A Comparison between Pre-training and Large-scale Back-translation for Neural Machine Translation

Dandan Huang[1,2*], **Kun Wang**[3*], **Yue Zhang**[1,2†]
[1] School of Engineering, Westlake University
[2] Institute of Advanced Technology, Westlake Institute for Advanced Study
[3] Microsoft STCA NLP Group
{huangdandan,zhangyue}@westlake.edu.cn, wangkun@microsoft.com

Abstract

BERT has been studied as a promising technique to improve NMT. Given that BERT is based on the similar Transformer architecture to NMT and the current datasets for most MT tasks are rather large, how pre-training has managed to outperform standard Transformer NMT models is underestimated. We compare MT engines trained with pre-trained BERT and back-translation with incrementally larger amounts of data, implementing the two most widely-used monolingual paradigms. We analyze their strengths and weaknesses based on both standard automatic metrics and intrinsic test suites that comprise a large range of linguistic phenomena. Primarily, we find that 1) BERT has limited advantages compared with large-scale back-translation in accuracy and consistency on morphology and syntax; 2) BERT can boost the Transformer baseline in semantic and pragmatic tasks which involve intensive understanding; 3) pre-training on huge datasets may introduce inductive social bias thus affects translation fairness.

1 Introduction

Neural machine translation (NMT) has shown promising results as an end-to-end approach to automatic translation (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017). One reason for its success is the availability of large amounts of training resources such as parallel corpora with high quality. For low-resource languages or domain-specific settings, monolingual data have also been effectively used by NMT systems (Zhang and Zong, 2016; Siddhant et al., 2020), providing rich linguistic features for translation.

Two lines of work have been done on leveraging monolingual corpora to improve translation quality. One approach is back-translation (Bojar and Tamchyna, 2011; Sennrich et al., 2016), in which an auxiliary target-to-source system is trained on genuine bitext, and then used to generate synthetic text from a large monolingual corpus on the target side. The synthetic and genuine pairs are then used together to train a source-to-target MT model.

An alternative method of using monolingual data is the pre-trained language model (Devlin et al., 2019; Radford et al., 2019), a neural network trained over large texts and can be incorporated into standard NMT encoder-decoder architectures (Jean et al., 2015; Gulcehre et al., 2015; Zhu et al., 2020). Pre-trained language models have led to improvements in NMT results across low-resource scenarios (Song et al., 2019), cross-lingual transfers (Conneau and Lample, 2019; Liu et al., 2020) and code-switching settings (Yang et al., 2020).

Among these two dominant monolingual paradigms, there has been relatively more work investigating how back-translation helps NMT. For example, initial studies show that back-translation is beneficial to machine translation by producing more fluent outputs (Edunov et al., 2020). However, relatively little work has focused on how pre-trained language models contribute to translation. We fill this gap by quantitatively comparing MT models trained with pre-trained language models and back-translation under a fair large-scale setting. Specifically, for pre-trained language models, we reimplement BERT-fused NMT (Zhu et al., 2020), and for back-translation, we use incrementally larger data amounts to train a range of systems, with the synthetic data being half, equal, twice and four times of the authentic data. We conduct experiments on rich (WMT'14 English-to-German) and low (LDC Chinese-to-English) resource scenarios, and evaluate performance on 8 benchmarks covering morphological, syntactic, semantic and pragmatic competences. Empirically, we find that:

* Equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1718–1732
August 1–6, 2021. ©2021 Association for Computational Linguistics

1. BERT yields improvement for standard NMT in BLEU but has no remarkable advantage compared with large-scale back-translation.
2. BERT has little effect on correcting smaller discrepancies in morphological and syntactic levels in NMT (Section 5.1& 5.2).
3. BERT features salient promotion for MT requiring heavy context understanding and intensive knowledge, but also brings concerns around bias and fairness (Section 5.3& 5.4).

To our knowledge, we are the first to detect the effectiveness of pre-training in NMT by a comparison with back-translation in a fair setting. We also contribute to the analysis of BERT in a bilingual situation.

2 Related Work

Pre-training in NMT Gulcehre et al. (2015) and Jean et al. (2015) are among the first to integrate language models into the decoder part of NMT. Subsequent work extends the studies by adding pre-trained representations in the encoder part (Edunov et al., 2019) or the both sides (Ramachandran et al., 2017) of NMT networks.

Recent research focused on leveraging the pre-trained BERT for NMT. Clinchant et al. (2019) utilize BERT on NMT's encoder. Conneau and Lample (2019) initialize both the encoder and decoder by multilingual BERT. Imamura and Sumita (2019) investigate a BERT fine-tuning method for NMT. Clinchant et al. (2019) compare different NMT architectures with BERT. Zhu et al. (2020) suggest using BERT as an extra memory. Specifically, they first encode the inputs by BERT and use the last layer's output as an extra memory. The Transformer NMT network uses an extra self-attention module to weigh the memory in each layer of both the encoder and decoder. The model shows a noticeable improvement in both supervised, semi-supervised and unsupervised tasks, leading to the new state-of-the-art results of using BERT in NMT. Given the significant improvements achieved by their work, we adopt this model in our experiments.

