Denoising Generation Denoising generation training aims to help UNMT models learn to generate fluent texts. Noise is introduced to input sentences via replace, delete, and shuffle functions, and then the UNMT model is tasked with encoding these noisy sentences and using the encoded noisy sentences to reconstruct the original sentences. The UNMT model is optimized by loss $\mathcal{L}_D$ during this training process:

$$\begin{aligned}\mathcal{L}_D = &\sum_{i=1}^{|X|} -\log P_{L_1 \to L_1}(X_i|N(X_i),\theta) \\ &+ \sum_{j=1}^{|Y|} -\log P_{L_2 \to L_2}(Y_j|N(Y_j),\theta),\end{aligned} \tag{1}$$

where $N(\cdot)$ refers to the noise functions and θ represents the UNMT model parameters. $P_{L_1 \to L_1}$ and $P_{L_2 \to L_2}$ denote the reconstruction probabilities in the languages L_1 and L_2, respectively.

Iterative Back-translation Back-translation (Sennrich et al., 2016a) was first proposed to boost translation performance using target-side monolingual data. By using symmetric models, it can boost translation in both directions. In UNMT, back-translation is used to synthesize pseudo parallel data from monolingual text, which alleviates the scarcity of true parallel data. This synthesis is performed repeatedly throughout the UNMT training. The loss, $\mathcal{L}_B$, is defined as follows:

$$\begin{aligned}\mathcal{L}_B = &\sum_{i=1}^{|X|} -\log P_{L_2 \to L_1}(X_i|S_{L_1 \to L_2}(X_i,\theta),\theta) \\ &+ \sum_{j=1}^{|Y|} -\log P_{L_1 \to L_2}(Y_j|S_{L_2 \to L_1}(Y_j,\theta),\theta),\end{aligned} \tag{2}$$

where $S_{L_1 \to L_2}$ and $S_{L_2 \to L_1}$ represent the translation processes from L_1 to L_2 and L_2 to L_1, respectively. $P_{L_1 \to L_2}$ and $P_{L_2 \to L_1}$ denote the translation probabilities between the two languages.

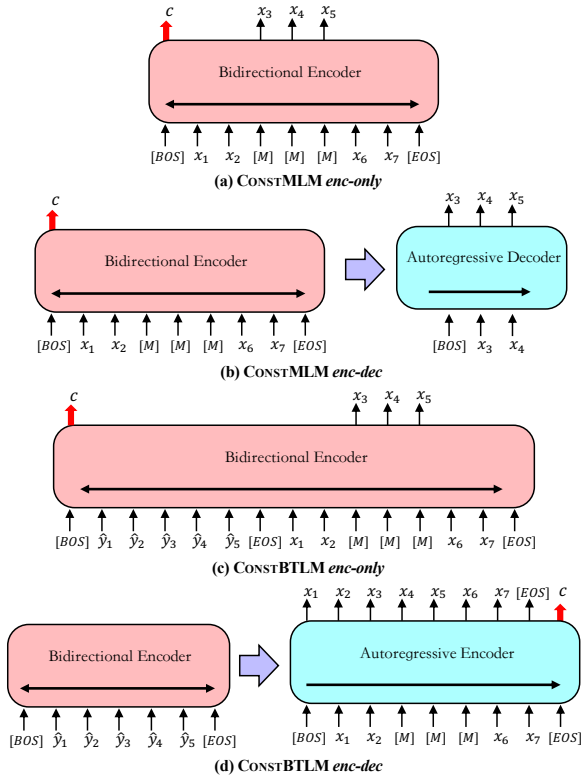

Figure 2: Schema of our proposed ConstMLM *enc-only*, ConstMLM *enc-dec*, ConstBTLM *enc-only*, and ConstBTLM *enc-dec*.

2.2 ConstMLM

We propose Constituent Masked Language Modeling (ConstMLM) in this section. ConstMLM is a variant of MLM that is enhanced with constituent syntax information. In traditional MLM, given a sentence $X = \{x_1, x_2, ..., x_n\}$, length of tokens n, and set of masked positions $\mathcal{M}$, the training loss $\mathcal{L}_{MLM}$ for the MLM training is:

$$\mathcal{L}_{MLM} = \sum_{i=1}^{|\mathcal{M}|} -\log P(x_{\mathcal{M}_i}|X_{\setminus\mathcal{M}},\theta) \tag{3}$$

where $|\mathcal{M}|$ is the size of set $\mathcal{M}$, and $X_{\setminus\mathcal{M}}$ indicates the sequence after masking. The masked positions

set $\mathcal{M}$ consists of randomly sampled discrete positions, that is, $\mathcal{M} = \mathbf{TopK}([\mathbf{rand}_i(0,1)]_{i=1}^n)$. Here, $\mathbf{TopK}$ is a function that selects positions by probability until the masking budget has been spent. In span-based MLM like (Joshi et al., 2020), a span of length ℓ is first sampled from a geometric distribution $\ell \sim \mathrm{Geo}(p)$, and the start position of a span is sampled in the same manner as in MLM, giving final masked span set $\mathcal{M}_S = \{(\mathcal{M}_i, \ell_i)\}$. In another linguistically guided language modeling approach, Zhou et al. (2020b) proposed Syntactic/Semantic Phrase Masking (SPM) for their model LIMIT-BERT. In SPM, the masked positions set consists of tuples randomly sampled from the linguistic span set instead of the discrete token position set. Only the span boundary information, however, is used in SPM; the linguistic label is ignored, so we remedy this and propose ConstMLM.

