Investigating Shared Representations in Translation Systems

Nathan Zeweniuk
260978250

Navid Hassan Zadeh
261019828

Yuhan Zhou
261037242

Abstract

This project investigates the feasibility of creating a language-independent representation for multilingual translation using a language-specific transformer-based encoder-decoder approach. We first select an initial set of languages and train language-specific encoders and decoders on all pairings. To examine the capability of the interlingua, we train a new encoder using an existing decoder for one of the initial languages and evaluate on translation to a different initial language. We found that a Chinese encoder trained with a decoder from a different language was able to achieve translation to english with performance only slightly worse than training a transformer directly on Chinese-English pairs. Additionally, we show that the choice of the initial languages affects downstream performance when adding a new encoder.

1 Introduction

An important challenge in the evolving field of Natural Language Processing (NLP) lies in crafting translation systems that are both effective and adaptable. Traditionally, these systems have centered around translating from a single source to a single target language. However, accommodating translation among all language pairs within a set demands a quadratic number of translation models and requires extensive translation data across all language combinations. To tackle this issue, a proposed solution involves splitting translation tasks into source analysis and target generation, leveraging an intermediary language known as the "interlingua." This language-independent framework provides an unambiguous representation of the input text's meaning. It has to fulfill a simple functional condition: the interlingua representation must be sufficient for accurate translation in a technical domain. (Lonsdale et al., 1994)

The key objectives of this project are to develop and investigate whether it is possible to create an interlingua by training language-specific encoders and decoders on all pairs of a small initial set of languages. We trained two models for the purpose of this project. The first model focused on three closely related languages that had shared some vocabulary, while the second model was trained with three more distinguished languages from distant language families. Then we added a pair of encoders and decoders of a new language to both models to investigate the effectiveness of the shared representation learned by the models.

2 Related Work

Encoder-decoder models are a common approach in modern machine translation. A typical example is the attentional RNN encoder-decoder approach proposed by Bahdanau et al. (2014) (Bahdanau et al., 2014). In this project, we use encoders and decoders based on the transformer architecture proposed by [transformer paper] which has led to advancements and state-of-the-art performance in machine translation and other areas of NLP.

There have been several attempts to do multilingual machine translation. One strategy is to translate to and from a pivot language in between the source and target languages, such as in (Cheng et al., 2016). Here, the pivot language acts as the interlingua between the other languages. However, the pivot language may not be the best or most efficient interlingual representation. This technique also requires more computational resources as there are essentially two passes per translation.

Another method is to use one large model that can translate between multiple languages such as the T5 model (Raffel et al., 2023). T5 uses transfer learning on applied to a number of different tasks including machine translation. With this approach, an interlingual language representation cannot easily be extracted from the model. Moreover, large models can suffer from increase memory and compute requirements.

Several other works combine language-specific encoders and decoders in different ways. Dong et al.(2015) (Dong et al., 2015), Zoph and Knight(2016) (Zoph and Knight, 2016), Luong et al.(2016) (Luong et al., 2015), Firat et al.(2016a) (Firat et al., 2016), etc. have explored the one source to many targets, many sources to one target, and many sources to many targets in multilingual MT settings.

Our work is closely related to the approach used by by Lu et al.(Lu et al., 2018). This technique uses language specific encoders and decoders which translate into and from a shared interlingua. However, our approach differs in a few ways. We use transformer based encoders and decoders rather than RNNs. Rather than using a shared encoder as part of the interlingua, our interlingua is based only on the encoder outputs and is entirely defined by the encoders and decoders that created it. Also, our training is done on pairings between all languages used to create the interlingua, not just pairings with English.

3 Method

The experimental phase of this study focused on evaluating the feasibility of creating a language-independent representation for multilingual translation. We mainly focused on two systems with the same architecture but different training:

- **Similar (Closely-linked) language family training** This model was trained on pairs from a set of languages containing English, French, and Spanish to develop a shared representation.
- **Different language family training** This model, however, was trained on pairs from a more diverse set of languages including English, Russian, and Arabic to develop a shared representation.

A new Chinese encoder is created for each set by training only on pairs with one of the initial languages (Spanish for the first set and Arabic for the second). To evaluate the performance of the new encoder, we also compare against a baseline transformer which was only trained on Chinese to English translations.