Back-translation Back-translation is a widely used data augmentation technology originally introduced for SMT (Bojar and Tamchyna, 2011) and then flourished in NMT (Sennrich et al., 2016). It has been studied with dual-learning frameworks (He et al., 2016), large-scale extensions (Edunov et al., 2018; Wu et al., 2019), iterative versions (Hoang et al., 2018), unsupervised scenarios (Artetxe et al., 2018; Lample et al., 2018), tagged back-translated sources (Caswell et al., 2019) as well as systematic analysis (Burlot and Yvon, 2018; Poncelas et al., 2018; Edunov et al., 2020). In line with Edunov et al. (2018), we aim to broaden understanding of back-translation in a large-scale manner. While their focus is on different methods that generate synthetic source sentences, ours is to investigate how large-scale pre-training compares with large-scale back-translation in boosting translation performance.

BERTology Much work has discussed BERT with respect to morphology (Edmiston, 2020; Haley, 2020), syntax (Hewitt and Manning, 2019; Lin et al., 2019; Goldberg, 2019), semantics (Ettinger, 2020; Warstadt et al., 2019; Tenney et al., 2019), and world knowledge (Poerner et al., 2019; Zhou et al., 2020). Both internal attention weights (Clark et al., 2019; Htut et al., 2019) and external task performances(Liu et al., 2019a; Zhou et al., 2020) have been used as means of investigation. Our work aligns with external evaluation. However, existing work considers a monolingual setting while we discuss these issues under a bilingual task.

3 Protocol for MT Evaluation

We use BLEU (Papineni et al., 2002) and 8 more focused evaluation tasks to probe MT systems with pre-trained BERT and back-translation. Below we introduce the error analysis protocols in detail.

3.1 Morphological Competence

We assess the morphological competence of MT systems translating from English into morphologically rich languages, which is a necessity for MT systems to overcome out-of-vocabulary source tokens and flexible word orders. We take Morpheval[1] (Burlot and Yvon, 2017; Burlot et al., 2018) as one of the representative test suits, consisting of a set of contrast pairs that can be triggered in the source language and evaluated in the target language (Table 1). This dataset describes three types of contrasts: the first evaluates one single morphological derivational feature such as number, gender, tense; the second evaluates agreement; the third concerns lexical replacements of the same category, testing whether morphological consistency still holds if a word is replaced by a hyponym.

[1] https://github.com/franckbrl/morpheval_v2

Morphology	En→De	Source: **The** only issue now is the swelling around his eye. **Her** only issue now is the swelling around his eye. Target: Das **einzige** Problem ist jetzt die Schwellung um sein Auge. Ihr **einziges** Problem ist jetzt die Schwellung um sein Auge. Target: Das **einzige** Problem ist jetzt die Schwellung um sein Auge. Das **einzige** Problem ist jetzt die Schwellung um sein Auge.	 ✔ ✗
Syntax	En→De	Source: Prague Stock Market **falls to minus** by the end of the trading day. Target: Die Prager Börse stürzt gegen Geschäftsschluss **ins Minus**. Target: Die Prager Börse stürzt gegen Geschäftsschluss **nicht ins Minus**.	 ✔ ✗
Homograph Translation	En→De	Source: I hope you didn't get distracted during your **watch**. Target: Ich hoffe, Sie wurden während Ihrer **Wache** nicht abgelenkt. (guard) Target: Ich hoffe, Sie wurden während Ihrer **Uhr** nicht abgelenkt. (clock)	 ✔ ✗
Conjunction Disambiguation	En→De	Source: Don't apologize to me, **but** to her. Target: Entschuldige dich nicht bei mir, **sondern** bei ihr. Target: Entschuldige dich nicht bei mir, **aber** bei ihr.	 ✔ ✗
Pronoun Coreference Resolution	En→De	Source: **It** could get tangled in your hair. Target: **Sie** könnte sich in deinem Haar verfangen. Target: **Er** könnte sich in deinem Haar verfangen. Target: **Es** könnte sich in deinem Haar verfangen.	 ✔ ✗ ✗
Idiom Translation	Zh→En	Source: 胸有成竹 Target: Be very ready; have a well-thought-out plan. Target: Have a well-formed **bamboo** in one's **chest**. (literal translation)	 ✔ ✗
Commonsense Reasoning	Zh→En	Source: 吃了游客的鳄鱼。 Target: The crocodile who ate the tourist. Target: Ate the tourist's crocodile.	 ✔ ✗
Gender Bias	En→De	Source: The janitor does not like **the baker** because **she** always messes up the kitchen. Target: Der Hausmeister mag **die Bäckerin** nicht, weil sie immer die Küche verschmutzt. Target: Die Hausmeisterin mag **der Bäcker** nicht, weil sie die Küche immer durcheinander bringt.	 ✔ ✗

Table 1: Test instances corresponding to each task. Key words are in bold. Elaboration is in Appendix A.

3.2 Syntactic Competence

We evaluate whether MT models can generate coherent and grammatical sentences. We adopt the LingEval97[2] (Sennrich, 2017), a test set of contrastive translation pairs for analysis of a number of syntactic phenomena including syntactic agreement over long distances, discontiguous verb-particle constructions, transliteration of names and faithful translation of polarity (Table 1).

