Revisiting Negation in Neural Machine Translation

Gongbo Tang[1] **Philipp Rönchen**[1] **Rico Sennrich**[2,3] **Joakim Nivre**[1]
[1]Department of Linguistics and Philology, Uppsala University, Sweden
[2]Department of Computational Linguistics, University of Zurich, Switzerland
[3]School of Informatics, University of Edinburgh
gongbo.tang,philipp.rönchen,joakim.nivre{@lingfil.uu.se}
rico.sennrich@ed.ac.uk

Abstract

In this paper, we evaluate the translation of negation both automatically and manually, in English–German (EN–DE) and English–Chinese (EN–ZH). We show that the ability of neural machine translation (NMT) models to translate negation has improved with deeper and more advanced networks, although the performance varies between language pairs and translation directions. The accuracy of manual evaluation in EN→DE, DE→EN, EN→ZH, and ZH→EN is 95.7%, 94.8%, 93.4%, and 91.7%, respectively. In addition, we show that under-translation is the most significant error type in NMT, which contrasts with the more diverse error profile previously observed for statistical machine translation. To better understand the root of the under-translation of negation, we study the model's information flow and training data. While our information flow analysis does not reveal any deficiencies that could be used to detect or fix the under-translation of negation, we find that negation is often rephrased during training, which could make it more difficult for the model to learn a reliable link between source and target negation. We finally conduct intrinsic analysis and extrinsic probing tasks on negation, showing that NMT models can distinguish negation and non-negation tokens very well and encode a lot of information about negation in hidden states but nevertheless leave room for improvement.

1 Introduction

Negation is an important linguistic phenomenon in machine translation, as errors in translating negation may change the meaning of source sentences completely. There are many studies on negation in statistical machine translation (SMT) (Collins et al., 2005; Li et al., 2009; Wetzel and Bond, 2012; Baker et al., 2012; Fancellu and Webber, 2014, 2015), but studies on negation in neural machine translation (NMT) are quite limited and results are partly conflicting. For example, Bentivogli et al. (2016) find that negation is still challenging, whereas Bojar et al. (2018) show that NMT models almost make no mistakes on negation using 130 sentences with negation from three language pairs as the evaluation set. Hence, it is still not clear how well NMT models perform on the translation of negation.

In this paper, we present both automatic and manual evaluation of negation in NMT, in English–German (EN–DE) and English–Chinese (EN–ZH). The automatic evaluation is based on contrastive translation pairs and studies translation from English into German/Chinese (EN→DE/ZH). The manual evaluation targets translation in all four translation directions. We find that the modeling of negation in NMT has improved with deeper and more advanced networks. The contrastive evaluation shows that deleting negation from references is more confusing to NMT models compared to inserting negation into references. For the manual evaluation, NMT models make fewer mistakes on negation in EN–DE, than in EN–ZH, and there are more errors on negation in DE/ZH→EN than in EN→DE/ZH. Moreover, under-translation is the most prominent error type in three out of four directions.

The black-box nature of neural networks makes it hard to interpret how NMT models handle the translation of negation. In Ding et al. (2017), neither attention weights nor layer-wise relevance propagation can explain why negation is under-translated. We are interested in whether the information about negation is not well passed to the decoder. Thus, we investigate the negation information flow in NMT models by raw attention weights and attention flow (Abnar and Zuidema, 2020). We demonstrate that the under-translation

Transactions of the Association for Computational Linguistics, vol. 9, pp. 740–755, 2021. https://doi.org/10.1162/tacl_a_00395
Action Editor: Mauro Cettolo. Submission batch: 11/2020; Revision batch: 2/2021; Published 8/2021.

of cues is not caused simply by a lack of negation information transferred to the decoder. We further explore the mismatch between source and target sentences—negation cues appearing only on the source side or only on the target side. We find that there are roughly 17.4% mismatches in the training data in ZH–EN. These mismatches could confuse NMT models and make the learning harder. We suggest to distill or filter training data by removing the sentence pairs with mismatches to make the learning easier. In addition, we conduct intrinsic analysis and extrinsic probing tasks, to explore how much information about negation has been learned by NMT models. The intrinsic analysis based on cosine similarity shows that NMT models can distinguish negation and non-negation tokens very well. The probing results on negation detection reveal that NMT can encode a lot of information about negation in hidden states but still leaves much room for improvement. Moreover, encoder hidden states capture more information about negation than decoder hidden states.

2 Related Work

2.1 Negation in MT

Fancellu and Webber (2015) conduct a detailed manual error analysis and consider three categories of errors, *deletion*, *insertion*, and *reordering*. They find that negation scope is most challenging and *reordering* is the most frequent error type in SMT. Here we study the performance of NMT models on translating negation.

Bentivogli et al. (2016) and Beyer et al. (2017) find that NMT is superior to SMT in translating negation. Bentivogli et al. (2016) observe that placing the German negation cue *nicht* correctly during translation is a challenge for NMT models, which is determined by the focus of negation and need to detect the focus correctly. Bojar et al. (2018) evaluate MT models on negation, translating from English into Czech, German, and Polish, using 61, 36, 33 sentences, respectively, with negation as the test sets. They find that NMT models almost make no mistakes on negation compared to SMT—NMT models only make two mistakes in the English–Czech test set. In this paper, we will conduct manual evaluation on four directions with larger evaluation sets, to get a more comprehensive picture of the performance on translating negation.

Sennrich (2017) evaluates subword-level and character-level NMT models on the *polarity* set of *LingEval97* and finds that negation is still a challenge for NMT, via scoring contrastive translation pairs. More specifically, the deletion of negation cues causes more errors. Ataman et al. (2019) show that character-level models perform better than subword-level models on negation. Instead, we evaluate NMT models with different neural networks to learn their abilities to translate negation, by scoring contrastive translate pairs.

Ding et al. (2017) find that neither attention weights nor layer-wise relevance propagation can explain under-translation errors on a negation instance. Thus, understanding the mechanism of dealing with negation is still a challenge for NMT. Most recently, Hossain et al. (2020) study the translation of negation on 17 translation directions. They show that negation is still a challenge to NMT models and find that there are fewer negation-related errors when the language is similar to English, with respect to the typology of negation. In our work, we conduct both automatic and manual evaluation on negation, and explore the information flow of negation to answer whether under-translation errors are caused by a lack of negation information transferred to the decoder.

