Numerous Multi-Pivot and Chained Pivot NMT for Low-Resource Language Translation

Stanford CS224N Custom Project

Cees Armstrong
Department of Mathematics
Stanford University
ceesa@stanford.edu

Kevin Reso
Department of Economics
Stanford University
kevreso@stanford.edu

Abstract

While NMT has been able to successfully render translation across many languages as accessible as ever, the issue of lacking parallel corpora remains an immense hurdle to accuracy. In this project, we aim to improve NMT methods for low-resource or even zero-resource languages that possess few or no large parallel corpora. We propose to enhance translation accuracy by implementing novel pivot translation architectures. Specifically, we develop a chained pivot approach, in which we link multiple pivots between source and target languages such that the linguistic distance, measured in perplexity, between each consecutive translation is minimised. We test the performance of our novel, chained pivot model against direct translation as a baseline. Additionally, we compare our chained pivot approach to multi-source pivot models using the same pivot languages. Specifically, although multi-pivot models have previously been implemented, we develop and test architectures with highly numerous pivots. Our experimental models trained on Europarl v8 data for English-Slovenian translation show that there are increasing returns, in terms of translation accuracy, to highly numerous multi-pivot methods, although significantly sized corpora are necessary for their efficacy. With respect to chained pivoting, our results show that, while the minimisation of perplexity distances in pivoting may play a role in improving BLEU scores, the successive translations required for chaining render it infeasible as a method for low-resource NMT improvement.

1 Key Information to include

- Mentor: Moussa Doumbouya
- External Collaborators (if you have any): No
- Sharing project: No
- Team Contributions: Kevin dealt with most of the pivoting code. Cees dealt with most of the experiments and the data processing. The writing of the paper was a largely joint effort.

2 Introduction

Neural machine translation (NMT) works well when supported by extensive parallel corpora, but many languages lack such data, making effective, traditional NMT models difficult to train. Regardless, the translation of such languages is a valuable pursuit. While many popular methods have been studied for low-resource translation, most fall under the category of supervised and semi-supervised methods (Haque et al., 2021). Pivot NMT (PNMT), however, is a method in which translation is learned without any direct corpora at all. Recently, it has been explored in the literature as a means of overcoming such a lack of data. PNMT uses an intermediary language to overcome the issue of lacking parallel data between language pairs, making it an interesting tool in the space of low-resource MT. At the same time, PNMT can degrade translation quality as a result of translation errors compounding through multiple consecutive iterations of translation. The upper bound of PNMT acucracy is known to be the accuracy of a direct translation model, seeing as it seeks to substitute the presence of extensive data (Dabre et al., 2021).

While PNMT has shown promise, in the recent literature, it has shown further promise when combined with multi-source NMT (MSNMT) methods. In MSNMT, multiple source languages are translated into a single target language to improve translation accuracy (Zoph and Knight, 2016). When aggregating the multiple target to source translations, there are N encoders for the N different inputs. Then, after in individual attentions A_i are computed, an individual decoder determines final attention A as each A_i is passed through a gating mechanism to determine its importance. An approach requiring multiple encoders consequently necessitates an N-way parallel corpus, limiting its effectiveness. Alternatively, certain ensembling approaches that do not require such N-way corpora achieve similar results when implemented within an MSNMT framework (Dabre et al., 2017), (Firat et al., 2016).

The combination of PNMT and MSNMT has recently yielded the advent of multi-source pivot NMT (MSPNMT), in which multiple pivot languages are used in the translation from a source language to a target. MSPNMT methods have yielded good results when it comes to the translation of low-resource languages (Dabre et al., 2021), (Mhaskar and Bhattacharya, 2022). In fact, MSPNMT may offer BLEU score improvements of up to 5.8 over baseline direct low-resource translations (Dabre et al., 2021). Although MSPNMT has been somewhat effective, as with most low-resource translation tools, it still struggles to approach the translation accuracy of direct models trained on full data. What's more, out of the few studies that assess its merits, seldom employ any more than four pivots total.