3.3 Semantic Competence

Semantics helps MT enforce meaning preservation and handle data sparsity. We measure semantic competence from the ambiguity of content words, conjunctions and pronouns, corresponding to tasks of *homograph translation*, *conjunction disambiguation*, and *pronoun coreference resolution*, respectively. First, homograph translation requires models to determine the intended sense of polysemous words in context. We adopt MUCOW[3] (Raganato et al., 2019), a lexical ambiguity benchmark in which a sentence containing an ambiguous word is paired with a correct reference and an incorrect modified translation with the ambiguous word being replaced by a word of a different sense. Second, NMT should theoretically be able to handle conjunctions with variant senses if the encoder captures clues from sentence structures. We use the test set of Popović (2019)[4], which translates the English conjunction *but* into two different German conjunctions *aber* or *sondern*. The former can be used after a positive or a negative clause, while the latter is only used after a negative clause when expressing a contradiction. Lastly, for coreference resolution, we adopt ContraPro[5] (Müller et al., 2018) to evaluate the accuracy when models translate the English pronoun *it* to its German counterparts *es* (it), *sie* (she) and *er* (he), based on a correct understanding of antecedents.

3.4 Pragmatic Competence

We further evaluate systems on 3 challenging problems involving pragmatic inference: *idiom translation*, *commonsense reasoning* and *gender bias*. First, idiom translation still presents a difficulty because the meaning of idioms is non-compositional and non-literal, making word-by-word translation incorrect. We use the CIBB dataset [6] (Shao et al., 2018), in which a blacklist consisting literal translation of idiom characters is constructed and once translations from NMT trigger the blacklist, the literal translation errors can be counted to score the systems. Another demanding competence for NMT is commonsense reasoning. He et al. (2020) build

[2] https: //github.com/rsennrich/lingeval97
[3] https://github.com/Helsinki-NLP/MuCoW
[4] https://github.com/m-popovic
[5] https://github.com/ZurichNLP/ContraPro
[6] https://github.com/sythello/CIBB-dataset

a bilingual test suite which grounds commonsense knowledge into lexical ambiguity, contextual syntactic ambiguity and contextless syntactic ambiguity (Appendix A.3). Each source sentence has one ambiguity type and corresponds to two contrastive translations. We use this test suite [7] to measure commonsense knowledge and inference of NMT outputs. Lastly, we estimate gender bias. Following Stanovsky et al. (2019), we use the WinoMT[8] dataset to extract gender features from translations and evaluate them against the gold annotations.

4 Experimental Setup

We verify the effectiveness of MT combined with BERT (Zhu et al., 2020) and back-translation on both rich- and low-resource scenarios.

4.1 Data and Baseline

For the rich-resource scenario, we take WMT'14 English-to-German (En→De) with a corpus size of 4.5M [9]. We use newstest2013 as the validation set and newstest2014 as the test set. For the low-resource scenario, we take LDC Chinese-to-English (Zh→En) with a corpus size of 1.25M. We use nist06 as the validation set and report an average score on nist02/03/04/05/08 test sets. We apply wordpieces (Wu et al., 2016) to preprocess data with a shared source and target vocabulary of 32K.

We train a standard Transformer NMT model (Vaswani et al., 2017) on `fairseq`[10] as a baseline. We adopt `transformer_big` for En→De and `transformer_base` for Zh→En with a 6-layer encoder-decoder network. We set the dropout ratio as 0.25 and use beam search with width 4 and length penalty 0.6 for inference.

4.2 BERT-fused NMT

BERT (Devlin et al., 2019) is composed of a layered self-attention Transformer network and is pre-trained on billions of unlabeled text to perform masked language modeling and next sentence prediction tasks. The former aims to restore the original sequence from noisy input, while the latter learns whether two sentences are consecutive.

Zhu et al. (2020) incorporate BERT into NMT systems. On the source side, given a language input x, the model first extracts the last layer's output of the context-aware representation from BERT encoder:

$$H_B = BERT(x), \tag{1}$$

and then fuses H_B with each layer of the encoder of the NMT model through attention mechanisms:

$$\begin{aligned} H_E^l = \frac{1}{2}\big(&attn_S(H_E^{l-1}, H_E^{l-1}, H_E^{l-1}) \\ &+attn_B(H_E^{l-1}, H_B, H_B)\big), \end{aligned} \tag{2}$$

where H_E^l refers to the hidden state after fusion of the l-th layer, $attn_S$ is the multi-head self-attention layer, and $attn_B$ is the BERT attention layer. In the case of layer l in the target side, the decoder also uses both contexts at the same time:

$$\begin{aligned} H_{DS}^l &= attn_{MS}(H_D^{l-1}, H_D^{l-1}, H_D^{l-1}), \\ H_D^l &= \frac{1}{2}\big(attn_B(H_{DS}^l, H_E^L, H_E^L) \\ &\quad +attn_E(H_{DS}^l, H_B, H_B)\big), \end{aligned} \tag{3}$$

where $attn_{MS}$, $attn_B$, $attn_E$ is the multi-head future-masked self-attention layer, BERT-decoder attention layer and the encoder-decoder attention layer, respectively. H_E^L is the output of the encoder.

Following Zhu et al. (2020), we first train a standard Transformer NMT and then initialize the weights of the BERT-fused model. We choose `bert_large_cased`[11] with 24 layers and 1024 hidden dimension for En→De and `bert_base_chinese`[12] with 12 layers and 768 hidden dimension for Zh→En, ensuring that the dimension of BERT and NMT model almost matches. BERT is fixed during training. The optimization algorithm is `Adam` in accordance with 0.0005 learning rate and the `inverse_sqrt` scheduler.