3.1 Dataset

The dataset utilized in this study comprises a multilingual parallel corpus acquired from the United Nations Parallel Corpus (UNPC) (Ziemski et al., 2016) via the Hugging Face datasets library. This corpus includes several translation pairs between English, French, Spanish, Russian, Arabic, and Chinese. Language pairs were tokenized using pretrained tokenizers from Huggingface for each of the specific languages: English (Raffel et al., 2020), French ((Labrak and Dufour, 2022), (UniversalDependencies), (Béchet, 2001), (Akbik et al., 2018)), Spanish ((Wolf et al., 2020), "Deepesp/gpt2-spanish"), Russian ((Wolf et al., 2020), "blinoff/roberta-base-russian-v0"), Arabic (Safaya et al., 2020), and Chinese ("bert-base-chinese", (Devlin et al., 2019)).

The training sets were created from samples of the first 3 million examples for each language pair. Many of the examples are very short phrases containing only dates, titles, abbreviations etc. To avoid the noise created by such examples, only examples with greater than 250 characters were used

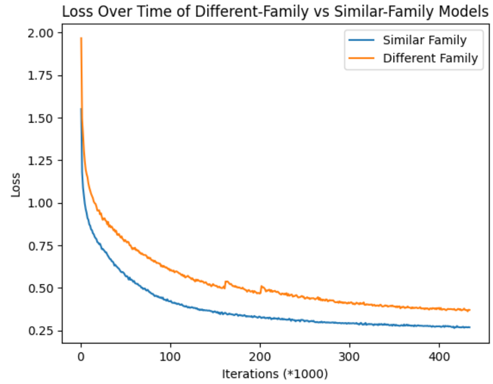

Figure 1: The comparison between training loss decreases over Iterations during training of the two models in this experiment.

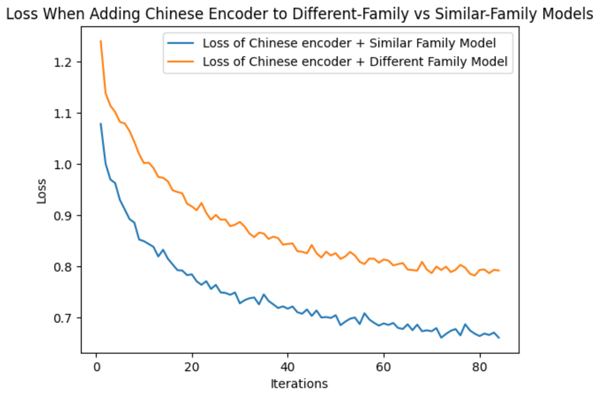

Figure 2: The comparison between training loss decreases when adding Chinese encoder to different-family vs similar-family models

for training, except some shorter examples which were added randomly with a probability of 0.1 to ensure the model sees the end of phrases. A further 10k examples from relevant pairs were set aside for testing. The test set was sampled in the same way (greater than 250 characters, and shorter with a probability of 0.1) leading to about 2k to 2.5k examples.

3.2 Model Architecture

Neural machine translation is implemented with a transformer-based (Vaswani et al., 2023) encoder-decoder architecture with an attention mechanism. We conducted this in supervised many-to-many settings with language-specific encoders and decoders. The both encoders and decoders consists of 6 layers, 512 expected features, and 8 attention heads. The feedforward network within the model uses 2048 hidden units, incorporating a dropout value of 0.2, and employs word embeddings sized at 512. Due to time and compute restrictions, no significant hyperparameter search was performed, but hyperparameters follow closely from (Vaswani et al., 2023).

3.3 Training

Each of the language pairs in the training set are put in batches of 16. The reverse of each language pair is also added to the training data so both an encoder and decoder can be created for each language. During each training step, a batch from a random pair of languages is selected. The specific encoder for the selected source language and specific decoder for the target language are used for a forward pass. The weights for the specific encoder and decoder are updated using the Adam optimizer (Kingma and Ba, 2017).

For the initial set of three languages, we trained for 2 epochs on 200k batches of 16. With 6 possible pairs, this results in 530k sentences per language pair. Figure 1 shows the training loss for this process. It should be noted that training had to be stopped and continue on two occasions for the set of dissimilar languages, as seen by the two blips in the training curve. When adding the new Chinese encoder, we trained on 20k batches, or about 320k sentences, for 4 epochs (Figure 2). The baseline Chinese to English transformer was trained in the same was as the added Chinese encoder.