2.2 Negation in Other Areas of NLP

Negation projection is the task of projecting negations from one language to another language, which can alleviate the workload of annotating negation. Liu et al. (2018) find that using word alignment to project negation does not help the annotation process. They also provide the *NegPar* corpus, an EN–ZH parallel corpus annotated for negation. Here we apply probing classifiers to directly generate negation annotations on Chinese using hidden states.

Negation detection is the task of recognizing negation tokens, which can estimate the ability of a model to learn negation. Fancellu et al. (2018) utilize LSTMs, dependency LSTMs, and graph convolutional networks (GCN) to detect negation scope, using part-of-speech tags, dependency tags, and negation cues as features. Recently, the pretrained contextualized representations have been widely used in various NLP tasks. Khandelwal and Sawant (2020) employ BERT (Devlin et al., 2019) for negation detection, including negation cue detection, scope detection, and event

detection. Sergeeva et al. (2019) apply ELMo (Peters et al., 2018) and BERT to negation scope detection and achieve new state-of-the-art results on two negation data sets. Instead of pursuing better results, here we aim to probe how much information about negation has been encoded in hidden states in a negation detection task.

3 Background

3.1 Negation

Negation in text generally has four components: cues, events, scope, and focuses. The cues are the words expressing negation. An event is the lexical component that a cue directly refers to. The scope is the part of the meaning that is negated and the focus is the most explicitly negated part of the scope (Huddleston and Pullum, 2002; Morante and Daelemans, 2012).

NegPar is a parallel EN–ZH corpus annotated for negation. The English part is based on *ConanDoyle-neg* (Morante and Daelemans, 2012), a collection of four Sherlock Holmes stories. Some scope-related phenomena are re-annotated for consistency. The annotations are extended onto its Chinese translations. Here are two annotation examples:

English: There was **no** response.
Chinese: **mei** you ren da ying.
[no have people answer reply.]

In these examples, ***no*** and ***mei*** marked in bold are the cues; *response* and *da ying* enclosed in boxes are the events; the underlined words belong to the negation scope. In *NegPar*, negation events are subsets of negation scope, and negation focuses are not annotated. Table 1 shows detailed statistics of *NegPar*. Note that a negation instance may not have all the three components. Moreover, not all parallel sentence pairs have negation in both source and target sentences. For more details, please refer to Liu et al. (2018).

Due to the lack of parallel data annotated for negation, most of the negated sentences in the previous studies are selected randomly. In *NegPar*, not only negation cues, but also events and scope are annotated, which is beneficial to evaluating NMT models on negation and exploring the ability of NMT models to translate negation.

		Train	Dev	Test	Total
English	Cue	984	173	264	1,421
	Event	616	122	173	911
	Scope	887	168	249	1,304
Chinese	Cue	1,209	231	339	1,779
	Event	756	163	250	1,169
	Scope	1,160	227	338	1,725

Table 1: Statistics of negation components in *NegPar*.

Deletion	Insertion
deleting *nicht* (not)	inserting *nicht*
replacing *kein* (no) with *ein* (a)	replacing *ein* with *kein*
deleting *un-*	inserting *un-*

Table 2: Six ways to reverse the polarity of sentences from the *polarity* category of *LingEval97*.

3.2 Contrastive Translation Pairs

Since we evaluate NMT models explicitly on negation, BLEU (Papineni et al., 2002) as a metric of measuring overall translation quality is not helpful. We conduct the targeted evaluation with contrastive test sets in which human reference translations are paired with one or more contrastive variants, where a specific type of error is introduced automatically.

NMT models are conditional language models that assign a probability $P(T|S)$ to a given source sentence S and the target sentence T. If a model assigns a higher probability to the correct target sentence than to a contrastive variant that contains an error, we consider it as a correct decision. The accuracy of a model on such a test set is the percentage of cases where the correct target sentence is scored higher than all contrastive variants.

LingEval97 (Sennrich, 2017) has over 97,000 EN→DE contrastive translation pairs featuring different linguistic phenomena. In this paper, we focus on the *polarity* category, which is related to negation and consists of 26,803 instances. For contrastive variants, the polarity of translations are reversed by inserting or deleting negation cues. Table 2 illustrates how the polarity is reversed.

3.3 Attention Flow

In Transformer models, the hidden state of each token is getting more contextualized as we move

to higher layers. Thus, the raw attention weights are not the actual attention to the input tokens.

Recently, Abnar and Zuidema (2020) have proposed attention flow to approximate the information flow. Attention flow considers not only the attention weights to the previous layer but also to all the lower layers. Formally, in the self-attention networks, given a directed graph $G = (V, E)$, where V is the set of nodes, and E is the set of edges; each hidden state or word embedding from different layers is a node; the attention weight is the value of an edge. Given a source node s and a target node t, the attention flow is the flow of edges between s and t, where the flow value should not exceed the capacity of each edge and input flow should be equal to output flow for the intermediate nodes in the path s to t. They apply a maximum flow algorithm to find the flow between s and t in a flow network.

In short, the attention flow utilizes the minimum value of the attention weights in each path, and also employs the residual connections of attention weights. They find that the patterns of attention flow get more distinctive in higher layers compared to the raw attention. Moreover, attention flow yields higher correlations with the importance scores of input tokens obtained by the input gradients, compared to using the raw attention weights. Abnar and Zuidema (2020) explore the attention flow of the encoder self-attention in the case of pre-trained language models. Here we compute the attention flow from decoder layers to source word embeddings, in the context of NMT.

4 Evaluation

In this section, we present the results of both automatic and manual evaluation on negation in EN–DE and EN–ZH, to get a more comprehensive picture of the performance on translating negation.

4.1 NMT Models

We use the *Sockeye* (Hieber et al., 2017) toolkit to train NMT models. For EN→DE, we train RNN-, CNN-, and Transformer-based models, following the settings provided by Tang et al. (2018). For the other directions, we only train Transformer models. Table 3 shows the more detailed settings.