In ConstMLM, we first extract and filter the constituent span set $\mathbf{CS} = \{(s, e, c)_i\}_{i=1}^m$, where s, e, and c represent the start position, end position, and syntactic category, respectively. During filtering, constituent parse trees with a span ratio greater than $\gamma = \ell/n$ are removed. Random sampling is also performed on this set to obtain the masked span set. Unlike SpanBERT and LIMIT-BERT, we only sample one span at a time because ConstMLM not only predicts the masked token in the sampled span but also predicts the syntactic category of the sampled span. ConstMLM sums the loss from both the span's syntactic category and the regular masked language model objective for each token in the masked span:

$$\begin{aligned}\mathcal{L}_{\text{ConstMLM}} = \sum_{e:s} &- \log P(x_i | X_{\setminus s:e}, \theta) \; + \\ &- \log P(c | X_{\setminus s:e}, \theta).\end{aligned} \tag{4}$$

Since the UNMT model architecture, which includes both an encoder and a decoder, is different from pre-trained language models in general, we provide two implementations of ConstMLM: *encoder-only* and *encoder-decoder*. In the *encoder-only* ConstMLM, the masked span's token and syntactic category prediction are both performed on the encoder side, which is no different from popular pre-trained language models such as BERT that only consist of encoders. Both target prediction probabilities are calculated using the following process:

$$\begin{aligned}P(x_i | X_{\setminus s:e}, \theta) =& \mathbf{Softmax}(\mathbf{MLP}(enc(X_{\setminus s:e}))), \\ P(c | X_{\setminus s:e}, \theta) =& \mathbf{Softmax}(\mathbf{MLP}(\\ & \mathbf{Pooling}(enc(X_{\setminus s:e})))),\end{aligned}$$

where $enc(\cdot)$ represents the encoding process, and $\mathbf{Pooling}(\cdot)$ is a pooling operation that uses a first-token pooling strategy.

In the *encoder-only* ConstMLM, only the encoder is updated by the loss; the decoder can not benefit from it. Using the same training method on the decoder as on the encoder is not viable; because the decoder uses incremental self-attention instead of full self-attention. To mitigate this, we propose an *encoder-decoder* ConstMLM, in which the masked token prediction probability is calculated as:

$$\begin{aligned}P(x_i | X_{\setminus s:e}, \theta) = \mathbf{Softmax}(&\mathbf{MLP}(\\ & dec([\langle \text{BOS} \rangle, X_{s:e-1}], enc(X_{\setminus s:e})))),\end{aligned}$$

where $dec(\cdot)$ represents the decoding process, and $[\langle \text{BOS} \rangle, X_{s:e-1}]$ is the operation of prepending a $\langle \text{BOS} \rangle$ token before sequence $X_{s:e-1}$. In *encoder-decoder* ConstMLM, the encoder still handles the incomplete sentence encoding, so the syntactic category prediction is consistent with that of the *encoder-only* version. This means that the weak alignment information brought by the syntactic category still directly trains the encoder, while the decoder is optimized by the span generation process.

2.3 BTLM and ConstBTLM

Whether in traditional MLM or span-based MLM, the number of tokens masked is limited to a certain ratio of the sentence. In BERT's implementation, at most 15% of the tokens are put up for masking. SpanBERT followed this practice and after obtaining span lengths by sampling a geometric distribution skewed towards shorter spans, removed spans with a length greater than $\ell_{max} = 10$. Skewing towards shorter spans is crucial because of an issue in MLM: if too many tokens are masked, it is difficult for the model to recover these tokens using the remaining incomplete sentences. Limiting the number of masked tokens is especially important for span-based MLM, as spans can compose much larger parts of the sentence.

We call this the difficulty of reasoning with insufficient information. This situation is still acceptable for language model pre-training, and limiting the maximum ratio of masked tokens in MLM and the span length in span-based MLM alleviates the issue, but for linguistically-guided span-based MLM, the length of the extracted span cannot be flexibly set because it contains specific grammatical information. Making the maximum span width

too small means too few spans or even no spans for some trees are extracted. To combat the difficulty of reasoning with insufficient information, we first propose Back-translation Language Modeling (BTLM), a training method that can use cross-lingual translation as a source of information for inference. It can be formally presented as:

$$\mathcal{L}_{\text{BTLM}} = \sum_{e:s} -\log P(x_i|X_{\setminus s:e}, S_{L_1\to L_2}(X), \theta), \quad (5)$$

In BTLM, the sentence X in language L_1 is first translated into language L_2 by $S_{L_1\to L_2}$ for use as cross-lingual context. Then, X is masked as in MLM. Finally, the target prediction is performed by combining and considering the cross-lingual context and the MLM context. Due to the existence of a complete (albeit noisy) cross-lingual context, the proportion of masked spans in a sentence can be significantly increased. In addition, this training forces the model to infer with a cross-language context, which implicitly promotes bilingual alignment.

Based on BTLM, as CONSTMLM was built on MLM, we propose Constituent Back-translation Language Modeling (CONSTBTLM). The loss of CONSTBTLM is calculated similarly to that of CONSTMLM:

$$\mathcal{L}_{\text{CONSTBTLM}} = \sum_{e:s} -\log P(x_i|X_{\setminus s:e}, S_{L_1\to L_2}(X), \theta) + -\log P(c|X_{\setminus s:e}, S_{L_1\to L_2}(X), \theta).$$

We also implemented *encoder-only* and *encoder-decoder* versions with CONSTBTLM for different purposes. In *encoder-only* CONSTBTLM, the target prediction probability becomes:

$$P(x_i|X_{\setminus s:e}, \hat{Y}, \theta) = \textbf{Softmax}(\textbf{MLP}(enc([\hat{Y}, X_{\setminus s:e}]))),$$
$$P(c|X_{\setminus s:e}, \hat{Y}, \theta) = \textbf{Softmax}(\textbf{MLP}(\textbf{Pooling}(enc([\hat{Y}, X_{\setminus s:e}])))),$$

where $\hat{Y} = S_{L_1\to L_2}(X)$, and $[\hat{Y}, X_{\setminus s:e}]$ indicates that the translated sequence $\hat{Y}$ is prepended to the rest of the sequence. Purely from an implementation perspective, the use of cross-lingual context here is consistent with the TLM proposed in (Conneau and Lample, 2019), but the difference is that we only mask the input monolingual sequence, while TLM masks both the input parallel sentences.

