Controlling the Complexity of Machine Translation Output

Kelly Marchisio and **Jialiang Craig Guo** and **Cheng-I Jeff Lai** and **Philipp Koehn**
Center for Language and Speech Processing
Johns Hopkins University

Abstract

Today's machine translation systems are one-size-fits-all regarding the text complexity of output. In this paper, we develop two methods for controlling the readability level of translations. In our first approach, source-side sentences in the training corpus are tagged based on the readability of the matching target sentences before being used in training. Our second approach alters the traditional encoder-decoder architecture by specifying a joint encoder and separate decoders for simple and complex decoding modes, with training data split by readability. We demonstrate effective control over output readability score on three Spanish-English test sets with little BLEU score degradation. One of our best-performing models translates newstest2013 to a Dale-Chall readability score of 5.93 in simple mode, and 9.36 in complex mode.

1 Introduction

Machine translation is concerned with generating a semantically accurate translation from another language. Apart from generating grammatically and semantically correct translations, though, there are other factors which affect whether a reader is able to understand a translation. One important and easily neglected factor of how well a machine translation system performs is the complexity of the text relative to the skill of the reader. For instance, in defining "machine translation" to a 7-year-old, one might say, "machine translation is a way to take a sentence from one language and turn it into a sentence in another language", whereas when conversing with an adult, one might explain, "machine translation is the automated process by which a sentence in a source language can converted into a sentence in a foreign language". Both sentences carry the same semantic meaning and do not require specialist technical knowledge, but some of the phrases in the second translation may be too advanced for a child, such as "automated", "process by which", and "converted".

In this paper, we develop two machine translation methodologies that can control the complexity and reading level of the output, because the reading skill of the user of the system can vary. For skilled readers, we aim to use complex words. For less-skilled readers, such as the average 7-year-old child, we aim to make the translation use difficult words rarely while keeping the core idea of the source sentence. Accordingly, we built a system where a user can specify the complexity/reading level of the translation they wish to be output.

2 Background: Readability Tests

To quantitatively evaluate the complexity of English sentences, we used the three commonly-used automated readability tests.

2.1 Dale-Chall Readability

The Dale-Chall Readability is a traditional readability score that relies on a list of common English words to assess readability (Chall & Dale, 1995). The score contains two variables: percent of number of words per sentence and percent of unfamiliar words, and the equation is given as:

$$0.1579(\frac{\text{difficult words}}{\text{words}} \times 100) + 0.0496(\frac{\text{words}}{\text{sentences}}) \quad (1)$$

2.2 Flesch-Kincaid Grade Level

Flesch-Kincaid Grade Level is one of the most widely used readability metric, which estimates the readability of text using cognitively motivated features (Kincaid, Fishburne Jr, Rogers, &

Chissom, 1975). Flesch-Kincaid Grade Level approximately corresponds to the US grade level. The score contains two variables: percent of number of words per sentence and percent of syllables per word, and the equation is given as:

$$0.39(\frac{\text{words}}{\text{sentences}}) + 11.8(\frac{\text{syllables}}{\text{words}}) - 15.59 \quad (2)$$

2.3 Flesch Reading Ease

We also evaluate the translated text against the less commonly used Flesch Reading Ease, which is computed as (Flesch, 1948):

$$206.835 - 1.015(\frac{\text{words}}{\text{sentences}}) - 84.6(\frac{\text{syllables}}{\text{words}}) \quad (3)$$

3 Factors Affecting the Complexity of the Output Translation

The complexity of the output translation Z of current MT systems is affected by the overall complexity of the target sentences in the training corpus Y.

To show this effect, we trained the OpenNMT default RNN model on four different training corpora and tested readability on each model's translation of WMT newstest2013. Examining Tables 1 and 2, we observe a relationship between the average Dale-Chall readability score of the training corpus and the readability of the output translation.

Corpus	Dale-Chall Score
OpenSubtitles	3.429
OS+Europarl	6.079
Paracrawl	7.924
Europarl	8.800

Table 1: Dale-Chall Readability of Training sets.

Test	DC	FKG	FRE	BLEU
gold	8.11	9.49	59.83	-
OpenSubtitles	7.09	8.25	67.52	18.33
OS+Europarl	7.61	9.15	63.40	24.79
Europarl	7.75	9.48	61.84	22.97
Paracrawl	7.92	9.36	61.11	27.38

Table 2: Effect of corpus on translation readability for newstest2013.

