Entity-aware Multi-task Training Helps Rare Word Machine Translation

Matīss Rikters[1]
[1]Artificial Intelligence
Research Center (AIRC)
National Institute of Advanced
Industrial Science and Technology
matiss.rikters@aist.go.jp

Makoto Miwa[1,2]
[2]Toyota Technological
Institute, Japan
makoto-miwa@toyota-ti.ac.jp

Abstract

Named entities (NE) are integral for preserving context and conveying accurate information in the machine translation (MT) task. Challenges often lie in handling NE diversity, ambiguity, rarity, and ensuring alignment and consistency. In this paper, we explore the effect of NE-aware model fine-tuning to improve the handling of NEs in MT. We generate data for NE recognition (NER) and NE-aware MT using common NER tools from Spacy and align entities in parallel data. Experiments with fine-tuning variations of pre-trained T5 models on NE-related generation tasks between English and German show promising results with increasing amounts of NEs in the output and BLEU score improvements compared to the non-tuned baselines.

1 Introduction

Machine translation (MT) of named entities (NEs) such as person or place names remains a significant challenge even for modern modelling architectures simply because they appear less frequently in training data than other words or phrases. Furthermore, new and unseen NEs get created every day like organization or product names, and even common nouns in certain contexts can become NEs. Meanwhile, the task of NE recognition (NER) has reached a fairly acceptable level for many languages with precision values of around 80–90%. Since most conventional MT models are trained to perform translation based only on the parallel training data and context provided, they still often struggle with rare NEs appearing less often during training or never at all. In such cases, the models tend to hallucinate by generating output comprised of tokens or subword units which are statistically close in the embedding space to the rare NE, but this can lead to the generation of a novel word or phrase instead of the proper acceptable translation.

In this work, we look into improving how the model handles NEs by highlighting them in the training data and training not only to translate but also to recognize NEs in plain input text. The motivation for this approach is for the model to form a more defined understanding of what certain NEs look like thus enabling it to handle them better when performing the MT task. We experiment with multi-task training and fine-tuning the T5 model (Raffel et al., 2020) for translation between English and German, as well as its multilingual counterpart mT5 (Xue et al., 2021) and the updated 1.1 version of T5. We compare the results with the non-modified versions of T5, mT5, and the instruction-tuned Flan-T5 (Chung et al., 2022).

Our contributions are 1) a novel, easily reproducible and further extensible method for fine-tuning transformer models in a multi-task fashion on named entity recognition and machine translation tasks; 2) empirical evaluation of the method on a recent shard task benchmark data set; 3) open-sourcing of data preparation and training scripts, and model checkpoints for reproducibility.

2 Related Work

T5 Fine-tuning Etemad et al. (2021) tune the model on abstractive summarisation using specific datasets. While the pre-trained model had already been exposed to this task, such fine-tuning led the authors to state-of-the-art results on several benchmarks. Zhuang et al. (2023) propose RankT5 to expand the capabilities of the T5 model into the text ranking task. They introduce ranking-specific losses for the task, significantly improving performance on select benchmarks. Tavan and Najafi (2022) participate in a SemEval shared task [1] on multilingual complex NER using the encoder from T5 for feature representation extraction.

[1] SemEval-2022: https://semeval.github.io/SemEval2022

Proceedings of the 17th International Natural Language Generation Conference, pages 47–54
September 23–27, 2024. ©2024 Association for Computational Linguistics

NE Translation Ugawa et al. (2018) encode NE tags alongside tokens and concatenate their embeddings. Modrzejewski et al. (2020) explore several methods for incorporating NE annotations into MT to improve NE translation. Their experiments with English-German and English-Chinese MT on WMT 2019 test sets demonstrate improvements over the baseline transformer models when using fine-grained NE annotations as input factors for MT training. Zeng et al. (2023) use a dictionary to look up translation candidates and prepend them to the decoder input. Hu et al. (2022) augment pre-training data with NEs replaced in the target language, pre-train the model to reconstruct such data to the original sentences and perform multi-task fine-tuning of the model on both the reconstruction task and MT. In contrast to related work, we aim to perform multi-task training on the monolingual NER tasks and the multilingual MT tasks.

