The Box is in the Pen: Evaluating Commonsense Reasoning in Neural Machine Translation

Jie He[†*], **Tao Wang**[‡*], **Deyi Xiong**[†], and **Qun Liu**[§]
[†] College of Intelligence and Computing, Tianjin University, Tianjin, China
[‡] School of Computer Science and Technology, Soochow University, Suzhou, China
[§] Huawei Noah's Ark Lab, Hong Kong, China
jieh@tju.edu.cn, rgwt1234@gmail.com
dyxiong@tju.edu.cn, qun.liu@huawei.com

Abstract

Does neural machine translation yield translations that are congenial with common sense? In this paper, we present a test suite to evaluate the commonsense reasoning capability of neural machine translation. The test suite consists of three test sets, covering lexical and contextless/contextual syntactic ambiguity that requires commonsense knowledge to resolve. We manually create 1,200 triples, each of which contain a source sentence and two contrastive translations, involving 7 different common sense types. Language models pre-trained on large-scale corpora, such as BERT, GPT-2, achieve a commonsense reasoning accuracy of lower than 72% on target translations of this test suite. We conduct extensive experiments on the test suite to evaluate commonsense reasoning in neural machine translation and investigate factors that have impact on this capability. Our experiments and analyses demonstrate that neural machine translation performs poorly on commonsense reasoning of the three ambiguity types in terms of both reasoning accuracy ($\leqslant 60.1\%$) and reasoning consistency ($\leqslant 31\%$). We will release our test suite as a machine translation commonsense reasoning testbed to promote future work in this direction.

1 Introduction

Sixty years ago, the pioneering machine translation researcher and linguist Bar-Hillel published his well-known argument on the non-feasibility of general-purpose fully automatic high-quality machine translation (FAHQT) due to the inevitable requirement of world knowledge to help machine translation to infer correct translations for ambiguous words or linguistic structures (Bar-Hillel, 1960a). The example that Bar-Hillel uses as an evidence for the need of commonsense knowledge in machine translation is "The box is in the pen", where machine translation is expected to perform reasoning on the relative sizes of "box" and "pen". Bar-Hillel also doubts that a machine, even equipped with extra-linguistic knowledge, would be able to reason with such knowledge spontaneously as human translators do (Bar-Hillel, 1960a; Macklovitch, 1995).

Modern natural language processing (NLP) has made tremendous progress, not only in building abundant resources to develop linguistic insights, but also in plenty of methodological practices. On the one hand, machine translation has been substantially advanced with large-scale parallel data and statistical models. Recent results even suggest that the quality of machine-generated translations is approaching professional human translators (Wu et al., 2016; Hassan et al., 2018). On the other hand, a wide variety of efforts have been conducted to either examine the commonsense reasoning capability of neural models in natural language understanding, establish commonsense reasoning challenges or enhance neural models in commonsense reasoning (Zhang et al., 2018; Talmor et al., 2018; Huang et al., 2019; Sap et al., 2019b).

Comparing with Bar-Hillel's doubts and recent progress on machine translation and commonsense reasoning, it is natural for us to ask questions: do we solve the machine translation impasse related to commonsense reasoning? Or particularly, are current neural machine translation models able to learn common sense? And if so, how much do they learn? Does neural machine translation acquire sufficient commonsense knowledge and have strong ability in commonsense reasoning to generate human-level high-quality translations? Methodological discussion on the feasibility of FAHQT given the recent progress is far beyond the scope of this work. Instead, we focus on empirically ana-

[*] Equal Contributions.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3662–3672

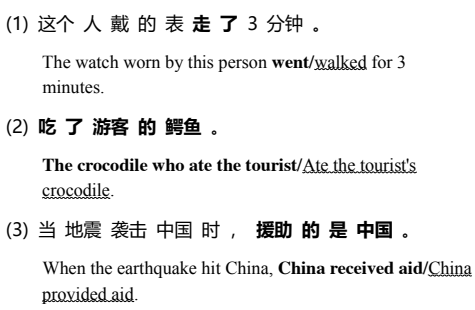

Figure 1: Examples of the lexical ambiguity (1), contextless syntactic ambiguity (2) and contextual syntactic ambiguity (3). English Translations in bold are correct while underlined translations are incorrect.

lyzing the capability of state-of-the-art neural machine translation models in using extra-linguistic commonsense knowledge to resolve ambiguity at different linguistic levels and select correct translations after disambiguation.