In this project, we develop two training methods which allow end-users some control over the the readability level of their output.

4 Proposed Architectural Approaches

4.1 Double-Decoder

The first approach is an encoder-decoder model with a shared encoder and two decoders – one for "complex" decoding, and another for "simple" decoding as seen in Figure 1. When training a complex sentence, the joint encoder is paired with the "complex" decoder and loss is calculated based on that encoder-decoder pair. For a simple sentence, the encoder is paired with the "simple" decoder. In this way, the encoder learns shared representations for all source sentences, while separate decoders tune themselves to sentences that conform to the desired complexity level. At inference time, we pass a flag indicating whether we want the output to be simple or complex. The corresponding decoder then translates the test set.

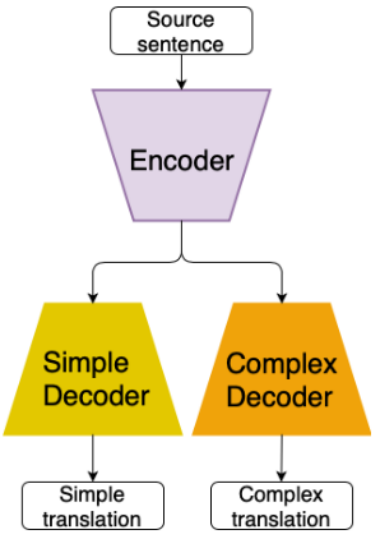

Figure 1: Encoder-decoder model with separate decoders for simple vs. complex output settings.

4.2 Tagged Data

Our second approach utilizes a short unique text string added to the end of each training sentence corresponding to that sentence's complexity. Inspired by (Sennrich, Haddow, & Birch, 2016), the intuition behind this simple method is that the attention mechanism learns to pay attention to the tag when decoding into the simple or complex setting. The approach requires no customization of model architecture or training procedure. At test time, source-side test sentences are augmented with the complex or simple string used during training to indicate the desired complexity of the

output. We chose unique text tags that were unlikely to appear elsewhere in the English corpus to avoid overloading the symbol with multiple meanings. A third unique tag was added to sentences that did not meet the chosen thresholds to be considered simple or complex, so the model might learn from these examples without detracting from the goal of keeping the mean readability score given a simple tag, versus given a complex tag, far apart.

4.3 Data Selection

We develop a novel method for data selection to generate our simple and complex training sets. We first score the readability of each target-side sentence in the corpus. Next, we select which sentences to include in the training sets based on their percentile rank for readability. For instance, for the double decoder architecture, we might choose to include the bottom 30% of available training sentences as the simple set, the top 30% as the complex set, and discard the remaining sentences. In the tagged data method, we equivalently tag the bottom and top 30th percentiles as simple/complex, and the remaining as neutral. We experiment with multiple thresholds, and report our results.

5 Technical Implementation

5.1 Datasets

We use three Spanish-English training sets: the European Parliament Proceedings (Europarl) (Koehn, 2005), OpenSubtitles2018 corpus (Lison & Tiedemann, 2016), and Paracrawl[1]. Europarl contains transcripts of European Parliamentary proceedings, OpenSubtitles2018 is a corpus of movie subtitles, and Paracrawl consists of aligned data scraped from the web. We chose Spanish to English translations for ease of qualitative assessment and corpus size. There are ∼2M Spanish-English sentence pairs in Europarl, 61.4M pairs in OpenSubtitles2018, and ∼16M aligned sentences in Paracrawl.

For corpus complexity experiments in Table 2 (hereafter, "baseline"), we use 2M randomly-selected lines from OpenSubtitles2018 as the OpenSubtitles training set. OS+Europarl consists of the OpenSubtitles training corpus concatenated with the full Europarl training set and shuffled, for ∼4M lines of Spanish-English text. The Paracrawl test set consists of 15 million randomly-selected lines from the Paracrawl corpus.

The development set for the OpenSubtitles baseline was 10K randomly-selected lines from OpenSubtitles. For the OS+Europarl baseline, it was the concatenation of newstest2012 (∼3000 lines) and 10K randomly selected lines from OpenSubtitles2018. Many more lines were chosen from OpenSubtitles2018 than the newstest sets for both dev and test because sentences in Open Subtitles tend to be shorter than in newstest, and we wanted to get a more representative sample of our performance by including more sentences. For the Europarl baseline, we used newstest2012, and for Paracrawl, 3000 randomly-selected lines from Paracrawl. Double decoder models were validated by assessing the performance of each decoder on the development set separately.