Correspondingly, in the *encoder-decoder* CONSTBTLM, the probabilities are calculated as:

$$P(x_i|X_{\setminus s:e}, \hat{Y}, \theta) = \textbf{Softmax}(\textbf{MLP}(dec([\langle\text{BOS}\rangle, \tilde{X}, \langle\text{EOS}\rangle], enc(\hat{Y})))),$$
$$P(c|X_{\setminus s:e}, \hat{Y}, \theta) = \textbf{Softmax}(\textbf{MLP}(\textbf{Pooling}(dec([\langle\text{BOS}\rangle, \tilde{X}, \langle\text{EOS}\rangle], enc(\hat{Y}))))),$$

where $\tilde{X}$ is the language sequence after being masked, which is equivalent to $X_{\setminus s:e}$ in meaning but keeps the same length as the original sentence. $[\langle\text{BOS}\rangle, \tilde{X}, \langle\text{EOS}\rangle]$ means that a $\langle\text{BOS}\rangle$ is prepended to $\tilde{X}$, and an $\langle\text{EOS}\rangle$ is appended. Due to the incremental attention adopted by the decoder, the **Pooling** here must choose the last-token pooling strategy.

In CONSTMLM, the encoder handles predicting syntactic categories. Although the *encoder-decoder* version is designed so that the decoder can also be updated during CONSTMLM training, it is still only responsible for the masked span sequence generation.

In the *encoder-only* version with CONSTBTLM, the encoder also handles the prediction of syntactic categories, but cross-lingual context is adopted to support larger span masking. As for the *encoder-decoder* version, the encoder handles the cross-lingual context and the decoder predicts syntactic categories and generates masked span text. In CONSTMLM and the *encoder-only* CONSTBTLM, the weak alignment training of the syntactic category is performed on the source side, while it is completed on the target side in the *encoder-decoder* CONSTBTLM. For detailed training process, please refer to Appendix A.1.

3 Empirical Evaluation

3.1 Setup

Following the XLM codebase[1] and model structure setup (6 stacked Transformer layers with hidden dimension size of 1024) of (Conneau and Lample, 2019), we train the baseline UNMT model with an embedding-shared Transformer encoder-decoder architecture. The UNMT model training is divided into two stages: pre-training and unsupervised training. Our method is only used in the second stage for fast convergence. In order to make the unsupervised training more sufficient, we used an epoch size of 400K instead of the original recommended 200K in XLM. The γ in CONSTMLM is set to 0.3, and 0.5 in CONSTBTLM.

As the source of monolingual corpus for training, we use the 2007-2018 News Crawl dataset for English (En), French (Fr), German (De), Romanian (Ro), and Chinese (Zh). Since the Chinese News Crawl data is relatively small, we extracted sentences from Wikipedia dumps and converted them from traditional Chinese to simplified Chinese for

[1] https://github.com/facebookresearch/XLM

Method / Data	En-Fr	Fr-En	En-De	De-En	En-Ro	Ro-En	En-Zh	Zh-En
Data Used	274M	274M	509M	509M	195M	195M	50M	50M
Results reported from previous papers on large scale datasets								
NMT (Lample et al., 2018c)	25.1	24.2	17.2	21.0	21.1	19.4	–	–
PBSMT (Lample et al., 2018c)	27.8	27.2	17.7	22.6	21.3	23.0	–	–
PBSMT + NMT (Lample et al., 2018c)	27.6	27.6	20.2	25.2	25.1	23.9	–	–
XLM (Conneau and Lample, 2019)	33.4	33.3	26.4	34.3	33.3	31.8	–	–
Results from our runs on large scale datasets								
XLM	36.3	33.8	26.8	34.1	33.9	32.0	25.2	15.4
+ ConstMLM *enc-only*	36.5	34.1	27.0	34.5	34.0	32.2	25.8	16.2
+ ConstMLM *enc-dec*	36.5	34.4	27.2	34.6	33.9	32.5	25.9	16.3
+ ConstBTLM *enc-only*	37.0	34.2	27.3	34.9	34.5	32.8	26.2	16.7
+ ConstBTLM *enc-dec*	**37.3**	**34.5**	**27.9**	**35.0**	**35.2**	**33.0**	**26.3**	**17.2**
Data Used	10M	10M	10M	10M	10M	10M	10M	10M
Results from our runs on smaller datasets								
XLM	33.3	31.2	24.5	29.7	31.2	28.4	19.3	11.0
+ ConstMLM *enc-only*	33.6	31.4	24.9	30.4	31.4	28.5	22.9	13.1
+ ConstMLM *enc-dec*	33.9	32.0	25.5	30.5	32.0	28.7	23.1	13.4
+ ConstBTLM *enc-only*	34.5	33.0	**26.3**	31.7	32.0	28.9	23.9	14.6
+ ConstBTLM *enc-dec*	**35.1**	**33.4**	26.0	**31.8**	**32.4**	**29.0**	**24.5**	**15.3**

Table 1: BLEU scores on WMT'14 English-French (En-Fr), WMT'16 English-German (En-De), WMT'16 English-Romanian (En-Ro), and WMT'20 English-Chinese (En-Zh) unsupervised translation tasks.

use. Joint Byte-Pair Encodings (BPE) (Sennrich et al., 2016a) with 60K merge operations were used in the translation experiments for all language pairs. We explored the role of UG at two different monolingual corpus sizes in UNMT. All monolingual data from the newstest 2008-2018 is combined for use in the large-scale setting, while a subset of 5M sentences per language was randomly sampled from this data in the smaller scale setting.

Our evaluations were mainly carried out under unsupervised and low-resource semi-supervised scenarios. In the unsupervised translation scenario, we reported results on WMT *newstest2014* for En-Fr and En-Ro, WMT *newstest2016* for En-De, and WMT *newstest2020* for En-Zh. In the low-resource semi-supervised translation scenario, the IWSLT'14 En-Fr and En-De parallel sentences were used for training. IWSLT14.TED.*dev2010*, *tst2010*, *tst2011*, and *tst2012* were merged to evaluate the En-Fr translation model and *dev2010*, *dev2012*, *tst2010*, *tst2011*, and *tst2012* in IWSLT14.TED to evaluate the En-De model.