3 Proposed Approach

Since the existing pre-trained T5 model versions have already been pre-trained on large multilingual corpora, the quality of the data used for fine-tuning on the resource-rich languages plays a more significant role than the quantity (de Gibert Bonet et al., 2022). We start with filtering out any critical noisy data from the WMT23[2] general translation shared task training set before tagging named entities in the form of XML boundary tags. Next, we prepend instructions to the source side of the training data as shown in Table 2 to indicate what we expect from the model in the output. Parallel data for the MT task have the source side enriched with NE tags where applicable, and the instruction for NE-MT at the beginning, while the target side remains as is. For the NER task, we have the NER instruction at the beginning followed by the text as is on the source side, and the text enriched with NE tags on the target side.

3.1 Training Setup

We combine and shuffle all training data for the tasks, and experiment with different quantities of data provided to the model during training in combination with the different model sizes. We tune the *small* size models using 100K examples, *base* with 1M, and *large* with 10M respectively. We base this choice on observations from preliminary experiments where small models often converged before reaching 1M examples and base models converged before seeing 10M. We apply this to the different T5 model variations (T5, T5 1.1, mT5, Flan-T5) with parameter ranges between around 60M to around 1B. We use the Adafactor optimizer with FP16 training, effective batch sizes of 256 for *large* models and 512 for *base* and *small* sized models, evaluation every 1000 steps, and early stopping set to 10 checkpoints of evaluation loss not improving.

4 Data Preparation

We use the English-German parallel data from the WMT 2023 shared task on general text translation for experimentation. To develop our models, we use the general test set from WMT22 and for evaluation and result reporting – general test set of WMT23. We first filter the data by removing noisy parallel segments. Then we populate the data with NE tags in either the source or target side, depending on the task. Finally, we prepend task-specific instructions to all source-side inputs. For the NER task training data, we use both source and target MT parallel sentences, essentially doubling the amount when compared to MT task data.

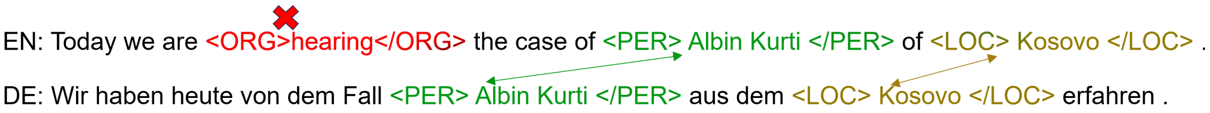
4.1 Dataset and Filtering

Since most training corpora are produced semi-automatically, errors such as misalignments between source and target sentences or direct copies of source to target can occur, as well as third-language data in seemingly bilingual data sets. To avoid such problems, we used data cleaning and pre-processing methods (Rikters, 2018) that include: 1) a unique parallel sentence filter; 2) equal source-target filter; 3) multiple sources - one target and multiple targets - one source filters; 4) non-alphabetical filters; 5) repeating token filter; and 6) correct language filter. We also perform pre-processing consisting of the standard Moses (Koehn et al., 2007) scripts for punctuation normalisation and cleaning. However, there is no separate tokenisation or splitting into subword units besides the tokeniser included with the model.

4.2 NE Tagging and Alignment

We use Spacy (Honnibal et al., 2020) to introduce NE tags for the source side of MT task training data and the target side of NER task data. Spacy was chosen mainly for its good balance of tagging accuracy, speed, and ease of use. As an additional quality assurance mechanism, we also tag the target side of MT data and keep only the NE tags that are

[2] WMT 2023 - http://www2.statmt.org/wmt23/



EN: Today we are <ORG>hearing</ORG> the case of <PER> Albin Kurti </PER> of <LOC> Kosovo </LOC> .

DE: Wir haben heute von dem Fall <PER> Albin Kurti </PER> aus dem <LOC> Kosovo </LOC> erfahren .

Figure 1: An example of alignment and misalignment between English and German entities. The NER model recognized “hearing” as an organisation entity for English, but there was no matching NE recognized for German, so this tag was dropped in the alignment process, while the person and location tags aligned correctly and were kept.