Neural network depth	8/6 (EN–DE/ZH)
Kernel size of CNNs	3
Trans. Att. head	8
Learning rate (initial)	2e-04
Embedding&hidden unit size	512
Mini-batch size (token)	4,096
Dropout (Trans./RNN&CNN)	0.1/0.2
RNN encoder	1 biLSTM + 6 uniLSTM
Optimizer	*Adam* (Kingma and Ba, 2015)
Checkpoint frequency	4,000
Label smoothing	0.1
Early stopping	32

Table 3: Settings for training NMT models.

EN→DE			DE→EN	EN→ZH	ZH→EN
RNN	CNN	Trans.	Trans.	Trans.	Trans.
25.2	25.3	27.6	34.3	33.9	23.5

Table 4: BLEU scores of NMT models with different architectures on the test sets (*newstest-2017*). *Trans.* is short for *Transformer*.

The training data is from the WMT17 shared task (Bojar et al., 2017).[1] There are about 5.9 million and 24.7 million sentence pairs in the training set of EN–DE and EN–ZH, respectively, after preprocessing with Moses scripts. Note that the training data on EN–ZH is from the official preprocessed data.[2] The Chinese segmentation is based on Jieba.[3] We learn a joint BPE model with 32K subword units (Sennrich et al., 2016) for EN–DE, and two BPE models with 32K subword units for Chinese and English, respectively. We employ the single model that has the best perplexity on the validation set for the evaluation, without any ensembles. Table 4 shows the BLEU scores of the trained NMT models on *newstest2017*, which are computed by *sacrebleu* (Post, 2018).[4]

Since these NMT models are trained with single sentences, feeding an input with multiple sentences into these models is likely to get an incomplete translation. To avoid these errors, we feed the sentence with negation cues into NMT models individually for the manual evaluation.

4.2 Automatic Evaluation

For the automatic evaluation, we let NMT models score contrastive translation pairs, in EN→DE and EN→ZH.

[1] http://www.statmt.org/wmt17/translation-task.html.

[2] http://data.statmt.org/wmt18/translation-task/preprocessed/zh-en/.

[3] https://github.com/fxsjy/jieba.

[4] https://github.com/mjpost/sacrebleu.

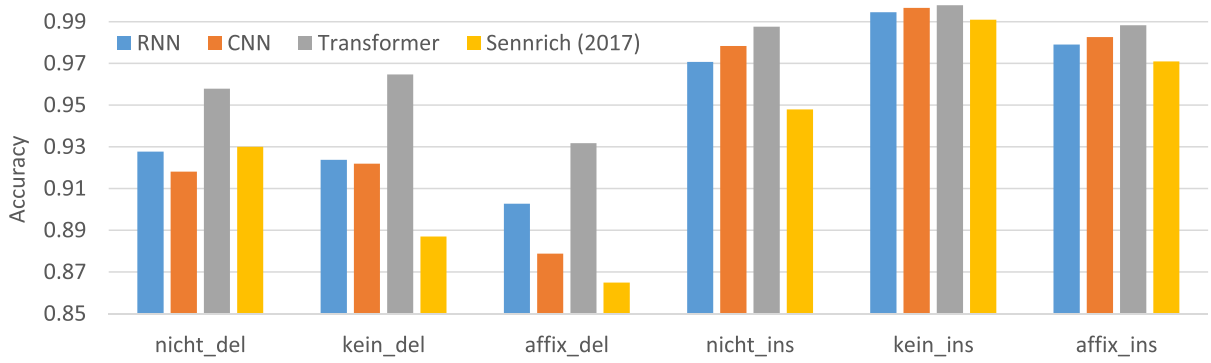

Figure 1: Performance of NMT models on scoring contrastive translations, in EN→DE, using the *polarity* category of *LingEval97*. The first three groups are on negation deletion, deleting *nicht*, *kein*, and affixes, while the last three groups are on negation insertion.

4.2.1 EN→DE

Sennrich (2017) has evaluated subword-level and character-level RNN-based models. Here we evaluate NMT models with different architectures, RNN-, CNN-, and Transformer-based models. The test set is the *polarity* category of *LingEval97*. Figure 1 displays the accuracy of NMT models.

Our NMT models are superior to the models in Sennrich (2017), except that *CNN* is inferior in the group *nicht_del*. Generally, we see that the performance on negation is getting better with the evolution of NMT models, with the *Transformer* consistently scoring best, and substantially better (by up to 8 percentage points) than the shallow RNN (Sennrich, 2017). The accuracy of the *Transformer* varies from 93.2% to 99.8%, depending on the group, which we consider quite strong.

It is interesting that NMT models make fewer mistakes when inserting negation cues into the reference compared to deleting negation cues from the reference, which means that positive contrastive variants are more confusing to NMT models. This is consistent with the results in Fancellu and Webber (2015), that SMT models make more errors when generating positive sentences than generating negative sentences, in terms of insertion/deletion errors. We will explore under-translation errors in the following sections.

4.2.2 EN→ZH

Following the *polarity* category in *LingEval97*, we create a contrastive evaluation set for negation on EN→ZH, using the development and test sets from the WMT shared translation task 2017–2020.[5] The contrastive evaluation set also has two sub-categories: negation deletion and negation insertion. We first select the five most popular Chinese negation cues – ‘‘bu’’, ‘‘mei’’, ‘‘wu’’, ‘‘fei’’, and ‘‘bie’’. Then, we manually delete the negation cue from the reference or insert a negation cue into the reference, without affecting the grammaticality. The negation deletion and negation insertion categories have 2,005 and 3,062 instances with contrastive translations, respectively.

As Transformer models are superior to RNN- and CNN-based models, here we only evaluate Transformer models. The accuracy on negation deletion and negation insertion categories is 92.1% and 99.0%, respectively. We can see that Transformer models perform quite well on EN→ZH, but not as well as on EN→DE. In accord with the finding in EN→DE, Transformer models here in EN→ZH also perform worse on the negation deletion category.

4.3 Manual Evaluation

We have evaluated NMT models on negation with contrastive translation pairs. However, scoring contrastive translation pairs is not the same as evaluating the translations directly. The contrastive translations only insert or delete a negation cue compared to the references, which is quite different from the generation of NMT models. In addition, the automatic evaluation only gives us the general performance on negation without any details on how negation is translated.