To acquire constituent parse trees for monolingual sentences, we adopted the current state-of-the-art Berkeley Neural Parser (Kitaev and Klein, 2018) as our parsing model and trained an En parser using PTB (Marcus et al., 1993), Fr and De parsers using the SPMRL14 multilingual constituent treebank (Seddah et al., 2014), and a Zh Parser using CTB (Xue et al., 2005). Since a constituent treebank is not available in Ro and for the consistency of the constituent trees used in En-Ro UNMT, we created En and Ro pseudo-constituent treebanks by converting their respective UD 2.7 treebanks using Head Feature Princinple (HFP) (Pollard and Sag, 1994), and trained En* and Ro* parsers using this. The processing and training details of each parser are presented in Appendix A.2. For each language, 500K sentences are parsed with these trained parsers for UNMT and low-resource semi-supervised NMT enhancement.

Method	En-Fr	Fr-En	En-De	De-En
Unsup. XLM	33.4	36.4	24.3	30.2
Semi. XLM	38.4	40.3	28.0	35.5
+ ConstMLM†	38.6	40.4	28.3	35.5
+ ConstMLM‡	38.7	40.5	28.4	35.7
+ ConstBTLM†	38.9	40.5	28.8	**36.1**
+ ConstBTLM‡	**39.0**	**40.7**	**28.9**	36.0

Table 2: BLEU scores on the semi-supervised IWSLT'14 En-Fr and En-De tasks. † means the encoder-only version is adopted, and ‡ means the encoder-decoder version is adopted.

3.2 Results and Analysis

The results of the UNMT experiment are mainly shown in Table 1. When a large-scale monolingual corpus is used, our baseline model outperforms XLM's reported results. This may be due to the use of the larger epoch size, which makes for more adequate training. Based on our strong baseline model, the four implementations of our ConstMLM and

CONSTBTLM approaches achieve consistent improvements in all language pairs, which demonstrates the effectiveness of universal grammar in UNMT. Based on the large-scale monolingual corpus scenario, comparing the four implementations of CONSTMLM and CONSTBTLM, we find that *enc-only* is generally weaker than the *enc-dec* implementation. This shows that training the model as a whole is better than training part of the model. This conclusion also partially explains the source of improvement of other *enc-dec* pre-training methods in UNMT like MASS (Song et al., 2019b) and BART (Lewis et al., 2020).

In the small-scale monolingual training data scenario, the performance of the baseline model has a large decline compared with the large-scale monolingual scenario, which shows that the size of monolingual data is still an important factor in the performance of the UNMT model. Similar to the large-scale monolingual scenario, our CONSTMLM and CONSTBTLM achieve improvements in translation performance, and the maximum increase is even greater than that in the large-scale monolingual scenario. This shows that in the case of relatively scarce training data, the introduction of universal grammar as a prior knowledge can effectively alleviate the performance loss.

Comparing the improved results in our approaches of each language pair horizontally, we find the average improvement of each language pair is basically consistent with the overlap of constituent labels between languages; that is, En-De, En-Zh, and En*-Ro* are more improved than is En-Fr (refer to Appendix 4.4 for the detailed statistics). This shows that the more grammatical commonalities two languages have, the greater their alignment's supervision will be. In addition, compared to the recent state-of-art work – MASS, due to their focus on pre-training, while ours concentrate on the NMT training with weak parallel information from universal grammar, our contribution is orthogonal to theirs.

In Table 2, we report the evaluation results of the low-resource semi-supervised scenario. We use a small-scale, monolingually trained UNMT model as the basis, so we also include the results of the UNMT model evaluated on the test datasets directly. After using the parallel data, the performance of our baseline model greatly improved, which reinforces our claim that UNMT models do not receive enough supervision in BT training.

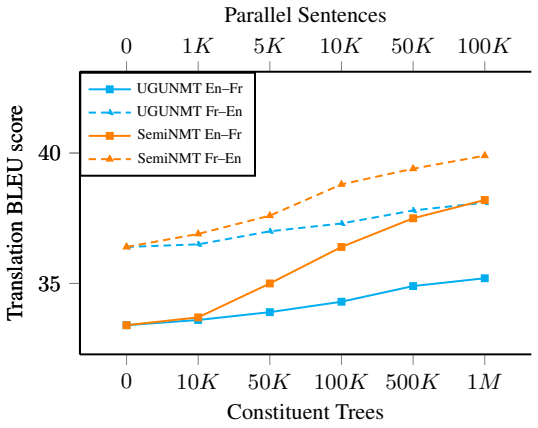

Figure 3: The Semi-supervised NMT and UNMT performance evaluated on IWSLT En-Fr benchmarks versus the number of parallel sentences and constituent trees used, respectively.

Method	γ	Phrase	En-De	De-En
XLM Baseline	–	–	24.5	29.7
+CONSTMLM‡	0.15	58.6%	25.1	30.0
	0.3	76.3%	25.5	30.6
	0.5	86.1%	24.1	29.3
+CONSTBTLM‡	0.15	58.6%	25.3	30.1
	0.3	76.3%	25.8	31.2
	0.5	86.1%	26.0	31.8
	0.6	88.4%	25.6	30.5

Table 3: Comparison of different maximum span ratios γ in CONSTMLM and CONSTBTLM for En-De UNMT. The **Phrase** column refers the proportions of the phrases kept under the maximum span ratio.

With the use of universal grammar for enhancement, the CONSTMLM and CONSTBTLM *enc-only* methods only achieved a slight improvement, which maybe suggest the training enhancement on the encoder side does not significantly improve the performance of translation after the introduction of parallel data. In the *enc-dec* approaches, the encoder and decoder are jointly optimized, and the performance improvement is greater, especially in CONSTBTLM *enc-dec* when larger and more spans can be leveraged.

4 Ablation Study

4.1 Constituent trees and parallel data size

To show that UG plays a similar role to the alignment information given by the parallel corpus, we compare the semi-supervised and UG-enhanced UNMT (UGUNMT) settings. The experimental results are evaluated on IWSLT'14 En-Fr. In the semi-supervised setting, we vary the amount of

parallel data, while we vary the number of monolingual parse trees in UGUNMT. The performance trend is shown in Figure 3.