	German	English
	LOC	LOC
	LOC	GPE
	MISC	-
	ORG	ORG
	PER	PERSON
Pr	0.85	0.90
Re	0.84	0.90
F_1	0.85	0.90
	English Only	
CARDINAL	DATE	EVENT
FAC	LANGUAGE	LAW
MONEY	NORP	ORDINAL
PERCENT	PRODUCT	QUANTITY
TIME	WORK_OF_ART	

Table 1: Entity alignment dictionary, and Spacy NER evaluation metrics - precision (Pr), recall (Re) and F_1. The bottom rows list NE types which are not available for German in Spacy.

symmetric between the two languages, as shown in Figure 1. The available classes of NEs to be recognized by NER tools depend highly on the language in question and available annotated training data for that language. Spacy supports recognition of only four classes in German - locations, organisations, persons, and miscellaneous. Meanwhile, for English, there are 18 different classes, and for other languages such as Japanese – even 22 NE classes. Furthermore, for English, there are two distinct granularities of location - GPE, which includes countries, cities, and states, and LOC, which covers all other non-GPE locations like mountain ranges, bodies of water, etc. To align recognized entities between English and German, we prepared an alignment dictionary as shown in Table 1.

Task	**Instruction**
T5 MT	translate English to German:
NER	recognize English entities:
NE-MT	entity translate German to English:

Table 2: Instruction examples for NE-aware T5 tuning. T5 MT represents instructions already in the pre-trained models. NER and NE-MT – our additions.

Model	**Size**	**EN-DE**	**DE-EN**
NE-T5	small	25.11	25.98
NE-T5	base	**26.29**	32.25
NE-T5	large	25.76	**32.45**
NE-T5 1.1	small	26.15	24.12
NE-T5 1.1	base	16.15	25.33

Table 3: MT evaluation results in BLEU for entity-aware fine-tuned models.

4.3 Instruction Formatting

The original T5 model was initially pre-trained using data prepared in the instruction-tuning format with instructions such as “translate English to German: ” or “summarize: ” prepended to each training data source input. Such instructions were also part of Flan-T5 training, but not mT5 or the 1.1 version of T5. We supplement these with instructions for NE-aware translation and the NER task as shown in Table 2.

In addition to the existing “translate” instruction, we add our custom “entity translate” instruction for input data with pre-annotated NEs. We also add fully custom instructions for recognising entities in English and German so that the model can learn NER for plain text inputs.

5 Results

We evaluate MT performance by computing BLEU (Papineni et al., 2002) scores using sacreBLEU (Post, 2018) and NER performance using

		NER		**NEs**	
Model	**Size**	**EN**	**DE**	**EN**	**DE**
NE-T5	small	86.86	82.70	333	450
NE-T5	base	84.31	85.21	320	458
NE-T5	large	**92.01**	**91.37**	308	447
NE-T5 1.1	small	88.93	85.18	331	451
NE-T5 1.1	base	80.59	81.42	329	495

Table 4: NER results for entity-aware fine-tuned models. The last two columns represent the number of NEs recognized in the generated translations.

Model	Size	EN-DE	DE-EN	EN	DE
T5	small	26.88	3.48	255	402
T5	base	29.83	3.27	265	415
T5	large	**30.23**	3.51	247	405
Flan-T5	small	6.48	15.01	281	436
Flan-T5	base	12.63	23.15	312	499
Flan-T5	large	15.31	**29.25**	318	446

Table 5: Baseline model results on MT for non-fine-tuned models. The last two columns represent the number of NEs recognized in the generated translations.

Model	Size	EN-DE	DE-EN	EN	DE
MT-T5	small	27.65	20.75	266	420
MT-T5	base	**30.40**	28.61	299	434
MT-T5 1.1	small	17.83	26.69	302	419
MT-T5 1.1	base	22.00	**30.72**	315	440
MT-mT5	small	16.09	23.50	252	402
MT-mT5	base	17.67	25.88	278	413

Table 6: Baseline results for models fine-tuned on only MT without entity-aware data. The last two columns represent recognized NE counts in the translations.

the F_1 score. An overview of the main automatic evaluation results is shown in Tables 3 and 4. By looking only at the BLEU scores, it does seem like DE-EN translation improves compared to baseline results in Tables 5 and 6 while EN-DE seems to be degraded. However, the amounts of recognized NEs in the generated translations are overall higher for the NE-aware models. Performance on the NER task is relatively low, aside from the T5 large model, but that is not our main focus.

5.1 Machine Translation

The highest-scoring NE-aware model for both English-German and German-English translation is the T5 base tuned with the 10M example data set, while overall including NER performance the T5 large model tuned with 10M examples seems better. Both of them fall behind the non-tuned baseline versions for EN-DE by 3.04 and 4.47 BLEU respectively, but both generate about 10% more NEs in the output than the baselines.

For a clearer comparison to the baselines we also evaluated the pure pre-trained models before any fine-tuning on the entity-aware data, as well as after fine-tuning only on MT data, but without any entity tags. Results of these experiments are shown in Table 5 and Table 6. Since none of the pre-training includes NE tasks, the NER part could not be evaluated. Furthermore, T5 was only pre-trained with instructions for translation from English into German, but not from German into English. This explains why the first three rows of the DE-EN column in Table 5 have such low scores. Meanwhile, mT5 and T5 1.1 cannot be evaluated without fine-tuning, since the instructions for translation or any other downstream task were not included in the model pre-training. As an alternative for mT5, we include evaluation results from Flan-T5 (Chung et al., 2022) in Table 5, which is a multilingual instruction-tuned version of T5.

For a more detailed look at the specific entity classes recognized by the models, Table 7 lists the recognized NE amounts in the source and reference files, baseline non-tuned T5 and Flan-T5 versions, as well as our NE-aware models. There are some differences between the recognized NEs in the source and target files, which is why we performed the NE alignment as mentioned in Section 4 to narrow them down to the lowest mutually matching amount. Out of all baselines, Flan-T5 large does generate a good amount of NEs in the output, but the small version and both T5 baselines noticeably fall behind. Both NE-aware T5 1.1 small and T5 large generate closer amounts of NEs in the output to the source and reference. These results show that the biggest improvements can be gained by fine-tuning the small versions of T5.

5.2 Named Entity Recognition

Given the overall low scores for NER in Table 4, we manually inspected the generated output files for the NER task. The most common critical errors for the small-size models were mismatching NE beginning and ending tags. Many lower-scored cases were also due to the entity not being tagged in the reference, but the model output correctly identified it. To further support this, we performed a manual evaluation included in the Appendix.

6 Conclusion

In this paper, we introduced a simple approach for fine-tuning sequence-to-sequence models that is effective at mitigating one of the commonly known drawbacks of MT - the translation of rare words and named entities. With a small training data modification, we were able to increase the amount of generated named entities in translations, and even achieve a higher BLEU score than the baselines when translating from English into German.

In future work, we plan to evaluate the approach on more languages and alternative NER taggers for training data generation. We are also eager to explore the applicability of the back-translation approach for incremental NER improvements, as well as an extension of our method to summarisation and question-answering tasks.

Acknowledgements

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Ethical Considerations

Our work fully complies with the ACL Code of Ethics[3]. We use only publicly available datasets and relatively low compute amounts while conducting our experiments to enable reproducibility. We do not conduct studies on other humans or animals in this research.

In this work, we only considered training our models on data that is publicly available to enable reproducibility. Also, since hyper-parameter tuning on training large models is computationally very costly, we opt for choosing mostly default parameters in our experiments.

Our proposed method is easily reproducible with publicly available model checkpoints, training data from previous shared tasks, and open-source software for data filtering and preparation cited in this paper. Our custom scripts prepared for NE-tagging and alignment, and T5 model fine-tuning are be released on GitHub[4] under the Apache-2.0 license. We also plan to release our best-performing model checkpoints on the Hugging Face Model Hub. The method is also not limited to the T5 model family in any way, so one could use another pre-trained model as the base, for example, NLLB (Costa-jussà et al., 2022).

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ona de Gibert Bonet, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estapé, and Maite Melero. 2022. Quality versus quantity: Building Catalan-English MT resources. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 59–69, Marseille, France. European Language Resources Association.