In order to achieve this goal, we manually build a machine translation commonsense reasoning test suite on Chinese-to-English translation with three types of commonsense-related ambiguities: lexical ambiguity, contextless and contextual syntactic ambiguity (see Section 3.1 for more details). Examples are shown in Figure 1. With this test suite, we thoroughly evaluate the commonsense reasoning ability of state-of-the-art neural machine translation models, e.g., LSTM- and Transformer-based NMT (Bahdanau et al., 2015; Vaswani et al., 2017). We also conduct analyses on the commonsense reasoning capability according to commonsense knowledge types, sentence length and reasoning consistency and the size of training data.

To the best of our knowledge, this is the first work to understand and measure the commonsense reasoning capability in neural machine translation. The contributions of this paper can be summarized as follows:

- We build a test suite[1] to examine the ability of neural machine translation in commonsense reasoning, which provides a benchmark testbed for tracking progress in this direction.
- Based on our experiments and analyses on evaluating commonsense reasoning in NMT, we find that: 1) commonsense reasoning related to lexical ambiguity and contextual syntactic ambiguity is more difficult than contextless syntactic ambiguity; 2) although the commonsense reasoning accuracy is higher than 50%, the reasoning consistency rate is far lower than 50% (random guess).

[1]The built commonsense test suite will be publicly available at https://github.com/tjunlp-lab/CommonMT.

2 Related work

We briefly review recent efforts related to commonsense reasoning in NLP. We refer readers to Storks et al. (2019)'s article for a thorough survey in this area.

Commonsense Datasets

According to Gunning (2018), commonsense knowledge normally consists of a general theory of how the physical world works and a basic understanding of human motives and behaviors. In recent years, a wide variety of datasets on the two kinds of commonsense knowledge have been proposed. Sap et al. (2019b) introduce Social IQA, containing 38k multiple choice questions for probing the commonsense reasoning about emotional and social in people's daily life. Similarly, Event2mind and Atomic (Rashkin et al., 2018; Sap et al., 2019a) focus on inferred knowledge in the form of *if-then* to reason about people's daily life behavior. For datasets on physical common sense, PIQA (Bisk et al., 2020) on commonsense phenomena in the physical world contains 21K QA pairs. SWAG and HellaSwag (Zellers et al., 2018, 2019) are datasets on commonsense NLI, where materials from video subtitles and wikihow articles are used to construct cloze tests. Bhagavatula et al. (2019) propose a dataset for abductive reasoning on events. The well-known Winograd Schema Challenge (WSC) test set (Levesque et al., 2012; Sakaguchi et al., 2020) focus on solving the commonsense problems in the form of coreference resolution. Different from them on monolingual data, we provide a bilingual commonsense test suite for machine translation.

Commonsense Reasoning in NLP

In addition to common sense datasets, we have also witnessed that commonsense knowledge has been recently explored in different NLP tasks. Just to name a few, Trinh and Le (2018), He et al. (2019) and Klein and Nabi (2019) use language models trained on huge text corpora to do inference on the WSC dataset. Ding et al. (2019) use commonsense knowledge in Atomic (Sap et al., 2019a) and Event2mind (Rashkin et al., 2018) on downstream tasks such as script event prediction. Bi et al. (2019) exploit external commonsense knowledge from ConceptNet (Speer et al., 2016)) in machine reading comprehension.

Commonsense Reasoning Evaluation
With pre-trained language models, like BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019) being widely used in various NLP tasks, studies have been performed to examine the commonsense reasoning capability in pre-trained neural language models. Wang et al. (2019) and Zhou et al. (2020) propose to measure the success rate of the pre-trained language models in commonsense inference by calculating LM probabilities. Two sentences which are used to test commonsense inference differ only in commonsense concepts. Feldman et al. (2019) further explore unsupervised methods to generate commonsense knowledge using the world knowledge of pre-trained language models. Our commonsense reasoning evaluation resonates with these evaluation efforts.

Commonsense Knowledge and Reasoning in Machine Translation
Commonsense knowledge has long been acknowledged as an indispensable knowledge source for disambiguation in machine translation (Bar-Hillel, 1960b; Davis and Marcus, 2015). Knowledge-based machine translation (KBMT), one of the popular machine translation paradigms in 1980s, lays much stress on extra-linguistic world knowledge in machine translation (Nirenburg, 1989). Large ontology that is constructed either manually or automatically to provide world knowledge is one of essential components in KBMT (Knight and Luk, 1994).