The trends in the figure demonstrate that the performance of the UNMT model steadily improved with the addition of parallel corpus. The performance changes for UGUNMT also had a similar trend with the increase in the constituent parse data. This suggests that UG information plays a role similar to that of parallel data; that is, it brings supervision signals. The demand for monolingual constituent parse data, however, is greater than that from parallel data, and the improvement of parallel data is greater than that from constituent parses, which shows that UG can only provide a weak signal of supervision. While UG cannot achieve the same effect as parallel data, it is quite useful when there is a lack of parallel data.

4.2 Different Maximum Span Ratios

As in our approach description, we propose BTLM and its variant with the goal of mitigating the difficulty of reasoning with insufficient information in MLM. Although this problem has been noted in the training of PrLMs such as SpanBERT, in order to verify this problem's presence in the UNMT model and show that our proposed BTLM alleviates this issue, we explored the effects of different maximum span ratios γ in UNMT training. The results are shown in Table 3.

The comparison shows that the higher γ is, the greater the utilization proportion of the phrases in the constituent trees is. In CONSTMLM and CONSTBTLM, when γ is small, the phrases for training are limited, and therefore, the performance gains are limited. With increased γ, the utilization proportion increases, but CONSTMLM struggles with reasoning with insufficient data because too many spans are masked, and the performance even declines compared to baseline. CONSTBTLM can adapt to larger γ and higher phrase utilization proportions, it achieves better results.

4.3 Cross-lingual Alignment Evaluation

In order to verify that better alignment in the UNMT model is obtained using UG and our proposed training approaches, we conducted an experimental exploration of embedding alignment according to the experimental settings of (Conneau and Lample, 2019) and evaluated models on the SemEval'17 En-De cross-lingual semantic word similarity task (Camacho-Collados et al., 2017).

Method	Cosine sim.	L2 dist.	Pearson cor.
Concat Fasttext	0.36	4.89	0.52
MUSE	0.38	5.13	0.65
XLM	0.55	2.64	0.69
+ CONSTBTLM‡	0.60	2.55	0.71

Table 4: Unsupervised cross-lingual alignment evaluation with word embedding Cosine similarity (Cosine sim.), L2 distance (L2 dist.), and Pearson correlation (Pearson cor.) between source words and their translations.

We adopted the same vocabulary size for Concat Fasttext (Bojanowski et al., 2016), MUSE (Alaux et al., 2018), and XLM baselines, and our best En-De UNMT model and extracted the embeddings for comparison. The results are shown in Table 4. As the results show, our method is not only better than pure embedding training methods, Concat Fasttest and MUSE, on the three evaluation metrics, but also surpasses our strong XLM baseline, which demonstrates that the alignment of the UGUNMT model is indeed improved with the weak alignment information from syntactic categories.

4.4 Universal Constituent Labels

To illustrate the universal nature of the phrase grammar, we calculate statistics on the labels of the constituents in the annotations of each language. Specifically, the proportions of shared and differing labels are also calculated. The statistics are shown in Table 7. The statistical data shows that most of the grammatical phenomena (constituent labels) of the three language pairs overlap, and distributions of these labels are also close across language. The proportions of common labels in En-De and En-Zh are greater than that in En-Fr. Although En, Fr, De, and Zh have their own unique grammatical phenomena, they have greater proportions of overlapping labels than differing labels. Since En and Ro are pseudo-constituent labels transformed from UD, they cannot be directly compared with En-Fr, En-De, and En-Zh, but they do also have many similar labels and comparable common label proportions, indicating the UD annotation's universality and the effectiveness of our conversion in preserving grammatical features. This does not explain more complicated issues such as language similarity or commonality but rather indicates the overlap of grammatical phenomena and universal features in the annotations and parser predictions.

4.5 Effects of SpanBERT, LIMIT-BERT, and CONSTBTLM for UNMT

From the main experiments, the UNMT performance is improved, especially for the small-scale data setting. To find out that if the improvements are caused by CONSTMLM/CONSTBTLM and the syntactic information is really necessary, we compare our approaches with LIMIT-BERT which apply a linguistically guided span based MLM objective during UNMT training, and SpanBERT which is with a non-syntax based span masking strategy. Compared with SpanBERT and LIMIT-BERT in our UNMT framework, the implementation is relatively simple. By removing the syntactic category prediction objective in the CONSTMLM *enc-only* variant, it is consistent with the objective of LIMIT-BERT, and further removes the use of the syntactic parse tree in the span sampling, the same objective of SpanBERT is achieved.

The results of the comparison are shown in Table 7. The use of SpanBERT and LIMIT-BERT training approaches has resulted in a performance improvement in translation over the XLM baseline, which indicates that additional span-based pre-training is helpful for UNMT. SpanBERT outperforms LIMIT-BERT because syntactic annotation is costly, the fixed-size syntactic parse tree used severely limits the pre-training with span boundaries considered only, while SpanBERT with dynamic span mask can get sufficient training. But in ConMLM, this disadvantage was mitigated by the introduction of additional syntactic label predictions, and when we used the *enc-dec* variant, which is more suitable for encoder-decoder structures, its performance exceeded SpanBERT. This suggests that it is not that syntactic information is useless. With the help of ConstBTLM, a stronger variant, the UNMT model achieves much better translation results. This demonstrates that in UNMT training on the one hand additional pre-training is helpful, on the other hand, the use of effective means to integrate the weak alignment information provided by syntactic parse tress is also beneficial to improve translation performance.

Method	En-Fr	Fr-En
XLM	33.3	31.2
SpanBERT	33.5	31.7
LIMIT-BERT	33.4	31.4
CONSTMLM *enc-only*	33.6	31.4
CONSTMLM *enc-dec*	33.9	32.0
CONSTBTLM *enc-dec*	35.1	33.4

Table 5: UNMT performance on WMT'14 En-Fr test set with small-scale data setting.

