Growing Together: Modeling Human Language Learning With n-Best Multi-Checkpoint Machine Translation

El Moatez Billah Nagoudi[1], **Muhammad Abdul-Mageed**[1], **Hasan Cavusoglu** [2]
[1] Natural Language Processing Lab
[2] Sauder School of Business
[1,2] The University of British Columbia
[1] {muhammad.mageed,moatez.nagoudi}@ubc.ca, [2] cavusoglu@sauder.ubc.ca

Abstract

We describe our submission to the 2020 Duolingo Shared Task on Simultaneous Translation And Paraphrase for Language Education (STAPLE) (Mayhew et al., 2020). We view MT models at various training stages (i.e., checkpoints) as human learners at different levels. Hence, we employ an ensemble of multi-checkpoints from the same model to generate translation sequences with various levels of fluency. From each checkpoint, for our best model, we sample n-Best sequences ($n = 10$) with a beam width $= 100$. We achieve 37.57 *macro* F_1 with a 6 checkpoint model ensemble on the official English to Portuguese shared task test data, outperforming a baseline Amazon translation system of 21.30 *macro* F_1 and ultimately demonstrating the utility of our intuitive method.

1 Introduction

Machine Translation (MT) systems are usually trained to output a single translation. However, many possible translations of a given input text can be acceptable. This situation is common in online language learning applications such as *Duolingo*,[1] *Babbel*[2], and *Busuu*.[3] In applications of this type, learning happens via translation-based activities while evaluation is performed by comparing learners' responses to a large set of human acceptable translations. Figure 1 shows an example of a typical situation extracted from the Duolingo application.

The main set up of the 2020 Duolingo Shared Task on Simultaneous Translation And Paraphrase for Language Education (STAPLE 2020) (Mayhew et al., 2020) is such that one starts with a set of English sentences (prompts) and then generates high-coverage sets of plausible translations in the five target languages: Portuguese, Hungarian, Japanese, Korean, and Vietnamese. For instance, if we want to translate the English (En) sentence *"is my explanation clear?"* to Portuguese (Pt), all the translated Portuguese sentences illustrated in Table 1 would be acceptable.[4]

Figure 1: Translations proposed by English language learners at various levels of fluency, from diverse backgrounds. Our multi-checkpoint ensemble models mimic learner fluency.[4]

Limited training data. One challenge for training a sufficiently effective model we faced is the limited size of the source training data released by organizers (4,000 source English sentences coupled with 226,466 Portuguese target sentences). We circumvent this limitation by training a model on a large dataset acquired from the OPUS corpus (as described in Section 3), which gives us a powerful MT system that we build on (see Section 4.2). We then exploit the STAPLE-provided training data in multiple ways (see Sections 4.3 and 4.4) to extend this primary model as a way to nuance the model to the shared task domain.

Paraphrase via MT. In essence, the shared task is a mixture of MT and paraphrase. This poses a second challenge: there is no paraphrase dataset

[1] https://www.duolingo.com/
[2] https://www.babbel.com/
[3] https://www.busuu.com/
[4] Examples taken from shared task description at: https://sharedtask.duolingo.com/.

to train the system on. For this reason, we resort to using outputs from the MT system in place of paraphrases. This required generating multiple sentences for each source sentence. To meet this need, we generate multiple translation hypotheses (n-Best) using a wide beam search (Section 5.1), perform 'round-trip' translations exploiting these multiple outputs (Section 5.2), and employ ensembles of checkpoints (Section 5.3).

Diverse outputs. A third challenge is that the target Portuguese sentences provided for training by organizers are produced by learners of English at various levels of fluency. This makes some of these Portuguese translations inarticulate (i.e., not quite fluent). MT systems are not usually trained to produce inarticulate translations (part of the time), and hence we needed to offer a solution that matches the different levels of language learners who produced the translations. Intuitively, we view MT systems trained at various stages (i.e., checkpoint) as learners with various levels of fluency. As such, we employ an ensemble of checkpoints to generate translations matching the different levels of learner fluency (Section 5.3). Ultimately, ***our contributions lie in alleviating the 3 challenges listed above***.

The remainder of the paper is organized as follows: Section 2 is a brief overview of related work. In Section 3, we describe the data we use for both training and fine-tuning our models. Section 4 presents the proposed MT system. Section 5 describes our different methods. We discuss our results in Section 6, and conclude in Section 7.