The test sets are newstest2013 (3000 lines), a combined test set of newstest2013 + a different 10K randomly-selected lines from OpenSubtitles2018, and different 3K randomly-selected lines from Paracrawl.

5.2 Data Preprocessing

All data were punctuation-normalized and tokenized using the standard Moses scripts (Koehn et al., 2007). Training data was then cleaned using Moses clean-corpus-n.perl using default parameters and a maximum sentence length of 100 words. All data were truecased and split into BPE (Sennrich, Haddow, & Birch, 2015) tokens using 32000 merge operations. After BPE processing, train and dev data were again cleaned with clean-corpus-n.perl using default parameters and a maximum length of 100 BPE tokens.

To select "simple" and "complex" data for the two approaches, we obtained the Dale-Chall readability score for each line in the training corpus and the average readability score for the corpus. We then selected percentile-based readability thresholds below which sentences would be labeled "simple", and above which they would be labeled "complex". We experimented with various thresholds.

5.3 Encoder/Decoder Models

The basic model architecture is the default RNN-based encoder-decoder model with attention (Luong, Pham, & Manning, 2015) from OpenNMT (Klein, Kim, Deng, Senellart, & Rush,

[1] `https://paracrawl.eu/releases.html`, version 1

2017). The encoder and decoder are two-layer LSTMs with hidden size = 500 and word embedding size = 500. The models were trained with stochastic gradient descent with the default learning rate of 1.0.

Each model was trained until performance on the validation set ceased to improve. For testing, we chose the model with lowest validation perplexity. In the case of double-decoder models, lowest perplexity did not typically occur at the same timestep for simple and complex decoders. In that case, we chose a model that had good performance on both validation sets.

5.4 Readability Scorers

Readability was scored using the textstat[2] implementations of the Dale-Chall (Chall & Dale, 1995) Flesch-Kincaid Readability Formula (Grade Level) (Kincaid et al, 1975), and Flesch Reading Ease (Flesch, 1948).

5.5 A Note about BLEU Score

Our goal is to have the output translation have reasonably high BLEU score (Papineni, Roukos, Ward, & Zhu, 2002) that aligns with translation performance by the baseline. In this way, we ensure that our system does not sacrifice much semantic translation quality. Without evaluating BLEU, one can envision how the "complex" system might outperform the simple system when evaluated only using readability tests by losing all meaning of the sentence and simply outputting complex words. This is not as trivial as with BLEU score evaluation under typical NMT conditions; We explicitly want translations to differ under our "complex"/"simple" experimental settings, but each sentence only has one gold translation. Thus, for instance, we would expect a Europarl source sentence evaluated as having high complexity by the Flesch-Kincaid readability to have some words modified when translated in the "simple" setting, therefore BLEU score based on matching ngram counts will suffer despite the fact that our target measure (lower complexity) may improve. We expect (and desire) a reduced BLEU score in this setting that reflects complex words being replaced by simpler ones. Even so, BLEU should not decrease severely since we do not expect our different experimental settings to completely rewrite sentences. Thus, we evaluate BLEU to ensure that translation quality does not suffer severely when adjusting output complexity.

[2] https://github.com/shivam5992/textstat

6 Results

Results of the experiments corresponding to each of the training and test sets are seen in the following tables. Baseline refers to the scores of the test set translation when translated with the baseline single encoder-decoder model trained on the same training set. We observe from Figure 2 that as the constraints for categorizing a sentence as simple or complex become more strict, the mean Dale-Chall score for newstest2013 trained in simple versus complex settings widens. Observing BLEU performance in Table 3, we see that BLEU does suffer in response to the more extreme translation changes brought about by increasingly divergent simple and complex training sentences. As mentioned previously, decreased BLEU in this setting may actually be a sign of improved performance towards our goal of controlling readability. Therefore to judge the quality of our translations, we must inspect manually. Results for all other training/test set combinations on both architectures are similar. Additional results are displayed in the appendix.