The research regarding the encouraging results of MSPNMT architectures motivates the study of how both a higher than typical number of pivots and alternative multi-pivot methods might effect low-resource translation. In particular, it begs the question of the most effective selection of possible pivots. We propose to train a direct translation baseline from English to Slovenian. Using this baseline as a reference, we will compare MSPNMT models with as many as six pivots and a novel multi-pivot approach. Our novel approach will involve chaining multiple pivots on the way from source to target such that the linguistic distance between each pivot is minimised according to the positions of each language in a perplexity space. We attempt to show that these implemented methods will demonstrate improved translation accuracy, as measured by BLEU score, in a low-resource setting.

3 Related Work

Most current research surrounding the concept of implementing multiple pivots for the sake of low-resource translation is confined to the use of two to four pivots. In no cases are there current results regarding any sort of linguistic distance minimising methods of pivot chaining.

Dabre et al. (2021) develop a simultaneous multi-pivot framework. By incorporating aspects of simultaneous NMT (SMNT) and PNMT into their architecture, the authors arrive at a SPNMT model. Incorporating, as described earlier, an attention-based multi-source approach into their SPNMT framework such that it might become a multi-pivot model, accuracy gains are made in the context of low-resource translation. Specifically, for translation models with two pivots (French, Spanish), a BLEU score improvement aboce low-resource direct translation of up to 5.8 is recorded.

The possibly positive results of multi-pivoting are shown to be reproducible by Mhaskar and Bhattacharya (2022). Using four languages (Tamil, Bengali, Gujurati, Hindi) as pivots at once, as well as a strategic decoder initialization in which the pivot-to-target decoders are initialized simultaneously with the source-to-pivot decoders. The authors observe similar benefits to such methods in the realm of translating low-resource languages. Similar results have also been shown for low-resource multi-pivot translation from Chinese to Lao through English and Thai (Wang and Li, 2023).

With respect to language distance for pivot selection in chaining, Gamallo et al. (2017) established a definition for linguistic distance using language identification methods. Based on perplexity, the authors plot the relative distances of all European languages such that is possible to determine the distance between any two languages through any others.

4 Approach

We use OpenNMT's sequence-to-sequence transformer architecture to train NMT models for various European language pairs (Klein et al., 2018). This gives us our baseline translation method, namely translating using the OpenNMT architecture (Klein et al., 2018). Using these pre-trained direct NMT models, we implement our own two approaches of interest:

1. Numerous Multi-Pivoting: We implement an MSPNMT architecture that configures translation from a source language through N pivots to a target language. Within our framwork, we deploy N pre-trained direct NMT models to facilitate translating from the source to each pivot. When going from each pivot to the target, we deploy another N pre-trained translation models 1. To return a final translation in the target language, we take the MSNMT approach of late-average ensembling where we take the mean of N translations at the output level to aggregate a final translation (Dabre et al., 2017), (Firat et al., 2016). In particular, using Hugging Face's AutoTokenizer, we determine probability distributions $p(y_t = w \mid y_{<t}, X_i)$ over the target vocabulary for each input (Hugging Face, 2024). We then take the average over the distributions to arrive at an output distribution that determines the final translation.

$$p(y_t = w \mid y_{<t}, X_1, \ldots, X_N) = \frac{1}{N} p(y_t = w \mid y_{<t}, X_1) + \cdots + \frac{1}{N} p(y_t = w \mid y_{<t}, X_N) \tag{1}$$

The architecture described can be seen in Figure 1, where we have a model using six pivot languages. Within our architecture, however, the same setup can be achieved with either a higher or lower number of pivot languages. This architecture is entirely coded ourselves, using models pre-trained using OpenNMT (Klein et al., 2018). It is motivated by the lack of study in using multiple-pivots past 4 languages.