4.3 Back-translation

For back-translation, we use the standard Transformer baseline with the method of Sennrich et al. (2016) to synthesize augmented data. Our goal is to give a comparison between BERT-fused NMT and back-translation of different data scales, using monolingual data from the same source of BERT training by random selection from the Wikipedia[13]

En→De		Zh→En	
Auth (M)	Synth (M)	Auth (M)	Synth (M)
4.500	2.250 4.500 9.000 18.00	1.250	0.625 1.250 2.500 5.000

Table 2: Corpora statistics of sentence pairs.

[7] https://github.com/tjunlp-lab/CommonMT
[8] https://github.com/gabrielStanovsky/mt_gender
[9] https://nlp.stanford.edu/projects/nmt/
[10] https://github.com/pytorch/fairseq
[11] https://huggingface.co/bert-large-cased
[12] https://huggingface.co/bert-base-chinese
[13] dumps.wikimedia.org/dewiki/latest

[14]. Previous work shows that data capacity for back-translation does not consistently improve performance beyond a threshold (Poncelas et al., 2018), therefore we choose a suitable amount and scale up the data from 625k to 18M with the ratio between authentic and synthetic data being 1:0.5, 1:1, 1:2 and 1:4, respectively (see Table 2). In total we have 18M monolingual sentences in German and 5M monolingual sentences in English. All datasets are preprocessed similarly to the training data.

4.4 Evaluation

We use the `multi-bleu.perl` from Moses on tokenized sentences for BLEU evaluation of all systems. The tasks of conjunction disambiguation and idiom translation are evaluated on the presence percentage of correct conjunction and pre-defined blacklist words, respectively. The task of gender bias is evaluated on morphological analysis from 3 aspects: overall accuracy calculated by the percentage of instances in which the translation preserved the gender of the entity from the original sentence, ΔG denoting the difference in performance between masculine and feminine scores, and ΔS indicating the difference in performance between pro-stereotypical and anti-stereotypical gender role assignments (see examples in Appendix A.4).

Other tests use a contrastive pair paradigm, which tests a model's ability to discriminate between given good and bad translations by exploiting the fact that NMT systems can be viewed as language models of the target language, conditioned on source texts. Similar to language models, NMT models can score a negative log probability for sentences. If the model score of the actual translation is smaller than the contrastive translation, we treat the decision as correct. We aggregate model decisions on the whole test set and report the overall percentage of correct decisions as results.

5 Results

The overall BLEU points are given in Table 3[15]. For both rich- and low-resource settings, the BERT-fused model demonstrates stronger performances than the baseline. However, systems augmented with back-translated data are better than the BERT-fused model, with the best score achieved by model trained with 2.25M synthetic data (1:0.5 setting) for En→De, and 1.25M synthetic data (1:1 setting) for Zh→En. This shows that in terms of BLUE, the advantage of large-scale pre-training is not obvious compared with large-scale back-translation, even though the latter requires far less training data and computational resources. Taking En→De as an example (Table 4), back-translation uses only 85% parameters compared to the BERT-fused method, while achieves higher BLEU points, 3.6 times faster decoding speed, and the same target/source length ratio which indicates an equivalent information richness in the target translation.

System	En→De	Zh→En
Standard Transformer	29.20	45.15
+ back translation (1:0.5)	**30.41**	46.70
+ back translation (1:1)	30.25	**47.23**
+ back translation (1:2)	30.18	47.04
+ back translation (1:4)	30.25	46.39
BERT-fused model	30.03	46.55

Table 3: Model performance in terms of BLUE scores (case-insensitive). The best scores are marked in bold.

System	Params	Speed (tok/sec)	Len% (tgt/src)
Back-translation	2.93B	1269.46	0.95
BERT-fused model	3.43B	355.24	0.95

Table 4: Model comparison in En→De. We list the results of baseline model and Zh→En in Appendix B.

5.1 Morphology

Table 5 shows the results for the morphology test in En→De translation. Generally, for derivational (Table 5a), agreement (Table 5b) and consistency (Table 5c) content, pre-training does not show prominent advantages over back-translation in helping the standard Transformer model convey correct morphology from source to target. Prior work on monolingual tasks (Hofmann et al., 2020; Edmiston, 2020; Haley, 2020) has shown that BERT is capable of encoding morphological information and many morphological features can be extracted by training a simple classifier on a BERT layer. In our bilingual task, however, BERT is trained in the source context and evaluated in the target language. The performance discrepancy shows that BERT's morphology prediction for novel words in mono language results from high-frequent morphological data during pre-training, which helps BERT to memorize the statistical connection over contextualized string cues. In contrast, NMT morphological rules involve both source and target languages, which is different from BERT training. Surface cues are not available for BERT in bilingual

[14] dumps.wikimedia.org/enwiki/latest

[15] We successfully reproduced the BLUE scores of the baseline and BERT-fused model as reported in Zhu et al. (2020).