[5] https://github.com/tanggongbo/negation-evaluation-nmt.

Category	Description
Correct	cues are translated into cues correctly
Rephrased	cues are translated correctly but not into a cue
Reordered	cues are translated but modify wrong constituents (incorrect scope/focus)
Incorrect	cues are translated but the event is translated incorrectly or the meaning is reversed
Dropped	cues are not translated at all

Table 5: Descriptions of the five translation categories.

	Correct	*Rephrased*	*Reordered*	*Incorrect*	*Dropped*	*Accuracy*
EN→DE	258 (92.8%)	8 (2.9%)	2 (0.7%)	3 (1.1%)	7 (2.5%)	95.7%
DE→EN	232 (92.8%)	5 (2.0%)	2 (0.8%)	11 (4.4%)	0 (0.0%)	94.8%
EN→ZH	393 (90.0%)	15 (3.4%)	3 (0.7%)	10 (2.3%)	16 (3.7%)	93.4%
ZH→EN	451 (80.1%)	65 (11.6%)	3 (0.5%)	21 (3.7%)	23 (4.1%)	91.7%

Table 6: Manual evaluation results in EN–DE and EN–ZH. Accuracy is the sum of *correct* and *rephrased*.

Thus, we further conduct manual evaluation on EN–DE and EN–ZH.

Due to the lack of parallel data annotated for negation, most of the negated sentences in previous studies have no annotations and are selected randomly. In *NegPar*, not only negation cues, but also events and scope, are annotated, which is beneficial for evaluating NMT models on negation and exploring the ability of NMT models to learn negation. These annotations allow us to evaluate negation from the perspectives of cues, events, and scope, rather than negation cues only. Thus, for EN–ZH, we conduct the manual evaluation based on *NegPar*, using both the development set and the test set. For EN–DE, we evaluate 250 sentences with negation cues that are randomly selected from *LingEval97* in each direction.

Given the strong performance of Transformer models in the automatic evaluation, we focus on this architecture for the manual evaluation. We classify the translations of negation into five categories: *Correct*, *Rephrased*, *Reordered*, *Incorrect*, and *Dropped*, depending on whether the cue, event and the scope are translated correctly. More detailed descriptions are provided in Table 5.

Table 6 gives the absolute frequency and percentage of each translation category in all the translation directions.[6] The accuracy of translating negation is the sum of *correct* and *rephrased*, and the accuracy in EN→DE, DE→EN, EN→ZH, and ZH→EN is 95.7%, 94.8%, 93.4%, and 91.7%, respectively. We can see that NMT models perform better at translating negation in DE–EN than in ZH–EN. In addition, under-translation errors are the main errors in three out of four directions while reordering errors only account for less than 1% in all directions. This contrasts with the results reported for SMT by Fancellu and Webber (2015), where reordering was a more severe problem than under-translation. It is reasonable because NMT models are conditional language models, and have fewer word order errors, compared to SMT models (Bentivogli et al., 2016), thus there are fewer reordering errors on translating negation. We can tell that the main error types with respect to negation have shifted from SMT to NMT.

4.3.1 EN–DE

As Table 6 shows, most of the translations belong to *correct*. The accuracy in EN→DE is 0.9% greater than that in DE→EN. 2.5% negation cues are not translated in EN→DE, while all the negation cues are translated by NMT models in DE→EN. However, there are more sentences where the negation events are not translated correctly in DE→EN. Compared to Bojar et al. (2018), our evaluation results for EN-DE are 4.3% lower. One possible reason for the difference is that our evaluation is based on a larger data set; another possible reason is that we also consider the translation of negation events and scope.

[6] `https://github.com/tanggongbo/negation-evaluation-nmt` provides the details.

Category	Source	Translation	Reference
Correct	would do him **no** harm	bu hui shang hai ta (not able to harm him)	dui ta bu hui you shen me hai chu (to him no able have any harm)
Rephrased	**bu** xi fei yong (no spare expense use)	able to spend enough money	spare no expense
Reordered	yi ge xing qi **bu** jian mian (a week no meet)	no one could meet for a week	be invisible for a week
Incorrect	spare **no** expense	bu yao hua qian mai (not spend money to buy)	bu xi fei yong (not spare expense)
Dropped	**bu** xing , Mo li luo zhi dao le (not fortunate, Murillo know truth already)	fortunately, Murillo knew that	Unhappily, Murillo heard of

Table 7: Translation examples (segments) from different categories. These segments are a subset of negation scope. The word in bold in the source is the cue. Words with dashed lines below are correct translations and words with wavy lines below are incorrect translations.

4.3.2 EN–ZH

Similar to the results in EN–DE, the accuracy in translating from English is greater than in translating into English. The accuracy in ZH→EN is 1.7% lower than in EN→ZH. There are more instances of negation that are rephrased in the translations in ZH→EN, without any negation cues in the translations. The NMT model in ZH→EN also makes more under-translation errors.

Table 7 further provides some translation examples. In the category *Rephrased*, negation cues are not directly translated into negation cues. Instead, the negation is paraphrased in a positive translation. In the *Rephrased* example, although there is no cue in the translation, the meaning is paraphrased by translating *bu xi* [no spare] into *spend*. In the *Reordered* example, the cue *bu* in the source is supposed to modify *jian* [meet], but the translation of the cue is placed before *one*, modifying the subject *one* instead of *meet*. In addition, even though the negation cues are translated, the negation events could be translated incorrectly, which can also have a severe impact on the translation. For the fourth example, there is a cue in the translation but *spare* in the source is translated into *spend*, which reverses the meaning completely. For the last example, the cue *bu* [no] is skipped and only the event *xing* [fortunate] gets translated.

We further check the under-translation errors of negation cues and find that some of them are caused by multi-word expressions (idioms), especially when translating Chinese into English. For example, *wu* [no] in *wu_bing_shen_yin* [no disease groan cry] is not translated. Fancellu and Webber (2015) have shown that the cues will not be under-translated if they are separate units in SMT. Thus, these words are then segmented into separate characters and the input is fed into NMT models again. This does fix a few errors. The *wu* [no] in *wu_bing_shen_yin* gets translated but the second *bu* [not] in *bu_gao_bu_ai* [not tall not short] is still not translated. Note that we only changed the segmentation during inference which is sub-optimal. We aim to show that the segmentation also could cause under-translation errors.