As data-driven machine translation, such as statistical machine translation (SMT) and neural machine translation, becomes de facto standard in machine translation, world knowledge has been less explicitly explored. Only a few studies have indirectly and partially exploited world knowledge in SMT or NMT, by incorporating linked open data resources such as DBpedia and BabelNet into SMT with modest improvements (Du et al., 2016; Srivastava et al., 2017; Moussallem et al., 2018).

3 Commonsense Reasoning Test Suite for Machine Translation

In this section, we discuss the design and construction of the test suite, including the rules and steps for building this test suite.

3.1 Test Suite Design

Different from commonsense reasoning in Winogram Schema Challenge (Levesque et al., 2012) or sentence reasonability judgment (i.e., "He put a turkey into the fridge" vs. "He put an elephant into the fridge") (Wang et al., 2019), where commonsense reasoning normally happens in one language, commonsense reasoning in NMT can be done either in the encoding of the source language (i.e., encoding reasonable source representations) or in the decoding of the target language (i.e., producing reasonable target outputs). As it is difficult to detect whether reasonable senses are identified and encoded in the encoder, we check target outputs from the decoder to test the commonsense reasoning capability of NMT. This is the first rule that we follow to design the test suite.

In the second rule for building the test suite, we manually create source sentences with ambiguity that requires commonsense reasoning. Inspired by Schwartz and Gomez (2009) and Ovchinnikova (2012), we ground the commonsense reasoning test on two types of ambiguity: lexical and syntactic ambiguity (LA and SA), which are common in machine translation. An example in LA is the "batter" in "she put the batter in the refrigerator" (food material vs. baseball player). SA relates to structures, for instance, "I saw a man swimming on the bridge" (I was standing on the bridge vs. The man was swimming on the bridge). We further refine SA into contextless (e.g., Example (2) in Figure 1) and contextual SA (e.g., Example (3) in Figure 1). The former can be correctly interpreted by resorting to commonsense knowledge while the latter cannot be interpreted uniquely if no more context is given.

The third rule that we conform to is to 1) create two contrastive source sentences for each lexical or syntactic ambiguity point, where each source sentence corresponds to one reasonable interpretation of the ambiguity point, and 2) to provide two contrastive translations for each created source sentence. This is similar to other linguistic evaluation by contrastive examples in the MT literature (Avramidis et al., 2019; Bawden et al., 2018; Müller et al., 2018; Sennrich, 2017). These two contrastive translations have similar wordings: one is correct and the other is not correct in that it translates the ambiguity part into the corresponding translation of the contrastive source sentence. This translation makes sense in the contrastive sentence but not in the sentence in question. Examples of contrastive source sentences and contrastive translations for each source sentence are shown in Figure 2, 3 and 4.

z_1 主力 部队 已经 对 敌人的 建筑 展开 了 **攻关** 。

e_1^r The main force has already launched an **attack** on the enemy's building.

e_1^c The main force has already launched a **research** on the enemy's building.

---

z_2 经过 两年 的 **攻关** ， 终于 解决 了 这道 技术 难题。

e_2^r After two years of **research**, this technical problem has finally been solved.

e_2^c After two years of **attack**, this technical problem has finally been solved.

Figure 2: An example block in the LA test set.

Finally, we have hired two linguistic experts to construct ambiguous source sentences and two professional human translators to provide contrastive translations for each source sentence. We ask them to create and translate with diverse words as much as possible and hire an extra linguistic expert and translator to review and double check source sentences and target translations after the two experts and translators cross check with each other.

3.2 Lexical Ambiguity Test Set

To construct this test set, we select words from a Chinese polysemous dictionary[2] so that the selected words have multiple interpretations. We avoid selecting words that are semantically close to one another in order to maintain diversity of the test set. We do not select words that are polysemous in Chinese but translated into the same words in English. Words that are translated into very different English words in different context and require commonsense knowledge to disambiguate are preferred.