(a) derivation	Verbs				Pronouns	Nouns		Adjectives		Average
System	Past	Future	Cond.	Neg.	Plur.	Compd.	Nbr.	Compar.	Superl.	
Standard Transformer	91.40	76.90	91.10	97.80	98.10	**63.80**	**66.40**	92.20	97.80	86.17
+ back translation (1:0.5)	92.90	77.90	89.10	97.60	98.80	57.10	62.80	**93.30**	98.40	85.32
+ back translation (1:1)	93.10	77.90	88.90	97.60	98.70	60.20	61.80	**93.30**	98.00	85.50
+ back translation (1:2)	94.70	76.80	93.80	97.60	98.10	58.80	63.80	92.40	**98.90**	86.10
+ back translation (1:4)	**95.80**	**79.20**	**95.40**	**98.40**	**98.90**	57.50	65.10	92.70	97.30	**86.70**
BERT-fused model	93.30	77.10	91.50	97.80	98.30	63.10	64.30	90.70	97.30	85.93

(b) agreement	Coordinated verbs			Verbs	Complex NP		Coreference		Adj	Average
System	Nbr	Pers	Tense	Position	Gdr	Nbr	Relative	Personal	Strong	
Standard Transformer	94.20	94.20	94.20	92.60	**100.0**	**100.0**	67.50	93.80	94.10	89.81
+ back translation (1:0.5)	96.20	96.20	96.00	95.50	**100.0**	**100.0**	67.30	94.30	**97.60**	**91.04**
+ back translation (1:1)	**96.70**	**96.70**	**96.50**	**95.70**	**100.0**	**100.0**	66.20	94.40	96.50	90.89
+ back translation (1:2)	95.00	95.20	95.20	94.70	99.80	100.0	67.40	91.90	96.70	90.33
+ back translation (1:4)	96.30	**96.70**	96.30	95.60	**100.0**	**100.0**	65.70	93.60	96.60	90.65
BERT-fused model	96.50	**96.70**	**96.50**	93.90	**100.0**	**100.0**	**67.70**	**95.00**	94.10	90.81

(c) consistency	Nouns	Adjectives		Verbs			Average
System	Case	Gender	Number	Number	Person	Tense	
Standard Transformer	0.019	0.010	0.008	0.034	0.020	0.070	0.027
+ back translation (1:0.5)	0.021	0.004	0.002	0.027	0.017	0.061	0.022
+ back translation (1:1)	0.016	0.005	0.004	**0.024**	0.013	0.050	0.019
+ back translation (1:2)	0.017	0.004	0.004	0.025	**0.012**	0.057	0.020
+ back translation (1:4)	**0.015**	**0.002**	**0.001**	0.028	0.018	**0.046**	**0.018**
BERT-fused model	0.024	0.010	0.007	0.027	0.014	0.064	0.024

Table 5: Performance on morphology tests. Parts **a** and **b** are evaluated by Accuracy values, while **c** by Entropy.

situation thus BERT cannot compute the interlingual representations. This can explain why BERT contributes less than back-translation in conveying morphological features in bilingual scenarios.

5.2 Syntax

The results for syntax tests in En→De are shown in Table 6. We find similar performances across all systems, indicating that solving problems regarding syntax is easy for the current standard Transformer since it has achieved a high accuracy close to 100. Neither back-translation nor pre-training brings significant benefits to the baseline. Initial work on monolingual tasks (Goldberg, 2019; Wolf, 2019) claims that BERT learns powerful syntactic representations and shows promise at agreement phenomena. However, our results show that in translation, BERT performs at best no better than the Transformer baseline and back-translation techniques in favoring the grammatical variants in the target sides. Inspired by the results of morphological and syntactic evaluations, we leave for future work to separately incorporate the source and target side pre-training in the encoder and decoder of NMT, with the aim to better leverage linguistic information contained in language models (Guo et al., 2020).

5.3 Semantics

Figure 1 shows results for translating sentences with ambiguous words in both the news domain (in-domain) and colloquial speech domain (out-of-domain). In the news domain, the F-score of the baseline is 0.715. With back-translation, the performance fluctuates but is worse than the BERT-fused model. The BERT-fused model performs the best of 0.735 in F-score and improves the baseline by 2.8%. In the colloquial speech domain where words are more frequent than news domains and thus have more senses, the BERT-fused model still maintains the top and surpasses the baseline by 11.7%. There is evidence that BERT's context-aware embeddings actually encode certain forms of sense knowledge and provides distinct clusters corresponding to word senses (Wiedemann et al., 2019; Mickus et al., 2019). Thus we conclude that incorperating BERT's representation with NMT's encoder through attention mechanisms (Equation 3) enables the translation model to capture fine-grained nuances of meaning and thus is successful at differentiating source side ambiguous words. However, when domain shifts, all models decline in performance and the BERT-fused model is no exception. Previous work has proven that pre-training on large scale datasets can improve out-of-domain model robustness (Hendrycks et al., 2019; Mathis et al., 2021). It seems that this poten-

	Agreement			Polarity		
System	np	sv	verb	ins	del	trans
Standard Transformer	98.70	98.23	98.53	99.41	95.10	98.45
+ back translation (1:0.5)	98.88	98.39	**99.18**	99.36	95.52	98.71
+ back translation (1:1)	98.92	98.49	99.10	**99.43**	95.08	98.54
+ back translation (1:2)	98.91	98.49	99.10	99.38	95.18	98.60
+ back translation (1:4)	**99.04**	**98.61**	99.06	99.41	95.05	**98.80**
BERT-fused model	98.57	98.13	98.82	99.41	**95.72**	98.54

Table 6: Accuracy values for syntax test suite.

