Controlling the Reading Level of Machine Translation Output

Kelly Marchisio[*] and **Jialiang Guo**[*] and **Cheng-I Lai** and **Philipp Koehn**
Center for Language and Speech Processing
Johns Hopkins University
{kellym,guo,clai24,phi}@jhu.edu

Abstract

Today's machine translation systems output the same translation for a given input, despite important differences between users. In practice, translations should be customized for each reader, for instance when translating for children versus in a business setting. In this paper, we introduce the task of reading level control to machine translation, and provide the first results. Our methods can be used to raise or lower the reading level of output translations. In our first approach, source-side sentences in the training corpus are tagged based on the reading level (readability) of the matching target sentences. 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 partitioned by reading level. We demonstrate control over output readability score on three test sets in the Spanish–English language direction.

1 Introduction

Though the goal of machine translation is to generate semantically accurate translations from one language to another, there are other factors which affect whether a translation is "good". One often-neglected factor is the reading level of the translation—different contexts require different reading levels. When translating for less-skilled readers, one may desire a translation with common vocabulary and simple sentence structures. In a professional setting, however, one often requires concise language with advanced vocabulary and syntactic structure.

For instance, when translating a Spanish web page about machine translation to an English-speaking 7-year-old, one might output, "machine translation is a way to take a sentence from one language and turn it into a sentence in another language". When advertising new machine translation software to a potential investor, one might explain, "machine translation is the automated process by which a sentence in a source language can be converted into a sentence in another language". Both sentences carry the same meaning and do not require specialist technical knowledge, but decreasing the complexity in the first makes it easier for a child to understand, and increasing the complexity in the second makes it sound more professional and sophisticated. Furthermore, for native speakers of low-resource languages where machine translation quality may currently be poor but who can read basic phrases in a second language where translation quality is high, they may prefer to read a lower complexity but semantically accurate translation in their second language over an inaccurate, garbled message in their native tongue.

In this paper, we introduce the task of reading level control (readability control) to machine translation. We develop two methodologies that control the reading level of a translation in the Spanish–English language direction, focusing on lexical complexity as a first step. For professional settings, we aim to produce advanced vocabulary. For less-skilled readers, the translation should use simple words while maintaining the meaning of the source sentence. Accordingly, we build a system where a user can specify the reading level ("sim-

[*] Equal contribution

ple" or "complex") of the translation they wish to be output. Future work should examine controlling other factors that affect the readability of a sentence, such as syntactic structure.

2 Background: Readability Tests

To quantitatively evaluate the reading level of English sentences, we use three commonly-used automated readability[1] tests.

2.1 Dale-Chall Readability

The Dale-Chall (DC) readability score utilizes a list of 3000 common English words, which captures lexical information of text (Chall and Dale, 1995). Words not in the list are considered "difficult". The metric is computed using the percentage of difficult words and the average number of words per sentence, as below:

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

2.2 Flesch-Kincaid Grade Level

One of the most widely-used readability metrics, Flesch-Kincaid Grade Level (FKG) approximately corresponds to grade level in the US schooling system (Kincaid et al., 1975). The score considers only two basic features of the text—the average number of words per sentence and the average number of syllables per word. It is computed as below:

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

2.3 Flesch Reading Ease

We also evaluate translations with Flesch Reading Ease (FRE) (Flesch, 1948), where higher scores indicate "easier" text. FRE was the basis for FKG, and is computed as:

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

2.4 Readability Tests for Other Languages

Apart from the three tests above for English, there are many readability tests available for other languages, such as Amstad readability index for German (Amstad, 1978), GulpEase index for Italian (Lucisano and Piemontese, 1988), and LIX for a wide range of languages (Björnsson, 1968). There are also various approaches to reading level scoring based on machine learning and natural language processing techniques (François and Miltsakaki, 2012).

In this work, we focus on the three traditional English readability tests mentioned above as a first step for the Spanish–English language direction. Though the readability tests aren't perfect, they achieve good results in our work and are easy to implement. We anticipate that our general frameworks will work with various target languages and readability scorers, provided the corresponding readability tests effectively estimate reading level.