5 Related Work

UNMT has been greatly developed in recent years (Artetxe et al., 2018b; Yang et al., 2018; Sun et al., 2019; Conneau and Lample, 2019; Ren et al., 2019). Syntax has been used extensively explored in supervised MT research field (Wu et al., 2018; Zhang et al., 2019; Currey and Heafield, 2019; Duan et al., 2019). Zhou et al. (2020b) leveraged syntactic and semantic spans for MLM to pre-train the language model and delivered promising results. Xu et al. (2020) incorporated syntax information into a UNMT model by leveraging linearized parse trees of the training sentences. In this work, we make the first attempt to use syntactic information as an auxiliary training objective for UNMT which differs from the motivation in syntax for supervised MT. A more detailed related work introduction and discussion is presented in the Appendix A.3.

6 Conclusion and Future Work

In this paper, we mine weak alignment information from universal grammar annotations and use it to improve unsupervised machine translation. Two specific training approaches, CONSTMLM and CONSTBTLM, are proposed to apply this weak supervision. Via empirical exploration on unsupervised and semi-supervised machine translation benchmarks, we verify that universal grammar will boost cross-lingual alignment for UNMT. Our analysis shows that using universal grammar, the reliance on parallel corpora can be reduced under the premise of achieving the same effect because the weak supervision signal based on universal grammar can play a similar role to the supervision signal of the parallel corpus.

In this work, we rely on the dependency syntax of 100+ languages provided by the universal dependency project for synthesizing pseudo-constituent syntax in some languages. In the future, we intend to train a multilingual parser based on the multilingual language model – XLM-R (with the training data as a combination of 10+ language constituent syntax), which has the ability to parse 100+ languages in a single model, further increasing the practicality of our method. In addition, we will examine more low-resource languages to verify the method's universality.

References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Roger Bacon. 1902. *The Greek grammar of Roger Bacon and a fragment of his Hebrew grammar*. The University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*.

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, Canada. Association for Computational Linguistics.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2018. Syntax-directed attention for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Noam Chomsky. 1965a. *Aspects of the Theory of Syntax*, volume 11. MIT press.

Noam Chomsky. 1965b. *Cartesian Linguistics: a chapter in the history of rationalist thought Harper and Row*. London.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Anna Currey and Kenneth Heafield. 2019. Incorporating source syntax into transformer-based neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 24–33, Florence, Italy. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sufeng Duan, Hai Zhao, Junru Zhou, and Rui Wang. 2019. Syntax-aware transformer encoder for neural machine translation. In *2019 International Conference on Asian Language Processing (IALP)*, pages 396–401. IEEE.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California. Association for Computational Linguistics.

John Edwards. 2002. *Multilingualism*. Routledge.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 823–833, Berlin, Germany. Association for Computational Linguistics.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 72–78, Vancouver, Canada. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29:820–828.

Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntax-aware multilingual semantic role labeling. 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 5350–5359, Hong Kong, China. Association for Computational Linguistics.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2061–2071, Melbourne, Australia. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Kevin Knight, Anish Nair, Nishit Rathod, and Kenji Yamada. 2006. Unsupervised analysis for decipherment problems. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 499–506, Sydney, Australia. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Shalom Lappin and Stuart M Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, pages 393–427.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. 2018. A unified syntax-aware framework for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2401–2411, Brussels, Belgium. Association for Computational Linguistics.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020a. Data-dependent gaussian prior objective for language generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zuchao Li, Zhuosheng Zhang, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2021a. Text compression-aided transformer encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zuchao Li, Hai Zhao, Shexia He, and Jiaxun Cai. 2021b. Syntax Role for Neural Semantic Role Labeling. *Computational Linguistics*, pages 1–46.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020b. Reference language based unsupervised neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4151–4162, Online. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *International Conference on Learning Representations*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Maria Nădejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Copenhagen, Denmark. Association for Computational Linguistics.

Thai Phuong Nguyen, Akira Shimazu, Tu-Bao Ho, Minh Le Nguyen, and Vinh Van Nguyen. 2008. A tree-to-string phrase-based model for statistical machine translation. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 143–150, Manchester, England. Coling 2008 Organizing Committee.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–16. Springer.

Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 241–248.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland. Dublin City University.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019a. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019b. Mass: Masked sequence to sequence pre-training for language generation. pages 5926–5936.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1235–1245, Florence, Italy. Association for Computational Linguistics.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wei Wang, Jonathan May, Kevin Knight, and Daniel Marcu. 2010. Re-structuring, re-labeling, and re-aligning for syntax-based machine translation. *Computational Linguistics*, 36(2):247–277.

Shuangzhi Wu, Dongdong Zhang, Zhirui Zhang, Nan Yang, Mu Li, and Ming Zhou. 2018. Dependency-to-dependency neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2132–2141.

Jia Xu, Na Ye, and GuiPing Zhang. 2020. Improving unsupervised neural machine translation with dependency relationships. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 429–440. Springer.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55, Melbourne, Australia. Association for Computational Linguistics.

Meishan Zhang, Zhenghua Li, Guohong Fu, and Min Zhang. 2019. Syntax-enhanced neural machine translation with syntax-aware word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1151–1161, Minneapolis, Minnesota. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL-08: HLT*, pages 559–567, Columbus, Ohio. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9628–9635. AAAI Press.