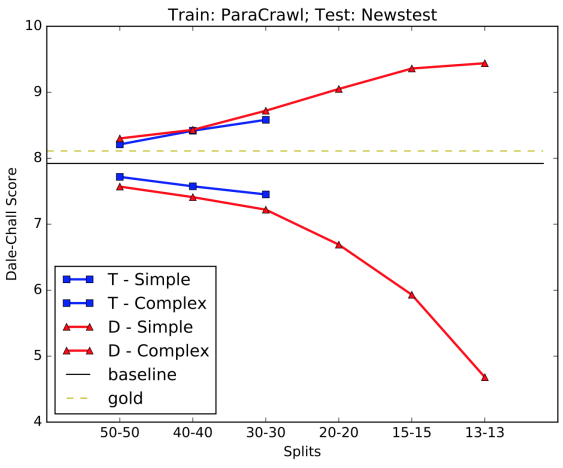

Figure 2: Results of double decoder and tagged models trained on Paracrawl data, tested on newstest2013.

6.1 Qualitative Results

The qualitative examples below were produced when translating newstest2013 using either of our architectures trained on Paracrawl data. Baseline translated sentence is italicised, the complex trans-

	DC	FKG	FRE	BLEU
gold	8.11	9.49	59.83	-
baseline	7.92	9.36	61.11	27.38
50-50	7.57	9.00	63.71	26.41
	8.30	9.59	59.16	26.71
40-40	7.41	8.82	64.93	26.19
	8.43	9.65	58.54	26.36
30-30	7.22	8.60	66.18	25.56
	8.72	9.84	56.79	25.89
20-20	6.69	7.97	69.75	23.51
	9.05	9.99	54.99	24.08
15-15	5.93	7.30	74.24	20.85
	9.36	10.16	53.19	22.04
13-13	4.68	6.43	80.02	18.52
	9.44	10.16	52.71	21.28

Table 3: Performance on newstest2013 of double-decoder models trained on Paracrawl data.

lation is in the middle, and the simplified translation appears after.

6.1.1 Double Decoder - 15/15 Split

But my provocations are directed to start a conversation.
But my provocations are **directed** to **initiate** a conversation.
But my provocations are **meant** to **start** a conversation.

Oh, that's going to be very difficult to recognize.
Oh, this will be **extremely difficult** to recognize.
Oh, that's going to be **very hard** to recognize.

You will speak and show it at the same time.
You will **discuss** and **display** the same time.
You will **speak** and **show** it at the same time.

Not everyone feels happy with the fact that...
Not all are satisfied with the fact that...
Not everyone feels happy with the fact that.

6.1.2 Tagged Model - 40/40 split

It hurts the health of people.
Greatly **impaired** people's health.
It **hurts** the health of people.

Anyway, we suppose that it is Higgs, as the possibilities of mistake seem few.
Anyway, let us **assume** that it is Higgs, as the **possibilities** of **error** seem to be few.
Anyway, let us **suppose** that it is Higgs, as the **chances** of **mistake** seem few.

..., because it will end up being hurt.
..., because it will end up **being damaged**.
..., because it will end up **getting hurt**.

gold: ..., where he arrived the previous day.
..., which **arrived** on **the previous day**.
..., which **came** to **the day before**.

7 Attention Visualization

In Figure 3, we see a heatmap of attention when the tagged data model was translating the same sentence in simple versus complex mode. When choosing the word "adversely" in complex mode versus "negatively" in simple mode, we see attention placed on the complexity indicator tags "czxc" and "szxc". This suggests that the model attended to the complexity tag when deciding which word to use.

8 Analysis and Discussion

We deem the overall best model to be the double decoder trained on Paracrawl with a 15/15 split of simple/complex data. This model translates newstest2013 to an overall Dale-Chall readability score of 5.93 in simple mode and 9.36 in complex mode, while retaining reasonable BLEU. Given the baseline readability score of 7.92 and the fact that the readability of the gold target sentences is 8.11, we have demonstrated success both raising and lowering the readability level of the test set. The results on the Paracrawl and OpenSubtitles2018+Europarl constructed test sets given Paracrawl or OpenSubtitles+Europarl training data suggests that our methods are general and applicable beyond the scope of the datasets we chose. Our qualitative examples demonstrate that though BLEU score depreciated, some of the decrease reflects correct text changes towards our goal of increasing or reducing text complexity.

We observed translations in "simple" mode sometimes ending early or producing short sentences (specifically, we observed Paracrawl 15-15 in the double decoder). Simple training sentences may tend to be shorter than complex training sentences, which may teach the simple decoder to produce short sentences.

