# Improving Similar Language Translation With Transfer Learning

**Ife Adebara** **Muhammad Abdul-Mageed**
Natural Language Processing Lab
The University of British Columbia
{ife.adebara,muhammad.mageed}@ubc.ca

## Abstract

We investigate transfer learning based on pre-trained neural machine translation models to translate between (low-resource) similar languages. This work is part of our contribution to the WMT 2021 Similar Languages Translation Shared Task where we submitted models for different language pairs, including French-Bambara, Spanish-Catalan, and Spanish-Portuguese in both directions. Our models for Catalan-Spanish (82.79 BLEU) and Portuguese-Spanish (87.11 BLEU) rank top 1 in the official shared task evaluation, and we are the only team to submit models for the French-Bambara pairs.

## 1 Introduction

We present the findings from our participation in the WMT 2021 Similar Language Translation shared task 2021, which focused on translation between similar language pairs in low-resource settings. The similar languages task focuses on building machine translation (MT) systems for translation between pairs of similar languages, without English as a pivot language.

Similarity between languages interacts with MT quality, usually positively (Adebara et al., 2020). Languages described as similar usually share certain levels of mutual intelligibility. Depending on the level of closeness, certain languages may share orthography, lexical, syntactic, and/or semantic structures which may facilitate translation between pairs. However, **(a) scarcity of parallel data** that can be used for training MT models remains a bottleneck. Even high-resource pairs can suffer from **(b) low-data quality**. That is, available data is not always guaranteed to be actual bitext with target standing as a translation of source. In fact, some open resources such as OPUS (Tiedemann, 2012a; Tiedemann et al., 2015) can suffer from noise such as when the source and target sentences belong to the same language. In this work, we tackle both **(a) scarcity** and **(b) low-data quality**. For **a**, we use simple knowledge transfer from already trained MT models to the downstream pair. For **b**, we use a simple procedure of language identification to remove noisy bitext where both the source and target are detected to be the same language or where source or target is identified as a different language from what it is expected to be.

The models we develop are for Spanish to Catalan (ES-CA), Catalan to Spanish (CA-ES), Spanish to Portuguese (PT-ES), Portuguese to Spanish (PT-ES), French to Bambara (FR-BM), and Bambara to French (BM-FR) language pairs[1]. Whenever possible, we choose an available MT model trained with the same source and target languages as our pair of interest. In cases where no such a model exists, we pick a model with either the source or the target language as our intended pair (Section 5). To show the utility of our transfer learning approach to the problem, we also train on one pair from scratch (which we treat as a *baseline*).

We experiment with tokenized **(primary models)** and untokenized **(contrastive models)** settings and compared the settings with models developed by fine-tuning pre-trained models as well as models trained from scratch. Our experiments show that the tokenized settings perform better than the untokenized settings for all language pairs. The model fine-tuned on top of the pre-trained MT model has higher performance than our baseline model from the first epoch compared with the model trained from scratch (for six epochs). Our models for the CA-ES and PT-ES language pairs *achieve top 1 rank in the offical shared task results*, with 82.96 and 47.71 BLEU scores respectively. In addition, we are the only team that submitted for the rest of the language pairs (i.e., ES-PT, FR-BM, and

[1] All models are available on https://github.com/fenimi/Similar-Languages-MT

*Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 273–278
November 10–11, 2021. ©2021 Association for Computational Linguistics

BR-FR).

The rest of our paper is organized as follows: we discuss related work in Section 2. We describe the data and pre-processing in Section 3. Next, we describe the data cleaning process in Section 4. In Section 5, we describe the models we developed for this task and we discuss the various experiments we perform. We also describe the architectures of the models we developed. Then we discuss the evaluation criteria in Section 6. Evaluation is done on both the validation and test sets. In section 7 we perform error analysis on the output of our models for some language pairs to determine the types of errors the models make. We conclude with discussion of the insights we gained from the shared task in Section 8.

## 2 Related Work

In recent times, there has been an increase of research interest in low-resource MT scenarios (Jawahar et al., 2021; Baziotis et al., 2020). NMT models, specifically those based on the Transformer architecture, have been shown to perform well when translating between similar languages (Przystupa and Abdul-Mageed, 2019; Adebara et al., 2020; Barrault et al., 2019, 2020), low resource scenarios (Adebara et al., 2021), and in contexts not involving English (Fan et al., 2021).