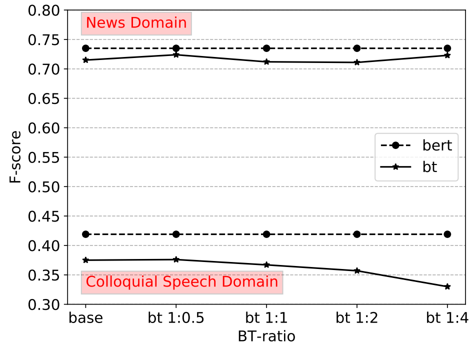

Figure 1: Results on homogragh translation test. We list specific data of each model in Appendix C.

tial is not fully exploited in cross-lingual settings. We plan to extend this point with the optimized model RoBERTa (Liu et al., 2019b) in future work.

Figure 2 shows the results for conjunction disambiguation. The accuracy of the BERT-fused model is 96.62, with which we identify a progress of the BERT-fused model over other systems. This shows that BERT's contextualized word embedding is useful to capture clues from sentence structures and form a generic idea of conjunctions. Conjunction can impact the structure of the surrounding sentences and is related more to fluency than to adequacy. Therefore it can be more difficult than content word ambiguity (Popović, 2019). We conclude that BERT can actually absorb fine-grained relevant sense information during pre-training, which helps learn meaningful conjunction sense distinctions.

Table 7 shows the results for coreference translation. The second column refers to the total accuracy of pronoun translation. The BERT-fused model achieves the score of 52.46, outperforming the others by 0.52-1.16 in accuracy. This corresponds to prior studies which show that BERT's attention matrices are able to do coreference resolution by effectively encoding coreference signal in deeper layers and at specific heads (Clark et al., 2019). The last two columns reflect the models' performance when antecedent location is inside or outside the current sentence. The accuracy of the BERT-fused model ranks the highest in short antecedent distance, outperforming others by 2-5 points, but deteriorates the most sharply as the distance between the pronoun and its antecedent increases. Though all models are ineffective in larger segments, the BERT-fused model even underperforms the baseline by 0.25 points. On the one hand, these observations prove the ability of BERT's deeply bidirectional representation con-

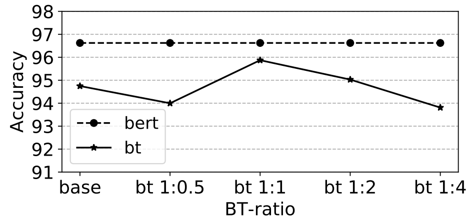

Figure 2: Results on conjunction disambiguation test. We list specific data of each model in Appendix C.

System	Total[1]	Intra[2]	External[3]
Standard Transformer	51.78	79.83	**44.76**
+ back translation (1:0.5)	51.30	82.33	43.54
+ back translation (1:1)	51.65	82.50	43.94
+ back translation (1:2)	51.64	82.08	44.03
+ back translation (1:4)	51.94	82.00	44.42
BERT-fused model	**52.46**	**84.25**	44.51

[1] Translating English pronoun *it* to German *es, sie, er*
[2] within segment [3] outside segment

Table 7: Accuracy values for reference pronoun translation(right part) and antecedent location (left part).

	Zh→En	En→De
System	Triggered	BLEU
Standard Transformer	377	29.54
+ back translation (1:0.5)	359	28.85
+ back translation (1:1)	306	27.53
+ back translation (1:2)	334	27.12
+ back translation (1:4)	344	26.76
BERT-fused model	**249**	**30.76**

Table 8: Results on idiom translation.

ditioned on both left and right context to capture intra-sentence dependency which is important for understanding coreferences. On the other hand, it also shows BERT's limitation on long-range features in document-level contexts, which is also observed by Joshi et al. (2019). As mentioned earlier in Section 4.2, one training task of BERT is to predict the next sentence. We assume that BERT is better than the standard Transformer to capture relation between two sentences and thus can improve performance on translation involving long-range features. Based on our results, however, seemingly BERT's potential in capturing sentence relations is not thoroughly exploited by NMT architectures.

5.4 Pragmatics

Table 8 shows results for idiom translation. Among all translations, the baseline triggers 377 literal errors. Back-translation makes progress on the basis of the baseline, while the BERT-fused model performs substantially better than all its counterparts, only triggering 249 literal errors in the blacklist. Regarding the effect of training data size, we find that from 377 errors with no back-translated sentence pairs to 306 with $1.25M$ sentence pairs, the

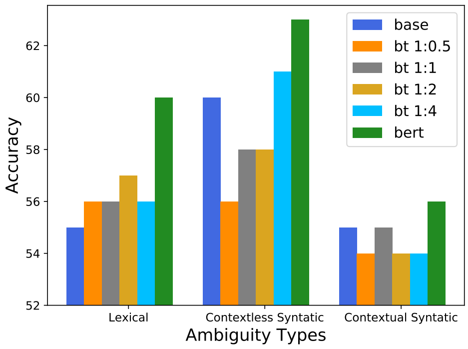

Figure 3: Results on commonsense reasoning.

errors continue to decrease as we add more synthetic data. However, it slightly rises when building systems with 2.5M synthetic data, showing that increasing data size is not the most useful to help idiom translation, while a better encoding of idiom expression via pre-training may help. The data size of Zh→En is relatively small, so we further verify BERT's effectiveness in the large-scale En→De experiment (elaborated in Appendix D). The BLEU results are summarized in the last column of Table 8. The BERT-fused model still gains the best performance among others with a score of 30.76. This shows that in addition to local syntactic properties, BERT's context-aware embedding based on previous and following context can help the encoder of NMT to capture global topical properties of words, thus making the model more expressive and understand the underlying meanings better.