This test set contains 200 example blocks. Each block is composed of two contrastive triples (z_1, e_1^r, e_1^c) and (z_2, e_2^r, e_2^c). As shown in Figure 2, z_1 and z_2 are contrastive with each other as they contain the same ambiguous word with different meanings. $e_\cdot^r$ and $e_\cdot^c$ are contrastive translations where the former is correct while the latter not. e_1^c and e_2^c are wrong translations in that they incorrectly interpret the ambiguous word in the way of e_2^r and e_1^r respectively. A selected polysemous word is used in only one example block.

3.3 Syntactic Ambiguity Test Sets

As mentioned before, we have two types of test sets for syntactic ambiguity: contextless and contextual

[2]Download link for the Chinese polysemous dictionary

z_1 维修 桌子 的 桌脚 。

e_1^r Repair the legs of the table.

e_1^c The leg that repairs the table.

---

z_2 维修 桌子 的 锤子 。

e_2^r The hammer that repairs the table.

e_2^c Repair the hammer of the table.

Figure 3: An example block in the contextless SA test set.

z_1 办公室 里 有 **两个 党 的 议员** ， 他们 互相 攻击 对方 党派 的 观点 。

e_1^r There are **members of two parties** in the office who are attacking each other's party views.

e_1^c There are **two members of the party** in the office who are attacking each other's party views.

---

z_2 办公室 里 有 **两个 党 的 议员** ， 他们 在 竞选 党 主席 。

e_2^r There are **two members of the party** in the office who are running for the party chairman.

e_2^c There are **members of two parties** in the office who are running for the party chairman.

Figure 4: An example block in the contextual SA test set.

SA. Before we construct the two test sets, we select Chinese structures that are typically ambiguous, just like PP attachment in English (e.g., "He ate the apple in the refrigerator" from Schwartz and Gomez (2009)).

Feng (1995) has deeply investigated syntactic ambiguity in Chinese and has found 26 structures that tend to generate sentences with different interpretations, such as "noun phrase + de (a Chinese particle) + shi (is) + noun phrase". From them, we use 12 structures to construct contrastive examples, where the subtle differences in Chinese can be clearly detected in English after translation.

With these 12 structure templates with potential syntactic ambiguity, we manually create 225 example blocks for the contextless SA test set and 175 blocks for the contextual SA test set. Examples of these two test sets are listed in Figure 3 and 4. Similar to the LA test set, each block is composed of two contrastive triples where two translations for each source sentence are also contrastive with each other in the way that we translate sentences in the LA test set. For the blocks in the contextless test set, we make sure that each ambiguous source sentence can be correctly interpreted with commonsense knowledge. We do not need extra context information for disambiguation. In con-

Test set	#triples	#unique tokens	Average tokens per sentence	total token numbers
LA	400	1,246/1,139/1,140	7.3/9.1/9.1	2,920/3,640/3,640
CL-SA	450	838/738/741	5.2/6.3/6.3	2,340/2,835/2,835
CT-SA	350	1,083/997/997	11.1/13.5/13.5	3,885/4,725/4,725
TOTAL	1,200	2,570/2,050/2,063	7.6/9.3/9.3	9,120/11,160/11,160

Table 1: Statistics on the test suite. Numbers a/b/c denote the corresponding number in source sentences/correct translations/incorrect translations. LA: lexical ambiguity; CL-SA: contextless SA; CT-SA: contextual SA.

Category	Descriptions	Examples	%
Properties	properties of objects	你/you 嘴/mouth 太快了/too fast	25.9
Behaviors	Behaviors that objects will take in a particular situation	鸡/ chicken 不/not 吃了/eat 因为/because 这只鸡/the chicken 已经/had already 吃了/eat 太多了/too much.	25.2
Taxonomy	Systematic classification of objects and concepts	今年/this year 风调雨顺/weather is good 农民的秋景/the harvest of the farmers' autumn 一定/must be 很好/very good.	21.1
Action	Some actions an object may be involved in	健康的/ healthy 医生/doctor 正在/is doing 手术/surgery.	15.8
Structures	Object A is part of Object B	削/Cut 西瓜的/the watermelon 皮/skin.	8.1
Emotions	Description of people's psychological activities and emotions	她/she 留下/leave 眼泪/tears 倾倒/pour out 她的/her 委屈/grievances.	2.6
Procedural	The type of common sense exercised in the performance of a task	学生/students 被调查/were investigated 因为/because 这些学生/these students 是/were 这个事件的/the incident 目击者/witnesses.	1.3

Table 2: Commonsense knowledge categories and their percentages in the test sets.

trast, we have to resort to additional context to interpret sentences in the contextual SA test set.

