Translating Similar Languages: Role of Mutual Intelligibility in Multilingual Transformers

Ife Adebara   **El Moatez Billah Nagoudi**   **Muhammad Abdul Mageed**

Natural Language Processing Lab

University of British Columbia

{ife.adebara,moatez.nagoudi,muhammad.mageed}@ubc.ca

Abstract

We investigate different approaches to translate between similar languages under low resource conditions, as part of our contribution to the WMT 2020 Similar Languages Translation Shared Task. We submitted Transformer-based bilingual and multilingual systems for all language pairs, in the two directions. We also leverage back-translation for one of the language pairs, acquiring an improvement of more than 3 BLEU points. We interpret our results in light of the degree of mutual intelligibility (based on Jaccard similarity) between each pair, finding a positive correlation between mutual intelligibility and model performance. Our Spanish-Catalan model has the best performance of all the five language pairs. Except for the case of Hindi-Marathi, our bilingual models achieve better performance than the multilingual models on all pairs.

1 Introduction

We present our findings from our participation in the WMT 2020 Similar Language Translation shared task, which focused on translation between similar language pairs in low-resource settings. Similar languages share a certain level of mutual intelligibility that may aid the improvement of translation quality. Depending on the level of closeness, certain languages may share similar orthography, lexical, syntactic, and or semantic structures which may make translation more accurate.

The level of mutual intelligibility is such that speakers of one language can understand another language without prior instruction in that other language. They can also communicate without the use of a lingua franca which is a link or vehicular language used for communicating between speakers of different languages (Gooskens, 2007). It is important to mention that, sometimes, the level of intelligibility varies in both directions. For instance, Slovene - Croatian intelligibility is said to

be asymmetric such that speakers or Slovene can understand spoken and written Croatian better than speakers of Croatian understand Slovene (Golubović and Gooskens, 2015).

Machine translation of similar languages has been explored in a number of works (Hajic, 2000; Currey et al., 2016; Dabre et al., 2017). This can be seen as part of a growing need to develop models that translate well in low resource scenarios. The goal of the current shared task is to encourage researchers to explore methods for translating between similar languages. We also view the shared task as useful context for studying interaction between degrees of similarity and mutual intelligibility on the one hand, and model performance on the other hand. We explore the use of bilingual and multilingual models for all the 5 shared task language pairs. We also perform back-translation for one language pair.

In the remainder of this paper, we discuss related literature in Section 2. We explain the methodology which includes a description of the Transformer model, back-translation and beam search in Section 3. In Section 4, 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 procedure in Section 6. Evaluation is done on both the validation and test sets. We conclude with discussion and the insights gained from this task in Section 7.

2 Related Work

Translation between similar languages has recently attracted attention. Different approaches have been adopted using state-of-the-art techniques, methods, and tools to take advantage of the similarity between languages even in low resource scenarios. Approaches that have been effective for other ma-

chine translation tasks have proven to achieve success in the context of similar language translation as well.

NMT models, specifically the Transformer architecture, has been shown to perform well when translating between similar languages (Baquero-Arnal et al., 2019; Przystupa and Abdul-Mageed, 2019). The use of in-domain data for fine-tuning has also proven to be of remarkable benefit for this task. This problem has also been tackled both by using character replacement to leverage the orthographic and phonological relationship between closely related mutually intelligible language pairs (Chen and Avgustinova, 2019). A new approach was also introduced for this task using a two-dimensional method that assumes that each word of the target sentence can be explained by all the words in the source sentence (Baquero-Arnal et al., 2019).