5 Interpretation

There are few studies on interpreting NMT models with respect to negation. Since Table 6 has shown that NMT models in EN–ZH suffer from more errors on negation, and since *NegPar* provides annotations of negation, we focus on interpreting NMT models in EN–ZH. NMT models consist of several components and we are interested in the information flow of negation to answer whether the under-translation is caused by not passing enough negation information to decoders, as well as exploring the ability of NMT models to learn negation.

5.1 Under-Translation Errors

Under-translation is the most frequent error type in our evaluation. If a negation cue is not translated by NMT models, either the negation information is not passed to the decoder properly, or the decoder does not utilize such information for negation generation. We employ raw attention weights and attention flow to explore the information flow.

5.1.1 Attention Distribution

Encoder-decoder attention weights can be viewed as the degree of contribution to the current word prediction. They have been utilized to locate unknown words and to estimate the confidence of translations (Jean et al., 2015; Gulcehre et al., 2016; Rikters and Fishel, 2017). However, previous studies have found that attention weights cannot explain the under-translation of negation cues (Ding et al., 2017). In this section, we first focus on the under-translated negation cues, checking the negation information that is passed to the decoder by the encoder-decoder attention. We compare the attention weights paid to negation cues, when they are under-translated and when they are translated into reference translations.

We extract attention distributions from each attention layer when translating sentences from the development set. Each attention layer has multiple heads and we average[7] the attention weights from all the heads. We utilize constrained decoding (Post and Vilar, 2018) to generate reference translations to get gold attention distribution. We find that source negation cues attract much less attention compared to when they are translated into references. Thus, we hypothesize that sufficient information about negation has not been passed to the decoder, and we can utilize the attention distribution to detect under-translated cues.

Now we further explore the attention distribution of under-translated and correctly translated cues, without using the gold attention distribution. We compute the Spearman correlation (ρ) between the weights and categories. If $|\rho|$ is close to 1, then categories have a high correlation with attention weights. However, the largest $|\rho|$ in EN→ZH and ZH→EN is 0.15 and 0.23, respectively, which means that there is almost no correlation between attention weights and categories. We inspect the weights and find that the weights to correctly translated cues range from 0.01 to 0.68, which cover most of the weights to dropped cues. This means that we cannot detect under-translated cues by raw attention weights.

As raw attention weights in Transformer are not the actual attention to input tokens, in the next section, we will apply attention flow, which has been shown to have higher correlation with the input gradients, to measure the negation flow.

[7]We also used maximum weights to avoid misleading conclusions when using average weights if the negation is modeled by a specific head, and we obtained the same conclusion.

		EN→ZH		ZH→EN	
Layer	Group	Attention flow	$\lvert\rho\rvert$	Attention flow	$\lvert\rho\rvert$
2	✓	0.89	0.04	0.80	0.15
	✗	0.90		0.70	
4	✓	0.89	0.06	0.85	0.08
	✗	0.91		0.84	
6	✓	0.77	0.06	0.82	0.07
	✗	0.78		0.72	

Table 8: Attention flow values from different decoder layers to source cues, and the absolute value of Spearman correlation (ρ) between attention flow and the cue's category. ✓ represents the correctly translated cues and ✗ represents the under-translated cues.

5.1.2 Attention Flow

We compute the attention flow to negation cues belonging to different groups; the input nodes are the hidden states from decoder layers; the output node is the word embedding of the negation cue. We utilize the maximum attention flow from the decoder to represent the attention flow to each source cue, and report the average value of all the attention flow. Table 8 shows the attention flow values from different decoder layers to source cues, and the absolute value of Spearman correlation (ρ) between attention flow and the cue's category. The attention flow values range from 0.70 to 0.91 for all the cues, which means that most of the cue information has been passed to the decoder and that the under-translation is not caused by not passing negation information to the decoder.

In addition, the attention flow values in *Dropped* and *Correct* are almost the same in EN→ZH and the correlation is smaller than 0.1. In ZH→EN, the attention flow is more distinct in the two cue groups, but the correlation values are still smaller than 0.15. Compared to raw attention weights, attention flow can provide more accurate information flow to the decoder, but neither raw attention weights nor attention flow exhibit any correlation between under-translation and the amount of negation information passed to the decoder.

Our analysis indicates that under-translation of negation cues may still occur even though there is information flow from the source negation cue

EN \ ZH	has_cue	no_cue
has_cue	2.60M (10.5%)	**0.15M (0.6%)**
no_cue	**4.16M (16.8%)**	17.84M (72.1%)

Table 9: Statistics of sentence pairs with and without cues in ZH–EN, including absolute number and ratio. ‘‘M’’ is short for million. Numbers in bold denote sentence pairs with cue-mismatch.

to the decoder. This indicates that methods to manipulate the attention flow, such as coverage models or context gates (Tu et al., 2016, 2017) may not be sufficient to force the model to produce negation cues. Our results also indicate that under-translation of negation cues may not be easily detectable via an analysis of attention.

5.1.3 Training Data Considerations

To further investigate why a model would fail to learn the seemingly simple correspondence (in the language pairs under consideration) between source and target side negation cues, we turn to an analysis of the parallel training data. Our manual analysis of the test sets has shown a sizeable amount (2–11%) of rephrasing where the translation of a negation is correct, but avoids grammatical negation. We hypothesize that such training examples could weaken the link between grammatical negation cues in the source and target, and favor their under-translation.

We perform an automatic estimate of cue-matches and cue-mismatches between source and target in the training data based on a short list of negation words.[8] Table 9 displays the amount of cue-match and cue-mismatch sentence pairs. There are 17.4% sentence pairs with cue-mismatch,[9] predominantly in ZH→EN, which agrees with the high amount of rephrasing we observed in our manual evaluation (Table 6). Such cue-mismatch sentence pairs, along with cue-match pairs, can make the learning harder and cause under-translation errors when there is no paraphrase to compensate for the dropped negation cue. Thus, one possible solution is to distill or filter training data to remove cue-mismatch sentence pairs to make the learning easier.