4 Test Suite Analysis

We provide statistical analyses on the built test suite, which cover its size, distribution of knowledge types and the reasoning accuracy of pretrained language models on target translations of target translations of this test suite.

4.1 General Statistics

Statistics on the built test suite are displayed in Table 1. We show the number of triples, the number of unique tokens, and the average number of tokens per sentence in each test set. Although sentences in the test suite are not very long, they are very challenging to be correctly translated as commonsense reasoning is involved, which will be verified in our experiments.

4.2 Commonsense Knowledge Type

Tandon et al. (2017) categorize commonsense knowledge into different types. Following their taxonomy of commonsense types, we compute the percentage of each type in our test suite, as shown in Table 2. Commonsense knowledge on properties, behaviors and taxonomy of objects/concepts are the top 3 commonsense knowledge types involved in our test sets.

	LA	CL-SA	CT-SA	Total
Random	0.500	0.500	0.500	0.500
GPT	0.775	0.650	0.597	0.678
GPT-2 base	0.803	0.642	0.606	0.688
GPT-2 medium	0.798	0.648	0.611	0.690
BERT-base	0.788	0.642	0.611	0.684
BERT-large	**0.818**	**0.682**	**0.623**	**0.712**

Table 3: Commonsense Reasoning accuracy of pretrained language models on the 1,124 instances of the test suite.

4.3 Evaluation of Pretrained Language Models on the Test Suite

In our test suite, we find that target translations of 93.7% instances (1,124 of 1200 test instances) can be determined if they are correct only from translations themselves (i.e., by performing commonsense reasoning), without reference to the corresponding source sentences. This is exactly what we want the test suite to be like as the purpose of this test suite is to evaluate commonsense reasoning rather than the ability of NMT in exploring source context for translation disambiguation not related to common sense. This is also consistent with the first rule for building the test suite: evaluating commonsense reasoning from the target side. Since the reasonability of these translations can be determined only from themselves, we want to know how challenging they are for pretrained language models in terms of commonsense reasoning. Hence, we evaluate state-of-the-art language models pretrained on large-scale data, including BERT (Devlin et al., 2019), GPT (Radford, 2018), and GPT-2 (Radford et al., 2019), on these 1,124 translation

pairs (pairs of reference and contrastive translations). For notational convenience, we still use the test suite to refer to these instances as only 76 cases are excluded for this evaluation.

Following Wang et al. (2019) and Zhou et al. (2020), for each pair (e^r, e^c), we use a pretrained language model to compute the language model score of the two translations. The translation with a higher score is labelled as the correct one by the language model. By comparing these labels with ground-truth labels, we can obtain the commonsense reasoning accuracy of the corresponding language model on these instances.

Results are shown in Table 3. All language models are better than random guess, validating the commonsense reasoning ability of them. They perform worse on the contextual SA test than on the other two test sets, demonstrating the difficulty in cross-sentence commonsense reasoning. BERT-large achieves the highest accuracy, 0.712. The number of parameters of BERT-large is equal to that of GPT2-medium, almost 3 times as large as that of GPT-2 base and BERT-base (340M vs. 117M). We conjecture that the reason for the superiority of BERT models over GPT/GPT-2 models is due to bidirectional context in BERT, which resonates with the findings of Zhou et al. (2020). The accuracies of all pretrained language models are all lower than 72%. This suggests that our test suite is very challenging in commonsense reasoning even for language models trained on an amount of data.

5 Experiments

In this section, we conducted extensive experiments to evaluate the commonsense reasoning capability of state-of-the-art neural machine translation on the built test suite.

5.1 Experimental setup

We adopted the CWMT Chinese-English corpus[3] of news domain as training data for NMT systems. This corpus contains 9M parallel sentences. We used byte pair encoding compression algorithm (BPE) (Sennrich et al., 2016) to process all these data and restricted merge operations to a maximum of 30k.

We trained two neural machine translation models on the training data: RNNSearch (Bahdanau et al., 2015) and Transformer (Vaswani et al., 2017).