2 Related Work

We focus our related work overview on the task of paraphrase generation and its intersection with machine translation. Paraphrasing is the task of expressing the same textual units (e.g. sentence) with alternative forms using different words while keeping the original meaning intact. [5] Over the last few years, MT has been the dominant approach for paraphrase generation. For instance, Barzilay and McKeown (2001); Pang et al. (2003) use multiple translations of the same text to train a paraphrase system. Similarly, Bannard and Callison-Burch (2005) use an MT phrase table to mapping an English sentences to various non-English sentences.

[5] https://dictionary.cambridge.org/dictionary/english/paraphrase

English sentence	is my explanation clear?
Accepted Portuguese Translations	- minha explicação está clara? - minha explicação é clara? - a minha explicação é clara? - está clara minha explicação? - minha explanação está clara? - é clara minha explicação?
English sentence	you look so pretty!
Accepted Portuguese Translations	- você está tão linda! - você está tão bonita! - você está muito linda! - você está muito bonita! - você parece tão linda! - você parece tão bonita!

Table 1: English sentences with their Portuguese translation samples from shared task training split.

More recently, advances in neural machine translation (NMT) have spurred interest in paraphrase generation (Sutskever et al., 2014; Luong and Manning, 2015; Aharoni et al., 2019). For example, Prakash et al. (2016) employ a stacked residual LSTM network to learn a sequence-to-sequence model on paraphrase data. A parpahrase model with adversarial training is presented by (Li et al., 2017). Wieting and Gimpel (2017); Iyyer et al. (2018) propose a translation-based paraphrasing system, which is based on NTM to translate one side of a parallel corpus. Paraphrase generation with pivot NMT is used by (Mallinson et al., 2017; Yu et al., 2018).

3 Data

3.1 Shared task data

As part of the STAPLE 2020 shared task, only `training` data were released. The target training split is a total of $526,466$ of learner translations of $4,000$ input (source) English sentences. We note that the number of translations of each English sentence varies, with an average of $\sim$ 132 Portuguese target sentences for each English source sentence. As shared task organizers point out, this training dataset can be used as a reference/anchor points, and also serves as a strong baseline. For `evaluation`, a sets of $60,294$ translations (learner-crafted sentences) of 500 input English sentences were available on Colab. `Test` data were also made available only via Colab and comprised 500 English sentences learner-translated

Corpus	Content	Documents	Sentences	En. Words	Pt. Words
ParaCrawl v5	Parallel corpora from Web Crawls collected in the ParaCrawl project	287	13.9M	341.4M	347.9M
TildeMODEL v2018	This is the Tilde MODEL Corpus – Multilingual Open Data for European Languages	6	3.6M	134.1M	100.4M
DGT	A collection of translation memories provided by the JRC	287	13.9M	341.4M	347.9M
SciELO	Parallel corpus of full-text articles in Portuguese, English and Spanish from SciELO	2	3.1M	92.8M	95.4M
OpenSubtitles	A new collection of translated movie subtitles	42,755	35.5M	283.4M	248.9M
Tanzil	A collection of Quran translations	15	0.1M	2.8M	2.4M
News Commentary	A parallel corpus of News Commentaries provided by WMT	7,185	0.6M	15.4M	15.5M
Europarl v8	A parallel corpus extracted from the European Parliament web site	10,344	2.0M	59.5M	6.1M
JW300 v1	JW300 is a parallel corpus of over 300 languages	26,991	2.2M	40.0M	40.8M
CAPES v1	Parallel corpus of theses and dissertation abstracts in Portuguese and English from CAPES	1	1.2M	38.4M	39.1M
EMEA v3	A parallel corpus from the European Medicines Agency	1,921	1.1M	12.0M	16.4M
QED v2.0a	Open multilingual collection of subtitles for educational videos and lectures	4,618	0.5M	8.7M	7.4M
JRC-Acquis v3.0	JRC-Acquis is a collection of legislative text of the European Union	20,507	1.7M	64.3M	64.8M
Wikipedia	A corpus of parallel sentences from Wikipedia	5	1.8M	47.0M	44.8M
TED2013	A parallel corpus of TED talk subtitles by CASMACAT	1	0.2M	3.1M	2.9M
GNOME.	A parallel corpus of GNOME localization files	1,307	0.6M	2.6M	3.7M
Tatoeba	A collection of translated sentences from Tatoeba	1	0.2M	11.0M	2.7M
ECB v1	Website and documentatuion from the European Central Bank	1	0.2M	5.8M	6.2M
bible-uedin v1	Multilingual parallel corpus created from translations of the Bible	2	62.2K	1.8M	1.7M
GlobalVoices	A parallel corpus of news from the Global Voices website	5,133	71.5k	2.3M	2.3M
KDE4	A parallel corpus of KDE4 system messages	2,136	0.2M	2.4M	2.7M
Ubuntu	A parallel corpus of the Ubuntu Dialogue Corpus	449	0.1M	0.7M	0.5M
EUconst v1	A parallel corpus collected from the European Constitution	47	10.9K	0.2M	0.2M
Books v1	A collection of copyright free book	1	1.4K	33.8K	32.3K
Total	All corpora extracted from OPUS	162,425	77.7M	1.5B	1.4B

Table 2: English-Portuguese datasets from Tiedemann (2012) used in our training.

into $67,865$ Portuguese sentences. For all training, development, and test data, these translations are ranked and weighted according to actual learner response frequency. We refer the reader to the shared task description for more information. [6]

3.2 OPUS data

In order to develop efficient English-Portuguese MT models that can possibly work across different text domains, we make use of a large dataset of parallel English-Portuguese sentences extracted from the Open Parallel Corpus Project (OPUS) (Tiedemann, 2012). OPUS[7] contains more than 2.7 billion parallel sentences in 90 languages. The specific corpus we extracted consists of data from multiple domains and sources including: ParaCrawl project (Esplà-Gomis et al., 2019), EUbookshop (Skadiņš et al., 2014), Tilde Model (Rozis and Skadinš, 2017), translation memories (DGT) (Steinberger et al., 2013), OpenSubtitles (Creutz, 2018), SciELO Parallel (Soares et al., 2018), JRC-Acquis Multilingual (Steinberger et al., 2006), Tanzil (Zarrabi-Zadeh, 2007), Europarl Parallel (Koehn, 2005), TED 2013 (Cettolo et al., 2012), Wikipedia (Wołk and Marasek, 2014), Tatoeba [8], QCRI Educational Domain (Abdelali et al., 2014), GNOME localization files, [9] Global Voices, [10] KDE4, [11], Ubuntu, [12] and Multilingual Bible (Christodouloupoulos and Steedman, 2015). To train our models, we extract more than $77.7M$ parallel (i.e., English-Portuguese) sentences from the whole collection. The extracted dataset comprises more than $1.5B$ English tokens and $1.4B$ Portuguese tokens. More details about the training dataset are given in Table 2.

3.3 Pre-Processing

Pre-processing is an important step in building any MT model as it can significantly affect the end results. We remove punctuation and tokenize all data with the Moses tokenizer (Koehn et al., 2007). We also use joint Byte-Pair Encoding (BPE) with 60K split operations for subword segmentation (Sennrich et al., 2016).

[6] https://sharedtask.duolingo.com/#data.
[7] http://opus.nlpl.eu/
[8] www.tatoeba.org
[9] www.10n.gnome.org
[10] www.globalvoices.org/
[11] www.i18n.kde.org
[12] www.translations.launchpad.net

4 Models

In this section, we first describe the architecture of our models. We then explain the different ways we train the models on various subsets of the data.

4.1 Architecture

Our models are mainly based on a Convolutional Neural Network (CNN) architecture (Kim, 2014; Gehring et al., 2017). This convolutional architecture exploits BPE (Sennrich et al., 2016). The architecture is as follows: 20 layers in the encoder and 20 layers in the decoder, a multiplicative attention (Luong et al., 2015) in every decoder layer, a kernel width of 3 for both the encoder and the decoder, a hidden size 512, and an embedding size of 512, and 256 for the encoder and decoder layers respectively. We use a Fairseq implementation (Ott et al., 2019).

4.2 Basic En↔Pt Models

We trained two MT models, English-to-Portuguese (En→Pt) and Portuguese-to-English (Pt→En), on 4 V100 GPUs, following the setup described in Ott et al. (2018). For both models, the learning rate was set to 0.25, a dropout of 0.2, and a maximum tokens of $4,000$ for each mini-batch. We train our models on the $77.7M$ parallel sentences of the OPUS dataset described in Section 3. Validation is performed on the development data from STAPLE 2020 (Mayhew et al., 2020).

4.3 En→Pt Extended Model

We use the training data of the STAPLE 2020 shared task[13] to create a new En-Pt parallel dataset. More specifically, at the *target* side, we use all the Portuguese gold translations while duplicating the same English source sentence at the *source* side. This results in a new training set of $251,442$ En-Pt parallel sentences. We refer to this training dataset as STAPLE-TRAIN, or simply *S-TRAIN*. We then merge OPUS and S-TRAIN to train an En→Pt model from scratch. We refer to this new model as the *extended model*.

4.4 En→Pt Fine-Tuned Model

Fine-tuning with domain-specific data, from a domain of interest, can be an effective strategy when it is desirable to develop systems for such a domain (Ott et al., 2019, 2018). Motivated by this, we experiment with using the STAPLE-based S-TRAIN parallel dataset from the previous subsection to fine-tune our En→Pt *basic* model for 5 epochs. [14] We will refer to the model resulting from this fine-tuning process simply as the *fine-tuned model*.

5 Model Deployment Methods

In order to enhance the 1-to-n En-Pt translation, we propose three methods based on the previously discussed MT models (see section 4). These methods are n*-Best prediction*, *multi-checkpoint translation*, and *paraphrasing*.

5.1 *n*-Best Prediction

We first use our three MT models (*basic*, *extended*, and *fine-tuned*) with a beam search size of 100 to generate n-Best translation hypotheses. We then use the average log-likelihood to score each of these hypotheses. Finally, we select the hypothesis with the n highest score as our output.

5.2 Paraphrasing

Paraphrasing is an effective data augmentation method which is commonly used in MT tasks (Poliak et al., 2018; Iyyer et al., 2018). In order to extend the list of accepted Portuguese translations, we use both of our En→Pt and Pt→En models, as follows:

1. Translate the English sentences using the En→Pt model. For instance, we generate n-Best ($n = 10$) Portuguese sentences for each English source sentence.

2. Then, we use the Pt→En model to get n'-Best English translations (we experiment with $n' = 1$, 3, and 5) for each of the 10 Portuguese sentence. At this point, we would have $10 * n'$ new English sentences (oftentimes with duplicate generations that we remove). These new sentences represent paraphrases of the original English sentence.

3. After de-duplication, the new English sentences are fed to the En→Pt model to get the 1-Best Portuguese translation.

[13] http://sharedtask.duolingo.com/#data

[14] We choose the number of epochs arbitrarily, but note that it is a hyper-parameter that can be tuned.

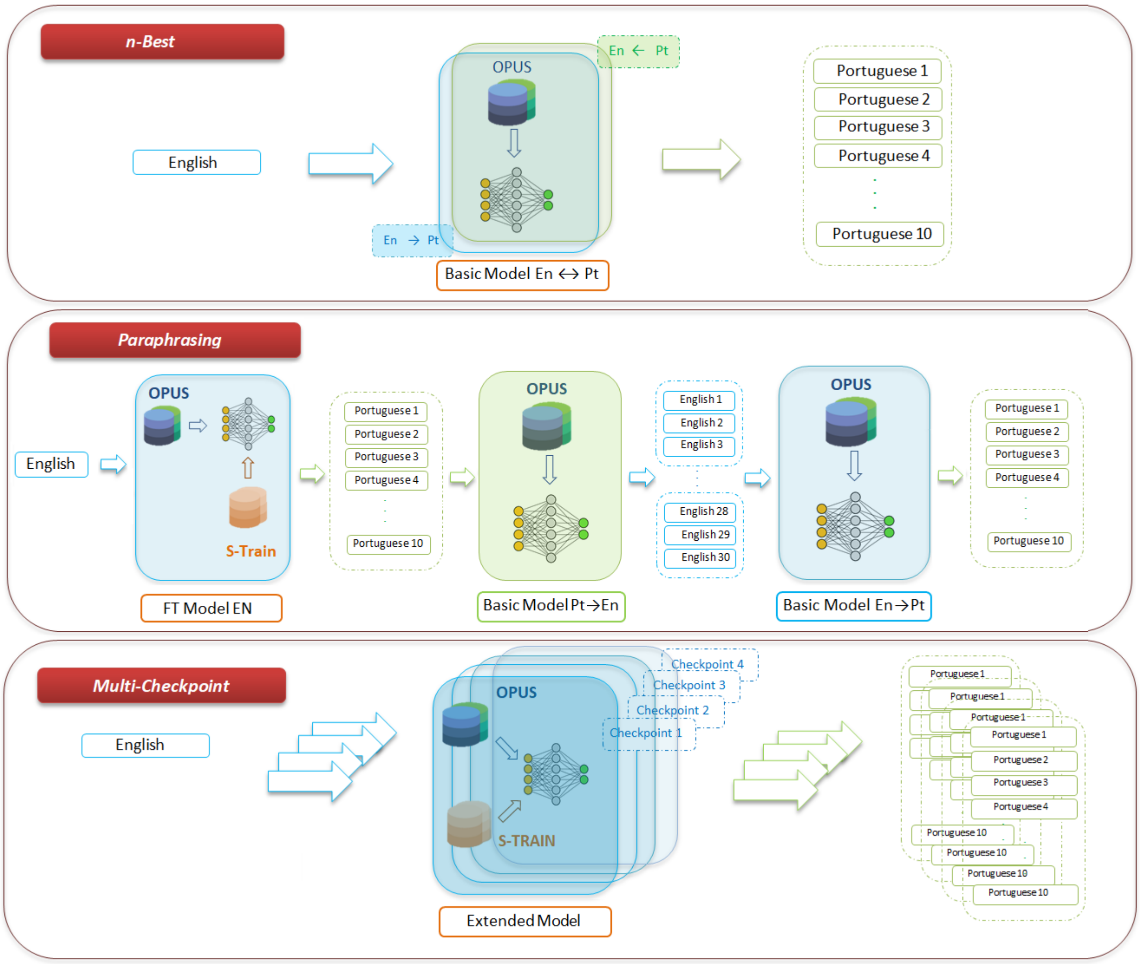

Figure 2: An illustration of our proposed models and methods: **(a)** ***n*-Best prediction method** with $n = 10$ resulting in the En→Pt *basic* model; **(b) paraphrasing method** with $n = 10$ and $n' = 3$ used in the En→Pt *fine-tuning* and the En↔Pt basic models, **(c) multi-checkpoint method** used with $n = 10$ and $m = 4$ for the En→Pt *extended* model.

5.3 Multi-Checkpoint Translation

Our third method is based on saving the models at given epochs (checkpoints) during training. We use the m last checkpoints (models) to generate the n-Best translation hypotheses (the same way as our n-Best prediction method). We then de-duplicate the outputs of all the m models and use them in evaluation. We now describe our evaluation.

6 Evaluation

In order to evaluate our methods, we carry out a number of experiments. First, we consider performance of each proposed method on the official `training` and `development` datasets of STAPLE (Mayhew et al., 2020). Our models were ultimately evaluated on the shared task `test` data. We now describe STAPLE evaluation metrics and baselines as provided by organizers, before reporting on our results on training, development, and test.

6.1 Evaluation Metrics & Baselines

Weights of Translation. We note that each Portuguese translated sentence has a weight as provided in the gold dataset. The weights of translations correspond to user (learner) response rates. These weights are used primarily for scoring. The STAPLE 2020 shared task data takes the format illustrated in Table 3.

Metrics. Performance of MT systems in the shared task is quantified and scored based on how well a model can return all human-curated acceptable translations, weighted by the likelihood that an English learner would respond with each translation (Mayhew et al., 2020). As such, the main scoring metric is the *weighted macro* F_1, with respect to the accepted translations. To

English Sentence : is my explanation clear?	
Weights	**Portuguese Translation**
0.26739	- minha explicação está clara?
0.16168	- minha explicação é clara?
0.11109	- a minha explicação é clara?
0.08778	- está clara minha explicação?
0.05717	- minha explanação está clara?
English Sentence : this is my fault.	
Weights	**Portuguese translation**
0.17991	- isto é minha culpa.
0.10664	- isso é minha culpa.
0.08944	- esta é minha culpa.
0.07794	- isto é culpa minha.
0.06803	- é minha culpa.