Within the realm of MT for low resource languages, recent work has focused on translation using large monolingual corpora due to the scarcity of parallel data for many language pairs (Lample et al., 2018, 2017; Artetxe et al., 2018b). These approaches have leveraged careful initialization of the unsupervised neural MT model using an inferred bilingual dictionary, sequence-to-sequence language models, and back-translation to achieve remarkable results. The bilingual dictionary is built without parallel data by using an unsupervised approach to align the monolingual word embedding spaces from each language (Conneau et al., 2017; Artetxe et al., 2018a). Since parallel data is not available in sufficiently large quantities, back-translation is used to create pseudo parallel data. The monolingual data of the target language is translated into the source using an existing translation system (e.g., one trained with available gold data). The output is then used to train a new MT model (Sennrich et al., 2015a). Weak supervision caused by back-translation results in a noisy training dataset. This eventually can affect translation quality.

More recent works adopt different approaches to manage noise in back-translation. For instance, phrase based statistical MT models are introduced as a posterior regularization during the back-translation process to reduce the noise and errors of the data generated (Ren et al., 2019). Another method (Artetxe et al., 2019b) uses cross lingual word embeddings incorporated with sub-word information. The weights of the log-linear model

is then tuned through an unsupervised process and the entire system is jointly refined in opposite directions to improve performance. This method outperforms previous SOTA model with about 5-7 BLUE points. A re-scoring mechanism that re-uses the pre-trained language model to select translations generated through beam search has also been found to improve fluency and consistency of translations (Liu et al., 2019). Yet another approach, combines cross-lingual embeddings with a language model to make a phrase-table (Artetxe et al., 2019a). The resulting system is then used to generate a pseudo parallel corpus with which a bilingual lexicon is derived. This approach can work with any word or cross-lingual embeddings techniques.

3 Methodology

Motivated by the success of Transformers and back-translation, we develope a sequence-to-sequence approach using the Transformer architecture perform back-translation for one language pair. For decoding, we use Beam Search (BS). BS is an heuristic decoding strategy based on exploring the solution space and selecting a sequence of words that maximize the overall likelihood of the target sentence. During the translation, we hold a beam of β sequences (*beam size*) which are iteratively extended. At each step, β words are selected to extend each of the sequences in the beam, so the output is β^2 candidate sequences (hypotheses), we retain only the β highest score hypotheses for the next step (top-β candidates) (Koehn, 2009). In all our experiments we use beam size of 5 whilst decoding.

3.1 Transformer

Our baseline models are based on the Transformer architecture. A Transformer (Vaswani et al., 2017) is a sequence-to-sequence model that does not have the recurrent architecture present in Recurrent Neural Networks (RNNs). It uses a positional encoding that can remember how sequences are fed into the model. These positions are added to the embedded representation (n-dimensional vector) of each word. Transformers have been shown to train faster than RNNs for translation tasks.

The encoder and decoder in a Transformer model have modules that consist mainly of multi-head attention and feedforward layers. The attention mechanism is based on a function that operates on Q (*queries*), K (*keys*), and V (*values*). The query

is a vector representation of one token in the input sequence, K refers to the vector representations of all the tokens in the input sequence. More information about the Transformer are in (Vaswani et al., 2017).

3.2 Back-translation

We perform back-translation using the monolingual model developed for the Croatian-Slovene (HR-SL) language pair. We use the best HR-SL model checkpoint that acquire the highest BLEU score on the DEV set to translate the monolingual HR data. This produces synthetic Slovene (SL) data which we then use as the source language while the original monolingual data is used as target when training the SL-HR model. We combine this data with the initial training data.

Due to time constraints, we used only a subset of the monolingual data with a beam size of 5.

3.3 Jaccard Similarity

Jaccard similarity compares similarity, diversity, and distance of data sets (Niwattanakul et al., 2013). It is calculated between two data sets in our case two languages) A and B by dividing the number of features common to the two sets (their intersection) by the union of features in the two sets, as in (1) below:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

We use tokens, identified based on white space, as features when we calculate Jaccard.