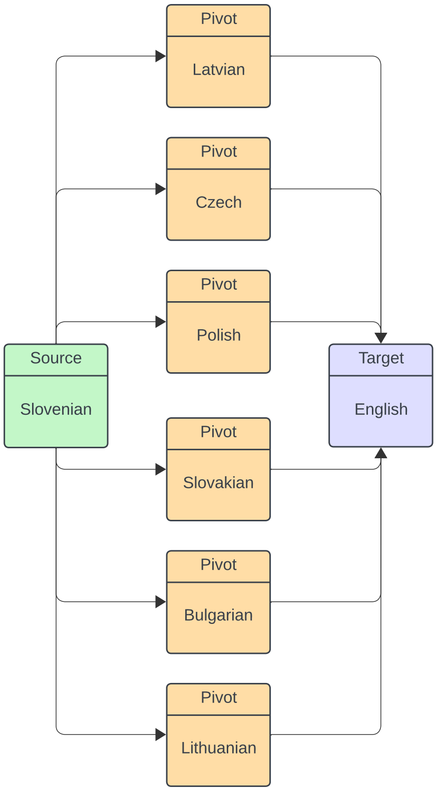

Figure 1: Numerous Multi-Pivoting

2. Chained Pivoting: We use the work of Gamallo et al. (2017) on perplexity-based linguistic distances to see how similar different languages are in a perplexity space. A graph of perplexity based distances can be seen in Figure 2. Using these distances, we implement a novel single pivot architecture with two chained pivots such that the linguistic distance between pivots is minimised. We deploy a pre-trained model from the source to the first pivot, a second pre-trained model from the first pivot to the second, and a final pre-trained model from the second pivot to the target. This architecture can be seen in Figure 3, with a sample pivoting chain already set up. This novel architecture is entirely coded ourselves, similarly using models pre-trained with OpenNMT (Klein et al., 2018).

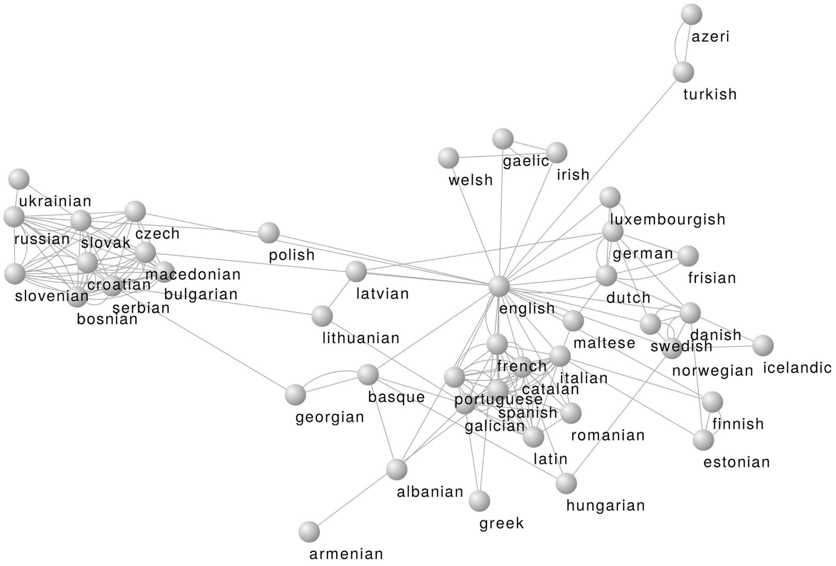

Figure 2: Perplexity Scores
(Gamallo et al., 2017)

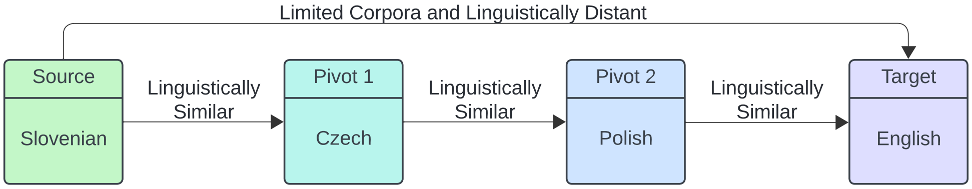

Figure 3: Chained Pivoting

5 Experiments

5.1 Data

We use the Europarl v8 dataset in this project (Koehn, 2005). The data consists of parallel corpora extracted from the proceedings of the European Parliament between 1996-2011 (Tiedemann, 2012). In particular, we use the following sub-datasets:

1. Slovenian-English (sl-en)
2. Slovenian-Polish (sl-pl)
3. Polish-English (pl-en)

4. Slovenian-Czech (sl-cs)
5. Czech-English (cs-en)
6. Slovenian-Bulgarian (sl-bg)
7. Slovenian-Lithuanian (sl-lt)
8. Bulgarian-English (bg-en)
9. Lithuanian-English (lt-en)
10. Slovenian-Slovakian (sl-sk)
11. Slovenian-Latvian (sl-lv)
12. Slovakian-English (sk-en)
13. Latvian-English (lt-en)
14. Czech-Polish (cs-pl)
15. German-Greek (de-el)
16. Slovenian-German (sl-de)
17. Greek-English (el-en)