Abdul Ghafoor Etemad, Ali Imam Abidi, and Megha Chhabra. 2021. Fine-tuned t5 for abstractive summarization. *International Journal of Performability Engineering*, 17(10).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2022. DEEP: DEnoising entity pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1753–1766, Dublin, Ireland. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alexander Waibel. 2020. Incorporating external annotation to improve named entity

[3] https://www.aclweb.org/portal/content/acl-code-ethics

[4] https://github.com/aistairc/instruction-ner-mt-t5

translation in NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 45–51, Lisboa, Portugal. European Association for Machine Translation.

Kishore Papineni, 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Matīss Rikters. 2018. Impact of Corpora Quality on Neural Machine Translation. In *In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018)*, Tartu, Estonia.

Ehsan Tavan and Maryam Najafi. 2022. MarSan at SemEval-2022 task 11: Multilingual complex named entity recognition using t5 and transformer encoder. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1639–1647, Seattle, United States. Association for Computational Linguistics.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Shufang Xie, Xu Tan, Tao Qin, and Tie-Yan Liu. 2023. Extract and attend: Improving entity translation in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1697–1710, Toronto, Canada. Association for Computational Linguistics.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313.

A Manual Evaluation

We performed a small-scale manual evaluation to further verify the effectiveness of our proposed approach. We randomly select 100 sentences from the evaluation data and manually judge the ability of different model variations to generate automatic translations and recognise named entities.

A.1 Machine Translation

Figure 2 shows one of the common examples where less common location names "Mazedonien" and "Nord-Mazedonien" are mistranslated or rather just simply copied over to the output in English without changing to the correct forms of "Macedonia" and "North Macedonia." The NE-aware model handles these entities better, while the full meaning of the sentence is perhaps not perfectly translated, but still better than the baseline model.

Meanwhile, Figure 3 shows an example where the NE-aware model generates an incorrect, but similarly sounding translation "Syria" to the German word "Sizilien," but the baseline model struggles with this even more by generating a complete hallucination "Sizii." In this case at least the NE-aware model was informed that it should be generating a location.

A.2 Named Entity Recognition

Figure 4 shows just one of many similar examples where one entity was indeed not recognized by the NE-T5 small model, however, two others were recognized by both models, but just not tagged in the reference we used for evaluation. Such cases may occur due to either the Spacy model failing to recognize them at all or on one of the source or target languages. Since in cases when the entity is recognized in one and not in the other language our NE alignment process may have dropped it.

B Recognized NEs in MT Output

Table 7 lists recognized NE amounts in source and reference files, baseline non-tuned T5 and Flan-T5 versions, as well as our NE-aware models.

Source: entity translate German to English: In <LOC>Mazedonien</LOC> stimmen heute rund 1,8 Millionen Bürger darüber ab, ob der Name ihres Landes in <LOC>Nord-Mazedonien</LOC> geändert werden soll.
Reference: In Macedonia around 1.8 million citizens will today agree whether the name of their country in North Macedonia should be changed.
Flan-T5 small: In Mazedonien, a total of 1.8 million people are voting against the name of their country in North-Mazedonien.
NE-T5 small: Around 1.8 million citizens in Macedonia today vote to change their country's name in North Macedonia.

Figure 2: An example of German to English translation output where the baseline model copies location names in German "Mazedonien" and "Nord-Mazedonien" to the English output while the NE-aware model generates correct translations "Macedonia" and "North Macedonia."

Source: entity translate German to English: Drei Männer sind in <LOC>Sizilien</LOC> festgenommen worden, sie sollen in libyschen Flüchtlingslagern vergewaltigt und gemordet haben.
Reference: Three men have been arrested in Sicily who are alleged to have tortured and murdered people in Libyan refugee camps.
Flan-T5 small: Three men are in Sizii, they should be in Libyan refugee camps and have been displaced.
NE-T5 small: Three men have been arrested in Syria, they are expected to have been raped and abused in Libyan refugee camps.

Figure 3: An example of German to English translation output where neither model produces the correct translation "Sicily," but our NE-aware model at least generates a valid location "Syria" while Flan-T5 hallucinates "Sizii."