4 Experiments

4.1 Model Architecture

Our neural network models are based on the Transformer architecture (as described in Section 3) implemented by Facebook in the Fairseq toolkit. The following hyper-parameter configuration was used: 6 attention layers in the encoder and the decoder, 4 attention heads in each layer, embedding dimension of 512, maximum number of tokens per batch was set to $4,096$, Adam optimizer with $\beta1 = 0.90$, $\beta2 = 0.98$, dropout regularization was set to 0.3, weight-decay was set at 0.0001, $label - smoothing = 0.1$, variable learning rate set at $5e-4$ with the inverse square root, lr-scheduler and $warmup - updates = 4,000$ steps. We used the label smoothed cross-entropy criterion, and gradient clip-norm threshold was set to 0.

5 Data

We used all the parallel data for all language pairs `http://www.statmt.org/wmt20/similar.html`. The task was constrained so we did not add any additional data to develop our models. We used the monolingual data for the SL-HR language pair for back-translation. Table 2 shows the size of the data in terms of the number of sentences and words for each language pair while Table 1 shows example source and corresponding outputs from our bilingual and multilingual models for each language pair. We also calculated the jaccard similarity for the training data we used for the tasks. "Jaccard similarity" measures the similarity between two text documents by taking the intersection of both and dividing it by their union. Linguists measure these intersections (Oktavia, 2019; Gooskens and Swarte, 2017) between languages to determine the level of mutual intelligibility as well as classify languages as dialects of the same language or different languages. We calculated Jaccard similarity for each language pair.

5.1 Pre-processing

Pre-processing was by a regular Moses toolkit (Koehn et al., 2007) pipeline that involved tokenization, byte pair encoding and removing long sentences. We applied Byte-Pair Encoding (BPE) (Sennrich et al., 2015b) operations, learned jointly over the source and target languages. For each language pair, we used 32k split operation for subword segmentation (Sennrich et al., 2016b). We run experiments with Transformers under three settings, as we explain next.

5.2 Models

We develop both bilingual and multilingual models using gold data for all pairs. For one pair, we also use back-translation with one bilingual model. We provide more details next.

5.2.1 Bilingual Models

In this setting, we build an independent model for each language pair. We develop models for both directions for all language pairs, thus ultimately creating 12 models (6 for each direction). We train each model on 1 GPU for 7 days.

5.2.2 Multilingual Models

We develop two multilingual models that translates between all languages; a model for each direction (2 models overall) (Johnson et al., 2017). We add

Model	Pair	Sentence	Translation
Bilingual Model	**Es-Ca**	Diseña stickers para soñar	Dissenya stickers per soyir
		Mueva el diez de corazones al nueve de corazones .	Mobles el deu de garrons al nou de garrons .
	Es-Pt	Diseña stickers para soñar	Design stickers para sonhar
		Mueva el diez de corazones al nueve de corazones .	Muda o dez corações para nove corações .
	Sl-Hr	Vesel pomladni pozdrav ob novi izdaji Bisnode novičk .	Sretan proljetni pozdrav na novom izdanju Bisnode Vijesti .
		Z lepimi pozdravi ,	S lijepim pozdravima ,
	Sl-Sr	Danes ni enostavno slediti vsem informacijam , ki so pomembne za poslovanje .	Danas nije lako pratiti sve informacije koje su važne za poslovanje .
		Iščete podatke za drugo državo ?	Tražite podatke za drugu zemlju ?
Multilingual Model	**Es-Ca**	Mueva el diez de corazones al nueve de corazones .	Muva el 10 de coração al 9 de coraons .
		el cuatro de diamantes	el quatre de diamants
	Es-Pt	Luche en el aire con un avión enemigo	Luche en l aire amb un avió enemigo
		Entonces , ¿ qué salió mal ?	Então , o que saiu mal ?
	Sl-Hr	Objašnjenje - Indeks plaćanja Datum :	Objašnjenje – Indeks plaćanja Datum :
		Poštovani ,	Poštovani ,
	Sl-Sr	Iščete podatke za drugo državo ?	Tražite podatke za drugu zemlju ?
		Vesel pomladni pozdrav ob novi izdaji Bisnode novičk .	Sretan proljetni pozdrav uz novo izdanje Bisnode novosti .