Regarding the difference between the double decoder and tagged data models, we observe that

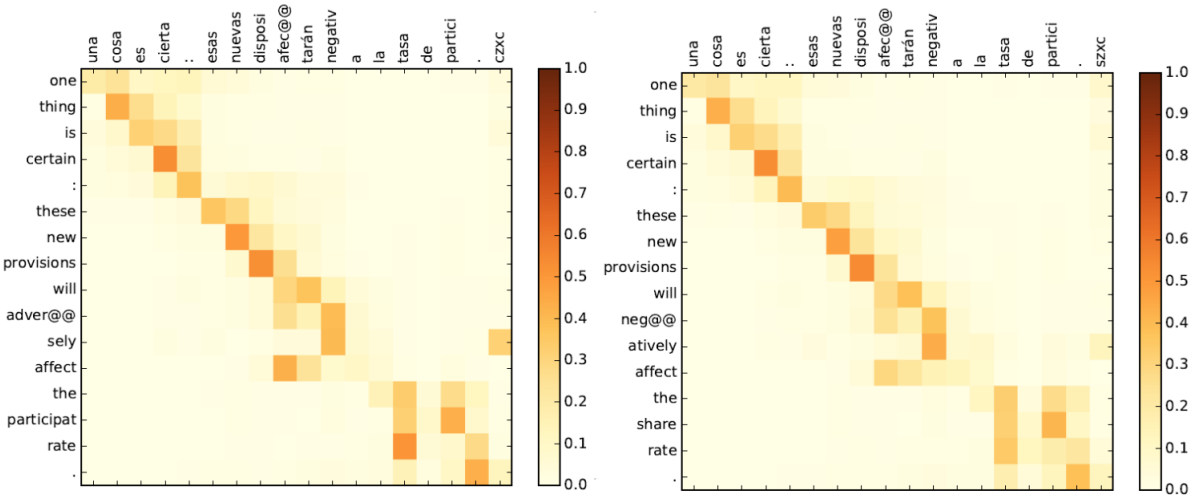

Figure 3: Attention visualisation in simple vs. complex mode of tagged model (40/40 split, trained on Paracrawl).

the double decoder is generally able to pull further apart the mean readability of sentences translated in simple vs. complex mode. The separated decoders may become more specialized towards creating sentences of particular relative readability levels, which may explain this observation.

We also observed the tagged model retaining higher BLEU in general than the double decoder. This could be related to the fact that the tagged model does not pull apart the means of simple and complex translations quite as far as the double decoder. That said, we suspect this phenomenon may be better explained by the fact that in the tagged model, we retain sentences of an intermediate complexity level and still use them in translation, but with an ambiguous complexity tag (in the double decoder model, we discard intermediate sentences). We suspect this extra data helped maintain high BLEU. The higher BLEU score also suggests that the tagged model may be preferable in low-resource settings.

9 Related Work

Prior work in machine translation and natural language processing primarily focus on readability assessment and text simplification. For readability assessment, a data-driven method is proposed in (Le, Nguyen, & Wang, 2018) for assessing the readability of document text, whereas (Ciobanu, Dinu, & Pepelea, 2015) investigated the readability of the MT system output with standard metrics. (Jones et al., 2005) also investigated the readability of MT and ASR systems output but with human evaluation. As for text simplification, (Hardmeier, Stymne, Tiedemann, & Nivre, 2013) proposes a document-level decoder for SMT and mentioned a case study that utilizes document-wide feature to improve the readability of text. Similarly, global features can be used for text simplification for SMT (Stymne, Tiedemann, Hardmeier, & Nivre, 2013). Contrary to (Stymne et al., 2013), (Xu, Napoles, Pavlick, Chen, & Callison-Burch, 2016) designed a new training objective for SMT text simplification.

Our work share similar grounds with these work such that it involves controlling the readability of machine translation output. Similar to [(Le et al., 2018),(Ciobanu et al., 2015),(Jones et al., 2005), we adopted evaluation metrics for assessing the MT output. However, the readability constraint is taken into account during training in our proposed approaches. (Stymne et al., 2013) introduces document-level feature such as type/token ratios and lexical consistency as input to MT system. On the other hand, our approaches at most require an additional simplicity/complexity tag. In addition, different from (Xu et al., 2016) in which new training objective is proposed for text simplification, our NMT training objective remains the same.

10 Conclusion

In this work, we developed two methods for gaining control of the readability level of output translations in neural machine translation. Both of our proposed models can significantly increase or de-

crease the readability levels of multiple test sets when trained on different corpora, and have good qualitative results. Notably, our tagged data model can be deployed immediately on existing NMT systems with no architectural changes.

