Using Whole Document Context in Neural Machine Translation

Valentin Macé, Christophe Servan

QWANT RESEARCH - 7 Rue Spontini, 75116 Paris, France

initial.lastname@qwant.com

Abstract

In Machine Translation, considering the document as a whole can help to resolve ambiguities and inconsistencies. In this paper, we propose a simple yet promising approach to add contextual information in Neural Machine Translation. We present a method to add source context that capture the whole document with accurate boundaries, taking every word into account. We provide this additional information to a Transformer model and study the impact of our method on three language pairs. The proposed approach obtains promising results in the English-German, English-French and French-English document-level translation tasks. We observe interesting cross-sentential behaviors where the model learns to use document-level information to improve translation coherence.

1. Introduction

Neural machine translation (NMT) has grown rapidly in the past years [1, 2]. It usually takes the form of an encoder-decoder neural network architecture in which source sentences are summarized into a vector representation by the encoder and are then decoded into target sentences by the decoder. NMT has outperformed conventional statistical machine translation (SMT) by a significant margin over the past years, benefiting from gating and attention techniques. Various models have been proposed based on different architectures such as RNN [1], CNN [3] and Transformer [2], the latter having achieved state-of-the-art performances while significantly reducing training time.

However, by considering sentence pairs separately and ignoring broader context, these models suffer from the lack of valuable contextual information, sometimes leading to inconsistency in a translated document. Adding document-level context helps to improve translation of context-dependent parts. Previous study [4] showed that such context gives substantial improvement in the handling of discourse phenomena like lexical disambiguation or co-reference resolution.

Most document-level NMT approaches focus on adding contextual information by taking into account a set of sentences surrounding the current pair [5, 6, 7, 8, 9, 10]. While giving significant improvement over the context-agnostic versions, none of these studies consider the whole document with well delimited boundaries. The majority of these approaches also rely on structural modification of the NMT model [7, 8, 9, 10]. To the best of our knowledge, there is no existing work considering whole documents without structural modifications.

Contribution: We propose a preliminary study of a generic approach allowing any model to benefit from document-level information while translating sentence pairs. The core idea is to augment source data by adding document information to each sentence of a source corpus. This document information corresponds to the belonging document of a sentence and is computed prior to training, it takes every document word into account. Our approach focuses on pre-processing and consider whole documents as long as they have defined boundaries. We conduct experiments using the Transformer base model [2]. For the English-German language pair we use the full WMT 2019 parallel dataset. For the English-French language pair we use a restricted dataset containing the full TED corpus from MUST-C [11] and sampled sentences from WMT 2019 dataset. We obtain important improvements over the baseline and present evidences that this approach helps to resolve cross-sentence ambiguities.

2. Related Work

Interest in considering the whole document instead of a set of sentences preceding the current pair lies in the necessity for a human translator to account for broader context in order to keep a coherent translation. The idea of representing and using documents for a model is interesting, since the model could benefit from information located before or after the current processed sentence.

Previous work on document-level SMT started with cache based approaches, [12] suggest a conjunction of dynamic, static and topic-centered cache. More recent work tend to focus on strategies to capture context at the encoder level. Authors of [6] propose an auxiliary context source with a RNN dedicated to encode contextual information in addition to a warm-start of encoder and decoder states. They obtain significant gains over the baseline.

A first extension to attention-based neural architectures is proposed by [7], they add an encoder devoted to capture the preceding source sentence. Authors of [8] introduce a hierarchical attention network to model contextual information from previous sentences. Here the attention allows dynamic access to the context by focusing on different sentences and words. They show significant improvements over a strong

SOURCE	*TARGET*
<doc1> Pauli is a theoretical physicist	Pauli est un physicien théoricien
<doc1> He received the Nobel Prize	Il a reçu le Prix Nobel
<doc2> Bees are found on every continent	On trouve des abeilles sur tous les continents
<doc2> They feed on nectar using their tongue	Elles se nourrissent de nectar avec leur langue
<doc2> The smallest bee is the dwarf bee	La plus petite abeille est l'abeille naine

Table 1: Example of augmented parallel data used to train the *Document* model. The source corpus contains document tags while the target corpus remains unchanged.

NMT baseline. More recently, [10] extend Transformer architecture with an additional encoder to capture context and selectively merge sentence and context representations. They focus on co-reference resolution and obtain improvements in overall performances.

The closest approach to ours is presented by [5], they simply concatenate the previous source sentence to the one being translated. While they do not make any structural modification to the model, their method still does not take the whole document into account.

3. Approach

We propose to use the simplest method to estimate document embeddings. The approach is called SWEM-aver (Simple Word Embedding Model – average) [13]. The embedding of a document k is computed by taking the average of all its N word vectors (see Eq. 1) and therefore has the same dimension. Out of vocabulary words are ignored.

$$Doc_k = \frac{1}{N} \sum_{i=1}^{N} w_{i,k} \tag{1}$$

Despite being straightforward, our approach raises the need of already computed word vectors to keep consistency between word and document embeddings. Otherwise, fine-tuning embeddings as the model is training would shift them in a way that totally wipes off the connection between document and word vectors.

To address this problem, we adopt the following approach: First, we train a baseline Transformer model (noted *Baseline* model) from which we extract word embeddings. Then, we estimate document embeddings using the SWEM-aver method and train an enhanced model (noted *Document* model) benefiting from these document embeddings and the extracted word embeddings. During training, the *Document* model does not fine-tune its embeddings to preserve the relation between words and document vectors. It should be noted that we could directly use word embeddings extracted from another model such as Word2Vec [14], in practice we obtain better results when we get these vectors from a Transformer model. In our case, we simply extract them from the *Baseline* after it has been trained.

Using domain adaptation ideas [15, 16, 17], we associate a tag to each sentence of the source corpus, which represents the document information. This tag takes the form of an additional token placed at the first position in the sentence and corresponds to the belonging document of the sentence (see Table 1). The model considers the tag as an additional word and replace it with the corresponding document embedding. The *Baseline* model is trained on a standard corpus that does not contain document tags, while the *Document* model is trained on corpus that contains document tags.

The proposed approach requires strong hypotheses about train and test data. The first downfall is the need for well defined document boundaries that allow to mark each sentence with its document tag. The second major downfall is the need to compute an embedding vector for each new document fed in the model, adding a preprocessing step before inference time.

4. Experiments

We consider two different models for each language pair: the *Baseline* and the *Document* model. We evaluate them on 3 test sets and report BLEU and TER scores. All experiments are run 8 times with different seeds, we report averaged results and p-values for each experiment.

Translation tasks are English to German, proposed in the first document-level translation task at WMT 2019 [18], English to French and French to English, following the IWSLT translation task [19].

4.1. Training and test sets

Table 2 describes the data used for the English-German language pair. These corpora correspond to the WMT 2019 document-level translation task. Table 3 describes corpora for the English-French language pair, the same data is used for both translation directions.

For the English-German pair, only 10.4% (3.638M lines) of training data contains document boundaries. For English-French pair, we restricted the total amount of training data in order to keep 16.1% (602K lines) of document delimited corpora. To achieve this we randomly sampled 10% of the ParaCrawl V3. It means that only a fraction of the source training data contains document context. The enhanced model learns to use document information only when it is available.

All test sets contain well delimited documents, *Baseline* models are evaluated on standard corpora while *Document* models are evaluated on the same standard corpora that have

Corpora	#lines	# EN	# DE
Common Crawl	2.2M	54M	50M
Europarl V9†	1.8M	50M	48M
News Comm. V14†	338K	8.2M	8.3M
ParaCrawl V3	27.5M	569M	527M
Rapid 19†	1.5M	30M	29M
WikiTitles	1.3M	3.2M	2.8M
Total Training	34.7M	716M	667M
newstest2017†	3004	64K	60K
newstest2018†	2998	67K	64K
newstest2019†	1997	48K	49K

Table 2: Detail of training and evaluation sets for the English-German pair, showing the number of lines, words in English (EN) and words in German (DE). Corpora with document boundaries are denoted by †.

Corpora	#lines	# EN	# FR
News Comm. V14†	325K	9.2M	11.2M
ParaCrawl V3 (sampled)	3.1M	103M	91M
TED†	277K	7M	7.8M
Total Training	3.7M	119.2M	110M
tst2013†	1379	34K	40K
tst2014†	1306	30K	35K
tst2015†	1210	28K	31K

Table 3: Detail of training and evaluation sets for the English-French pair in both directions, showing the number of lines, words in English (EN) and words in French (FR). Corpora with document boundaries are denoted by †.

been augmented with document context. We evaluate the English-German systems on newstest2017, newstest2018 and newstest2019 where documents consist of newspaper articles to keep consistency with the training data. English to French and French to English systems are evaluated over IWSLT TED tst2013, tst2014 and tst2015 where documents are transcriptions of TED conferences (see Table 3).

Prior to experiments, corpora are tokenized using Moses tokenizer [20]. To limit vocabulary size, we adopt the BPE subword unit approach [21], through the SentencePiece toolkit [22], with 32K rules.

4.2. Training details

We use the OpenNMT framework [23] in its TensorFlow version to create and train our models. All experiments are run on a single NVIDIA V100 GPU. Since the proposed approach relies on a preprocessing step and not on structural enhancement of the model, we keep the same Transformer architecture in all experiments. Our Transformer configuration is similar to the baseline of [2] except for the size of word and document vectors that we set to $d_{model} = 1024$, these vectors are fixed during training. We use $N = 6$ as the number of encoder layers, $d_{ff} = 2048$ as the inner-layer dimensionality, $h = 8$ attention heads, $d_k = 64$ as queries and keys dimension and $Pdrop = 0.1$ as dropout probability. All experiments, including baselines, are run over 600k training steps with a batch size of approximately 3000 tokens.

For all language pairs we trained a *Baseline* and a *Document* model. The *Baseline* is trained on a standard parallel corpus and is not aware of document embeddings, it is blind to the context and cannot link the sentences of a document. The *Document* model uses extracted word embeddings from the *Baseline* as initialization for its word vectors and also benefits from document embeddings that are computed from the extracted word embeddings. It is trained on the same corpus as the *Baseline* one, but the training corpus is augmented with (see Table 1) and learns to make use of the document context.

The *Document* model does not consider its embeddings as tunable parameters, we hypothesize that fine-tuning word and document vectors breaks the relation between them, leading to poorer results. We provide evidence of this phenomena with an additional system for the French-English language pair, noted *Document+tuning* (see Table 5) that is identical to the *Document* model except that it adjusts its embeddings during training.

The evaluated models are obtained by taking the average of their last 6 checkpoints, which were written at 5000 steps intervals. All experiments are run 8 times with different seeds to ensure the statistical robustness of our results. We provide *p-values* that indicate the probability of observing similar or more extreme results if the *Document* model is actually not superior to the *Baseline*.

4.3. Results

Table 4 presents results associated to the experiments for the English to German translation task, models are evaluated on the newstest2017, neswtest2018 and newstest2019 test sets. Table 5 contains results for both English to French and French to English translation tasks, models are evaluated on the tst2013, tst2014 and tst2015 test sets.

En→De: The *Baseline* model obtained State-of-The-Art BLEU and TER results according to [24, 25]. The *Document* system shows best results, up to 0.85 BLEU points over the *Baseline* on the newstest2019 corpus. It also surpassed the *Baseline*e by 0.18 points on the newstest2017 with strong statistical significance, and by 0.15 BLEU points on the newstest2018 but this time with no statistical evidence. These encouraging results prompted us to extend experiments to another language pair: English-French.

En→Fr: The *Document* system obtained the best results considering all metrics on all test sets with strong statistical evidence. It surpassed the *Baseline* by 1.09 BLEU points and 0.85 TER points on tst2015, 0.75 BLEU points and 0.76 TER points on tst2014, and 0.48 BLEU points and 0.68 TER points on tst2013.

Fr→En: Of all experiments, this language pair shows the most important improvements over the *Baseline*. The

Model	newstest2017		newstest2018		newstest2019	
En→De	BLEU	TER	BLEU	TER	BLEU	TER
Baseline	26.78	54.82	40.61	41.02	35.67	46.80
Document	**26.96****	**54.76**	**40.77**	**40.97**	**36.52***	**46.36***

Table 4: Results obtained for the English-German translation task, scored on three test sets using BLEU and TER metrics. p-values are denoted by * and correspond to the following values: $^{*}< .05$, $^{**}< .01$, $^{***}< .001$.

Translation direction	**Model**	tst2013		tst2014		tst2015	
		BLEU	TER	BLEU	TER	BLEU	TER
En→Fr	Baseline	46.05	37.83	43.38	39.71	41.41	42.18
	Document	**46.53***	**37.15****	**44.14****	**38.95****	**42.50*****	**41.33*****
Fr→En	Baseline	45.99	34.64	42.96	37.30	39.91	39.06
	Document+tuning	45.94	34.42	43.16	36.93	40.14	38.70
	Document	**47.28*****	**33.80*****	**44.46*****	**36.34*****	**41.72*****	**38.04*****

Table 5: Results obtained for the English-French and French-English translation tasks, scored on three test sets using BLEU and TER metrics. p-values are denoted by * and correspond to the following values: $^{*}< .05$, $^{**}< .01$, $^{***}< .001$.

Document model obtained substantial gains with very strong statistical evidence on all test sets. It surpassed the *Baseline* model by 1.81 BLEU points and 1.02 TER points on tst2015, 1.50 BLEU points and 0.96 TER points on tst2014, and 1.29 BLEU points and 0.83 TER points on tst2013.

The *Document+tuning* system, which only differs from the fact that it tunes its embeddings, shows little or no improvement over the *Baseline*, leading us to the conclusion that the relation between word and document embeddings described by Eq. 1 must be preserved for the model to fully benefit from document context.

4.4. Manual Analysis

In this analysis we present some of the many cases that suggest the *Document* model can handle ambiguous situations. These examples are often isolated sentences where even a human translator could not predict the good translation without looking at the document, making it almost impossible for the *Baseline* model which is blind to the context. Table 6 contains an extract of these interesting cases for the French-English language pair.

Translation from French to English is challenging and often requires to take the context into account. The personal pronoun *"lui"* can refer to a person of feminine gender, masculine gender or even an object and can therefore be translated into *"her"*, *"him"* or *"it"*. The first example in Table 6 perfectly illustrate this ambiguity: the context clearly indicates that *"lui"* in the source sentence refers to *"ma fille"*, which is located three sentences above, and should be translated into *"her"*. In this case, the *Baseline* model predict the personal pronoun *"him"* while the *Document* model correctly predicts *"her"*. It seems that the *Baseline* model does not benefit from any valuable information in the source sentence. Some might argue that the source sentence actually contains clues about the correct translation, considering that *"robe à paillettes"* (*"sparkly dress"*) and *"baguette magique"* (*"magic wand"*) probably refer to a little girl, but we will see that the model makes similar choices in more restricted contexts. This example is relevant mainly because the actual reference to the subject *"ma fille"* is made long before the source sentence.

The second example in Table 6 is interesting because none of our models correctly translate the source sentence. However, we observe that the *Baseline* model opts for a literal translation of *"je peux faire le poirier"* (*"I can stand on my head"*) into *"I can do the pear"* while the *Document* model predicts *"I can wring"*. Even though these translations are both incorrect, we observe that the *Document* model makes a prediction that somehow relates to the context: a woman talking about her past disability, who has become more flexible thanks to yoga and can now twist her body.

The third case in table 6 is a perfect example of isolated sentence that cannot be translated correctly with no contextual information. This example is tricky because the word *"Elle"* would be translated into *"She"* in most cases if no additional information were provided, but here it refers to *"la conscience"* (*"consciousness"*) from the previous sentence and must be translated into *"It"*. As expected the *Baseline* model does not make the correct guess and predicts the personal pronoun *"She"* while the *Document* model correctly predicts *"It"*. This example present a second difficult part, the word *"son"* from the source sentence is ambiguous and does not, in itself, inform the translator if it must be translated into *"her"*, *"his"* or *"its"*. With contextual information we know that it refers to *"[le] monde physique"* (*"[the] physical world"*) and that the correct choice is the word *"its"*. Here the *Baseline* incorrectly predicts *"her"*, possibly because of its earlier choice for *"She"* as the subject. The *Document* model makes again the correct translation.

According to our results (see Table 5), the English-French language pair also benefits from document-level information but to a lesser extent. For this language pair, ambiguities

Fr-En	
Context	[...] et quand ma fille avait quatre ans, nous avons regardé "Le Magicien d'Oz" ensemble. Ce film a complètement captivé son imagination pendant des mois. Son personnage préféré était Glinda, bien entendu.
Source	Ça lui donnait une bonne excuse pour porter une robe à paillettes et avoir une baguette magique.
Ref.	It gave her a great excuse to wear a sparkly dress and carry a wand.
Baseline	It gave **him** a good excuse to wear a glitter dress and have a magic wand.
Document	It gave **her** a good excuse to wear a glitter dress and have a magic wand.
Context	Mon père passait souvent les grandes vacances à essayer de me guérir ... Mais nous avons trouvé un remède miracle : le yoga. [...] j'étais une comique de stand-up qui ne tenait pas debout.
Source	Maintenant, je peux faire le poirier.
Ref.	And now I can stand on my head.
Baseline	Now I can **do the pear**.
Document	Now, I can **wring**.
Context	C'est le but ultime de la physique : décrire le flux de conscience. Selon cette idée, c'est donc la conscience qui met le feu aux équations. Selon cette idée, la conscience ne pendouille pas en dehors du monde physique ...
Source	Elle siège bien en son cœur.
Ref.	It's there right at its heart.
Baseline	**She** sits well in **her** heart.
Document	**It** sits well in **its** heart .