[8]English negation words: *no, non, not, ’t, nothing, without, none, never, neither*. Chinese negation characters: *bu, mei, wu, fei, bie, wei, fou, wu*.

[9]Note that this is only a simple approximation. We aim to demonstrate the sizeable mismatched training data rather than the accurate distribution. We manually checked 100 randomly selected sentence pairs, of which 30% are classified incorrectly. These errors are caused by ignoring English words with negative prefixes/suffixes or viewing any Chinese words with negative characters as negative words, such as *unknown* in English and *nan fei* [South Africa] in Chinese.

5.2 Intrinsic Investigation

We are also interested in exploring whether NMT models can distinguish negation and non-negation tokens, and therefore conduct an intrinsic investigation on hidden states—by computing the cosine similarity between tokens with different negation tags. Since NMT models can translate most negation instances correctly, we hypothesize that the hidden states are capable of distinguishing negation from non-negation tokens. We investigate hidden states from both encoders and decoders. As the hidden state in the last decoder layer is used for predicting the translation, we only explore the decoder hidden states at the 6th layer. We use Sim_{ce} to represent the cosine similarity between negation cues and negation events, Sim_{cs} to represent the cosine similarity between negation cues and tokens belonging to negation scope, and Sim_{co} to represent the cosine similarity between negation cues and non-negation tokens. We simply use the mean representation for tokens that are segmented into subwords.

Figure 2 shows the cosine similarity between negation cues and events, scope, and non-negation tokens, using hidden states from encoders and decoders. Sim_{ce} is substantially higher than Sim_{cs}, and Sim_{cs} is higher than Sim_{co}. This result reveals that negation events are closer to negation cues compared to tokens belonging to the negation scope. We can also infer that NMT models can tell negation and non-negation tokens apart as Sim_{co} is distinctly lower than Sim_{ce} and Sim_{cs}. However, even the highest Sim_{ce} is only around 0.5, which means that the representations of negation components are quite different.

In the encoder, Sim_{ce}, Sim_{cs}, and Sim_{co} have the same trend that the similarity is higher in upper layers. In addition, we can tell that negation cues interact with events and scope, but also non-negation tokens. Compared to the negation representations from encoders, the negation representations from decoders are less distinct because they are closer to each other. Sim_{ce}, Sim_{cs}, and

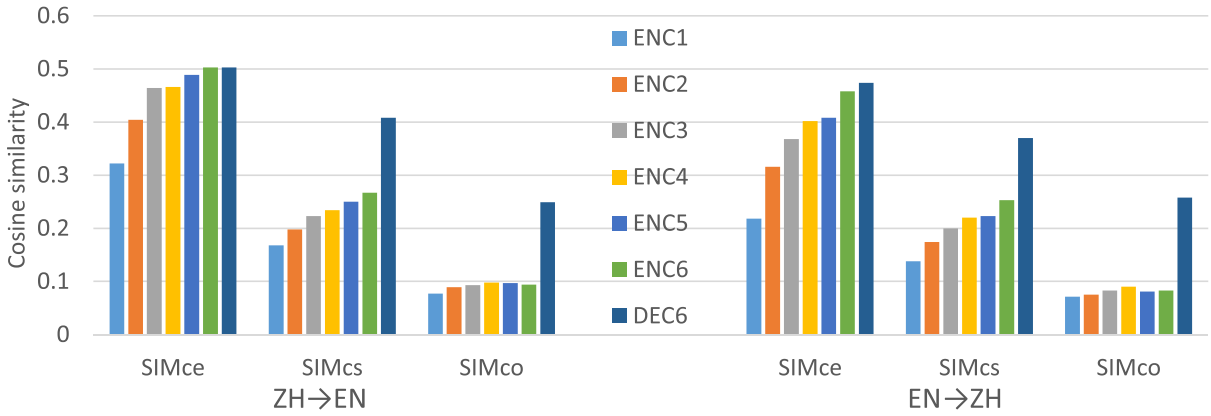

Figure 2: Cosine similarity between negation cues and events, scope, and non-negation tokens in ZH–EN, using hidden states from different layers. *ENCi* represents hidden states from the ith encoder layer and *DEC6* denotes hidden states from the 6th decoder layer.

Sim_{co} are higher when using the hidden states from the 6th decoder layer (*DEC6*) than when using the 6th encoder layer (*ENC6*). We attribute this to the fact that hidden states in decoders are more contextualized because they consider contextual information from both the source and the target.

5.3 Probing NMT Models on Negation

We have shown that NMT models can distinguish negation and non-negation tokens in the previous section, but how much information about negation has been captured by NMT models is still unclear. In this section we will investigate the ability to model negation in an extrinsic way, namely, probing hidden states on negation in a negation projection task (Liu et al., 2018) and a negation detection task (Fancellu et al., 2018). In the negation projection task, instead of projecting English negation annotations to Chinese translations using word alignment, we use probing classifiers trained on Chinese to directly generate the negation annotations. In the negation detection task in English, we employ simple classifiers rather than specifically designed models to detect each token. In brief, given a hidden state, we train classifiers to predict its negation tag, *cue*, *event*, *scope*, or *others*.

5.3.1 Settings

The probing task on negation cues is a binary classification task, the output space is $\{cue, others\}$, while the classifiers for event and scope are tri-class classification tasks with an output space $\{cue, event/scope, others\}$, because only predicting event/scope is challenging to these classifiers.

The probing classifiers in this section are feed-forward neural networks (MLP) with only one hidden layer, using ReLU non-linear activation. The size of the hidden layer is set to 512 and we use the Adam learning algorithm. The classifiers are trained using cross-entropy loss. Each classifier is trained on the training set for 100 epochs and tuned on the development set. We select the model that performs best (F1 score) on the development set and apply it to the test set. In addition, we train 5 times with different seeds for each classifier and report average results. We use precision, recall, and F1 score as evaluation metrics.

5.3.2 Negation Projection

Table 10 shows the projection results of negation cues, scope, and events, on both development and test sets. *ENC*/*DEC* refers to using hidden states from encoders or decoders. *ENC* achieves the best result on all the negation projection tasks and is significantly better than the word alignment based method in Liu et al. (2018). *ENC* also performs better than *DEC*, which means that negation is better modeled in encoder hidden states than in decoder hidden states.