The commonsense reasoning results are shown in Figure 3. The results clearly show that the BERT-fused model is better than the baseline and back-translated models in all three reasoning types, with the largest superiority on lexical ambiguity, a smaller gap on contextless syntactic ambiguity, and the weakest gap on context syntactic ambiguity. The performance of back-translation shows that incrementally larger amounts of training data do not consistently improve the commonsense reasoning performance of NMT, therefore it is likely the knowledge implied in the pre-trained language model that enhances commonsense reasoning ability of MT systems. Prior work (Zhou et al., 2020) has proven BERT's effectiveness in promoting commonsense ability in monolingual tasks. We further find that in bilingual scenario, BERT can also help model utilize knowledge via injecting prior information on the encoder part of NMT.

The results for gender translation are presented in Table 9. With BERT, gender bias in MT is not

System	Accuracy	ΔG	ΔS
Standard Transformer	71.2	3.9	9.3
+ back translation (1:0.5)	67.0	7.8	11.8
+ back translation (1:1)	71.6	2.7	10.6
+ back translation (1:2)	**75.1**	**0.1**	**5.2**
+ back translation (1:4)	72.1	2.0	5.5
BERT-fused model	71.4	3.2	14.6

Table 9: Performance on gender bias test suite. For ΔG and ΔS, higher numbers indicate stronger biases.

mitigated. The best performance is achieved by the model trained with back-translation data in a 1:2 setting, scoring 75.1, 0.1 and 5.2 in Accuracy, ΔG and ΔS, respectively. The scores of the BERT-fused model are 71.4, 3.2, 14.6, respectively, not competitive with the baseline on Accuracy and ΔG, and even much poor on ΔS. On the one hand, this further indicates that BERT may encode unintended social correlations during pre-training (May et al., 2019; Tan and Celis, 2019), and will propagate bias to downstream MT application. On the other hand, the poor ΔS score shows that the BERT-fused model is prone to translate based on gender stereotypes, and suffer deteriorated performance when translating antistereotypical assignments. This is in line with prior observations in QA and relation classification (Poerner et al., 2019) which shows that BERT's knowledge can come from learning stereotypical associations.

6 Conclusion

We presented a quantitative study of BERT in NMT as compared with large-scale back-translation. With 8 intrinsic evaluation tasks which cover a large range of linguistic phenomena, our observations suggest that BERT's bi-directional architecture, contextualized representation and knowledge learned from pre-training can help NMT manage semantic and pragmatic difficulties, but BERT-style representations may additionally introduce artifacts undesired in MT. For morphological and syntactic problems in which BERT does well in monolingual tasks, there is still limitation under the bilingual setting, requiring breakthroughs in BERT-fused modeling. Our findings about BERT are largely in line with research in monolingual setting, while we broaden the analysis under bilingual situations.

Acknowledge

We thank all anonymous reviewers for their constructive comments. This work is supported by a research grant from Sichuan Lan-bridge Information Technology Co., Ltd.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 546–560, Belgium, Brussels. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong. Association for Computational Linguistics.

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

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.

Daniel Edmiston. 2020. A systematic analysis of morphological content in bert models for multiple languages. *arXiv preprint arXiv:2004.03032*.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. 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 4052–4059, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. 2020. Incorporating bert into parallel sequence decoding with adapters. *arXiv preprint arXiv:2010.06138*.

Coleman Haley. 2020. This is a BERT. now there are several of them. can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341, Online. 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.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. 2019. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020. DagoBERT: Generating derivational morphology with a pretrained language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3848–3861, Online. Association for Computational Linguistics.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. 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.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W Mathis. 2021. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1859–1868.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2019. What do you mean, bert? assessing bert as a distributional semantics model. *arXiv preprint arXiv:1911.05758*.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa. *arXiv preprint arXiv:1911.03681*.

A Poncelas, D Shterionov, A Way, GM de Buy Wenniger, and P Passban. 2018. Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*.

Maja Popović. 2019. Evaluating conjunction disambiguation on English-to-German and French-to-German WMT 2019 translation hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. 2018. Evaluating machine translation performance on Chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

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

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *arXiv preprint arXiv:1911.01485*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Thomas Wolf. 2019. Some additional experiments extending the tech report`` assessing berts syntactic abilities`` by yoav goldberg. Technical report, Technical report.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. 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 4207–4216, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Details on Test Suites

For your reference, below we make more elaborations on evaluation test suites.

A.1 Morphology test

This test set is structured in the form of contrastive pairs. In accordance with Table 5, we have:

1. Verbs-past: differ in the tense of the main verb (present in one source sentence while past in the other).

2. Verbs-future: differ in the tense of the main verb (present in one source sentence while future in the other).

3. Verbs-cond.: a verb in future tense is turned into its conditional form.

4. Verbs-neg.: differ in the polarity of the main verb (affirmative in one source sentence while negative in the other).

5. Pronouns-plur.: differ in the number of the pronoun (a singular pronoun in one source sentence while a plural form in the other).

6. Nouns-compd.: the first source sentence contains a multiword expression that is most likely translated by a compound in German. The other is modified by one single English word in the multiword expression, such that the new German translation should result in a compound that has at least one morpheme in common with the one seen in the first translation.

7. Nouns-nbr.: differ in the number of the noun (a singular noun in one sentence while a plural form in the other).

8. Adjectives-compar.: differ in the form of the adjective (the bare adjective in one sentence while the comparative form in the other).