[3]Available at: http://nlp.nju.edu.cn/cwmt-wmt

We used the Transformer base model with 6 layers and 8 self-attention heads per layer. As for RNNSearch, we employed neural architecture with 4 layers of LSTM and 512-dimension hidden states. We used Adam (Kingma and Ba, 2015) to train both NMT models. $\beta 1$ and $\beta 2$ of Adam were set to 0.9 and 0.999, the learning rate was set to 0.0005, and gradient norm was set to 5. To take full advantage of GPUs, we batched sentences of similar lengths. We trained both models on a single machine with 8 1080Ti cards. Each mini-batch contained 32,000 tokens. During decoding, we employed the beam search algorithm and set the beam size to 5.

5.2 Evaluation Metrics

For translation performance evaluation, we used sacrebleu (Post, 2018) to calculate case-sensitive BLEU-4 (Papineni et al., 2001).

To evaluate the commonsense reasoning accuracy of NMT on the test suite, we applied NMT models to score each pair (s, t) as follows:

$$Score(t|s) = \frac{1}{|t|}\sum_{i=0}^{|t|} \log p(t_i|t_{<i}, s) \quad (1)$$

where $p(t_i|t_{<i}, s)$ is the probabilty of the target word t_i given the target history and source sentence. Given a triple (z, e^r, e^c), if an NMT model scores the reference translation higher than the contrastive translation (i.e., $Score(e^r|z) > Score(e^c|z)$), the NMT model is believed to make a correct commonsense reasoning prediction. This is reasonable as e^r and e^c are only different in words or structures related to the lexical or syntactical commonsense ambiguity point as described in Section 3.1. By scoring each triple with an NMT model, we can measure the commonsense reasoning accuracy of the model on our test suite.

5.3 Results

BLEU scores for the two NMT models are given in Table 4. Commonsense reasoning results on the test suite are provided in Table 5.

From the table and figure, we can observe that

- Both BLEU and commonsense reasoning accuracy clearly show that Transformer is better than RNNSearch.
- Both RNNSearch and Transformer perform better on the contextless SA than on the contextual SA according to the commonsense reasoning accuracy. This is consistent with the results of pretrained language models shown in

	LA	CL-SA	CT-SA	Total
RNNSearch	25.82	21.59	27.98	25.86
Transformer	31.97	27.84	31.30	30.75

Table 4: BLEU scores on the test sets.

	LA	CL-SA	CT-SA	Total
RNNSearch	0.543	0.569	0.551	0.555
Transformer	0.565	0.656	0.571	0.601

Table 5: Commonsense Reasoning accuracy on the test sets.

Table 3, suggesting that cross-sentence commonsense reasoning is also challenging for NMT. Notice that the commonsense reasoning accuracy of pretrained language models cannot be directly compared to that of NMT models due to different evaluation procedure, mechanisms for commonsense reasoning and different test data. The BLEU scores on the contextless SA test set are lower than those on the contextual SA. We conjecture that this is because the contextless SA test set consists of very short sentences. Wrongly translated words therefore have a very big impact on BLEU scores.

- The performance gap between Transformer and RNNSearch on the CL-SA test set is larger than that on the other two test sets. The reason might be that the self-attention mechanism allows Transformer to more easily detect collocations (e.g., "leg" and "table" in Figure 3) for disambiguation on the CL-SA test set. Many CL-SA cases can be disambiguated by collocations according to our observation on this test set.
- Compared with the relative BLEU improvement of Transformer over RNNSearch, the relative improvement in terms of commonsense reasoning accuracy is smaller (8.2% vs. 18.91% in BLEU), indicating that more efforts are expected to not only improve translation quality in terms of BLEU but also to enhance commonsense reasoning ability in NMT.

5.4 Effect of the Size of Training Data

We conducted experiments to investigate the impact of the amount of training data on the commonsense reasoning performance of the state-of-the-art NMT model Transformer. Results are displayed in Figure 5. Generally, with the increase of training data, The common-sense reasoning ability of NMT systems

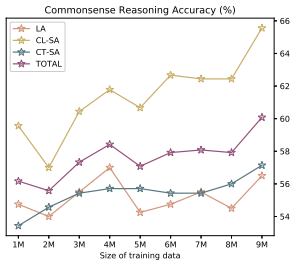

Figure 5: Commonsense Reasoning accuracy of the Transformer on the test sets with different size of training data.