3 Factors Affecting the Reading Level of the Output Translation

At test time, it is reasonable to anticipate that advanced vocabulary and phrases in a source sentence will be translated into advanced vocabulary and phrases in a target sentence, and simple lexical features of a source to simple lexical features in a target. This leads to a problem in the typical setting where there is a single source document at test time. Since the source has fixed complexity, users do not have control over the reading level of the output. As a result, we must find other ways of controlling output reading level besides altering the source.

In this section, we demonstrate that the reading level of output translations is also affected by the overall reading level of target-side sentences during training. We train four OpenNMT (Klein et al., 2017) default RNN models on four separate training corpora in the Spanish–English language direction. The corpora have different overall target-side readability (Table 1). We then test the readability of each model's translation of WMT newstest2013[2] (Table 2). Please see Section 5 for implementation details and description of datasets.

Corpus	DC	FKG	FRE
OpenSubtitles	3.43	2.28	89.39
OpenSubtitles+Europarl	6.08	7.27	69.43
ParaCrawl	7.92	11.17	56.43
Europarl	8.80	12.41	48.94

Table 1: Overall readability scores of the target-side sentences in different training corpora. Lower DC score, lower FKG score, and higher FRE score indicate simpler sentences.

[1] Throughout this paper, we use the terms "readability", "reading level", and "text complexity" interchangeably.

[2] http://www.statmt.org/wmt13/translation-task.html

	DC	FKG	FRE
gold	8.11	9.49	59.83
OpenSubtitles	7.09	8.25	67.52
OpenSubtitles+Europarl	7.61	9.15	63.40
Europarl	7.75	9.48	61.84
ParaCrawl	7.92	9.36	61.11

Table 2: Effect of the training corpus on translation readability for newstest2013. Lower DC score, lower FKG score, and higher FRE score indicate simpler sentences.

Examining Tables 1 and 2, we observe that the readability of the translation tends to mimic the readability of the target sentences in the training corpus. This effect inspired us to partition the training data into "simple" and "complex" subsets so the model can learn how sentences of lower and higher reading level should look.

4 Proposed Approaches

In this paper, we develop two training methods which allow some control over the reading level of machine translation output.

4.1 Data Tagging

Inspired by Sennrich et al. (2016)'s work controlling politeness, our first approach utilizes a short text token added to the end of each source-side training sentence, which corresponds to the matching target-side sentence's readability. The intuition behind this method is that the attention mechanism will learn to pay attention to the complexity token when decoding in the simple or complex setting.

A token indicating whether each training sentence pair is of low or high reading level is used if the target sentence meets a preset readability threshold. A third token indicating intermediate reading level is added to sentences that do not meet the chosen thresholds, so that the model can learn other knowledge—such as a better language model and alignment—from these examples.

The data tagging approach requires no customization of model architecture or training procedure. At test time, we append a "simple" or "complex" token to the test source sentences to specify the desired reading level of the output. We choose tokens that are unlikely to appear in the target language to avoid overloading the symbols with multiple meanings.

4.2 Double-Decoder

The second approach is an encoder-decoder model with a shared encoder and two decoders—one for "complex" decoding, and the other for "simple" decoding as shown 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 a shared representation for all source sentences, while separate decoders tune themselves to sentences that have the desired reading 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.

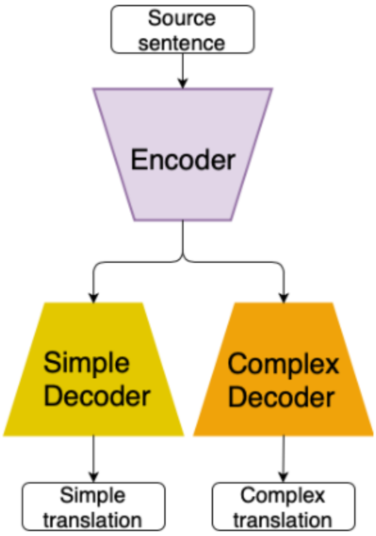

Figure 1: Encoder-decoder model with separate decoders for simple vs. complex output settings.