Table 6: Translation examples for the French-English pair. We took the best models of all runs for both the *Baseline* and the *Document* enhanced model

En-Fr	
Context	[The speaker in this example is an old police officer saving a man from suicide] But I asked him, "What was it that made you come back and give hope and life another chance ?" And you know what he told me ?
Source	He said "You listened."
Ref.	Il a dit : "Vous avez écouté."
Baseline	Il a dit : "**Tu** as écouté."
Document	Il a dit : "**Vous** avez écouté."

Table 7: Translation example for the English-French pair.

about personal pronouns are less frequent. Other ambiguous phenomena like the formal mode (use of *"vous"* instead of *"tu"*) appear. Table7 presents an example of this kind of situation where the word *"You"* from the source sentence does not indicate if the correct translation is *"Vous"* or *"Tu"*. However it refers to the narrator of the story who is an old police officer. In this case, it is very likely that the use of formal mode is the correct translation. The *Baseline* model incorrectly predicts *"Tu"* and the *Document* model predicts *"Vous"*.

5. Conclusion

In this work, we presented a preliminary study of a simple approach for document-level translation. The method allows to benefit from the whole document context at the sentence level, leading to encouraging results. In our experimental setup, we observed improvement of translation outcomes up to 0.85 BLEU points in the English to German translation task and exceeding 1 BLEU point in the English to French and French to English translation tasks. Looking at the translation outputs, we provided evidence that the approach allows NMT models to disambiguate complex situations where the context is absolutely necessary, even for a human translator.

The next step is to go further by investigating more elaborate document embedding approaches and to bring these experiments to other languages (e.g.: Asian, Arabic, Italian, Spanish, etc.). To consider a training corpus with a majority of document delimited data is also very promising.

6. References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[3] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1243–1252.

[4] R. Bawden, R. Sennrich, A. Birch, and B. Haddow, "Evaluating discourse phenomena in neural machine translation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018.

[5] J. Tiedemann and Y. Scherrer, "Neural machine translation with extended context," in *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 82–92.

[6] L. Wang, Z. Tu, A. Way, and Q. Liu, "Exploiting cross-sentence context for neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2826–2831.

[7] S. Jean, S. Lauly, O. Firat, and K. Cho, "Does neural machine translation benefit from larger context?" *arXiv preprint arXiv:1704.05135*, 2017.

[8] L. Miculicich, D. Ram, N. Pappas, and J. Henderson, "Document-level neural machine translation with hierarchical attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2947–2954.

[9] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, and Y. Liu, "Improving the transformer translation model with document-level context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 533–542.

[10] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov, "Context-aware neural machine translation learns anaphora resolution," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1264–1274.

[11] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 2012–2017. [Online]. Available: https://www.aclweb.org/anthology/N19-1202

[12] Z. Gong, M. Zhang, and G. Zhou, "Cache-based document-level statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 909–919.

[13] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[15] R. Sennrich, B. Haddow, and A. Birch, "Controlling politeness in neural machine translation via side constraints," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 35–40.

[16] C. Chu, R. Dabre, and S. Kurohashi, "An empirical comparison of domain adaptation methods for neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 385–391.

[17] C. Chu and R. Wang, "A survey of domain adaptation for neural machine translation," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1304–1319.

[18] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri, "Findings of the 2019 conference on machine translation (wmt19)," in *Proceedings of the Fourth Conference on Machine Translation*

(Volume 2: Shared Task Papers, Day 1). Florence, Italy: Association for Computational Linguistics, August 2019, pp. 1–61.

[19] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and F. Marcello, "The iwslt 2015 evaluation campaign," in *Proceedings of the twelfth International Workshop on Spoken Language Translation*, Da Nang, Vietnam, December 2015, pp. 2–10.

[20] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, 2007, pp. 177–180.

[21] R. Sennrich, B. Haddow, and A. Birch, "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)*, vol. 1, 2016, pp. 1715–1725.

[22] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.

[23] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proceedings of ACL 2017, System Demonstrations*, Vancouver, Canada, July 2017, pp. 67–72.

[24] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, and M. Turchi, "Findings of the 2017 conference on machine translation (wmt17)," in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 169–214.

[25] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, and C. Monz, "Findings of the 2018 conference on machine translation (wmt18)," in *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, October 2018, pp. 272–307.