In addition, we investigate hidden states from different encoder layers. Figure 4 shows the F1 scores on the development set, using hidden states from different encoder layers. We can see that hidden states from lower layers perform better in negation cue projection, while hidden states from upper layers are better in negation event/scope projection. One possible explanation is that negation cues in upper layers are fused with other negation information, which confuses

Data	Model	Cues			Scope			Events		
		P	R	F1	P	R	F1	P	R	F1
Dev	Liu et al. (2018)	0.49	0.42	0.45	0.64	0.44	0.50	0.40	0.27	0.32
	ENC	**0.915**	**0.665**	**0.770**	**0.814**	**0.530**	**0.642**	**0.598**	**0.335**	**0.429**
	DEC	0.754	0.488	0.592	0.738	0.489	0.588	0.487	0.272	0.348
Test	Liu et al. (2018)	0.478	0.382	0.425	0.583	0.312	0.406	0.338	0.180	0.235
	ENC	**0.892**	**0.581**	**0.704**	**0.743**	**0.496**	**0.595**	**0.496**	**0.285**	**0.362**
	DEC	0.686	0.362	0.474	0.656	0.456	0.538	0.470	0.225	0.304

Table 10: Precision (P), recall (R), and F1 scores of the negation projection tasks in EN→ZH, using NMT hidden states, comparing with the word alignment based method (Liu et al., 2018). *ENC* represents the hidden states from the 1st encoder layer in cue projection, and represents the hidden states from the 6th encoder layer in scope/event projection. *DEC* denotes the hidden states from the 6th decoder layer.

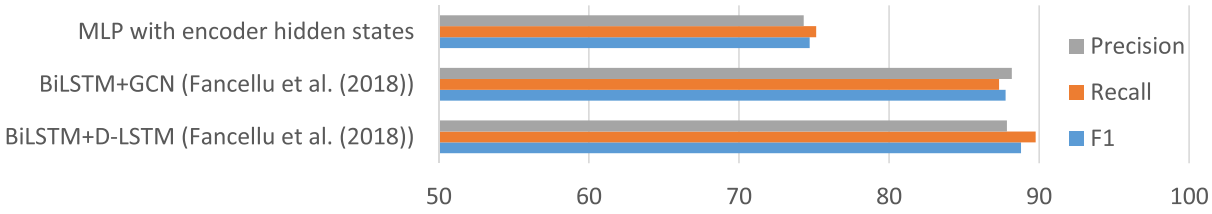

Figure 3: Results (%) on negation scope detection in English. *MLP* is the probing classifier; *GCN* is graph convolutional network; *D-LSTM* is bidirectional dependency LSTM.

the classifier. However, negation events/scope in upper layers interact more with negation cues and non-negation tokens, which makes them more distinctive.

5.3.3 Negation Scope Detection

Figure 3 shows the results of the negation scope detection task. We only report the results of using encoder hidden states that perform the best. The MLP classifier trained on encoder hidden states achieves 74.31%, 75.14%, and 74.72% on precision, recall, and F1, respectively,[10] and it is distinctly inferior to the other two models. However, methods from Fancellu et al. (2018) are specifically designed for negation scope detection and add extra information (negation cues, POS tags) to supervise the model, while the MLP classifier is designed to jointly predict negation cues as well, only using hidden states. We can conclude that some information about negation scope is well encoded in hidden states, but there is still room for improvement.

[10]Here we only report the result of using hidden states from the 6th encoder layer. We also tried hidden states from other encoder layers and decoders and obtained similar results as in the negation projection task.

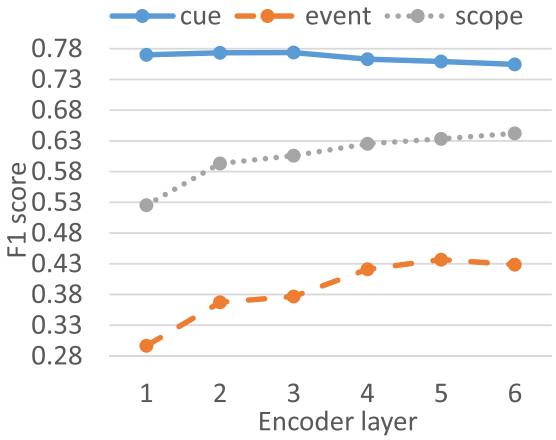

Figure 4: F1 scores of the negation projection tasks, on the development set, using hidden states from different encoder layers.

5.3.4 Incorrectly Translated Sentences

We further probe encoder hidden states from correctly and incorrectly translated sentences on negation cues and scope, to explore the quality of hidden states from incorrectly translated sentences. Note that we do not consider the under-translated cues. Figure 5 exhibits the performance

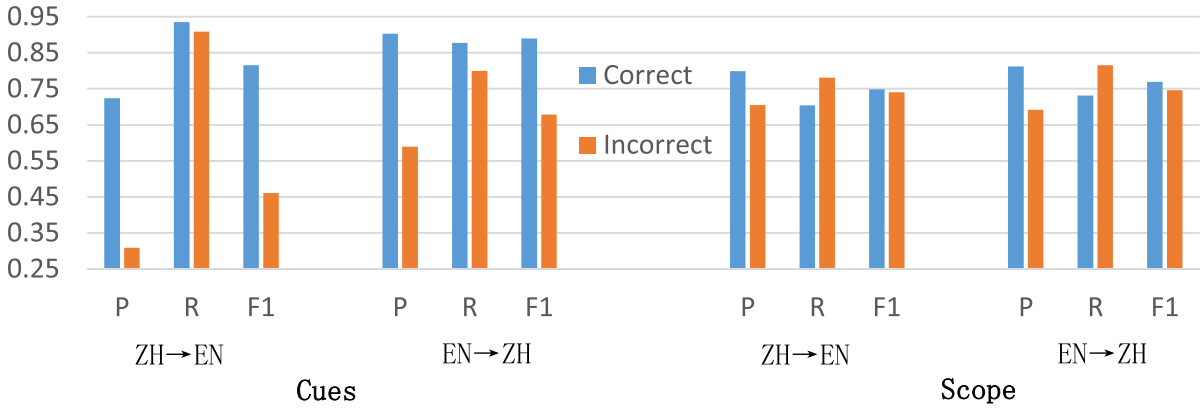

Figure 5: Results on negation cue/scope detection in ZH–EN, using encoder hidden states from sentences where the negation is correctly translated (*Correct*) and incorrectly translated (*Incorrect*).

of negation detection on cues and scope. *Correct* represents hidden states from correctly translated sentences and *Incorrect* stands for hidden states from incorrectly translated sentences. *Incorrect* performs worse than *Correct*, especially on the negation cue detection task, which confirms the effectiveness of using probing tasks to explore the information about negation in hidden states.