Furthermore, pre-training techniques have been successful for many NLP tasks (Zoph et al., 2016; Durrani et al., 2021) including NMT (Aji et al., 2020; Weng et al., 2020). Self-supervised pre-training acquires general knowledge from a large amount of unlabeled monolingual or multilingual data to improve the training process of downstream tasks (Aji et al., 2020; Devlin et al., 2018). The pre-trained model acquires some syntactic and semantic knowledge which can be transferred as initialized parameters to improve NMT models and translation quality (Goldberg, 2019; Jawahar et al., 2019; Aji et al., 2020). The intuitive justification for using pre-trained models is that the embedding space becomes more consistent, with semantically similar words closer together.

The knowledge from pre-trained language models (LMs) can be used to initialize the NMT model before training it on parallel data. However, there are certain limitations for MT tasks. First, LMs cannot be easily fine-tuned for MT tasks. Second, there is a discrepancy between pre-training objectives for LMs and the training objective in MT. Existing pre-training approaches such as mBART rely on auto-encoding objectives to pre-train the models, which are different from MT. Furthermore, LMs learn to reconstruct all source tokens with some noises, while NMT learns to translate most source tokens and copy only a few of them. LM pre-training is said to copy about 65% of tokens, while NMT training needs to copy less than 10% (Knowles and Koehn, 2018). The unexpected knowledge/bias can be therefore propagated to the NMT model via pre-training, which may result in NMT models mistakenly copying source tokens to the target side (Liu et al., 2021). For instance, because copying behaviours can be learned, a source word such as “shoe” may be copied to the target by pre-training based NMT models instead of providing a translation. Therefore, fine-tuning MT models on pre-trained LMs still do not achieve adequate improvements.

In order to address the difference in training objectives that using pre-trained language models results in, we use pre-trained MT models to initialize our models. This is still a type of *transfer learning*.

Following the justification for pre-trained models, we hypothesize that two linguistically similar languages will share closer semantic and syntactic relationships. This is based on the assumption that the more similar the source and target languages, the more similar the syntax and semantic properties and the higher the gains from using pre-trained models will be. We now introduce our data.

## 3 Data

For our experiments, we use parallel data from OPUS (Tiedemann, 2012b). Our data are from the following language pairs Spanish and Catalan, Spanish and Portuguese, and French and Bambara. We use data in the two directions from each of these three pairs. Details about our data is in Table 1.

| Pair | Lang | Sent | Words |
|---|---|---|---|
| **ES-CA** | **ES** | 10M | 284.6M |
| | **CA** | 10M | 273.3M |
| **ES-PT** | **ES** | 4.1M | 86.6M |
| | **PT** | 4.1M | 82.7M |
| **FR-BM** | **FR** | 9.9K | 179.3K |
| | **BM** | 9.9K | 202.9K |

Table 1: Number of sentences and words for the training data used for each language pair. We also report the type token ratio (TTR) before and after tokenization.

### 3.1 Pre-Processing

We perform pre-processing using the Moses toolkit (Koehn et al., 2007). For each language not supported by Moses, we use the tokenization setting of the language it is translated to. This applies only to Bambara, for which we used tokenization for the French language. We perform data cleaning, as we explain next.

## 4 Data Cleaning & Analysis

We perform data cleaning on the ES-CA, CA-ES, ES-PT, and PT-ES language pairs. We do not clean the French and Bambara pairs because we had very few training sentences for these. For cleaning, we run the langid tool (Lui and Baldwin, 2012) on the concatenation of the source and target and remove sentences that are not identified as belonging to one or both of the language pair. In Table 2, we provide some examples of data points we remove from the training data during data cleaning. These examples are removed because the claimed language is different from the language predicted by langid. After cleaning, we are left with $\sim$ 10M clean sentences out of $\sim$ 18.3M sentences for the Spanish and Catalan pair, and $\sim$ 3.1M clean sentences out of $\sim$ 4.2M sentences for the Spanish and Portuguese pair, respectively. We note that removed data comprise large portions of each dataset, thus confirming our concerns about data quality.

| Sentence | Claimed | Predicted |
|---|---|---|
| *Animal Crossing:* | Spanish | English |
| *Pico de Santo Tomés* | Spanish | Portug. |
| *Quinto Sereno Sammonico* | Portug. | Italian |
| *La sombra del caudillo* | Catalan | Spanish |
| *Cultura del Nepal* | Catalan | Spanish |
| *Morts a Rāwalpindi* | Catalan | French |

Table 2: Examples removed from our training data. "Claimed" refers to the expected language as coming from source, while "predicted" is what langid.py identified.