Junru Zhou, Zuchao Li, and Hai Zhao. 2020a. Parsing all: Syntax and semantics, dependencies and spans. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4438–4449, Online. Association for Computational Linguistics.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020b. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Appendix

A.1 Training Procedure

Algorithm 1: Training Procedure

Input: Source Monolingual Data D_S,
Source Monolingual Data D_T,
Source Parse Tree TD_S,
Target Parse Tree TD_T;
Model Parameters θ;
Training Epochs N;

for $t = 1, 2, \cdots, N$ **do**
  Back-translation Step
    $\mathrm{B}_S \leftarrow \mathit{Sample}(\mathrm{D}_S)$;
    $\hat{\mathrm{B}}_T \leftarrow \mathit{MT}_{\mathrm{S2T}}(\mathrm{B}_S)$;
    $\mathcal{L}_{\mathrm{T2S}} \leftarrow \mathit{Likelihood}(\mathit{MT}_{\mathrm{T2S}}(\hat{B}_T), \mathrm{B}_S)$;
    $\theta \overset{update}{\longleftarrow} \mathcal{L}_{\mathrm{T2S}}$;
    $\mathrm{B}_T \leftarrow \mathit{Sample}(\mathrm{D}_T)$;
    $\hat{B}_S \leftarrow \mathit{MT}_{\mathrm{T2S}}(\mathrm{B}_T)$;
    $\mathcal{L}_{\mathrm{S2T}} \leftarrow \mathit{Likelihood}(\mathit{MT}_{\mathrm{S2T}}(\hat{B}_S), \mathrm{B}_T)$;
    $\theta \overset{update}{\longleftarrow} \mathcal{L}_{\mathrm{S2T}}$;
  CONSTMLM / CONSTBTLM *Step*
    $\mathrm{B}_S, \mathrm{Tr}_S \leftarrow \mathit{Sample}(\mathrm{TD}_S)$;
    $\mathrm{B}_S^M, \mathrm{B}_S^L \leftarrow \mathit{Mask}(\mathrm{B}_S, \mathrm{Tr}_S)$;
    if CONSTBTLM **then**
      $\hat{\mathrm{B}}_T \leftarrow \mathit{MT}_{\mathrm{S2T}}(\mathrm{B}_S)$;
      $\mathcal{L}_{\mathrm{S}} \leftarrow \mathit{Likelihood}(\mathit{MLM}([\hat{\mathrm{B}}_T \odot \mathrm{B}_S^M]), \mathrm{B}_S) + \mathit{Likelihood}(\mathit{LabelPred}([\hat{\mathrm{B}}_T \odot \mathrm{B}_S^M]), \mathrm{B}_S^L)$;
    else
      $\mathcal{L}_{\mathrm{S}} \leftarrow \mathit{Likelihood}(\mathit{MLM}(\mathrm{B}_S^M), \mathrm{B}_S) + \mathit{Likelihood}(\mathit{LabelPred}(\mathrm{B}_S^M), \mathrm{B}_S^L)$;
    $\theta \overset{update}{\longleftarrow} \mathcal{L}_{\mathrm{S}}$;
    $\mathrm{B}_T, \mathrm{Tr}_T \leftarrow \mathit{Sample}(\mathrm{TD}_T)$;
    $\hat{\mathrm{B}}_T \leftarrow \mathit{MT}_{\mathrm{S2T}}(\mathrm{B}_T)$;
    if CONSTBTLM **then**
      $\mathrm{B}_T^M, \mathrm{B}_T^L \leftarrow \mathit{Mask}(\mathrm{B}_T, \mathrm{Tr}_T)$;
      $\mathcal{L}_{\mathrm{T}} \leftarrow \mathit{Likelihood}(\mathit{MLM}([\hat{\mathrm{B}}_T \odot \mathrm{B}_T^M]), \mathrm{B}_T) + \mathit{Likelihood}(\mathit{LabelPred}([\hat{\mathrm{B}}_T \odot \mathrm{B}_T^M]), \mathrm{B}_T^L)$;
    else
      $\mathcal{L}_{\mathrm{T}} \leftarrow \mathit{Likelihood}(\mathit{MLM}(\mathrm{B}_T^M), \mathrm{B}_T) + \mathit{Likelihood}(\mathit{LabelPred}(\mathrm{B}_T^M), \mathrm{B}_T^L)$;
    $\theta \overset{update}{\longleftarrow} \mathcal{L}_{\mathrm{T}}$;

A.2 Parser Training and Evaluation

In this section, we evaluate the performance of parsers used in this paper on their respective test sets. Our parsing model is based on the architecture described in (Kitaev and Klein, 2018), a state-of-the-art multilingual parser. We trained our En constituent parser with Penn Treebank (Marcus et al., 1994), Zh parser with Chinese Penn Treebank (Xue et al., 2005), and the Fr and De parsers with the SPMRL 2013/2014 shared task (Seddah et al., 2013, 2014). Thus, these parsers are also evaluated on the test datasets of these treebanks or shared tasks.

Language	P	R	F_1	EM
En	95.54	95.12	95.33	53.24
Fr	87.63	87.03	87.33	24.24
De	91.51	88.71	90.09	54.06
Zh	92.16	91.88	92.02	43.97
En*	81.69	80.03	80.85	43.81
Ro*	79.28	78.69	78.98	24.42

Table 6: Constituent parsing performance on the test datasets. * indicates models trained using UD-transformed constituent data.

Some languages lack well-annotated constituent treebanks, which adds some difficulty to our research in using universal grammar for UNMT. Universal Dependencies (UD), however, is a consistent dependency syntactic annotation on more than 100 languages. Dependency treebanks are usually converted from constituent treebanks, though they may be independently annotated as well for the same languages. Constituent trees can be accurately converted to dependency representations using grammatical rules or machine learning methods (de Marneffe et al., 2006). Such convertibility shows a close relation between constituent and dependency representations. Therefore, we consider transforming the widely annotated UD treebank[2] into a constituent treebank for languages that lack constituent annotations. It is not hard to obtain an approximate constituent structure from the dependency structure, but the labels change a lot, and it is also very difficult to train a machine learning conversion model when the original constituent annotations are lacking.