4.3 Data Selection

4.3.1 Partitioning by Readability Level

We use a method of data selection to partition our data into "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, in the 30-30 setting for the double-decoder architecture, we 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 data tagging approach, we equivalently tag the bottom and top 30% as simple/complex, and the remaining as neutral. We experiment with multiple thresholds.

4.3.2 Oversampling

Though more extreme data partitioning endows more effective control over output readability, it also brings potential problems; the data tagging approach has limited "simple" and "complex" examples from which to gain knowledge about reading level, and the double-decoder approach discards so much data that it could suffer translation quality degradation. We therefore use oversampling to reinforce the effect of data with extreme readability. For the data tagging approach, we use an extreme data partition (e.g., 15-15) and oversample all examples tagged as "simple" or "complex"; for the double-decoder approach, we use the 50-50 data partition but oversample the extreme parts (top 15% and bottom 15%).

5 Technical Implementation

5.1 Datasets

We use three Spanish–English training sets: the European Parliament Proceedings (Europarl) (Koehn, 2005), OpenSubtitles2018 (OS) corpus (Lison and Tiedemann, 2016), and ParaCrawl[3]. Europarl contains transcripts of European Parliamentary proceedings, OpenSubtitles2018 is a corpus of movie subtitles, and ParaCrawl consists of data scraped from the web.

For training each model and for the preliminary experiments in Table 2, we use either: ~2 million randomly-selected lines from OpenSubtitles2018, the ~2 million line Europarl training set, a concatentaion of the aforementioned two corpora (OS+Europarl), or 14.7 million randomly-selected lines from ParaCrawl.

Development sets are: 10,000 held-out lines from OpenSubtitles2018 for the OpenSubtitles baseline, newstest2012 for the Europarl baseline, the concatenation of newstest2012 and the OpenSubtitles development set for the OS+Europarl baseline, and 3,000 held-out lines from ParaCrawl for the ParaCrawl baseline. Double-decoder models are validated by assessing the performance of each decoder separately on the development set.

The test sets are newstest2013 (3,000 lines), a combined test set of newstest2013 plus 10,000 held-out lines from OpenSubtitles2018, and 3,000 held-out lines from ParaCrawl.

5.2 Data Preprocessing

All data were punctuation-normalized, tokenized, truecased, and cleaned to a maximum sentence length of 100 words using the standard Moses scripts (Koehn et al., 2007). We applied BPE (Sennrich et al., 2015) to all data using 32,000 merge operations. Training and development 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 apply the data selection method of Section 4.3 using the Dale-Chall readability score. All readability scores in this work were calculated after removing BPE, detruecasing, and detokenizing the data.

5.3 Models & Training

The basic model architecture is the default RNN-based encoder-decoder model with attention (Luong et al., 2015) from OpenNMT. The encoder and decoder are two-layer LSTMs (Hochreiter and Schmidhuber, 1997) with a 500-dimension hidden size and 500-dimension word embeddings. The models were trained with batch size 64 using stochastic gradient descent with the default initial learning rate of 1.0. We decay the learning rate by a factor of 0.5 starting at 50,000 steps, and further decay every subsequent 10,000 steps.

Each model was trained until performance on the validation set ceased to improve. For testing, we chose the model with lowest validation perplexity. For 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.

Readability was scored using the textstat[4] implementations of the Dale-Chall, Flesch-Kincaid (Grade Level), and Flesch Reading Ease formulas. BLEU was scored using `multi-bleu-detok.perl` from the Moses toolkit (Koehn et al., 2007). Statistical significance was assessed using SciPy (Jones et al., 2001).

6 Results

6.1 Quantitative Results

Tables 3 and 4 show the readability performance of data tagging and double-decoder approaches on

[3] https://ParaCrawl.eu/releases.html, version 1

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

newstest2013 at different levels of data partitioning. (For example, a 30-30 partition corresponds to the case where the bottom/top 30% of data are labeled as simple/complex.) “Baseline” hereafter refers to the single encoder-decoder model trained on the original, unpartitioned dataset. These tables demonstrate effective control over average output readability for both approaches. We also conducted two-tailed paired samples t-tests[5] which demonstrated that DC, FKG, and FRE results in both decoding modes are significantly different from the baseline ($p<0.001$).