## 5 Models

### 5.1 Primary and Contrastive Models

We developed our *primary* and *contrastive* models using Transformers from HuggingFace library (Wolf et al., 2019). The primary models were developed using tokenized data while the contrastive models employed untokenized data. For the tokenized setting, we used Moses tokenizer (as explained earlier) while the untokenized setting used the data just as they were made available to us by shared task organizers.

| Hyperparameter | Values |
|---|---|
| encoder layers | 6 |
| decoder layers | 6 |
| attention heads | 8 |
| hidden layers | 6 |
| embedding dimension | 512 |
| dropout | 0.0 |
| vocab size | 49,621 |

Table 3: Hyperparameter settings for the HuggingFace Marian Transformer models.

| | Model | #Epochs | #Highest |
|---|---|---|---|
| **FR-BM** | tok | 100 | 55 |
| **BM-FR** | tok | 100 | 60 |
| **ES-CA** | untok | 6 | 3 |
| | tok | 8 | 3 |
| **CA-ES** | untok | 7 | 7 |
| | tok | 8 | 8 |
| **ES-PT** | untok | 17 | 15 |
| | tok | 35 | 13 |
| **PT-ES** | untok | 18 | 16 |
| | tok | 34 | 23 |

Table 4: Description the number of epochs for training each model and the epoch with the highest BLEU score.

| Pair | Untokenized | Tokenized |
|---|---|---|
| es-ca | 77.3 | 86 |
| ca-es | 87.7 | 87.8 |
| es-pt | 46.9 | 47.3 |
| pt-es | 52.9 | 53.6 |
| fr-bm | - | 6.6 |
| bm-fr | - | 6.07 |

Table 5: BLEU scores on our Dev set.

We used the pre-trained NMT models developed by Helsinki-NLP on HuggingFace. We used pre-trained models closest to the language pairs we trained. For language pairs without existing pre-trained models, we used a close language pair with either the source or target matching one of our downstream task languages in a given pair. Specifically, we used the following Marian models released by Helsinki-NLP: es-ca (for ES-PT), ca-es (for CA-ES and PT-ES), fr-en (for FR-BM), and en-fr (for BM-FR).

As an example, we report the hyperparameters for the CA-ES *primary* model in Table 3. This model had the best BLEU and RIBES score for this

| Pre-Trained Model | Pair | Untokenized System | | | Tokenized System | | |
|---|---|---|---|---|---|---|---|
| | | BLEU | RIBES | TER | BLEU | RIBES | TER |
| ca-es | ca-es | 76.8 | 95.19 | 15.421 | **82.79** | **96.98** | **10.918** |
| es-ca | es-pt | 35.61 | 82.48 | 52.61 | 38.10 | 85.35 | 46.556 |
| ca-es | pt-es | 43.86 | 85.10 | 43.801 | **47.71** | **87.11** | **39.213** |
| fr-en | fr-bm | - | - | - | 1.32 | 24.79 | 97.89 |
| en-fr | bm-fr | - | - | - | 3.62 | 36.17 | 101.52 |

Table 6: BLEU, RIBES and TER Scores on the Test set for the Tokenized (primary) and untokenized (contrastive) configurations. The models for CA-ES and PT-ES language pairs were the best performing models highlighted in bold type.