9. Adjectives-superl.: one sentence contains an adjective while the other contains its superlative form.

10. Coordinated verbs: one sentence contains a simple verb while the other contains a coordinated VP in the form of "verb and verb".

11. Verb position: the sentence pairs are generated by locating complex sentences where the principal clause can be omitted and the subordinate clause leads to a German translation where the verb should be located at the end of the clause.

12. Complex NP: one sentence contains a personal pronoun while the other contains a complex NP in the form of "adj+noun".

13. Coreference: one sentence contains a coreference link involving a personal pronoun (it) or a relative pronoun (that, which, who, whom, whose). The antecedent noun of the pronoun is changed to a synonym in the other sentence.

14. Strong adjective: one sentence contains a subject noun phrase with a definite article, an adjective and a noun. The other simply replaces the article by a possessive determiner. In German, an adjective following a definite article does not contain any gender marker in its ending, whereas it does contain it when following a possessive determiner.

15. Nouns: one sentence contains a noun while the other with hyponyms.

16. Adjectives: one sentence contains an adjective while the other with hyponyms.

17. Verbs: one sentence contains a verb while the other with hyponyms.

A.2 Syntax test

This test set is structured in the form of contrastive pairs. In accordance with Table 6, we have:

1. Noun-phrase agreement: the determiners agree with their head noun in number and gender in one sentence, while the other sentence randomly changes the gender of a singular definite determiner to introduce an agreement error.

2. Subject-verb agreement: subjects and verbs agree with one another in grammatical number and person in one sentence, while the other swaps the grammatical number of a verb to introduce an agreement error.

3. Separable verb particle: verbs and their separable prefix form a semantic unit in one sentence, while the other sentence replaces a separable verb particle with one that has never

been observed with the verb in the training data.

4. Polarity-inserting: one sentence remains the right polarity, while in the other sentence we reverse polarity by inserting the negation particle nicht (not) or the negation prefix -un.

5. Polarity-deleting: one sentence remains the right polarity, while in the other sentence we reverse polarity by deleting the negation particle nicht (not) or the negation prefix -un.

6. Transliteration: one sentence maintains a right name, while in the other sentence, two adjacent characters of the name are swapped.

A.3 Pragmatics test: Commonsense

In accordance with Figure 3, we have:

1. Lexical ambiguity: relates to word meanings which can be disambiguited by resorting to commonsense knowledge.

2. Contextless syntactic ambiguity: relates to sentence structures which can be correctly interpreted by resorting to commonsense knowledge.

3. Context syntactic ambiguity: relates to sentence structures which cannot be interpreted uniquely if no more context is given.

A.4 Pragmatics test: Gender bias

In accordance with Table 9, we have:

1. Masculine and feminine gender role: e.g., a male doctor versus a female nurse.

2. Stereotypical and anti-stereotypical gender role: e.g., a female nurse versus a female doctor.

B Model comparison

Below we list supplement results of model comparison in Zh→En (Table 10) and En→De (Table 11).

C Data of experiment results

Below we list specific data of each model in the tests of homograph translation (Table 12), conjunction disambiguation (Table 13) and commonsense reasoning(Table 14).

D Idiom translation in En→De

Fadaee et al. (2018) build a bilingual data set for idiom translation in En→De. It consists of 1500 parallel sentences whose English side contains an idiom and the German side refers to a proper reference translation. The evaluation method is BLEU. We adopt this data set in our experiment.

Zh→En	Params	Speed (tok/sec)	Len% (tgt/src)
Transformer	2.69B	1533.02	1.3
Back-translation	2.69B	1533.02	1.3
BERT-fused model	3.13B	732.07	1.3

Table 10: Supplement of Zh→En Model comparison.

En→De	Params	Speed (tok/sec)	Len% (tgt/src)
Transformer	2.93B	1269.46	0.95

Table 11: Supplement of En→De Model comparison.

	News Domain			Colloquial Speech Domain		
System	Precision	Recall	F-score	Precision	Recall	F-score
Standard Transformer	0.781	0.659	0.715	0.442	0.326	0.375
+ back translation (1:0.5)	0.788	**0.670**	0.724	0.447	0.325	0.376
+ back translation (1:1)	0.792	0.647	0.712	0.430	0.321	0.367
+ back translation (1:2)	0.794	0.644	0.711	0.437	0.303	0.357
+ back translation (1:4)	0.796	0.662	0.723	0.427	0.270	0.330
BERT-fused model	**0.816**	0.669	**0.735**	**0.510**	**0.356**	**0.419**

Table 12: Results on homograph translation test.

System	Total
Standard Transformer	94.74
+ back translation (1:0.5)	94.00
+ back translation (1:1)	95.87
+ back translation (1:2)	95.03
+ back translation (1:4)	93.81
BERT-fused model	**96.62**

Table 13: Accuracy for conjunction disambiguation test.

System	LA[1]	CL_SA[2]	CT_SA[3]
Standard Transformer	55	60	55
+ back translation (1:0.5)	56	56	54
+ back translation (1:1)	56	58	55
+ back translation (1:2)	57	58	54
+ back translation (1:4)	56	61	54
BERT-fused model	**60**	**63**	**56**

[1] lexical ambiguity [2] contextless syntactic ambiguity
[3] contextual syntactic ambiguity

Table 14: Accuracy for commonsense reasoning test.