Table 1: Examples sentences from the various pairs and corresponding translations based on the bilingual and multilingual models. Examples are from the DEV set.

	Language	#sentences	#words
Hi-Mr	hi	43.2K	829.9K
	mr	43.2K	600K
	mono-hi	113.5M	4.74B
	mono-mr	4.9M	112.6M
Es-Ca	es	11.3M	150.4M
	ca	11.3M	163M
	mono-es	58.4M	1.5B
	mono-ca	28M	763.7M
Es-pt	es	4.15M	86.6M
	pt	4.15M	82.5M
	mono-es	58.4M	1.47B
	mono-pt	11.4M	233.9M
Sl-Hr	sl	17.6M	113.09M
	hr	17.6M	117.73M
	mono-sl	46.25M	770.6M
	mono-hr	64.5M	1.24B
Sl-Sr	sl	14.1M	79.1M
	sr	14.1M	86.1M
	mono-sl	46.2M	770.6M
	mono-sr	24M	489.9M

Table 2: Number of sentences and words for the training data used for each language pair

a language code representing the target language as the start token for each line of the source data. We train each model on 4 GPUs for 7 days. We use the same hyper-parameters values set for the bilingual models. Multilingual models enable us to determine the impact of learning similar languages with a shared representation.

5.2.3 Bilingual Model with Back-translation

For the third approach, we combine back-translation with the bilingual translation model for the SL-HR language pair. We incorporated the monolingual data to do this. This was influenced by report (Sennrich et al., 2016a) in literature on the significant improvement of translation quality when monolingual data is incorporated into training data through back-translation. We were able to test the effect of back-translation on one model.

6 Evaluation

We evaluated both the DEV and TEST sets. We used the best-checkpoint metric with BLEU score to evaluate the validation set at each iteration. We used a beam size of 5 during the evaluation . We de-tokenized from BPEs back into words.

6.1 Evaluation on DEV set

We report the results on the DEV sets for each language pair in Table 3. These models were trained without the monolingual data except the SL-HR pair with the asterisk.

Pair	Bilingual Models	Multiling. models
hi-mr	12.14	16.35
mr-hi	10.63	01.02
es-ca	74.85	16.13
ca-es	74.24	64.57
es-pt	46.71	26.41
pt-es	41.12	05.81
*sl-hr	36.89	-
sl-hr	33.28	09.25
hr-sl	55.51	07.94
sl-sr	40.80	32.80
sr-sl	39.80	06.97

Table 3: Evaluation in BLEU on the development set for the different language pairs. The asterisk shows the model with back-translation

The bilingual models outperform the multi-lingual models for all language pairs except the hi-mr language pair.

6.2 Evaluation on TEST

In order to evaluate the test data, we removed the byte-pair code from the test set. We used the fairseq-generate mode while translating the test set. We show results on the test set in Table 4.

Pair	Bilingual Models	Multiling. models
hi-mr	0.49	-
es-ca	41.74	8.49
ca-es	45.23	45.86
es-pt	23.35	17.06
pt-es	24.26	21.55
*sl-hr	20.92	22.26
hr-sl	14.94	7.37
sl-sr	14.7	20.18
sr-sl	19.46	11.37

Table 4: The BLEU scores for some of the models on the test set. The language pair with the asterisks has back-translation [1]

6.3 Discussion

We used the Jaccard similarity to measure the level of mutual intelligbility. Figure 1 shows a positive correlation between the BLEU scores and the Jaccard similarity between each language pair. [2] This relationship hold both for the bilingual and multilingual models. One exception is the Slovene-Serbian pair (SL-SR) where higher similarity does not translate into higher BLEU. For example, the SL-SR BLEU is below the SL-HR BLEU even though the latter pair has a higher similarity score. Interpreting Jaccard to mean mutual intelligibility, our findings imply a higher intelligibility is correlated with higher BLEU scores. However, there is a need to further investigate this relationship due to the SL-SR we observe.