| Pair | Category | Text |
|---|---|---|
| ES-CA | Source | Por consiguiente, el Fondo debe movilizarse para aportar una contribución financiera en favor de Bulgaria, Grecia, Lituania y Polonia. |
| | Untok Output | Por lo tanto , el Fondo debe movilizarse para que se conceda una contribución financiera a Bulgaria, Grecia, Lituania y Polonia. |
| | Tok Output | Per tant, el Fons s'ha de mobilitzar per aportar una contribució financera a favor de Bulgaria, Grècia, Lituània i Polònia. |
| | Reference | Per tant, el Fons s'ha de mobilitzar per aportar una contribució financera en favor de Bulgària, Grècia, Lituània i Polònia. |
| CA-ES | Source Text | A fi de reduir al mínim el temps necessari per mobilitzar el Fons, aquesta Decisió s'ha d'aplicar a partir de la data de la seva adopció, |
| | Untok Output | A fin de reducir al mínimo el tiempo necesario para movilizar el Fondo, esta Decisión debe aplicarse a partir de la fecha de su adopción, |
| | Tok Output | Con el fin de reducir al mínimo el tiempo necesario para movilizar el Fondo, esta Decisión se ha de aplicar a partir de la fecha de su adopción. |
| | Reference | Con el fin de reducir al mínimo el tiempo necesario para movilizar el Fondo, la presente Decisión debe aplicarse a partir de la fecha de su adopción, |
| ES-PT | Source Text | Posición del Parlamento Europeo de 6 de abril de 2017 (pendiente de publicación en el Diario Oficial) y Decisión del Consejo de 11 de mayo de 2017. |
| | Untok Output | Posição do Parlamento Europeu de 6 de A bril de 2017 ( pendente de publicação no Jornal Oficial) e D ecisão do Conselho de 11 de Maio de 2017. |
| | Tok Output | Posição do Parlamento Europeu de 6 de A bril de 2017 ( indiferente à publicação no Jornal Oficial da União Europeia e decisão do Conselho de 11 de Maio de 201 ). |
| | Reference | Posição do Parlamento Europeu de 6 de abril de 2017 (ainda não publicada no Jornal Oficial) e decisão do Conselho de 11 de maio de 2017. |
| PT-ES | Source Text | Os Estados-Membros transmitem os dados referentes ao transporte por vias navegáveis interiores no seu território nacional à Comissão (Eurostat). |
| | Untok Output | Los Estados miembros transmitirán a la Comisión (Eurostat) los datos relativos al transporte por vías navegables interiores en su territorio nacional. |
| | Tok Output | Los Estados miembros transmitirán a la Comisión los datos relativos al transporte por vías navegables interiores en su territorio nacional (convenientREAT) |
| | Reference | Los Estados miembros transmitirán los datos relativos al transporte por vías navegables interiores en su territorio nacional a la Comisión (Eurostat). |
| FR-BM | Source Text | vous pourriez peut-être organiser de petits groupes pour lire et discuter de ce livre, chapitre par chapitre. |
| | Tok Output | - An y'a mɛn kura in mɛn: Mama denmuso, an ka jamana de wa k'a furu min ma kɔnɔ. |
| | Reference | aw bɛse ka to ka mɔgɔw dalajɛka gafe in kalan; ani ka hakilina falenfalen kɛsigidaw kan kelen kelen. |
| BM-FR | Source Text | Hakilijigin ka ɲɛsin kɛnɛyabaarakɛlaw ma |
| | Tok Output | cher agent de santé villageois, |
| | Reference | cher agent de santé villageois, |

Table 7: Examples sentences from the various pairs and corresponding translations based on the untokenized and tokenized models. Examples are from the Dev set. We highlight the differences between the outputs from the untokenized model and the reference text with blue highlights and the differences between the tokenized model and the reference text in red highlights . It can be observed that the number of errors in the untokenized model (based simply on the number of blue highlights here) is larger than that in the tokenized model (less errors/red highlights

language pair. We trained each model for different number of epochs due to time and GPU constraints. We show the number of epochs each model is trained for and the epoch with the highest BLEU score in Table 4. We did not train any *contrastive* models for FR-BM and BM-FR pairs, so we report the training for the primary (tokenized) models only.

### 5.2 Baseline

We developed a single baseline model based on Transformers as implemented in Fairseq. This model does not use any pre-trained MT models nor tokenization. This model was developed for the ES-PT pair for six epochs and it achieved a BLEU score of 37.6. For comparison, we developed a model for the same pair (i.e., ES-PT) based on an already available pt-es pre-trained MT model. After six epochs, this ES-PT model employing transfer learning achieved 52.18 BLEU (thus significantly outperforming our baseline). Based on this result, we resumed with experiments for all other language pairs *without* including a baseline. Ideally, we would train such baseline models for all the pairs. However, due to limited time and GPU resources, we only trained a baseline for a single pair.