Source: recognize English named entities: Frankfurt speculations that the Bank of England (BoE) will soon be reducing its interest rates are putting pressure on the pound sterling. On Friday, the British currency dropped by up to 0.4 percent down to 1.2269 dollars.
Reference: <LOC> Frankfurt </LOC> speculations that the Bank of England (BoE) will soon be reducing its interest rates are putting pressure on the pound sterling. On Friday, the British currency dropped by up to 0.4 percent down to 1.2269 dollars.
NE-T5 small: Frankfurt speculations that <ORG> the Bank of England </ORG> (<ORG> BoE </ORG>) will soon be reducing its interest rates are putting pressure on the pound sterling. On Friday, the British currency dropped by up to 0.4 percent down to 1.2269 dollars.
NE-T5 large: <LOC> Frankfurt </LOC> speculations that <ORG> the Bank of England </ORG> (<ORG> BoE </ORG>) will soon be reducing its interest rates are putting pressure on the pound sterling. On Friday, the British currency dropped by up to 0.4 percent down to 1.2269 dollars.

Figure 4: An example of English NER output where the two NE-aware models recognize "the Bank of England" and "BoE" as entities, which were not marked in the reference. The small model does fail to recognize "Frankfurt" as a location, but the large one succeeds.

			(DE→) EN				**(EN→) DE**		
Model	**Size**	**PER**	**LOC**	**ORG**	**Total**	**PER**	**LOC**	**ORG**	**Total**
Reference		126	98	89	313	169	179	107	455
T5	small	128	70	57	255	141	183	78	402
T5	large	121	60	66	247	146	182	77	405
Flan-T5	small	117	83	81	281	145	182	109	436
Flan-T5	large	124	93	101	318	151	195	100	446
NE-T5 1.1	small	138	97	96	331	172	184	95	451
NE-T5	large	122	91	95	308	161	187	99	447

Table 7: Recognized NE counts in the evaluation sets for English ↔ German translation.

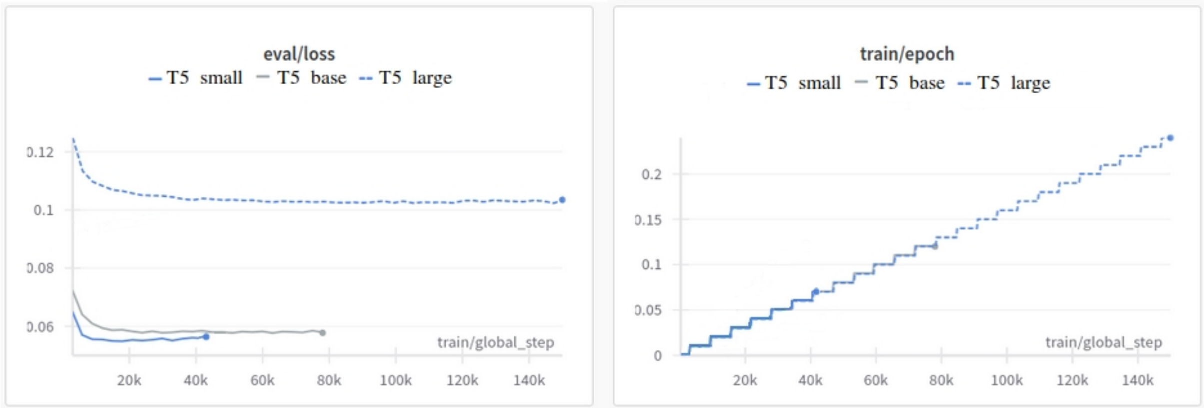


Figure 5: Training progress for T5 models using the 10M example-sized training data set.

	EN-DE	DE-EN
T5-small	25.03±0.09	26.11±0.15
T5-base	26.10±0.21	31.77±0.48

Table 8: Average machine translation experiment results in BLEU scores for small and base models with different random seeds.

C Preliminary Experiments

Figure 5 shows results from our preliminary experiments where we performed fine-tuning on *small*, *base*, and *large* versions of T5 using the 10M version of the training data set. The *small* model converged after seeing just over 6% of the data, the *base* – around 13%, and the *large* – 24% of the training data. Therefore, we chose to limit the data amounts for experiments to 100K for *small* size models, 1M for *base*, and 10M for *large* versions of the T5 family models.

We also experimented with runs on the small and base models with 100K and 1M training data sizes respectively using three random seeds (347155, 42, 9457). The final results from these experiments are shown in Table 8. Since the variance for each was relatively low, we limited our further experiments to use only the first of the three random seeds.