In order to address this inconvenience, we propose converting the dependency structure to the constituent structure using the HFP. Our UNMT model does not need a genuine constituent label; rather, it only needs labels to be consistent across corpora in different languages. As a result, we use the relationship between the head word of a constituent and its dependency head as a constituent label, resulting in a complete annotated constituent parse tree. Like (Kitaev et al., 2019), we use the pre-trained language model BERT to enhance the parser. En uses *bert-base-cased*, Zh uses *bert-base-chinese*, and Fr, De, and Ro use *bert-base-*

[2]http://hdl.handle.net/11234/1-3424

$\mathbf{L_1}, \mathbf{L_2}$	$\mathbf{L_1} \cap \mathbf{L_2}$	$\mathbf{L_1} - \mathbf{L_2}$	$\mathbf{L_2} - \mathbf{L_1}$
En,Fr	2 (57.84%, 57.50%)	24 (42.16%)	30 (42.50%)
En,De	4 (87.74%, 72.57%)	22 (12.26%)	21 (27.43%)
En,Zh	11 (81.10%, 75.68%)	15 (18.90%)	15 (24.32%)
En*,Ro*	39 (95.69%, 95.15%)	9 (4.31%)	10 (4.85%)

Table 7: Statistics of common and different constituent labels in different language pairs. * indicates that the statistics are based on the dataset transformed from UD. The $\mathbf{L_1} \cap \mathbf{L_2}$ column refers to the number of common constituent labels for languages L_1 and L_2, and the proportions of these labels appearing in the respective datasets are in parentheses. $\mathbf{L_1} - \mathbf{L_2}$ refers to the number and proportions of constituent labels that only exist in language L_1, $\mathbf{L_2} - \mathbf{L_1}$ refers to the number and proportions of constituent labels that only exist in language L_2.

multilingual-cased. The results of the evaluation on each language data set are shown in Table 6.

A.3 Related Work

Unsupervised machine translation systems have been developed since Knight et al. (2006). Ravi and Knight (2011) framed the unsupervised MT problem as a decipherment task between two languages. With the development of deep end-to-end neural network translation and language models, UNMT has begun to be competitive in translation benchmarks. Before this development, unsupervised cross-lingual embeddings (Artetxe et al., 2017; Zhang et al., 2017) and word translation with parallel data (Lample et al., 2018b) were alternative approaches to unsupervised machine translation. (Artetxe et al., 2018a; Lample et al., 2018c) studied unsupervised training using phrase-based translation systems. Recently, UNMT has been a hot research topic in machine translation (Artetxe et al., 2018b; Yang et al., 2018; Sun et al., 2019; Conneau and Lample, 2019; Ren et al., 2019; Sun et al., 2020; Li et al., 2020b). Our work builds on part of these works in unsupervised machine translation, but we focus on improving by leveraging universal grammar.

Grammar information, especially syntax information, has always been the focus of research in the field of machine translation. In statistical machine translation, syntactic trees were used as the basis for re-structuring, re-labeling, and re-aligning (re-ordering) sentences to improve the translation accuracy (Wang et al., 2010). Based on the type of linguistic information used, the syntactic SMT can be divided into four types: tree-to-string, string-to-tree, tree-to-tree, and hierarchical phrase-based (Zhang et al., 2008; Nguyen et al., 2008). Our use of universal grammar to enhance UNMT, from a motivation perspective, is similar to a tree-to-tree approach in SMT. Parallel syntactic trees are used to obtain structure alignment information in tree-to-tree SMT, while our approach leverages non-parallel syntactic parsing trees to obtain weak alignment information based on our proposed training objectives in UNMT. In NMT, syntactic information is mainly used as features and/or constraints (regularization). (Eriguchi et al., 2016; Bastings et al., 2017) augmented the RNN encoder for feature extraction with an additional syntactic encoder as in Tree-LSTM (Tai et al., 2015) and GCN (Kipf and Welling, 2016); and combined this with a standard RNN decoder. Chen et al. (2018) extended the local attention in RNN-based NMT with a syntax-distance constraint that makes the model focus more on syntactically related source words. (Wu et al., 2018; Zhang et al., 2019; Currey and Heafield, 2019; Duan et al., 2019) explored the role of explicit syntactic information in Transformer-based NMT. In addition, He et al. (2018); Li et al. (2018); Zhou et al. (2020a); He et al. (2019); Li et al. (2021b) also shown a positive effect for other downstream tasks.

Sharing NMT model parameters with a syntactic parser for multi-task learning is also a popular approach to obtaining syntactically-aware representations (Luong et al., 2016; Dyer et al., 2016; Eriguchi et al., 2017; Nădejde et al., 2017). The use of syntax in UNMT research is relatively rare. Xu et al. (2020) incorporated syntax information into a UNMT model by leveraging linearized parse trees of the training sentences. Although all these works use syntactic information, our motivation is very different. Unlike other approaches that use syntax information as a feature or constraint, we use syntax information to produce a form of weak supervision that can guide model training. We differ from multi-task learning approaches combining syntax and machine translation in that our purpose is not to predict the syntactic tree but to align text across languages using syntactic categories, and we do this through a masking-prediction process of syntactic constituents.

Pre-trained language models like BERT (Devlin et al., 2019; Zhang et al., 2020), XLM (Conneau and Lample, 2019), ALBERT (Lan et al., 2020), and ELECTRA(Clark et al., 2020) have shown

strong performance gains in various NLP tasks by using a self-supervised training task, masked language modeling. SpanBERT (Joshi et al., 2020) was designed to better represent and predict spans of text and masked random contiguous spans of text rather than random individual tokens. LIMIT-BERT (Zhou et al., 2020b) introduced a Syntactic/Semantic Phrase Masking (SPM) for language pre-training that used linguistically-guided masking, meaning the spans masked were ensured to be valid language components. The CONSTMLM encoder-only version we proposed is essentially the same as the LIMIT-BERT, but we further proposed the CONSTMLM encoder-decoder version in order to adapt to training a UNMT model.

There is an issue with span-based MLM. When the span selected for masking is too long, the remaining words in the sentence are not enough for the model to infer the masked part, and this training will be ineffective. SpanBERT and LIMIT-BERT only account for this issue by limiting the maximum length of spans. Inspired by Translation Language Modeling (TLM) (Conneau and Lample, 2019), we propose BTLM to address this issue. Though both BTLM and TLM consider cross-lingual context for inference, TLM uses parallel corpora for cross-lingual alignment training, while BTLM bypasses the need for parallel corpora, uses its translation as cross-lingual context, and only selects input sentences for MLM. MASS (Song et al., 2019b) and BART (Lewis et al., 2020) adopted encoder-decoders for model pre-training, and the encoder-decoder versions of our approaches follow this schema but with a different aim and motivation.