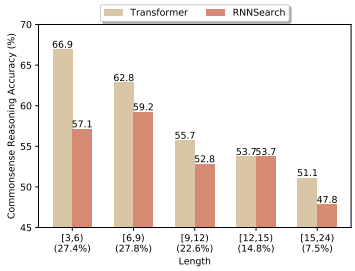

Figure 6: Commonsense Reasoning accuracy against the length of source sentences. The percentage of each group is shown under the corresponding length interval.

rises too. Although we used all CWMT Chinese-English training data to train NMT, we didn't have a chance to see that the commonsense reasoning accuracy tends to level off. We conjecture that the growth has the potential to continue. We leave using more data to measure the growth momentum of NMT commonsense reasoning to our future work.

Yet another finding from Figure 5 is that the commonsense reasoning performance on the contextless SA test set is always higher that the other two test sets. As shown in the last subsection, the reasons for this may be due to shorter sentences and collocations in this test set.

5.5 Effect of Sentence Length

We carried out an analysis on the impact of the length of source sentences on commonsense reasoning. We divided the test suite into 5 groups according to the length of source sentences. The results are shown in Figure 6. Generally, Transformer is better than RNNSearch in almost all length groups. With the length of source sentences increasing, the commonsense reasoning performance tends to go down. This may suggest that long-distance or cross-sentence commonsense reasoning is more challenging for NMT than short-distance reasoning, which is consistent with our finding on the CL-SA test set.

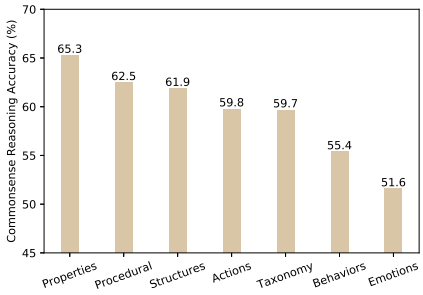

Figure 7: Commonsense Reasoning accuracy of the Transformer on the different commonsense knowledge types.

	RNNSearch	Transformer
LA	0.24	0.26
CL-SA	0.31	0.39
CT-SA	0.27	0.27
Total	0.27	0.31

Table 6: Rates of Reasoning consistency on the three test sets.

Error type	Example	%
Common sense errors	Origin: 公园里有三个幼儿园的孩子，总共有6个孩子在做游戏。 Reference: There are children of three kindergartens in the park, and a total of six children are playing games. Transformer: There are three kindergarten children in the park, a total of 6 children are playing games.	71.6%
Ordinary meaning errors	Origin: 这个工程已经下马。 Reference: This project has been abandoned. Transformer: The factory is already off.	22.7%
Other errors	Origin: 我写了六天字帖。 Reference: I wrote copybooks for six days. Transformer: I wrote six days.	5.7%

Table 7: Translation error types. Words related to translation errors are underlined.

5.6 Effect of Commonsense Knowledge Types

Finally, we analyzed the commonsense reasoning capability of Transformer on different commonsense knowledge types. Studying different types of common sense can help us understand what kind of commonsense knowledge is more needed to solve commonsense reasoning problems in NMT. The results are shown in Figure 7. Transformer-based NMT obtains relatively good results on commonsense reasoning on properties, structures, actions, but performs badly on reasoning on behaviors and emotions.

6 Further Analysis

6.1 Analysis on Reasoning Consistency

Our test suite contains 600 example blocks, each of which focuses on only one LA/SA ambiguity point. For the two reasonable interpretations (z_1, z_2) of a given ambiguity point, NMT models need to make two translation predictions: one for (e_1^r, e_1^c) and the other for (e_2^r, e_2^c). If they choose e_1^r and e_2^r (both right reasoning predictions) or e_1^c and e_2^c (both wrong reasoning predictions), we treat this as a consistent reasoning, otherwise inconsistent. Partially inspired by Zhou et al. (2020), we conducted an analysis on reasoning consistency.

We counted the times that a tested NMT model made consistent reasoning predictions and calculated the consistency rate on the three test sets. Results are shown in Table 6. Disappointingly, the reasoning consistency rates for both RNNSearch and Transformer are lower than random guess (0.5). On the contextless SA test set where both NMT models have higher reasoning accuracies, the rates of reasoning consistency are also higher than those of the other two test sets.