11 Future Work

The bottom $\sim$10% of Paracrawl data had very low readability score, which we believe may have severely negatively-impacted results as we made stricter requirements for qualification as a "simple" sentence to the double decoder. In the future, we plan to experiment with cleaning the corpus and discarding "junk" sentences from either extreme of the readability spectrum before training. We suspect doing so will help us retain better BLEU and have more accurate translations, rather than learning from too much "junk" data. Additionally, we plan to experiment using the "intermediary" sentences in the double decoder to see if doing so retains BLEU as we observed in the tagged model. We might split the intermediary sentences into two groups and alternate training of each decoder with these extra sentences. We suspect this may pull the readability means of simple vs. complex translations closer together, but produce better translations overall.

Finally, we observed exciting effects related to formality which are outside the scope of this paper. Particularly when training on Europarl and OpenSubtitles2018 data, we often observed that sentences trained in "complex" mode appeared more formal than those trained in "simple" mode; most contractions were removed, and word selection appeared more formal (in line what the data we might expect to be produced during European Parliamentary Proceedings). In the future, we plan repeat these experiments with the goal of increasing/decreasing formality of output translations, and have already observed promising first results within these experiments.

Acknowledgments

Arya, Shouyang(sp?)

References

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new dale-chall readability formula*. Brookline Books.

Ciobanu, A. M., Dinu, L. P., & Pepelea, F. (2015). Readability assessment of translated texts. In *Proceedings of the international conference recent advances in natural language processing* (pp. 97–103).

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Hardmeier, C., Stymne, S., Tiedemann, J., & Nivre, J. (2013). Docent: A document-level decoder for phrase-based statistical machine translation. In *Acl 2013 (51st annual meeting of the association for computational linguistics); 4-9 august 2013; sofia, bulgaria* (pp. 193–198).

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., & Weinstein, C. (2005). Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *Acoustics, speech, and signal processing, 2005. proceedings.(icassp'05). ieee international conference on* (Vol. 5, pp. v–1009).

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proc. acl.* Retrieved from `https://doi.org/10.18653/v1/P17-4012` doi: 10.18653/v1/P17-4012

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Mt summit* (Vol. 5, pp. 79–86).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... others (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the acl on interactive poster and demonstration sessions* (pp. 177–180).

Le, D.-T., Nguyen, C.-T., & Wang, X. (2018). Joint learning of frequency and word embeddings for multilingual readability assessment. In *Proceedings of the 5th workshop on natural language processing techniques for educational applications* (pp. 103–107).

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sennrich, R., Haddow, B., & Birch, A. (2016). 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* (pp. 35–40).

Stymne, S., Tiedemann, J., Hardmeier, C., & Nivre, J. (2013). Statistical machine translation with readability constraints. In *Proceedings of the 19th nordic conference of computational linguistics (nodalida 2013); may 22-24; 2013; oslo university; norway. nealt proceedings series 16* (pp. 375–386).

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, *4*, 401–415.

12 Appendix

	DC	**FKG**	**FRE**	**BLEU**
gold	4.49	3.89	83.00	-
baseline	4.52	4.01	81.85	27.30
50-50	4.03	3.52	85.34	26.70
	6.35	5.00	74.41	24.19
40-40	3.90	3.46	85.75	26.56
	6.60	5.18	72.90	23.45
30-30	3.70	3.30	86.86	26.28
	7.59	5.75	68.57	22.16

Table 4: Performance on combined test set of double-decoder models trained on Paracrawl data.

	DC	**FKG**	**FRE**	**BLEU**
gold	8.01	9.77	56.06	-
baseline	7.83	9.57	56.56	31.68
50-50	7.41	9.08	60.21	30.77
	8.26	9.98	53.77	30.53
40-40	7.19	8.86	61.69	29.95
	8.43	9.96	53.68	29.63
30-30	6.88	8.94	60.48	28.58
	8.73	10.24	51.32	28.50

Table 5: Performance on the Paracrawl test set of double-decoder models trained on Paracrawl data.