We pre-process the data in three stages, as per the OpenNMT framework (Klein et al., 2018):

- Filtering: we filter the datasets to remove low-quality segments, including misalignment, empty segments and duplicates.
- Subword Tokenization: we split the data into words and then into sub-words, so that when the model sees a word that looks similar to a sub-word in its vocabulary, then it can continue its translation instead of marking the word as unknown. Due to our small corpora sizes, we still end up with some unknown pieces of data.
- Data Splitting: we split the data into a train, development, and test set. We have augmented the OpenNMT code here. Our code allows for assigning a size to the training data set, beyond just assigning development and test set sizes. As such, we are able to simulate low-resource language translation by cutting corpora sizes.

5.2 Evaluation method

We use two main evaluation methods to gauge translation quality:

- BLEU Score: We use Sacre-BLEU, a BLEU score implementation that reports a translation quality metric, scoring translations between 0-100. The Sacre-BLEU framework provides more hassle-free computation of scores and a comparable benchmark. Naturally, however, all the limitations of BLEU scores still carry over, as it is at heart just calculating BLEU scores (Post, 2018).
- Qualitative Evaluation: We manually observed a random sample of translated sentences from the corpora to see how good the translations to English actually were. This helped to augment our understanding of the efficacy of our methods beyond BLEU score, which frequently fails to capture some of the contextual and idiosyncratic facets of translation in addition to basic grammatical accuracy and intelligibility.

5.3 Experimental details

We varied the learning rate as per Vaswani et al. (2023). This meant that the learning rate increased during the warm-up steps and decreased proportionally thereafter. Our dynamic learning rate started at 0.000028, reached 0.00279 at the end of the warm-up, and then decreased all the way down to 0.00091 by the end of training. We trained our models on an NVIDIA L4 with 10,000 steps, of which 1,000 were warm-up. Each model took between 60 and 120 minutes to train.

We varied model configurations to reach optimum translation accuracy. Primarily, our configurations took the form of varying dropout. As we trained models with varying dataset sizes, this informed how much dropout we applied. We applied dropout and attention dropout in the following manner after carrying out tests on translation accuracy after training with different amounts of dropout.

Corpora Size	Dropout Rate	Attention Dropout Rate
50,000	0.2	0.2
250,000	0.1	0.1

Table 1: Dropout Rates for Different Corpora Sizes

We found that this was able to effectively strike a balance between overfitting and underfitting to result in optimal BLEU score results. Increasing attention and normal dropout together meant that we had higher dropout both within the attention mechanisms and in the outputs of the feed-forward layers, improving overall training accuracy and hence model quality when corpora sizes were so low.

5.4 Results

- We began by training a direct translation model on every language pair listed above, each on a corpus of 250,000 parallel words. Each direct model that we trained achieved a BLEU score greater than 10.000 with a maximum of 25.571 for our Polish-English model. First, we trained a direct Slovenian-English baseline model on a significantly cut corpora of only 50,000 in order to simulate a low-resource setting. Then, for Slovenian to English we implementing multi-pivot models up to six pivots with two single pivot models as baselines. Lastly, we implement two chained pivot models with the first minimising linguistic distance between pivots and the second maximising it.

Translation Method	Languages	BLEU Score
Direct	sl-en	13.091
Single-Pivot	sl-pl-en	2.285
Single-Pivot	sl-cs-en	2.111
Multi-Pivot	sl-pl/cs-en	3.395
Multi-Pivot	sl-pl/cs/bg/lt-en	3.561
Multi-Pivot	sl-pl/cs/bg/lt/sk/lv-en	3.622
Chained Min-Pivot	sl-cs-pl-en	1.161
Chained Max-Pivot	sl-de-el-en	0.861

Table 2: Performance of Pivot Methods

- Across the board, our implemented methods, despite employing direct models trained on five times the parallel data, achieved less than half the accuracy of the baseline model with respect to BLEU scores. A qualitative assessment of randomly sampled sentences also results in similar conclusions as the reported BLEU scores with respect to the efficacy of our methods.
- Regarding numerous multi-pivoting, our results are neither better nor worse than expected. To overcome the decreased accuracy inherent to pivoting methods, it may have been the case that our pre-trained models would have had to have been trained on much larger parallel corpora. Regardless of these concerns, increasing the number of pivots all the way to six results in improved accuracy.
- As far as our chained pivoting results are concerned, while there clearly exists a BLEU score preference toward the chain that minimises linguistic distance, both chained pivot models yield results significantly worse than not only the direct translation baseline but any of our implemented multi-pivot approaches. It is likely that in our implementation, the minimisation of linguistic distance does not overcome the compounding of inaccuracies resulting from successive pivoting.