Partition	Mode	DC	FKG	FRE	BLEU
-	gold	8.11	9.49	59.83	-
-	baseline	7.92	9.36	61.11	27.38
50-50	simple	7.72	9.15	62.87	27.32
	complex	8.21	9.53	59.72	27.27
30-30	simple	7.45	8.98	64.41	27.14
	complex	8.58	9.79	57.57	27.09
15-15	simple	7.26	8.80	65.60	26.74
	complex	8.72	9.83	56.57	26.62
15-15*	simple	6.96	8.45	67.78	25.91
	complex	8.96	9.93	55.42	25.47
13-13	simple	7.23	8.82	65.69	26.71
	complex	8.69	9.82	56.68	26.74

Table 3: Performance on newstest2013 of data tagging approach trained on ParaCrawl. DC, FKG, and FRE are readability measures (lower indicates simpler for DC/FKG, and higher for FRE). e.g., 7.72 is the average DC score of the output in simple mode using a 50-50 partition. 15-15* means oversampling the top/bottom 15% of data (3x). All DC/FKG/FRE results are significant ($p<0.001$).

For all models, translations in complex mode are slightly shorter than in simple mode, and have slightly more bytes per word. In the data-tagging 15-15 mode, complex mode translations averaged 18.4 words per line, versus 19.3 in simple mode. The bytes-per-word were 6.0 and 5.7 for complex and simple mode, respectively. As data splits became less aggressive, the difference decreased. This suggests that in complex mode, the models attempt to be more concise while using longer words.

Partition	Mode	DC	FKG	FRE	BLEU
-	gold	8.11	9.49	59.83	-
-	baseline	7.92	9.36	61.11	27.38
50-50	simple	7.57	9.00	63.71	26.41
	complex	8.30	9.59	59.16	26.71
50-50*	simple	7.41	8.87	64.59	25.71
	complex	8.43	9.66	58.46	26.01
30-30	simple	7.22	8.60	66.18	25.56
	complex	8.72	9.84	56.79	25.89
20-20	simple	6.69	7.97	69.75	23.51
	complex	9.05	9.99	54.99	24.08
15-15	simple	5.93	7.30	74.24	20.85
	complex	9.36	10.16	53.19	22.04

Table 4: Performance on newstest2013 of double-decoder models trained on ParaCrawl data. In the 50-50* setting, 50% of data is designated “simple”, 50% “complex”, and the most extreme 15% of simple/complex data are oversampled (3x). All DC/FKG/FRE results are significant ($p<0.001$).

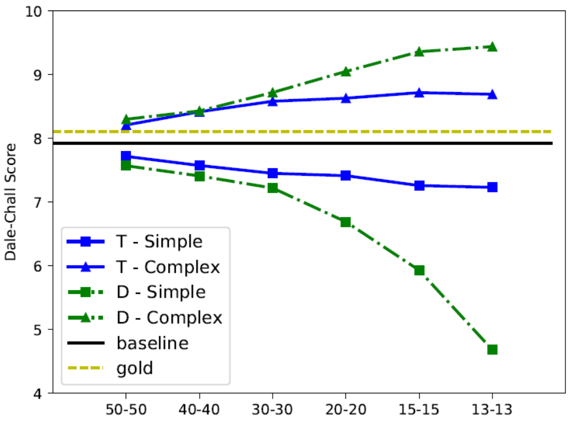

Figure 2: Readability results of newstest2013 translation in simple and complex mode for data tagging (T) and double-decoder (D) models trained on ParaCrawl.

Figure 2 demonstrates that as the constraints for categorizing a sentence as “simple” or “complex” become more strict, the gap widens between the mean readability score in simple mode and complex mode. This holds for all three readability metrics, on all three test sets and the two training corpora that with which we experimented.

In Table 3, we see little negative effect on BLEU (Papineni et al., 2002) for the data tagging approach. In Table 4, however, we see that BLEU suffers as the double-decoder model receives less data. In the 13-13 partition, BLEU drops to 18.52 in simple mode, when the simple decoder receives only ∼1.9 million sentences, many of which are very short.