6.2 Analysis on Translation Errors

We have already automatically evaluated commonsense reasoning in NMT with both reasoning accuracy and reasoning consistency rate. We further manually analyzed the translation errors of Transformer on the entire test suite. We roughly grouped translation errors into three categories: common sense errors (translations that are not consistent with common sense), ordinary meaning errors (wrong translations of sources words that are not commonsense ambiguity points) and other errors (e.g., missing words). These errors were manually detected and labeled by two annotators. They checked all examples in the test suite. The inter-annotator agreement, measured as the rate of the number of labels that the two annotators annotate consistently against the total number of labels from the two annotators, is 92%.

Results are reported in Table 7. The majority of translation errors are indeed related to common sense (71.6%). This suggests that our test suite is a suitable and challenging testbed for evaluating commonsense reasoning in NMT.

7 Conclusion

In this paper, we have presented a test suite, including a lexical ambiguity test set and two syntactic ambiguity test sets, to evaluate the commonsense reasoning capability of state-of-the-art neural machine translation models. We elaborate the rules of building this test suite and conduct statistical analyses on it. Our evaluation experiments and analyses on this test suite suggest that commonsense reasoning in modern machine translation models is still in its infant stage and that more efforts are to be expected to advance NMT in this direction.

Acknowledgments

The present research was supported by the National Natural Science Foundation of China (Grant No. 61861130364), Natural Science Foundation of Tianjin (Grant No. 19JCZDJC31400) and the Royal Society (London) (NAF\R1\180122). We would like to thank the anonymous reviewers for their insightful comments. The corresponding author is Deyi Xiong (dyxiong@tju.edu.cn).

References

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yehoshua Bar-Hillel. 1960a. A demonstration of the nonfeasibility of fully automatic high quality machine translation. *Appendix III of 'The present status of automatic translation of languages', Advances in Computers*, 1:158–163.

Yehoshua Bar-Hillel. 1960b. The present status of automatic translation of languages. *Advances in Computers*, 1:91–163.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *ArXiv*, abs/1908.05739.

Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2019. Incorporating external knowledge into machine reading for generative question answering. 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 2521–2530, Hong Kong, China. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

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.

Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, and Junwen Duan. 2019. Event representation learning enhanced with external commonsense knowledge. 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 4896–4905, Hong Kong, China. Association for Computational Linguistics.

Jinhua Du, Andy Way, and Andrzej Zydron. 2016. Using BabelNet to improve OOV coverage in SMT. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 9–15, Portorož, Slovenia. European Language Resources Association (ELRA).

Joshua Feldman, Joe Davison, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models. *IJCNLP*, abs/1909.00505.

Zhiwei Feng. 1995. On the potential nature of chinese ambiguous constructions. *In Chinese. Journal of Chinese Information Processing*, 9(4):14–24.

David Gunning. 2018. Machine common sense concept paper. *CoRR*, abs/1810.07528.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Pengcheng He, Xiaodong Liu, Weizhu Chen, and Jianfeng Gao. 2019. A hybrid neural network model for commonsense reasoning. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 13–21, Hong Kong, China. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4831–4836, Florence, Italy. Association for Computational Linguistics.

Kevin Knight and Steve K Luk. 1994. Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Elliott Macklovitch. 1995. The future of mt is now and bar-hillel was (almost entirely) right. In *The Fourth Bar-Ilan Symposium on Foundations of Artificial Intelligence*.

Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. 2018. Machine translation using semantic web technologies: A survey. *ArXiv*, abs/1711.09476.

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.

Sergei Nirenburg. 1989. Knowledge-based machine translation. *Machine Translation*, 4(1):5–24.

Ekaterina Ovchinnikova. 2012. Integration of world knowledge for natural language understanding. In *Atlantis Thinking Machines*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WINOGRANDE: an adversarial winograd schema challenge at scale. *In AAAI*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. 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 4462–4472, Hong Kong, China. Association for Computational Linguistics.

Hansen A. Schwartz and Fernando Gomez. 2009. Acquiring applicable common sense knowledge from the web. In *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, pages 1–9, Boulder, Colorado, USA. 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. 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)*, volume 1, pages 1715–1725.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Ankit Srivastava, Georg Rehm, and Felix Sasaki. 2017. Improving machine translation through linked data. *Prague Bull. Math. Linguistics*, 108:355–366.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *CoRR*, abs/1904.01172.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*.

Niket Tandon, Aparna S. Varde, and Gerard de Melo. 2017. Commonsense knowledge in machine intelligence. *SIGMOD Record*, 46(4):49–52.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. 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*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *ArXiv*, abs/1810.12885.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *AAAI*.