4 Results

We evaluate our method on the test set using Bleu and Rouge scores. The results in Table 1 show that Chinese to English translation using the interlingua only performs slightly worse than directly trained transformer despite not being trained on any Chinese-English pairs. However, the overall performance is fairly low, and it is uncertain whether this trend would continue with better performing models. The similar family interlingua seems to lead to slightly higher quality Chinese to English translations than the dissimilar family interlingua. This may be because the similar family interlingual representations are closer to English, so the training process is similar to training directly from Chinese to english. In both cases, the added encoder performs worse than the reference encoder that was trained in the initial interlingua. While this may indicate the interlingua is not capable enough, the fact that the baseline is also low indicates the challenge comes from encoding Chinese at all rather than losing information when encoding to the interlingua.

Table 2 shows an example translation to and from the interlingua in English. The translation seems to capture the essence of the phrase and includes keywords like ’translation’ and ’system’, but suffers from grammatical errors and repetition. The repetition is a known problem and has been tackled by works such as (Zhang et al., 2021). It has also been shown that transformer decoders are capable of generating meaningful and grammatically correct outputs (Radford and Narasimhan, 2018).

5 Discussion and Conclusion

Our experiments show there is potential for this method to be used to create an interlingua for multilingual translation. We were able to achieve similar performance for Chinese to English translation training only one encoders for other languages in the interlingua when compared to a transformer trained directly on Chinese to English translation. However, the overall translation quality between dissimlar languages is fairly low. This indicates that our model architecture or training procedure could be improved. We suspect training on a more diverse dataset with more examples for a longer time would yield better translation quality, and allow us to fully assess the capability of our method. Further improvements upon this method could lead to modular, adaptable multilingual translation sys-

	Bleu	Rouge1	Rouge2	RougeL
Zh-En (similar)	12.2	38.4	16.0	31.4
Es-En (similar)	30.0	60.8	40.0	55.0
Zh-En (dissimilar)	10.0	37.3	16.8	30.8
Ru-En (dissimilar)	15.4	45.6	25.7	40.2
Zh-En (baseline)	12.9	35.6	13.2	27.6

Table 1: Bleu and Rouge scores for translations to English using the similar and dissimilar language family interlingua.

Similar Family Interlingua	Different Family Interlingua
"<s> - this is a test test of the new translation programme - the new translation system.</s>"	"<s> This is testing of a new translation translation system test this new translation system translated automation system.</s>"

Table 2: Example translation using english encoder and decoder for the phrase "This is a test of the new translation system."

tems.

5.1 Future Work

Further experiments are required to fully understand the capability of the interlingua. We would like to explore alternate model architectures and hyperparameters. With access to a larger dataset, we could examine an interlingua created with a larger number of initial languages. Further analysis could look at the behavior of adding new encoders to the interlingua as well. The results from this study do not present a strong conclusion on how structural differences in language affect the interlingua created by the initial set.

One of the benefits of this method is direct access to interlingual language representations. It is known that LLMs tend to suffer from worse performance in languages that do not make up a significant portion of their training data (Huang et al., 2023). With an expressive enough interlingua, generative language models could theoretically be trained on these interlingual representations even if the majority of the data is in one language.

6 Statement of Contributions

Please refer to Table 3.

7 Code files and model Weights

We uploaded all our code files and the model weights after training to the following GitHub page: https://github.com/navidhsnz/comp550-Final-Project.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Frédéric Béchet. 2001. Lia$_tagg : astatisticalpostagger + syntacticbracketer. Technicalreport, Aix - MarseilleUniversityCNRS.$

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016. Neural machine translation with pivot languages. *CoRR*, abs/1611.04928.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei.

Yuhan	Navid	Nathan
Research and planning, Evaluation and evaluation metrics, Report writing and editing, Analysis	Research and planning, Initial training method, Exploring alternative architectures, Model training and evaluation, Report writing and editing	Research and planning, Model architecture, Data processing and loading, Report writing and editing

Table 3: Contributions of the members.

2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Yanis Labrak and Richard Dufour. 2022. ANTILLES: An Open French Linguistically Enriched Part-of-Speech Corpus. In *25th International Conference on Text, Speech and Dialogue (TSD)*, Brno, Czech Republic. Springer.

Deryle W Lonsdale, Alexander Franz, and John RR Leavitt. 1994. Large-scale machine translation: An interlingua approach. In *Iea/aie*, pages 525–530.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

UniversalDependencies. Universaldependencies/ud$_{f}rench-gsd$.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Ying Zhang, Hidetaka Kamigaito, Tatsuya Aoki, Hiroya Takamura, and Manabu Okumura. 2021. Generic mechanism for reducing repetitions in encoder-decoder models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1606–1615, Held Online. INCOMA Ltd.

Micha
l Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.