Table 5 compares the two approaches with different training and test sets, reporting the difference between the baseline Dale-Chall score and the readability of translated test sets in simple and complex modes. We observe that both methods

[5] https://docs.scipy.org/doc/scipy-1.1.0/reference/generated/scipy.stats.ttest_rel.html

Training Corpus	Approach	Test Set ParaCrawl	 OpenSubtitles+Europarl	 Newstest2013
ParaCrawl	T-15/15	-0.84 / +0.73	-0.74 / +2.79	-0.67 / +0.79
	D-15/15	**-2.41 / +1.53**	**-1.74 / +4.36**	**-1.99 / +1.44**
OpenSubtitles+Europarl	T-40/40	-0.90 / +0.61	-0.64 / **+2.67**	-0.80 / **+0.70**
	D-40/40	**-1.99 / +0.66**	**-1.11** /+2.53	**-1.69** / +0.67

Table 5: Performance of high-performing models with double-decoder (D) and data tagging (T) approaches on three test sets. The left/right number is the difference in Dale-Chall score between the baseline and the simple/complex translation. Model with the larger difference is bolded.

System	BLEU	Human Eval
Baseline	27.38	6.54
Weaker Baseline	24.34	5.64
Complex	24.08	5.75
Simple	23.51	5.84

Table 6: Average model score in human evaluation for models trained on Paracrawl. Complex and Simple represent complex and simple modes for the double-decoder approach with a 20-20 data partition.

work well for all six train-test pairs, and that the double-decoder method generally makes the simple translations simpler and the complex translations more complex, than the data tagging approach.

6.2 Qualitative Results

The qualitative examples in Tables 7 were produced when translating newstest2013 using the data tagging and double-decoder approaches.

We observe from the examples that both approaches successfully control the complexity of output sentences. Furthermore, the baseline appears an appropriate intermediary between the two complexity levels; For the baseline translation "This attitude is a deplorable vision of future.", the data tagging approach changes *"deplorable vision"* to *"terrible view"* in simple mode to decrease complexity. In complex mode, however, the model changes *"is"* to *"implies"* to make the sentence even more complex, keeping *"deplorable vision"*. We also see change in sentence structure. For example, in simple mode the data tagging approach produces, *"there is..."*, while in complex mode it produces, *"...occurred"* or *"...existed"*.

In the double-decoder approach we observe some loss in meaning for certain sentences as the threshold for training sentences to be qualified as simple or complex becomes more restrictive.

6.3 Human Evaluation

We performed human evaluation to determine whether the lower BLEU score observed in more extreme data-partitioning conditions in the double-decoder approach was the result of true loss in translation quality, or desirable swapping of simple/complex words. We randomly sampled 50 translations from newstest2013 and obtained the translations from the double-decoder 20-20 partition setting, along with the baseline model and a weaker baseline trained to achieve comparable BLEU to that of the double-decoder approach. Nine English-speaking adults each scored approximately one-third of the sampled translations on a 10-point scale so that each translation received three scores. Reviewers were instructed to score how well each translation matched the meaning of the reference, along with the fluency of the translation. Examples were presented in blocks with the reference translation followed by the four system translations in a random order for each block. Participants each scored 15 or 20 blocks.

In Table 6, we show the average score that translations from each system received. We observe that while the drop in BLEU in Table 4 reflects some lowered translation quality as judged by human reviewers, the loss in quality is smaller than the BLEU depreciation makes it seem. When compared to a baseline model with comparable BLEU to that of the "simple" and "complex" modes (the "weaker" baseline), the double-decoder approaches fair better in human evaluation despite having lower BLEU scores. This indicates that BLEU over-penalizes models trained to control readability level, and that readability-controlled translations are better than they appear based on BLEU alone.

Note that in this section we only performed human evaluation on outputs from the double-

Src	*Por este motivo, no creo que se haya producido una ruptura tan drástica como en el caso de Libia.*
Ref	*Therefore I do not think that this was as dramatic a fracture as was the case in the matters regarding Libya.*
Baseline	For this reason, I don't believe that there was such a drastic rupture as in the case of Libya.
Simple	For this reason, I don't believe that **there is** a drastic **break** as in Libya.
Complex	For this reason, I don't believe that a drastic **rupture occurred** as in Libya's case.
Src	*Esta actitud supone una deplorable visión de futuro.*
Ref	*This is woefully short-sighted.*
Baseline	This attitude is a deplorable vision of future.
Simple	This attitude **is** a **terrible view** of the future.
Complex	This attitude **implies** a **deplorable vision** of future.
Src	*Pero mis provocaciones están dirigidas a que se inicie una conversación.*
Ref	*But my provocations are aimed at starting conversation.*
Baseline	But my provocations are directed to start a conversation.
Simple	But my provocations are **meant** to **start** a conversation.
Complex	But my provocations are **directed** to **initiate** a conversation.
Src	*No todos se sienten contentos con el hecho de que...*
Ref	*Not everyone is happy that...*
Baseline	Not everyone feels happy with the fact that...
Simple	**Not everyone feels happy** with the fact that...
Complex	**Not all are satisfied** with the fact that...

Table 7: Example translations of newstest2013 in simple/complex mode from models trained on ParaCrawl (15-15). The first two examples come from the data tagging approach (15-15), and the second two come from the double-decoder approach (15-15).

decoder 20-20 model (which had a ∼3–4 BLEU drop compared to the baseline) whereby we do observe some loss in translation quality from the baseline. However, for the models which achieve very similar BLEU scores to the baseline, such as the data-tagging 50-50 and 30-30 model, there may be no loss in translation quality. Human evaluation could verify this notion.

6.4 Attention Visualization

In Figure 3, we see a heatmap of attention when the data tagging approach translated the same sentence in simple and complex modes. 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.

In many cases, however, the difference in word choice is not reflected by attention to the complexity tag. This could be because the difference in attention values is too small for humans to detect the color difference in the heatmap. A more plausible explanation is that information about the reading level has been passed to the hidden states at all positions by the bi-LSTM, so that the decoder doesn't need to pay attention to the complexity token (the last hidden state) to make different word choices.

6.5 Adaption to Multiple Reading Level Setting

Our approaches can be adapted to the multiple reading level setting. We experimented using the data tagging approach with the data equally-partitioned into five reading levels (Reading Level A-E), with A being the lowest and E the highest. The results are given in Table 8. We observe effective control over reading level at this finer level of granularity. Similar BLEU scores to that of the baseline indicate that different modes maintain translation quality.

6.6 Analysis and Discussion

We have demonstrated success both raising and lowering the reading level of test sets using two different methods. The results on multiple test sets and training corpora suggest that our methods are general and applicable beyond the scope

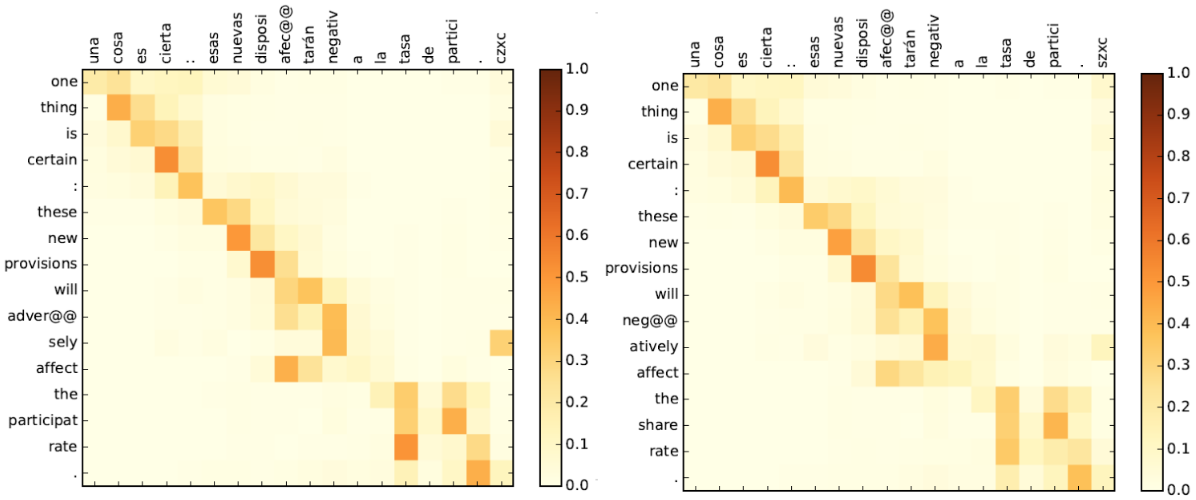

Figure 3: Attention visualization in simple vs. complex mode of data tagging approach (40-40 partition, trained on ParaCrawl).

Specified Readability	FKG	DC	BLEU
Baseline	9.36	7.92	27.38
Reading Level A	8.66	7.22	26.81
Reading Level B	9.01	7.67	27.14
Reading Level C	9.34	8.10	27.29
Reading Level D	9.61	8.49	27.12
Reading Level E	10.03	9.06	26.17

Table 8: Readability performance of the data tagging method at five levels of readability, trained on ParaCrawl and tested on newstest2013.

of the datasets we chose. Our qualitative examples demonstrate that though BLEU score depreciates, some of the decrease reflects correct changes towards our goal of adjusting reading level.

Translations in "simple" mode sometimes end early or are too short (in the double-decoder ParaCrawl 15-15 model, specifically). Simple training sentences tend to be shorter than complex training sentences, which may teach the simple decoder to produce short sentences.

We observe that the double-decoder is generally able to pull the mean readability of sentences translated in simple vs. complex mode farther apart than the data tagging approach. The separated decoders become more specialized towards creating sentences of particular relative readability levels, which may explain this observation.

We also observed the data tagging approach retaining higher BLEU than the double-decoder. We suspect this is because in the data tagging approach, we retain sentences of an intermediate complexity level during training, and this extra data helps maintain high BLEU. On the other hand, the double-decoder model with a 15-15 data partition receives ∼2.2 million simple sentences and ∼2.2 million complex sentences. This means that the encoder is trained on less than 30% of the data as the baseline, and each decoder is trained on approximately 15% of the data. This lower-data condition likely contributes to the lower BLEU score for double-decoder models, and explains why the data tagging approach does not suffer the same loss in BLEU. This also suggests that the data tagging approach may be preferable in low-resource settings. That said, human evaluators rated translations from the double-decoder approach higher than a baseline with similar BLEU performance.

7 Related Work

Our work is similar to style transfer and work controlling style during natural language generation (e.g., (Carlson et al., 2017; Fu et al., 2017; John et al., 2018; Ficler and Goldberg, 2017)), and to the text simplification literature (e.g., (Napoles and Dredze, 2010; Nisioi et al., 2017)). In style transfer, NMT methods using double-decoder architectures have been used, for instance, to output formal vs. informal or positive vs. negative versions of a source sentence (e.g., (Fu et al., 2017; Prabhumoye et al., 2018)). Sennrich et al. (2016) use tokens similar to our complexity tags in NMT to specify politeness in their English-German output. Vanmassenhove et al. (2018) and Kobus et al. (2016) retain gender information and domain information, respectively, in NMT through a tag to improve the translation quality.

As far as we are aware, we are the first authors to use NMT to both reduce and increase the complexity of translations. Unlike most of the text simplification literature, we simplify output cross-linguistically and also increase text complexity. In statistical machine translation, Stymne et al. (2013) translate and simplify output, while Niu et al. (2017) control formality in French–English translation. Štajner and Popović (2016) investigate how simplifying source-side sentences affects adequacy and fluency in English–Serbian translation. Interestingly, we notice qualitative similarities between our "complex" translations and the formal output of Niu et al. (2017), though the authors did not frame these qualitative differences as increases in complexity.

Prior work in machine translation and NLP has focused on readability assessment and text simplification. For readability assessment, a data-driven method is proposed in Le et al. (2018) for assessing the readability of document text, whereas Ciobanu et al. (2015) investigated the readability of the MT system output with standard metrics. Jones et al. (2005) also investigated the readability of MT and ASR system output but with human evaluation. As for text simplification, Hardmeier et al. (2013) proposes a document-level decoder for SMT and mentioned a case study that utilizes document-wide features to improve the readability of text. Contrary to Stymne et al. (2013), Xu et al. (2016) designed a new training objective for SMT text simplification. Similar to Le et al. (2018), Ciobanu et al. (2015), and 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 features such as type/token ratios and lexical consistency as input to the MT system. On the other hand, our approaches at most require an additional simplicity/complexity tag. Different from Xu et al. (2016) in which new training objective is proposed for text simplification, our NMT training objective remains the same.

8 Conclusion

In this work, we are the first authors to address the important task of controlling the reading level of machine translation output, and provide the first results. This work is important for practitioners who wish to control the simplicity or complexity of text that their machine translation system produces.

We develop two methods for controlling the reading level of output translations in NMT. Both of our proposed models successfully increase or decrease the reading level of multiple test sets when trained on different corpora, and have good qualitative results. Furthermore, our human evaluation indicates that the readability-level controlled translations are better than a baseline which had higher BLEU.

Notably, our data tagging approach can be deployed immediately on existing NMT systems with no architectural changes. We demonstrate a trade-off between more effective control of reading level and BLEU score, particularly with the double-decoder approach. As the data partition becomes more aggressive, the difference in reading level between the two modes increases, but BLEU score drops. We show that this effect can be mitigated by oversampling.

In the future, we plan to experiment with different language pairs and readability scorers. We also plan to discard sentences with very low readability scores and filter training corpora to exclude low-quality examples, which Junczys-Dowmunt (2018) demonstrated can severely degrade model performance. We expect these methods will help us retain better BLEU. Furthermore, we will use the state-of-the-art transformer model which we expect to provide improved BLEU and greater control over reading level in the data tagging method, because the complexity tag will contribute to each word's representation via self-attention (Vaswani et al., 2017).

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 observed that sentences trained in "complex" mode appeared more formal than those trained in "simple" mode; most contractions were removed, and words appeared more formal. We plan to repeat these experiments, and have observed promising first results.

Acknowledgements

The authors would like to thank Jason Eisner, Huda Khayrallah, Shuoyang Ding, Arya McCarthy, and Rebecca Knowles for their insightful comments and advice. We also thank our anonymous reviewers for their helpful comments.

References

Amstad, Toni. 1978. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service.

Björnsson, Carl-Hugo. 1968. *Läsbarhet: hur skall man som författare nå fram till läsarna?* Bokförlaget Liber.

Carlson, Keith, Allen Riddell, and Daniel Rockmore. 2017. Zero-shot style transfer in text using recurrent neural networks. *arXiv preprint arXiv:1711.04731*.

Chall, Jeanne Sternlicht and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Ciobanu, Alina Maria, Liviu P Dinu, and Flaviu Pepelea. 2015. Readability assessment of translated texts. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 97–103.

Ficler, Jessica and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. *arXiv preprint arXiv:1707.02633*.

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

François, Thomas and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57. Association for Computational Linguistics.

Fu, Zhenxin, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. *arXiv preprint arXiv:1711.06861*.

Hardmeier, Christian, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 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*, pages 193–198. Association for Computational Linguistics.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

John, Vineet, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python.

Jones, Douglas, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 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*, volume 5, pages v–1009. IEEE.

Junczys-Dowmunt, Marcin. 2018. Microsoft's submission to the WMT2018 news translation task: How I learned to stop worrying and love the data. *CoRR*, abs/1809.00196.

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

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Kobus, Catherine, Josep Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180. Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Le, Dieu-Thu, Cam-Tu Nguyen, and Xiaoliang Wang. 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*, pages 103–107.

Lison, Pierre and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Lucisano, Pietro and Maria Emanuela Piemontese. 1988. Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 3(31):110–124.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Napoles, Courtney and Mark Dredze. 2010. Learning simple wikipedia: A cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 42–50. Association for Computational Linguistics.

Nisioi, Sergiu, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 85–91.

Niu, Xing, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Prabhumoye, Shrimai, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 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*, pages 35–40.

Štajner, Sanja and Maja Popović. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Stymne, Sara, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 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*, number 085, pages 375–386. Linköping University Electronic Press.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xu, Wei, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.