	DC	**FKG**	**FRE**	**BLEU**
gold	4.49	3.89	83.00	-
baseline	4.20	3.77	83.63	29.00
50-50	3.63	3.26	86.95	27.34
	6.23	4.95	74.84	26.07
40-40	3.09	2.90	89.22	25.41
	6.73	5.22	72.62	24.41
30-30	1.02	0.92	98.47	15.35
	7.58	5.78	68.19	22.26

Table 6: Performance on combined test set of double-decoder models trained on EuroParl+OpenSubtitles2018 data.

	DC	FKG	FRE	BLEU
gold	8.11	9.49	59.83	-
baseline	7.61	9.15	63.40	24.79
50-50	6.76	8.09	69.51	22.29
	8.06	9.51	60.65	24.57
40-40	5.92	7.24	74.20	19.21
	8.28	9.66	59.47	23.95
30-30	1.59	1.58	98.00	3.95
	8.65	9.93	57.17	23.30

Table 7: Performance on newstest2013 of double-decoder models trained on EuroParl+OpenSubtitles2018 data.

	DC	FKG	FRE	BLEU
gold	8.01	9.77	56.06	-
baseline	7.32	9.20	61.59	22.80
50-50	6.34	7.73	69.39	18.95
	7.86	9.62	58.51	22.37
40-40	5.33	6.84	73.92	15.25
	7.98	9.89	57.26	22.59
30-30	1.14	1.10	99.36	2.32
	8.39	10.08	55.04	21.61

Table 8: Performance on Paracrawl test set of double-decoder models trained on EuroParl+OpenSubtitles2018 data.

	DC	FKG	FRE	BLEU
gold	4.487	3.888	83.002	-
baseline	4.523	4.006	81.847	27.30
50-50	4.203	3.759	83.677	27.45
	6.032	4.885	75.315	25.97
40-40	4.059	3.661	84.466	27.43
	6.435	5.068	73.882	25.38
30-30	3.994	3.565	85.183	27.49
	6.936	5.265	72.372	24.71

Table 9: Performance on combined test set for tagged model trained on ParaCrawl.

	DC	FKG	FRE	BLEU
gold	8.009	9.774	56.060	-
baseline	7.826	9.572	56.558	31.68
50-50	7.539	9.390	58.329	31.84
	8.173	9.855	54.677	31.66
40-40	7.412	9.121	60.303	31.92
	8.280	9.830	54.772	31.69
30-30	7.309	9.158	60.145	31.71
	8.486	9.970	53.667	31.51

Table 10: Performance on ParaCrawl test set for tagged model trained on ParaCrawl.

	DC	FKG	FRE	BLEU
gold	8.105	9.491	59.825	-
baseline	7.924	9.356	61.115	27.38
50-50	7.717	9.149	62.865	27.32
	8.210	9.530	59.718	27.27
40-40	7.574	9.052	63.759	27.09
	8.417	9.697	58.414	27.24
30-30	7.451	8.977	64.406	27.14
	8.582	9.790	57.566	27.09

Table 11: Performance on newstest2013 for tagged model trained on ParaCrawl.

	DC	FKG	FRE	BLEU
gold	4.487	3.888	83.002	-
baseline	4.198	3.766	83.631	29.00
50-50	3.828	3.519	85.541	29.10
	5.708	4.624	77.323	27.93
40-40	3.555	3.355	86.668	28.33
	6.867	5.247	72.647	26.34
30-30	3.199	3.023	88.246	27.13
	6.476	5.007	74.149	26.69

Table 12: Performance on combined test set for tagged model trained on Europarl+OpenSubtitles2018.

	DC	FKG	FRE	BLEU
gold	8.009	9.774	56.060	-
baseline	7.321	9.201	61.592	22.80
50-50	6.816	8.840	64.595	22.61
	7.758	9.588	59.159	23.18
40-40	6.421	8.335	67.609	21.84
	7.929	9.686	58.142	23.08
30-30	5.814	7.223	71.665	19.01
	8.064	9.592	57.830	22.32

Table 13: Performance on Paracrawl test set for tagged model trained on Europarl+OpenSubtitles2018.

	DC	FKG	FRE	BLEU
gold	8.105	9.491	59.825	-
baseline	7.614	9.151	63.399	24.79
50-50	7.215	8.890	65.653	24.77
	7.967	9.433	61.415	24.99
40-40	6.819	8.486	68.314	24.03
	8.311	9.655	59.520	24.73
30-30	6.190	7.647	71.990	22.63
	8.298	9.566	59.721	24.69

Table 14: Performance on newstest2013 for tagged model trained on Europarl+OpenSubtitles2018.