7 Conclusion

We described our contribution to the WMT2020 Similar Languages Translation Shared Task. We developed both bilingual and multilingual models for all pairs, in both directions. We showed back-translation to help improve performance on one pair. We also showed how mutual intelligibility between a pair of languages (measured by Jaccard similarity) positively correlate with model

---

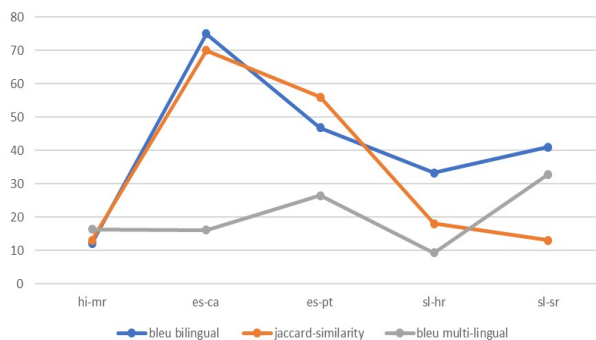[2]We multiplied the jaccard similarity by 100 to reduce the range of values on the y axis.

Figure 1: Interaction between performance in BLEU and Jaccard similarity.

performance (in BLEU). Future work can focus on exploiting other similarity metrics and providing a more in-depth study of mutual intelligibility between similar languages and how it interacts with MT model performance both in bilingual and multilingual models. The utility of back-translation on pairs we have not studied can also be fruitful.

Acknowledgements

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In Thirty-Second AAAI Conference on Artificial Intelligence.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019a. Bilingual lexicon induction through unsupervised machine translation. arXiv preprint arXiv:1907.10761.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019b. An effective approach to unsupervised machine translation. arXiv preprint arXiv:1902.01313.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041.

Pau Baquero-Arnal, Javier Iranzo-Sánchez, Jorge Civera, and Alfons Juan. 2019. The mllp-upv spanish-portuguese and portuguese-spanish machine translation systems for wmt19 similar language translation task. In Proceedings of the 4th Conference on MT (Volume 3: Shared Task Papers, Day 2), pages 179–184.

Yu Chen and Tania Avgustinova. 2019. Machine translation from an intercomprehension perspective. In Proceedings of the 4th Conference on MT (Volume 3: Shared Task Papers, Day 2), pages 192–196.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. arXiv preprint arXiv:1710.04087.

Anna Currey, Alina Karakanta, and Jon Dehdari. 2016. Using related languages to enhance statistical language models. In Proceedings of the NAACL Student Research Workshop, pages 116–123.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation, pages 282–286.

Jelena Golubović and Charlotte Gooskens. 2015. Mutual intelligibility between west and south slavic languages. Russian Linguistics, 39(3):351–373.

Charlotte Gooskens. 2007. The contribution of linguistic factors to the intelligibility of closely related languages. Journal of Multilingual and multicultural development, 28(6):445–467.

Charlotte Gooskens and Femke Swarte. 2017. Linguistic and extra-linguistic predictors of mutual intelligibility between germanic languages. Nordic Journal of Linguistics, 40(2):123–147.

Jan Hajic. 2000. Machine translation of very close languages. In 6th Applied NLP Conference, pages 7–12.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

Philipp Koehn. 2009. Statistical machine translation. Cambridge University Press.

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.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. arXiv preprint arXiv:1804.07755.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. arXiv preprint arXiv:1908.05925.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In Proceedings of the international multiconference of engineers and computer scientists, volume 1, pages 380–384.

Diana Oktavia. 2019. Understanding new language: Mutual intelligibility in romance language pair. Journal Of Language Education and Development (JLed), 2(1):180–187.

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.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with smt as posterior regularization. arXiv preprint arXiv:1901.04112.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.