6 Analysis

Our model's success is tied to the quality of its underlying, pre-trained direct NMT models. As a multi-pivot architecture, when using individual models that register low BLEU scores, the entire pivot model achieves correspondingly low BLEU scores. On the flip side, the better individual pivot points are, in terms of BLEU scores, the better the entire multi-pivot system becomes. As much is to say that if the underlying direct models are trained on significantly large corpora such as to allow high accuracy, the pivot model will be improved.

For these underlying models, there were two factors that really impacted the quality of their BLEU scores. First, consistent data availability was an issue. The larger we allowed the corpora to be in training, the greater the BLEU scores become. For example, in initial trials we tested English to Polish translation with a training set of 250,000 and again with a training set of 50,000. On the 250,000 iteration of training, we achieved a BLEU score of 25.571, while on the 50,000 iteration, we only got a BLEU score of 13.712. Second, the amount of dropout used had real impacts on the quality of our training due to model underfitting and overfitting. In particular, as we were dealing with small datasets, overfitting was a real issue, meaning that we had to increase dropout and attention dropout to cope. When running tests to find optimum dropout rates for small datasets, we saw accuracy beginning to dip after only 1,000 steps when dropout rates were low.

Working within the chaining model architecture, linking more pivots together in a row was seen to have largely negative effects on our translation accuracy. This was somewhat expected, as introducing pivots generally decreases translation quality over direct baselines. At the same time, however, the extent to which this was true was surprising, as two chained pivots decreased BLEU scores significantly as compared to only one pivot. In fact, the decrease between the two models was more than fifty percent.

Within our multi-pivot architectures, increasing the number of pivots, on the other hand, produced improvements in BLEU scores. This was most pronounced when adding the second pivot, and then became less pronounced afterwards as more pivots were thereafter added on, perhaps indicating a function of diminishing returns to more pivots.

The last key takeaway was regarding the linguistic similarity of languages and how it affected our models. We saw benefits to using languages that were linguistically similar based on perplexity scores as our pivots in not only our chaining methodology but our generally multi-pivot methodology as well. When using languages that linguistically were nearer to one another in the pivoting process, translation quality was generally higher for the same corpora sizes.

We will go through a few interesting translations to get a more holistic picture of translation quality past BLEU scores:

1. Direct translation from Slovenian to English:
   Reference Sentence: It is extremely important to stick to the proper order: Planet, people and product.
   Translated Sentence: It is very important to say that we are appropriate: order, people and production.
   The similarity in translation here is indicative of how good direct translation methods are. Even with a very small corpora of 50,000 parallel words, the meaning is largely delivered, even if there are some discrepancies such as 'Planet' being translated to 'order'.

2. Chained translation from Slovenian to Czech to Polish to English:
   Reference Sentence: The Group of the European People's Party (Christian Democrats) has not negotiated with any individual groups.
   Translated Sentence: Let me give you a great deal.
   The stark contrast in meaning displays the sheer inaccuracy of the translation. However, it is interesting that the translation still in large part makes English sentence as a series of words, even if the meaning is entirely lost, as it may suggest that there are grammatical links that have remained through the translation, something lost on a BLEU score.

3. Multi-Pivot translation from Slovenian to Polish/Czech to English:
   Reference Sentence: The Group of the European People's Party (Christian Democrats) has not negotiated with any individual groups.
   Translated Sentence: It leads to an extraordinary closure of genetically modified medicines as withdrawn.
   This translation is almost entirely inaccurate. This highlights the issues with pivot-based translation as a means of translation sizes, as the original poor translation is degraded when going through two repetitions of translation.

From these set of translations, it is evident that the quality of translation through pivot-based architecture can be exceedingly poor relative to direct models, especially when dealing with small corpora sizes across each pivot.