Table 3: English sentences with their Portuguese translation and Weights samples from shared task train data.

compute *weighted macro* F_1 (see formula 6), the weighted F_1 for each English sentence (s) is calculated and the average over all the sentences in the corpus is computed. The weighted F_1 (see formula 5) is computed using the unweighted precision (see formula 1) and the weighted recall (see formulas 2, 3 and 4).

$$Precision\ (s) = \frac{TP_s}{TP_s + FN_s} \quad (1)$$

$$WTP_s = \sum_{s \in TP_s} weight(t) \quad (2)$$

$$WFN_s = \sum_{s \in FN_s} weight(t) \quad (3)$$

$$Weighted\ Recall\ (s) = \frac{WTP_s}{WTP_s + WFN_s} \quad (4)$$

$$Weighted\ F1(s) = \frac{2 \cdot Prec.\ (s) \cdot W.\ Recall\ (s)}{Prec.\ (s) + W.\ Recall\ (s)} \quad (5)$$

$$Weighted\ Macro\ F_1 = \sum_{s \in S} \frac{Weighted\ F1(s)}{|S|} \quad (6)$$

Baselines. We adopt the two baselines offered by the task organizers. These are based on Amazon and Fairseq translation systems and are at 21.30% and 13.57%, respectively. More information about these baselines can be reviewed at the shared task site listed earlier.

6.2 Evaluation on TRAIN and DEV

In this section, we report the results of our 3 proposed methods, *(a) n-Best prediction*, *(b) paraphrasing*, and *(c) multi-checkpoint translation* using the MT models presented in section 4.

Evaluation on TRAIN. For (a) the n-Best prediction method, we explore the 4 different values of n in the set $\{5, 10, 15, 20\}$. For (b) the paraphrase method, we set the number of Portuguese sentences to $n' = \{1, 3, 5\}$. Finally, (c) the multi-checkpoint method was tested with 4 different values for the number of checkpoints $m = \{2, 4, 6, 8\}$.
For paraphrasing and multi-checkpoint translation, we fix the number of n-best translations n to 10, varying the values of n' and m only when evaluating our *extended* model. This leads us to identifying the best evaluation values of $n' = 3$ and $m = 6$, which we then use when evaluating our *basic* and *fine-tuned* models.

Evaluation on DEV. For evaluation on the STAPLE development data, we adopt the same procedure followed for evaluation on the train split. Table 4 summarizes our experiments with different configurations (i.e., values of n, n', and m) on train and development task data, respectively.

Discussion. Results presented in Table 4 demonstrate that all the models with the different methods and configurations outperform the the official shared task baseline with $macro\ F_1$ scores between 27.41% and 40.78%. As expected, fine-tuning the En→Pt basic model with the S-TRAIN data-set improves the results with a mean of $+1.46\%$ on the training data. We also observe that training on the concatenated OPUS and S-TRAIN data-sets *from scratch* leads to better results compared to the *exclusive fine-tuning* method.

Based on these results, we can see that the best configuration is the multi-checkpoint method used with the *extended* MT model. This configuration obtains the best $macro\ F_1$ score of 40.78% and 39.21% on the training and development STAPLE data splits, respectively.

		Train Data								
		Basic Model			**Extended Model**			**Fine-Tuned Model**		
Method	**n**	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1
***n*-Best Prediction**	5	55.44	23.87	27.41	45.51	28.24	29.43	44.68	26.91	28.38
	10	42.78	29.65	28.47	46.02	34.18	**33.51**	41.81	32.19	30.33
	15	37.42	27.09	29.17	39.25	35.50	31.80	45.51	28.24	29.43
	20	29.68	38.24	27.48	39.22	35.49	31.79	46.23	27.04	30.27
	n′	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1
Paraphrasing	1	-	-	-	40.24	35.01	31.68	-	-	-
	3	40.24	35.01	31.68	46.45	35.08	**34.39**	39.98	37.27	32.89
	5	-	-	-	40.44	39.20	34.16	-	-	-
	m	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1
Multi Checkpoint	2	-	-	-	58.81	31.57	36.57	-	-	-
	4	-	-	-	50.53	44.22	40.76	-	-	-
	6	44.44	45.52	39.46	49.58	44.92	**40.78**	36.77	52.73	38.54
	8	-	-	-	42.16	44.96	39.28	-	-	-
		DEV Data								
		Basic Model			**Extended Model**			**Fine-Tuned Model**		
Method	**n**	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1
***n*-Best Prediction**	5	-	-	-	52.48	26.27	29.87	-	-	-
	10	32.56	36.83	29.33	36.52	41.09	**32.96**	35.39	37.84	31.30
	15	-	-	-	38.62	37.46	32.36	-	-	-
	20	-	-	-	36.03	37.44	31.33	-	-	-
	n’	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1
Paraphrasing	1	-	-	-	45.77	33.31	33.05	-	-	-
	3	48.66	31.17	32.43	46.34	33.85	33.17	39.98	37.27	32.89
	5	-	-		46.14	34.26	**33.40**	-	-	-
	m	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1	Prec.	W. Recall	W. F_1
Multi Checkpoint	2	-	-	-	55.88	32.16	35.26	-	-	-
	4	-	-	-	52.27	37.35	38.25	-	-	-
	6	45.35	43.20	38.04	56.42	37.31	39.16	45.01	41.23	37.26
	8	-	-	-	53.83	38.85	**39.21**	-	-	-

Table 4: Performance on the STAPLE 2020 Train and Dev data splits.

		Extended Model		
Method	**m**	Prec.	W. Recall	W. F_1
Aws Baseline	-	87.80	13.98	21.29
Fairseq Baseline	-	28.25	11.70	13.57
Multi-Checkpoint	4	60.14	33.14	37.06
	6	53.83	36.50	**37.57**
	8	49.94	38.27	37.21

Table 5: Results on STAPLE 2020 Test Data.

6.3 Evaluation on TEST

In test phase, we submitted translations from 3 systems for the STAPLE English-Portuguese sub-task. The 3 systems are based on our *multi-checkpoint translation* with the *extended* model. The number of checkpoints used was $m = \{4,\ 6,\ 8\}$, and n is fixed to 10 (i.e., the best value of n identified on training data with our *extended* model). Table 5 shows the results of our 3 final submitted systems as returned by the shared task organizers.

Discussion. Our results indicate that when the multi-checkpoint method with the extended model and only two last checkpoints ($m = 4$) is used, the $macro\ F_1$ score reaches 37.07% (with a best precision of 60.14%). This method with $m = 6$ represents our best $macro\ F_1$ score 37.57% for the English-Portuguese translation sub-task. We note that with this configuration we outperform the Amazon and Fairseq translation baseline systems (at $+15.92\%$ and $+23.99\%$, respectively) provided by the task organizers. We also observe that when m is set to 8, the $macro\ F_1$ slightly decreases to 37.21%. Ultimately, our findings show the utility of using multiple checkpoint ensembles as a way to mimic the various levels of language learners. Simple as this approach is, we find it quite intuitive.

7 Conclusion

In this work, we described our contribution to the 2020 Duolingo Shared Task on Simultaneous Translation And Paraphrase for Language Education (STAPLE) (Mayhew et al., 2020). Our system targeted the English-Portuguese sub-task. Our models effectively make use of an approach based on n-Best prediction and multi-checkpoint translation. Our use of the OPUS dataset for training proved quite successful. In addition, based on our results, our intuitive deployment of a multi-checkpoint ensemble coupled with n-Best decoded translations seem to mirror leaner proficiency. As future work, we plan to explore other methods on new language pairs.

Acknowledgements

MAM gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences Research Council of Canada (SSHRC), and Compute Canada (www.computecanada.ca).

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The amara corpus: Building parallel language resources for the educational domain. In *LREC*, volume 14, pages 1044–1054.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 50–57.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. *arXiv preprint arXiv:1809.06142*.

Miquel Esplà-Gomis, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. Paracrawl: Web-scale parallel corpora for the languages of the eu. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

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

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, et al. 2007. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In *Final Report of the Johns Hopkins 2006 Summer Workshop*.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2017. Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279*.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

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

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893.

Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. 2020. Simultaneous translation and paraphrase for language education. In *Proceedings of the ACL Workshop on Neural Generation and Translation (WNGT)*. ACL.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 102–109. Association for Computational Linguistics.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018. On the evaluation of semantic phenomena in neural machine translation using natural language inference. *arXiv preprint arXiv:1804.09779*.

Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual lstm networks. *arXiv preprint arXiv:1610.03098*.

Roberts Rozis and Raivis Skadinš. 2017. Tilde model-multilingual open data for eu languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, 131, pages 263–265. Linköping University Electronic Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of LREC*.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A large parallel corpus of full-text scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2013. Dgt-tm: A freely available translation memory in 22 languages. *arXiv preprint arXiv:1309.5226*.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. 2012:2214–2218.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Krzysztof Wołk and Krzysztof Marasek. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. *Procedia Technology*, 18:126–132.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Hamid Zarrabi-Zadeh. 2007. Tanzil project. *URL: http://tanzil. net/wiki/Tanzil_Project*.