6 Conclusion

In this paper, we have explored the ability of NMT models to translate negation through evaluation and interpretation. The accuracy of manual evaluation in EN→DE, DE→EN, EN→ZH, and ZH→EN is 95.7%, 94.8%, 93.4%, and 91.7%, respectively. The contrastive evaluation shows that deleting a negation cue from references is more confusing to NMT models than inserting a negation cue into references, which indicates that NMT models have a bias against sentences with negation. We show that NMT models make fewer mistakes in EN–DE than in EN–ZH. Moreover, there are more errors in DE/ZH→EN than in EN→DE/ZH.

We also have investigated the information flow of negation by computing the attention weights and attention flow. We demonstrate that the negation information has been well passed to the decoder, and that there is no correlation between the amount of negation information transferred and whether the cues are under-translated or not. Thus, we consider attempts to detect or even fix under-translation of cues via an analysis or manipulation of the attention flow to have little promise. However, our analysis of the training data shows that negation is often rephrased, leading to cue mismatches which could confuse NMT models. This suggests that distilling or filtering training data to make grammatical negation more consistent between source and target could reduce this under-translation problem.

In addition, we show that NMT models can distinguish negation and non-negation tokens very well, and NMT models can encode substantial information about negation in hidden states but nevertheless leave room for improvement. Moreover, encoder hidden states capture more information about negation than decoder hidden states; negation cues are better modeled in lower encoder layers while negation events and tokens belonging to negation scope are better modeled in higher encoder layers.

Overall, we show that the modeling of negation in NMT has improved with the evolution of NMT – with deeper and more advanced networks; the performance on translating negation varies between language pairs and directions. We also find that the main error types on negation have shifted from SMT to NMT—under-translation is the most frequent error type in NMT while other error types such as reordering were equally or more prominent in SMT.

We only conduct evaluation in EN–DE and EN–ZH, and German/Chinese and English are very similar in expressing negation. It will be interesting to explore languages have different characteristics on negation in the future, such as Italian, Spanish, and Portuguese, where double negation is very common.

Acknowledgments

We thank all reviewers, and the action editors (Mauro Cettolo and Chris Quirk), for their valuable and insightful comments. We also thank Qianchu Liu for providing the *NegPar* data set. We acknowledge the computational resources provided by CSC in Helsinki and Sigma2 in Oslo through NeIC-NLPL (www.nlpl.eu). GT was mainly funded by the Chinese Scholarship Council (no. 201607110016).

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.385

Duygu Ataman, Orhan Firat, Mattia A. Di Gangi, Marcello Federico, and Alexandra Birch. 2019. On the importance of word boundaries in character-level neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 187–193, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5619

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Modality and negation in SIMT use of modality and negation in semantically-informed syntactic MT. *Computational Linguistics*, 38(2):411–438. https://doi.org/10.1162/COLI_a_00099

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: A case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1025

Anne Beyer, Vivien Macketanz, Aljoscha Burchardt, and Philip Williams. 2017. Can out-of-the-box NMT beat a domain-trained Moses on technical data? In *The 20th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 41–46, Prague, Czech Republic. Charles University.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Philip Williams, David Mareček, Martin Popel, Rudolf Rosa, Josef Jon, and Michal Kašpar. 2018, Final report on employing semantic role labelling and shallow proxies for negation and fidelity checking in MT, The University of Edinburgh.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics)*, pages 531–540, Ann Arbor, USA. Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219906

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, USA. Association for Computational Linguistics.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159,

Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1106

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2018. Neural networks for cross-lingual negation scope detection. *CoRR*, cs.CL/1810.02156v1.

Federico Fancellu and Bonnie Webber. 2014. Applying the Semantics of Negation to SMT Through N-best List Re-ranking. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 598–606, Gothenburg, Sweden. Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1063

Federico Fancellu and Bonnie Webber. 2015. Translating negation: A manual error analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11, Denver, USA. Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-1301

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1014

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, cs.CL/1712.05690v1.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020. It's not a Non-Issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.345

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge, UK. Cambridge University Press. https://doi.org/10.1017/9781316423530

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, California, USA.

Jin-Ji Li, Jungi Kim, Dong-Il Kim, and Jong-Hyeok Lee. 2009. Chinese syntactic reordering for adequate generation of Korean verbal phrases in Chinese-to-Korean smt. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 190–196, Athens, Greece. Association for Computational Linguistics.

Qianchu Liu, Federico Fancellu, and Bonnie Webber. 2018. NegPar: A parallel corpus annotated for negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan. European Language Resources Association.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation in Conan Doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*,

pages 1563–1568, Istanbul, Turkey. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for 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 1314–1324, New Orleans, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1119

Matīss Rikters and Mark Fishel. 2017. Confidence through attention. In *Proceedings of the 16th Machine Translation Summit (MT Summit 2017)*, pages 299–311. Nagoya, Japan.

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. https://doi.org/10.18653/v1/E17-2060

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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Elena Sergeeva, Henghui Zhu, Amir Tahmasebi, and Peter Szolovits. 2019. Neural token representations and negation and speculation scope detection in biomedical and general domain text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 178–187, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-6221

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? A targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4263–4272, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1458

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99. https://doi.org/10.1162/tacl_a_00048

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Dominikus Wetzel and Francis Bond. 2012. Enriching parallel corpora for statistical machine translation with semantic negation rephrasing. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 20–29, Jeju, Republic of Korea. Association for Computational Linguistics.