## 6 Evaluation

We evaluated our models on both the Dev and Test sets. We used the checkpoint with the best BLEU score as evaluated on DEV as our best model. We used a beam size of four during evaluation on both Dev and Test and evaluated on de-tokenized data.

**Evaluation on Dev set.** We report the results on the Dev sets for each language pair in Table 5. As explained, the models were trained with both tokenized and untokenized data. As Table 5 shows, the tokenized setting yielded the highest performance for *all* language pairs. We show sample outputs from our tokenized models (from Dev data) in Table 7.

**Evaluation on Test set.** Our Test set performance was evaluated by the shared task organizers using BLEU (Papineni et al., 2002), RIBES, and TER (Snover et al., 2009). We report the scores in Table 6. Each of our CA-ES and PT-ES models ranked top 1 based on the official shared task results. In addition, we were the only team to submit models in the official competition for the French-Bambara pairs. As with the Dev set, the tokenized setting gave the highest performance for all language pairs.

## 7 Effect of Language Similarity

In order to gain some insight into the interference of similarity between languages of a given pair, we performed an analysis based on Levenstein distance that allows us to identify the percentage of cognates shared between the languages. We then compared system output to the reference sentences, trying to quantify how much the system is able to translate cognates correctly (in this case the correct translation will have the same cognate word in the target as it is in the source). We performed this analysis for one language pair: CA-ES and found that the model learned the cognates correctly up to 80% of the time.

## 8 Conclusion

We describe our contribution to the WMT2021 Similar Languages Translation Shared Task. We develop models for ES-CA, CA-ES, ES-PT, PT-ES, FR-BM, and BM-FR and show the improvement our models make with tokenized data when compared to untokenized data. We also show the utility of transfer learning based on fine-tuning NMT pre-trained models. Future work can investigate how the choice of pre-trained models affects the downstream tasks.

## Acknowledgements

We gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences Research Council of Canada (SSHRC), Compute Canada (www.computecanada.ca), and UBC ARC–Sockeye (https://doi.org/10.14288/SOCKEYE).

## References

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2021. Translating the unseen? yoruba-english mt in low-resource, morphologically-unmarked settings. *arXiv preprint arXiv:2103.04225*.

Ife Adebara, El Moatez Billah Nagoudi, and Muhammad Abdul Mageed. 2020. Translating similar languages: Role of mutual intelligibility in multilingual transformers. In *Proceedings of the Fifth Conference on Machine Translation*, pages 381–386, Online. Association for Computational Linguistics.

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710.

Loïc Barrault, Magdalena Biesialska, Ondrej Bojar, M. Costa-jussà, C. Federmann, Yvette Graham, Roman Grundkiewicz, B. Haddow, M. Huck, E. Joanis, Tom Kocmi, Philipp Koehn, Chi kiu Lo, Nikola Ljubesic, Christof Monz, Makoto Morishita, M. Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *WMT@EMNLP*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. *arXiv preprint arXiv:2004.14928*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep nlp models? *arXiv preprint arXiv:2105.15179*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-Centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Yoav Goldberg. 2019. Assessing Bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2021. Exploring text-to-text transformers for English to Hinglish machine translation with synthetic code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 36–46, Online. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does Bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Rebecca Knowles and Philipp Koehn. 2018. Context and copying in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3034–3041.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL companion volume proceedings of the demo and poster sessions*, pages 177–180.

Xuebo Liu, Longyue Wang, Derek F Wong, Liang Ding, Lidia S Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the copying behaviors of pre-training for neural machine translation. *arXiv preprint arXiv:2107.08212*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

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 on ACL*, pages 311–318. ACL.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the 4th Conference on MT (Volume 3: Shared Task Papers, Day 2)*, pages 224–235.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268.

Jörg Tiedemann. 2012a. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Jörg Tiedemann. 2012b. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Languages Resources Association (ELRA).

Jörg Tiedemann, Filip Ginter, and Jenna Kanerva. 2015. Morphological segmentation and opus for finnish-english machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 177–183.

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2020. Acquiring knowledge from pre-trained model to neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9266–9273.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.