EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

EAMT2018

Proceedings of the
21st Annual Conference of the European Association for Machine Translation

28–30 May 2018
Universitat d'Alacant
Alacant, Spain

Edited by

Juan Antonio Pérez-Ortiz
Felipe Sánchez-Martínez
Miquel Esplà-Gomis
Maja Popović
Celia Rico
André Martins
Joachim Van den Bogaert
Mikel L. Forcada

Organised by

Universitat d'Alacant
Universidad de Alicante

transducens
research group

ISBN: 978-84-09-01901-4

An Analysis of Source Context Dependency in Neural Machine Translation

Xutai Ma
Electrical and Computer Engineering
Johns Hopkins University
xutai_ma@jhu.edu

Ke Li
Electrical and Computer Engineering
Johns Hopkins University
kli26@jhu.edu

Philipp Koehn
Computer Science Department
Johns Hopkins University
phi@jhu.edu

Abstract

The encoder-decoder with attention model has become the state of the art for machine translation. However, more investigations are still needed to understand the internal mechanism of this end-to-end model. In this paper, we focus on how neural machine translation (NMT) models consider source information while decoding. We propose a numerical measurement of source context dependency in the NMT models and analyze the behaviors of the NMT decoder with this measurement under several circumstances. Experimental results show that this measurement is an appropriate estimate for source context dependency and consistent over different domains.

1 Introduction

Neural machine translation (NMT) with encoder-decoder structure and attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) has achieved great success on several machine translation tasks. Different from phrase based systems, neural machine translation is trained end-to-end and learns the alignment and translation jointly.

At each decoding step, the alignment is predicted in the attention layer and represented as a distribution over words in a source sequence. Then the source context information, which is an attention-weighted sum over encoder hidden states, is fed into the decoder for the prediction of the next word.

The decoder in a NMT model is similar to a recurrent neural network language model (RNNLM) (Mikolov et al., 2010), with additional input from the source side. It takes the previous hidden state, the previous predicted target word embedding, and source context information as inputs and produces a distribution over the next target words.

This end-to-end approach can achieve state-of-the-art performance on several machine translation tasks. Joint training of the translation model and alignment gives a soft alignment between source side and target side. However, some of its flaws are observed under certain settings (Koehn and Knowles, 2017). One of the most common and important issue of neural machine translation is that it often generates fluent but inadequate translations especially under domain mismatch conditions.

An example[1] is shown in Figure 1. Here, the translation generated by the NMT models — while being fluent English — has no semantic connection to the source sentence. Moreover, out-of-domain models cause even more severe inadequacy.

An intuitive explanation for this observation is that the NMT decoder lacks effective attention to the source information. Because of the similarity between the NMT decoder and an RNNLM, it is possible that NMT models generate sentence based on its internal language model without properly taking advantage of the source information. While several researchers have explored the atten-

[1] In this example, we choose a sentence that has been processed by byte pair encoding (BPE) (Sennrich et al., 2016) and "@@" is used as a splitter token. The reason for this is that BPE has become a standard pre-processing step, which helps reducing the vocabulary size. Long words in original text will be split into sub-word "phrases". It is also very interesting to investigate how sub-word prediction related to the source information.

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 189–197
Alacant, Spain, May 2018.

Source	*der hat also die Was@@ er@@ stoff@@ emission bei verschiedene Frequ@@ enzen aufgenommen .*
Reference	*it recorded the hydrogen radio emission at different frequencies .*
In-domain Translation	*so he recorded the security clearance on several frequencies .*
Out-of-domain Translation	*indeed , He has [more] example in suc@@ cession .*

Figure 1: An example of an inadequate translation of a Germany to English NMT model. In domain data is subtitle dataset and out of domain data is koran dataset

tion mechanism in NMT, none of them have numerically analyzed whether an NMT decoder sufficiently utilizes the source information.

In this paper, we propose a numerical approach for source context dependency analysis in NMT models. We list some reasons why we should care this dependency.

1. While translation of content words, such as nouns and verbs, highly depends on source information, function words, such as determiners and prepositions, depend more on language-internal properties. We want to investigate whether NMT models are able to learn this difference.

2. The NMT decoder functions similarly to a RNN language model. It takes both previous hidden state, the previous predicted word embedding, and the source context vector as inputs for every recurrent neural network (RNN) cell. It is possible that under certain circumstances the source context vector has little impact on updating the state in the decoder. That means the decoder may fail to use sufficient information from the source sentence. This could be one of the reasons why NMT models sometimes generate fluent but inadequate sentences.

3. As observed by Koehn and Knowles (2017), under some data conditions such as domain mismatch, some failed translations seem to ignore the source sentence. By analyzing source context dependency, we can gain insight into the reason of the failure.

Our contributions in this paper include:

- We propose a numerical measurement for source context dependency in NMT models. It is based on the distribution of words generated from the decoder. The measurement is very general to sequence to sequence models and their variations.
- We carried out a series of experiments under different settings to analyze the behavior of NMT models with this measurement. Moreover, we numerically analyze source context dependency related to part of speech categories, domain mismatch and translation length.

2 Related Work

A number of researchers have been working on exploring the "black box" of neural machine translation models. Belinkov et al. (2017a) investigated how NMT models learn word structure and representation quality on part-of-speech and morphological tags. Belinkov et al. (2017b) and Dalvi et al. (2017) explored the capability of representation in NMT hidden layers of part-of-Speech and semantic tagging in neural machine translation using multi-task training.

There is some research focusing on the attention mechanism in NMT. Liu et al. (2016) proposed a training scheme to learn attention under the guidance from conventional alignment models. Cohn et al. (2016) incorporates structural alignment biases to improve the alignment quality learned in the attention layer. Ghader and Monz (2017) proposed a numerical approach for analyzing the capability of attention.

Some research also focuses on analysis and visualization of NMT models for better understanding. Visualization of attention weights is a common tool for NMT analysis (Ding et al., 2017). Moreover, Shi et al. (2016) correlated activation values of individual LSTM nodes in the translation model with the length of the translated sentences.

Some research has addressed a similar topic as we tackle in our paper. Instead of doing numerical analysis, they proposed new structures to improve both the adequacy and fluency in NMT. Tu et al. (2017) proposed a context gate structure in NMT decoders to control the portion of source or target side information fed into the decoder and they ob-

tained a 2.3 BLEU points improvement compared a standard attention based NMT baseline. Zheng et al. (2018) introduced a novel mechanism to separate the source information into two parts: translated past context and untranslated future context. They fed the two parts to both the attention model and the decoder states and reported improvement on several translation tasks compared with the conventional coverage model.

3 Methodology

3.1 Neural Machine Translation

A variety of alternative neural machine translation approaches have been recently proposed (Gehring et al., 2017; Vaswani et al., 2017). In this paper, we will focus on the most common model used today, the encoder-decoder based NMT model with an attention layer (Bahdanau et al., 2015; Luong et al., 2015).

The encoder in the neural machine translation model is a bi-directional recurrent neural network structure which encodes the source tokens sequence into a sequence **H** of context-related vector representations h_j upon an embedding layer.

$$\mathbf{H} = h_0, h_1, \ldots, h_{n-1}, h_n \tag{1}$$

The decoder of the NMT model is a recurrent neural network (Elman, 1990). There are several widely used variations, such as Long Short Time Memory (Gers et al., 1999) and Gated Recurrent Unit (GRU) (Chung et al., 2014) In this paper, we choose to use GRU for analysis.

Let us now introduce the structure of the NMT decoder. In decoder, the distribution for next possible words at each step is generated by:

$$P(y_i \mid y_{<i}, \mathbf{x}) = g(y_{i-1}, s_i, c_i) \tag{2}$$

where **x** is a sequence of vectors representing the source sentence, and s_i is RNN hidden state and calculated by:

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \tag{3}$$

g and f are some nonlinear functions.

The context vector c_i at step i comes from:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{4}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{5}$$

where

$$e_{ij} = a(s_{i-1}, h_j) \tag{6}$$

is an alignment model which scores how well inputs around position j and the output at position i match. h_j is the encoder hidden state at step j.

3.2 Source Context Dependency Measurement

If an NMT model properly considers source context information, a significant difference should be observed between distributions with and without the source context vector. Considering this, we propose a distribution distance based method to calculate the source context dependency in an NMT model.

We first train an attention based NMT model. During decoding we have two decoders, a main decoder and an auxiliary decoder as shown in Figure 2. The main decoder is a normal NMT decoder with the source context vector computed by a weighted sum of encoder hidden states. The auxiliary decoder shares parameters with the NMT decoder but zeros out the source context vector at each decoding step. The previous predicted target word embedding for the auxiliary decoder is from the main NMT decoder. The hidden states of the NMT and auxiliary decoders are denoted separately as s_i and s_i^{aux} in Figure 2 at each step i while they are the same indeed.

We then introduce the source context dependency measure. At i-th decoding step, we have two distributions for predicting the next translated word given history and context from the NMT decoder and the auxiliary decoder denoted as $P_{main}(y_i)$ and $P_{aux}(y_i)$ in Figure 2, where y_i is the i-th predicted word. We then define the source context dependency measure of word y_i as

$$D_{y_i}^{p} = d_{KL}\left(P_{main}(y_i), P_{aux}(y_i)\right) \tag{7}$$

$$= d_{KL}\left(P(y_i \mid y_{<i}, c_i), P(y_i \mid y_{<i}^{aux}, \vec{\mathbf{0}})\right) \tag{8}$$

$$= d_{KL}\left(g(y_{i-1}, s_i, c_i), g(y_{i-1}, s_i^{aux}, \vec{\mathbf{0}})\right) \tag{9}$$

where

1. d_{KL} is a function to calculate the KL-divergence between the two distributions.

2. $P_{main}(y_i) = P(y_i \mid y_{<i}, c_i)$ is the distribution over the next word given history information and source context vector c_i at step i.

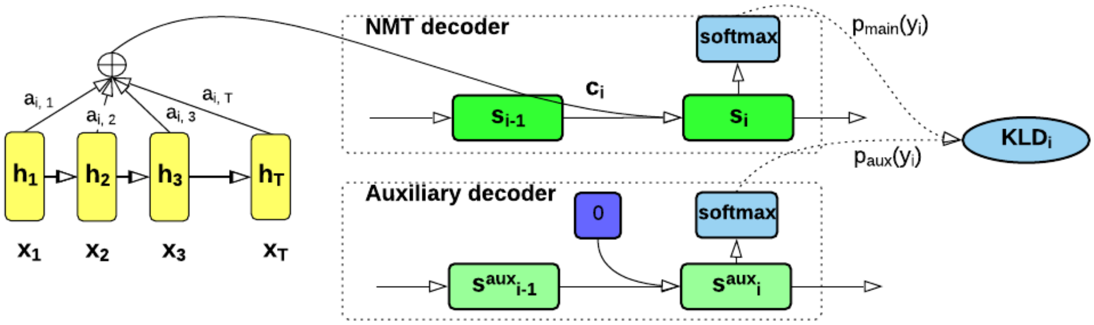

Figure 2: Context dependency measure: Both standard NMT and an auxiliary decoder that ignores the source context make word predictions. We measure the KL divergence between these predictions.

3. $P_{aux}(y_i) = P(y_i \mid y^{aux}_{<i}, \vec{0})$ is the distribution over the next word given history information and a zeroed out source context vector at step i. Notice that $y^{aux}_{<i}$ and $y_{<i}$ are actually the same sub-sequence. However, we are using an "aux" superscript here to emphasize that their representations, which are the hidden states, are different when predicting the next word.

The first distribution comes from main decoder and second distribution comes from auxiliary decoder.

Notice that we compute source context dependency scores during decoding, not training. Furthermore, it is also compatible with beam search. In addition to main decoder hidden states and previous predictions, hidden states from auxiliary decoder and source context dependency scores of previous words are also tracked for each hypothesis in the beam. Since the two decoders share parameters, no additional training is needed for the source context dependency calculation given a trained NMT model.

An alternative implementation for computing the source context dependency score would be to only use one decoder. At each decoding step, we can calculate the distance between distributions from the main decoder with and without the source context vector. However, the previous hidden state potentially contains both history and previous source context information. Thus, source context creeps into the decoder state. With a auxiliary decoder, we can completely eliminate the influence of source context.

4 Experimental Setup

We use the toolkit Nematus (Sennrich et al., 2017) for training and decoding. We use the gated recurrent unit (GRU) (Chung et al., 2014) in both encoder and decoder with a dimension of 1024. The dimension of embedding layer is 500. For optimizer, Adadelta (Zeiler, 2012) with learning rate 0.0001 is used. Dropout (Srivastava et al., 2014) with 0.2 probability was used to prevent overfitting. For decoding, we use beam search with a beam width 12.

Byte pair encoding (BPE) (Sennrich et al., 2016) is used for processing training data to fit a 50,000 subwords vocabulary limit. We use BPE since it has been a very popular preprocessing procedure for machine translation, so that our evaluation method can be used in more general cases.

In part-of-speech (POS) analysis, we use Stanford POS tagger (Toutanova et al., 2003) with a universal POS tagset. We first convert the translated subwords to complete words and tag the sequences with the Stanford POS tagger We find that the amount of subwords is significantly smaller than complete words. So we let each subword inherit the tag from the corresponding complete word[2].

We carried out our experiments on German–English translation tasks. We used five corpora in five domains from OPUS (Tiedemann, 2012), which is briefly described in Table 1. We use five corpora because we want to show that our metric and its analysis are general and consistent over different domains. Moreover, we would like to know

[2]An alternative would be to distinguish tags for split and unsplit words. We did this as well, but found no significant difference.

Dataset	Abbreviates	Descriptions	Size(English)
OpenSubtitle2016	subtitles	Translatio Movie subtitles	118.8M
JCR-Acquis	acquis	Legislative text of the European Union	34.1M
EMEA	emea	Documents from the European Medicines Agency	12.0M
Tanzil	koran	Translations of Koran	11.3M
IT	it	Documents of GNOME, OpenOffice, KDE, PHP, Ubuntu	2.6M

Table 1: Summary of five corpora from OPUS

how domain mismatch affects the source content dependency.

5 Analysis

5.1 Auxiliary Decoder

We briefly described the auxiliary decoder above in Section 3.2. The basic assumption is that an auxiliary decoder contains history information and behaves similar to a recurrent neural language model. The difference between an auxiliary decoder and a standard RNNLM (for the target language) is in the aspect of training. The auxiliary decoder sharing parameters with the main NMT decoder is trained with source side information while the standard RNNLM is only trained on the target corpus. Considering the mismatch of training and testing situation for the auxiliary decoder, the performance on language modeling task of it can be worse than a standard RNNLM.

To demonstrate the similarity with a standard language model and verify our assumption about the performance of the auxiliary decoder, we evaluate the auxiliary decoder and several standard language models on the language model task. The standard language models include two n-gram models and a RNN based language model. A n-gram model is a statistical model that predicts the probability of the next word given the previous $n-1$ history words. We use a 2-gram and a 3-gram model with Kneser-Ney smoothing (Kneser and Ney, 1995). The two n-gram models are trained using the toolkit SRILM (Stolcke, 2002). The RNN based language model is a two-layer LSTM with both embedding and hidden dimensions 500. We trained the LSTM LMs using Pytorch. The optimization method is Adam with initial learning rate 0.001. The architecture and optimization settings of the LSTM LM are the same as the auxiliary decoder.

The perplexity results on four datasets by the two n-gram models, the LSTM-LM, and the auxiliary decoder are in Table 2. We can see that although the auxiliary decoder has worse perplexity than a standard LSTM language model, its performance are similar to a n-gram language model for most situations. This observation is consistent with our assumption.

Model	Acquis	EMEA	Koran	IT
2-gram	72.3	86.7	86.3	120.0
3-gram	38.1	44.6	61.4	46.7
LSTM-LM	19.8	19.2	19.6	30.9
Auxiliary Decoder	44.0	51.9	103.7	57.1

Table 2: Perplexities from different models on four test corpora from different domains.

5.2 Part of Speech

Different words have different dependencies on source context. It is very natural to assume that content words, such as nouns and verbs, tend to have higher dependencies on the source context, while function words like adpositions depend more on the target side.

Figure 4 is an example of the source context dependency measurement on a translated English sentence from Germen meaning *"after my study of electronics, I came here in 1954."*[3]. We can see that content words such as *"study"*, *"electr@@"*, *"19@@"* and *"54"* have relatively high dependency scores. Meanwhile, functional words like *"of"*, *","*, and *"in"* have lower scores.[4]

[3] This sentence comes from subtitle dataset

[4] One might notice that it is not always true that all content words scores are high and all function words scores are low in this sentence. For example, word *"after"* has a very high score, and word (or sub-word) *"onics"* has very low score. The reason for the first case is that while predicting the first word, internal language model in decoder will always prefer the most common word in training data since there is no history. So even *"after"* is a functional word, the most of the information decoder needs to generate *"after"* comes from attention vector, which results into a very high score according to our metric. As to the second case, *"onics"* is a sub-word of *"electronics"*. Since the sub-word phrase (*"electro@@"*, *"onics"*) is relatively frequent and these two sub-words are highly unlikely to appear independently, the decoder can be confident to predict *"onics"* given previous prediction *"elec-*

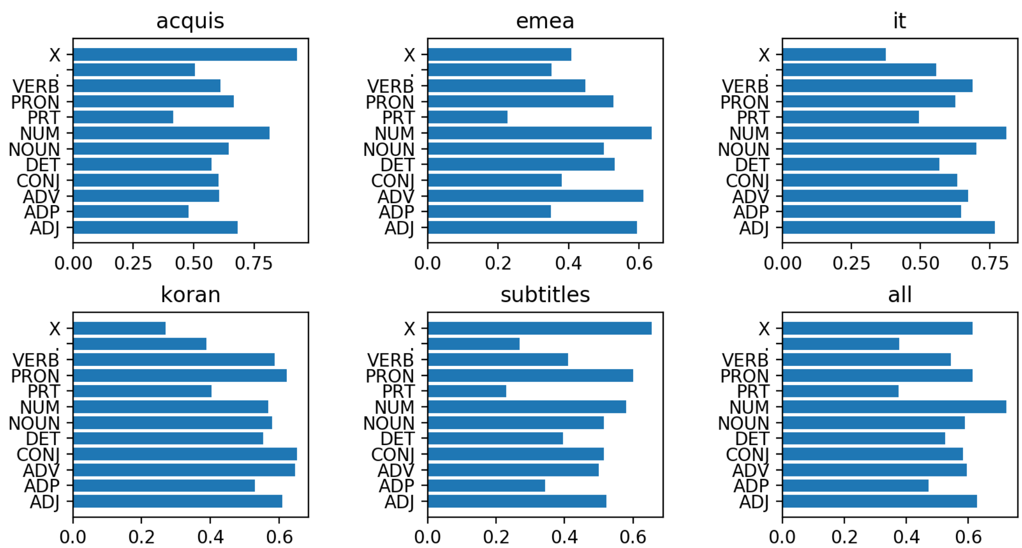

Figure 3: Scores for different categories of part of speech (POS).

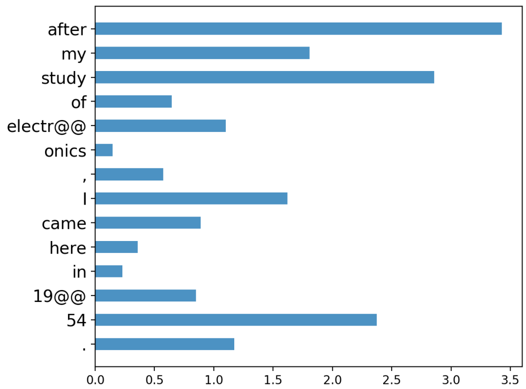

Figure 4: English translation and source context dependency of Germen sentence *"nach meinem Studium in Elektroni@@ k kam ich hier in 19@@ 54 ."*.

We then compare the source context dependency of translated words with part-of-speech (POS) tags. We calculate the average source context dependency score for each POS category over test sets from five corpora, shown in Figure 3. We can observe that although the distribution of scores are different among domains, they all have a similar tendency. Adpositions and particles have lower source context dependency than other categories, especially numbers. This observation is consist with our intuition.

Another interesting observation is that the average score of functional words in certain situations can be high. For example, determinators in EMEA dataset is even higher than nouns. There are two reasons for this observation. First, EMEA is a highly structured and repetitive corpus. The NMT models can generate nouns with little context information since noun phrases in this corpus are frequent. The decoder can easily determine the remaining words given the first word of a phrase. The second reason is that although functional words seems to rely more on decoder, some of them still need context information. For example, if a sentence contains the noun phrase *"an apple"*, the model will generate the correct determiner *"an"* rather than *"a"* from source information — the determiner *"a"* is highly dispreferred by the language model.

tro@@" with very limited source side information. These two cases are actually quite rare in our corpora, so we did not use them for POS tagging analysis.

5.3 Domain Mismatch

Domain mismatch is a major challenge for NMT. Training an NMT model in one domain can make the decoder overfit that particular domain. Thus, during decoding the decoder can produce fluent but inadequate sentences on out-of-domain test data. Therefore, we wondered if domain mismatch can cause less source context dependency.

We calculate source context dependency scores under domain mismatch settings. Five NMT models were trained on five datasets shown in Table 1. We then apply them on five test datasets and calculate source context dependency scores, respectively. Next, means and variance of scores among test sets with different models are calculated.

The results are shown in Table 3 and Table 4. Domains of training data are in columns and do-

Train \ Test	law	medical	it	koran	subtitles
all	2.594	2.831	2.283	2.418	2.372
law	**3.956**	3.695	3.463	3.382	3.818
medical	1.694	**2.186**	1.615	1.383	1.414
it	3.597	3.536	**4.312**	3.423	3.937
koran	1.965	1.765	2.024	**3.14**	1.93
subtitles	0.955	0.982	0.971	1.021	**1.489**

Table 3: Means of source context dependency scores on different test datasets translated by different models.

Train \ Test	law	medical	it	koran	subtitles
all	3.5	3.44	2.07	1.725	2.061
law	**15.06**	10.9	9.93	8.83	13.5
medical	3.74	**5.26**	2.94	1.89	2.8
it	8.61	8.13	**14.15**	8.46	11.3
koran	3.74	3.37	4.02	**7.87**	4.2
subtitles	0.676	0.619	0.66	0.571	**1.521**

Table 4: Variances of source context dependency scores on different test datasets translated by different models.

Train \ Test	law	medical	it	koran	subtitles
all	31.1	45.1	35.3	17.9	26.4
law	**31.1**	12.1	3.5	1.3	2.8
medical	3.9	**39.4**	2.0	0.6	1.4
it	1.9	6.5	**42.1**	1.8	3.9
koran	0.4	0.0	0.0	**15.9**	1.0
subtitles	7.0	9.3	9.2	9.0	**25.9**

Table 5: BLEU scores of source context dependency scores on different test datasets translated by different models, reported by Koehn and Knowles (2017).

mains of test data are in rows. "all" in the Table 3 and Table 4 means the model was trained on a combination of the five datasets. Since we care more how one certain model behaves on test sets from different domains, we compare the scores along the rows. It is noticeable that all the five models have highest source context dependency scores when translating in-domain test data. Higher means indicate that in-domain models depend more on source information. Higher variances show that in-domain models are also better at learning differences among different word, because we expect a good model has more context dependency on content related words and less on history related words. This can be one of the reasons why NMT models often generate fluent but inadequate translations in domain mismatch settings.

We also list the BLEU score reported by Koehn and Knowles (2017) on the same task, shown as Table 5. We can see that inability of incorporating context information into the decoder can be a main reason for the failure in domain mismatch setting.

5.4 Sentence Length

The translation quality is sensitive to the lengths of the sentences. Moreover, for longer sentences, it is possible that a NMT model considers more history information rather than source context information. We want to know how sentence length affects source context dependency.

We calculate source context dependency for sentences with different lengths. Results are shown in Figure 5. We find that longer sentences have lower source context dependency, which is consistent with our hypothesis[5].

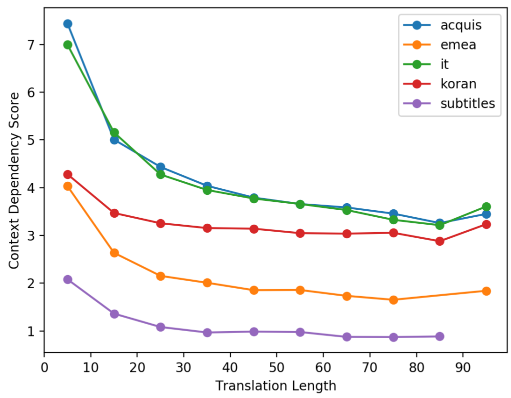

Figure 5: Source dependency scores with length of sentence

However, we detect a different tendency compared with the analysis by Koehn and Knowles (2017) which show lower translation quality for longer sentences. However, source context dependency is not the only factor that determines translation quality. When the length of translation increases, history information from the language model is increasingly informative (and hence predictive).

[5]One can notice that there are some small fluctuations in sentence length from 70 to 90. This can be caused by a small percentage of sentences in that length range (70-80: ∼ 4%, 80-90: ∼ 3%).

6 Conclusion and Future Work

In this paper, we proposed a measurement of source context dependency in neural machine translation models. With our measurement, we analyzed source context dependency with different POS tags, domains and sentence lengths. From the analysis, we can see our measurement is a good estimation of source context dependency.

In the future, we plan to extend our research in two directions. One is to investigate the relationship between source context dependency and word level translation quality, so that we can immediately detect when the system goes off track. The other is to improve the performance of NMT models. Since our measurement is differentiable, we can use it as an auxiliary term of the training objective function.

Acknowledgement

We would like to appreciate Center of Language and Speech Processing, Johns Hopkins University for providing hardwares for our experiments. We also appreciate valuable suggestions and advice from people in Johns Hopkins University machine translation group.

This work was partially supported by the IARPA MATERIAL program.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR*.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017a). What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2017b). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10. Asian Federation of Natural Language Processing.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS Deep Learning Workshop.*

Cohn, T., Hoang, C. D. V., Vymolova, E., Yao, K., Dyer, C., and Haffari, G. (2016). Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885. Association for Computational Linguistics.

Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., and Vogel, S. (2017). Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 142–151. Asian Federation of Natural Language Processing.

Ding, Y., Liu, Y., Luan, H., and Sun, M. (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.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: Continual prediction with lstm.

Ghader, H. and Monz, C. (2017). What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39. Asian Federation of Natural Language Processing.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acous-*

tics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on, volume 1, pages 181–184. IEEE.

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

Liu, L., Utiyama, M., Finch, A. M., and Sumita, E. (2016). Neural machine translation with supervised attention. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3093–3102.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (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.

Shi, X., Knight, K., and Yuret, D. (2016). Why neural translations are the right length. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2278–2282, Austin, Texas. Association for Computational Linguistics.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180.

Tu, Z., Liu, Y., Lu, Z., Liu, X., and Li, H. (2017). Context gates for neural machine translation. *TACL*, 5:87–99.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zheng, Z., Zhou, H., Huang, S., Mou, L., Dai, X., Chen, J., and Tu, Z. (2018). Modeling past and future for neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:145–157.