7 Conclusion

Our project, while not beating baseline OpenNMT translation, showed that adding more pivots improved translation quality and that minimising perplexity distances between languages showed improvements in BLEU score translations. Such results indicate that future research on low-resource NMT improvements might focus on these two aspects of NMT methodology. This demonstrated the importance in choosing languages for pivots and not just using data availability, but also looking at perplexity distance when determining pivots. Our implementation of an additional, chained pivot was largely ineffective, and we have learned that additional pivots are ineffective translation strategies if parallel corpora exist between target and source languages.

Our work is limited by two major factors. First, our findings are all on the Europarl dataset. Our research is, therefore, confined to European languages. Further work would be needed to see if they can be replicated on other datasets and languages. Second, our findings on chaining rely on only a few pivot chains. It is possible that these are outliers, and further research would need to be done

Further research looking into incorporating pivots into our chaining architecture to see if this improves the accuracy of a chain based method would be fascinating. This would entail a source language reaching multiple pivots, and those pivots could thereafter reach a further series of pivots, and then all the second layer of pivoted languages could rejoin to reach the target, which one would expect to improve translation accuracy. Research on varying corpora sizes further to see how much larger pivot corpora need to be for them to begin to overtake direct translation methods would also be a valuable avenue of further study, in order to determine at what point pivot architectures can be useful in low-resource language translation.

8 Ethics Statement

Inaccurate translations can have major impacts on our perceptions of both other peoples and other cultures. When we get inaccurate translations, our ability to communicate with others is significantly hampered, which poses risks with respect to people misjudging each other, as people are fed inaccurate representations of what is being said.

Performing a translation, therefore, carries significant societal risks associated with these inaccuracies. These can be largely grouped into two camps:

1. Potential for Bias: Language is often filled with cultural nuance. As translation models are reliant on their input data, they often carry forward the biases reflected in this data. In large part, this entails creating a system that perpetuates stereotypes which can be incredibly harmful to those whom the model is biased against. Our proposed translation methodology would carry with them these biases from its data too.
2. Potential for Errors: Error-ridden translation perpetuates misunderstandings between peoples. This can cause embarrassment and confusion, and in extreme cases can even create enmity as people think that the other is saying something which they are not. Our translation methodology, when it makes errors, could cause this exact kind of problem.

There are principally two ways to best mitigate these issues:

1. Transparency within the translation: It needs to be clear that a piece was machine translated and can, as a result, carry the associated inaccuracies in which machine translation often results. In particular, for low-resource languages, this needs to be made very clear as translation quality is often lower. This helps when dealing with biases and errors present in the translation, as it makes those using the translator more aware that the translation might be poor and therefore on the lookout for it.
2. Ongoing certification of translation accuracy: Certain phrases should be pulled out at random and tested for translation accuracy by professional translators. This will allow us to get a

better picture of our translation accuracy from the model and where it goes wrong, so that certain types of phrases can perhaps be highlighted as problem phrases, and the model can be improved accordingly. In particular, it means that certain biases that the model expresses can be identified and noted so that people recognise the limitation. Similarly, it means that common errors can be noted in the translation model. These pieces of feedback can also be used to update the model to make it more accurate.

References

Raj Dabre, Fabien Cromieres, and Sadao Kurohashi. 2017. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages.

Raj Dabre, Aizhan Imankulova, Masahiro Kaneko, and Abhisek Chakrabarty. 2021. Simultaneous multi-pivot neural machine translation.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation.

Pablo Gamallo, Jose Ramom Pichel, and Iñaki Allegria. 2017. From language identification to language distance.

Rejwanul Haque, Chao-Hong Liu, and Andy Way. 2021. Recent advances of low-resource neural machine translation.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. Opennmt: Neural machine translation toolkit.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation.

Shivam Mhaskar and Pushpak Bhattacharya. 2022. Multiple pivot languages and strategic decoder initialization helps neural machine translation.

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, Belgium, Brussels. Association for Computational Linguistics.

J. Tiedemann. 2012. Parallel data, tools and interfaces in opus. Association for Computational Linguistics.

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

Liqing Wang and Jiaquan Li. 2023. Research on chinese-lao neural machine translation based on multi-pivot.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation.