Can We Use Word Embeddings for Enhancing Guarani-Spanish Machine Translation?

Santiago Góngora, Nicolás Giossa, Luis Chiruzzo
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
{sgongora,nicolas.giossa,luischir}@fing.edu.uy

Abstract

Machine translation for low-resource languages, such as Guarani, is a challenging task due to the lack of data. One way of tackling it is using pretrained word embeddings for model initialization. In this work we try to check if currently available data is enough to train rich embeddings for enhancing MT for Guarani and Spanish, by building a set of word embedding collections and training MT systems using them. We found that the trained vectors are strong enough to slightly improve the performance of some of the translation models and also to speed up the training convergence.

1 Introduction

In recent years the performance of machine translation systems has grown alongside with the rise of neural architectures (Zhang and Zong, 2020; Castilho et al., 2017) that infer the translation patterns while consuming a huge amount of data at training time. However, this high performance is hard to achieve when one (or both) of the languages is considered a low-resource language (Mager et al., 2018). That is the case for Guarani, an indigenous language spoken by nearly 10 million people in South America. It has the characteristic of being one of the few indigenous languages used for daily communication, both by people who identify with indigenous ethnicity as well as people who do not. According to the Paraguayan census office almost 70% of Paraguayans speak some form of Guarani at home[1], but despite this, it remains a low-resource language in the NLP community (Joshi et al., 2020), and the existing attempts at building machine translation systems for this language have not achieved very high results yet.

Qi et al. (2018) found that using pretrained word embeddings could be useful when building machine translation systems for low-resource scenarios. Considering the scarcity of Guarani-Spanish parallel text, the aim of this work is to evaluate if it is possible to enhance a MT system by incorporating word embeddings built with the available monolingual data. In order to do this, we first trained a set of word embedding collections and selected the best of these models according to some intrinsic tests. Finally we trained machine translation experiments using the different embeddings and compared them to the base scenario where no pretrained embeddings were used.

The intrinsic tests and other resources used in this paper are available on GitHub[2].

2 Related work

Although there have been some efforts on developing resources for Guarani, it remains largely under-explored in NLP. The current reference corpus for Guarani is COREGUAPA (Secretaría de Políticas Lingüísticas del Paraguay, 2019), it can be queried online but not be downloaded. Other resources include a Spanish-Guarani parallel corpus built from news sites and blogs (Chiruzzo et al., 2020), two corpora for sentiment analysis (Rios et al., 2014; Agüero-Torales et al., 2021), and a small Universal Dependencies corpus of the Mbya Guarani dialect (Thomas, 2019; Dooley, 2006). Except COREGUAPA, which cannot be downloaded, all of these resources are rather small for building accurate statistical models.

Interest towards machine translation for indigenous languages of the Americas has increased lately. An important antecedent is the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP) (Mager et al., 2021), which organized a shared task on MT from Spanish to several indigenous languages, including Guarani, with several participants. The test set for this shared

[1] https://www.ine.gov.py/news/news-contenido.php?cod-news=505

[2] https://github.com/sgongora27/Guarani-embeddings-for-MT

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 127 - 132
May 26-27, 2022 ©2022 Association for Computational Linguistics

task was a subset of the XNLI corpus (Conneau et al., 2018) translated to all languages. However, Guarani-Spanish machine translation still remains under-explored. There are some works that take into account the lack of available data (Alcaraz and Alcaraz, 2020; Gasser, 2018; Rudnick et al., 2014; Abdelali et al., 2006), or try to use the rich Guarani morphology to enhance the translation results Borges et al. (2021).

The use of word embeddings to enhance machine translation in low-resource scenarios has been previously explored (Qi et al., 2018), obtaining good results overall. They report that using pre-trained embeddings for both the source and target languages seem to improve results for translating low-resourced languages, but the improvement is much lower for languages with large amounts of data. Furthermore, (Shapiro and Duh, 2018) explores alternatives to include pre-trained embeddings in MT systems for a morphologically rich language, and (Nguyen and Chiang, 2017) uses a transfer learning approach for enhancing translation for a low-resource pair, but considering data from other related low-resources pairs as well.

3 Word embeddings

In a previous work (Góngora et al., 2021) we carried a first round of experiments with Guarani word embeddings, collecting text from news sites, tweets and the Guarani Wikipedia[3]. We classified each tweet in one of three categories (A: very reliable, B: reliable, and C: unreliable) according to the probability of being in Guarani using a heuristic based on the number of Guarani tokens from a frequent words list. Finally, for evaluating the then trained embeddings, we also presented two sets of intrinsic tests based on the original tests from Mikolov et al. (2013). One of them is a translation of the original *capital-common-countries* (*ccc*) set, while the other is a new set for *family* relations, inspired in the original one.

In the current work, we collected more data from the different sources and added datasets such as *The Bible*[4] and *The book of Mormon*[5]. We also translated the classic similarity test MC-30 (Miller and Charles, 1991) to Guarani in order to have another intrinsic test to perform (in addition to the *family* and *capital-common-countries* tests).

We trained a set of 24 different word embedding models in Guarani with different configurations. All of them were built using the gensim library (Řehůřek and Sojka, 2010) implementation of the word2vec C-Bow algorithm (Mikolov et al., 2013). The configurations differ in how much text was used (see below), the embeddings size (150 or 300) and the window size (6, 7 or 8). The number of tokens used in the different experiments varies between 1.9M and 2.7M depending on the different data sets we use, as shown in table 1. The *base text* set is used in all models, while some models also include the A, A+B, or A+B+C tweet sets.

Set	**Tokens**	**Sentences (s)** or **Tweets (t)**
The Bible	760,697	99,689 s
The Book of Mormon	204,434	58,995 s
Guarani Wikipedia	504,730	28,123 s
News	433.134	51,753 s
Base text (the four sets above)	1,902,995	238,560 s
Very reliable tweets (A)	11,791	811 t
Reliable tweets (B)	75,493	6,498 t
Unreliable tweets (C)	706,907	71,767 t
Total	2,697,186	

Table 1: Number of tokens for each of the sets used for training the word embedding models.

3.1 Analogy and Similarity tests

In order to perform a preliminary evaluation of these models we used the previously mentioned analogy (*family* and *ccc*) and similarity (*MC-30*) tests. Table 2 shows the results for these tests, indicating the configuration of each of the twenty-four models. The results of the analogy tests (*family* and *ccc*) are precision using top 1 (T1) or top 5 (T5) matches, while the similarity test (*MC-30*) is Spearman's rank correlation. In order to compare the performance we also include a row for a *baseline* consisting of the best result for each of the intrinsic tests achieved by the models in our previous work (Góngora et al., 2021), which were trained with size 150, window 7 and did not use any of the tweet sets.

Overall we can see a great improvement over the results of the analogy tests reported in the previous work (*baseline*), which can be explained in part because we are using a larger amount of text for training the models. However, there is a noticeable gap between the results for *family* and the *ccc* tests. This difference may be due to the type and style of texts used during training: neither the Bible nor

[3] https://dumps.wikimedia.org/gnwiki/ - February 2021.

[4] https://biblics.com/gn - July 2021.

[5] https://www.churchofjesuschrist.org/study/scriptures/bofm?lang=grn - July 2021.

Size	W	Tweets	family T1	family T5	ccc T1	ccc T5	MC-30
150	6	none	42.86	52.38	6.52	18.58	0.515
150	6	A	45.24	57.14	7.11	17.39	0.527
150	6	AB	42.86	52.38	7.71	18.77	0.530
150	6	ABC	45.24	52.38	4.15	15.42	0.500
150	7	none	**54.76**	54.76	9.09	18.77	0.440
150	7	A	50.00	52.38	7.11	15.61	0.556
150	7	AB	40.48	54.76	8.10	18.38	0.499
150	7	ABC	45.24	54.76	4.35	14.43	0.502
150	9	none	45.24	54.76	9.09	**21.34**	0.495
150	9	A	45.24	54.76	6.92	18.38	0.475
150	9	AB	50.00	54.76	7.31	17.19	0.449
150	9	ABC	42.86	52.38	6.52	19.17	0.460
300	6	none	45.24	47.62	7.91	17.59	**0.569**
300	6	A	42.86	54.76	8.10	17.79	0.473
300	6	AB	40.48	50.00	5.93	17.00	0.552
300	6	ABC	40.48	47.62	4.74	17.98	0.541
300	7	none	42.86	52.38	7.71	20.95	0.403
300	7	A	45.24	52.38	7.51	20.16	0.511
300	7	AB	50.00	**59.52**	**9.49**	18.97	0.512
300	7	ABC	40.48	52.38	8.70	17.79	0.538
300	9	none	50.00	54.76	6.52	17.59	0.519
300	9	A	45.24	57.14	7.71	18.38	0.521
300	9	AB	47.62	52.38	8.10	19.76	0.543
300	9	ABC	38.10	54.76	6.32	20.16	0.513
Baseline			41.27	48.41	5.53	13.37	-

Table 2: Results for the intrinsic evaluation of the 24 models trained. Maximum scores in bold, minimum scores underlined. *Baseline* refers to the best result for each test reported in our previous work (Góngora et al., 2021).

the Book of Mormon include modern countries and cities in their sentences. Also the Guarani Wikipedia is really small, even having some articles containing just a single line, so the occurrence of these kind of words is pretty low. Lastly the *ccc* test does not take into account South American countries, which might be the more likely ones to appear in our news set.

The results for the similarity test (*MC-30*) are good enough, ranging from 0.403 to 0.569, even compared to the state of the art for English[6] which ranges from 0.618 to 0.92 but trained with much larger resources. For this test we could not compare the results with a previous baseline since it was not used in our previous work.

4 Machine translation experiments

We carried a series of machine translation experiments to compare the use of randomly initialized embeddings with the use of different pretrained embedding configurations. All experiments were done using OpenNMT[7] with its default configuration, an encoder-decoder model implemented with stacked LSTMs and an attention model, so that the difference between experiments would only be the embeddings initialization.

For those models using pre-trained word embeddings we had to choose both the Spanish embeddings and the Guarani embeddings. For Spanish we chose a collection of size 300 trained by Azzinnari and Martínez (2016) using a corpus of 6 billion words. Due to limitations of OpenNMT, the Guarani embeddings size must also be 300. Therefore we chose some of the twenty-four models trained according to their size (300), their Spearman's correlation score for the *MC-30* test (see table 2) and the subsets of tweets used for training them:

- `s300w6none`: size 300, window 6, no tweets
- `s300w9ab`: size 300, window 9, tweets A+B
- `s300w7abc`: size 300, window 7, tweets A+B+C

We trained three translation models in each direction (Guarani-Spanish and Spanish-Guarani) using them as pre-trained word embeddings. We also trained an additional model in each direction without using pre-trained word embeddings (i.e. using *randomly* initialized embeddings). In all cases the models were trained for 80K steps — saving a checkpoint every 5K steps — using the training set from Chiruzzo et al. (2020) (*Train2020*) and the training set from the parallel data we presented in our previous work (Góngora et al., 2021) plus 383 new parallel sentences collected for this work (we call this union *Train2021*).

We then chose, for each model, the checkpoint that maximized the ChrF metric for the dev set (*Dev2020+Dev2021*). The test results will be reported over the test set from (Chiruzzo et al., 2020) (*Test2020*), the test partition of our own parallel set (*Test2021*), and the dev and test sets from (Mager et al., 2021) (*ANLP Dev* and *ANLP Test*), using the BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) scores. Table 3 shows the size of all the aforementioned datasets.

Corpus	Set Name	Sentences	Guarani Tokens	Spanish Tokens
Our parallel set	*Train 2021*	12,129	274,734	528,018
	Dev 2021	1,514	34,238	65,940
	Test 2021	1,532	34,597	68,805
(Chiruzzo et al., 2020)	*Train 2020*	11,501	214,727	304,012
	Dev 2020	1,481	26,606	37,355
	Test 2020	1,549	27,351	38,908
(Mager et al., 2021)	*ANLP Dev*	996	7,216	11,180
	ANLP Test	1,004	6,501	10,074

Table 3: Size of the parallel corpora partitions.

[6] `https://aclweb.org/aclwiki/MC-28_Test_Collection_(State_of_the_art)`

[7] `https://opennmt.net/`

Test Set	Test2020		Test2021		ANLP Dev		ANLP Test	
Models `Gn-Es`	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF
`random`	21.90	37.26	15.12	37.71	0.41	12.22	0.37	11.75
`s300w6none`	22.64	**38.63**	15.75	**39.13**	0.48	13.44	0.51	12.85
`s300w9ab`	22.49	38.32	15.85	38.76	0.44	13.52	0.44	**12.93**
`s300w7abc`	22.54	38.46	15.75	38.94	0.57	**13.65**	0.50	12.75
(Borges et al., 2021)	20.30	-	-	-	-	-	-	-
Models `Es-Gn`	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF
`random`	20.55	36.52	20.59	**37.08**	0.27	12.77	0.49	12.91
`s300w6none`	20.19	**36.95**	17.33	35.42	0.32	**13.10**	0.45	12.72
`s300w9ab`	19.75	35.13	20.24	36.23	0.36	12.49	0.17	**13.00**
`s300w7abc`	18.44	33.74	19.81	35.98	0.23	11.98	0.12	12.06
ANLP first place	-	-	-	-	-	-	6.13	33.6
ANLP *baseline*	-	-	-	-	-	-	0.12	19.3
ANLP last place	-	-	-	-	-	-	0.13	10.8

Table 4: BLEU and ChrF results of the translation experiments over the different test sets.

4.1 Guarani-Spanish

Figure 1 shows how BLEU and ChrF scores change at each checkpoint. We observe that, in general, models that use pretrained embeddings tend to converge earlier. This is particularly important when experimenting with several models and having little computing power available.

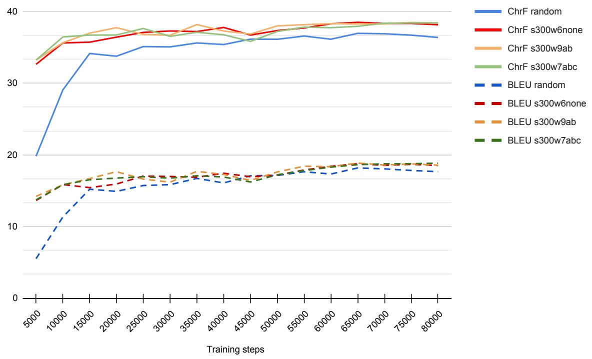

Figure 1: BLEU and ChrF evolution on the dev set for each checkpoint while training the `Gn-Es` models.

The top rows of table 4 shows the results over the test sets for the best model in each configuration. We also show the only result available for comparison in the direction `Gn-Es` (Borges et al., 2021), which used the (Chiruzzo et al., 2020) test corpus. We outperformed their results, which probably is because our models use more training data (they used only the train partition from Chiruzzo et al. (2020)).

We can also see that using pretrained word embeddings improved the performance with respect to the randomly initialized model on every test set. However, notice that the performance for the ANLP sets (Mager et al., 2021) drops dramatically. We think this could be explained by the more varied text styles present in these test sets, in contrast with the more uniform news text used for training.

4.2 Spanish-Guarani

Regarding the translation in the `Es-Gn` direction, figure 2 shows the results over the dev set and we can see the behavior is different. Although the faster convergence is observed again, the randomly initialized model performs as high as the pretrained ones. We can also see some performance stability problems as peaks in the graph. This behavior could be due to the target language embeddings being trained with fewer data, which is in line with what (Qi et al., 2018) reported.

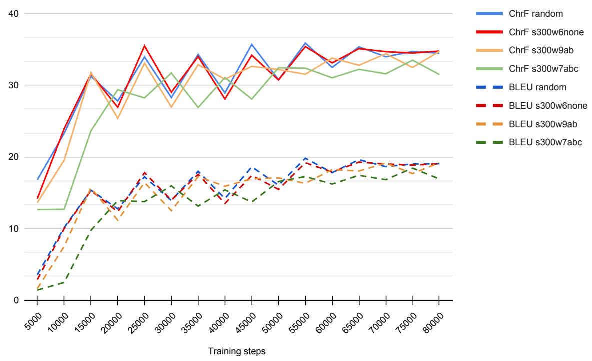

Figure 2: BLEU and ChrF evolution on the dev set for each checkpoint while training the `Es-Gn` models.

As can be seen in table 4 the results in this case are mixed, since the pretrained models do not outperform the randomly initialized model in all cases. Furthermore, the performance over the AmericasNLP sets also drops significantly, which probably has the same cause as the performance difference on the opposite direction.

In this direction it was possible to compare our best models with the performance obtained by AmericasNLP shared task participants (Mager et al., 2021). As shown in the bottom rows of table 4, our models perform between the bottom participants and the baseline. However, we did not aim to optimize the performance for this scenario: in this work we tried to focus only on analyze the use of pretrained word embeddings, and further work is needed to improve the training configurations with parameter tuning or different preprocessing techniques.

5 Conclusions

The results obtained in our experiments show that — with the currently available data — we can start to see some improvements when using pre-trained embeddings; at least in the `Gn-Es` direction. The performance of the `Gn-Es` models that used pretrained embeddings was slightly better than the performance of the one that did not use them. Additionally, the developed systems converge faster when using pretrained embeddings, which is especially useful in the scenario that is common for low-resource research labs, that of having little computing power. However, in the `Es-Gn` direction the results were more mixed, which is aligned with the conclusions of Qi et al. (2018).

There are still many lines to explore. First, trying other methods and algorithms for building embeddings such as FastText, which could be better for morphologically rich languages such as Guarani (Bojanowski et al., 2017; Shapiro and Duh, 2018). Second, we must explore the different OpenNMT configuration possibilities. We could also use back-translation techniques as well, such as the approach explored by (Vázquez et al., 2021), the winning system in AmericasNLP shared task. Finally more diverse text is needed, considering the difference observed while evaluating over the AmericasNLP sets. This diversity is also needed for improving the word embeddings performance. The great differences between both analogy tests suggests that the words in the *capital-common-countries* test might not be suitable for Guarani, perhaps due to the topics covered in Paraguayan news which refer mainly to countries in the region.

References

Ahmed Abdelali, James Cowie, Steve Helmreich, Wanying Jin, Maria Pilar Milagros, Bill Ogden, Hamid Mansouri Rad, and Ron Zacharski. 2006. Guarani: a case study in resource development for quick ramp-up mt. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,"Visions for the Future of Machine Translation"*, pages 1–9.

Marvin Agüero-Torales, David Vilares, and Antonio López-Herrera. 2021. On the logistical difficulties and findings of jopara sentiment analysis. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 95–102, Online. Association for Computational Linguistics.

NB Alvarenga Alcaraz and PR Alvarenga Alcaraz. 2020. Aplicación web de análisis y traducción automática guaraní–español/español–guaraní. *Revista Científica de la UCSA*, 7(2):41–69.

Agustín Azzinnari and Alejandro Martínez. 2016. Representación de Palabras en Espacios de Vectores.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Yanina Borges, Florencia Mercant, and Luis Chiruzzo. 2021. Using guarani verbal morphology on guarani-spanish machine translation experiments. *Procesamiento del Lenguaje Natural*, 66:89–98.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Robert A Dooley. 2006. Léxico guarani, dialeto mbyá com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística. *Cuiabá, MT: Sociedade Internacional de Lingüística*, 143:206.

Michael Gasser. 2018. Mainumby: un ayudante para la traducción castellano-guaraní. *arXiv preprint arXiv:1810.08603*.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. pages 202–217.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Adolfo A. Rios, Pedro J. Amarilla, and G. Giménez-Lugo. 2014. Sentiment categorization on a creole language with lexicon-based and machine learning techniques. *2014 Brazilian Conference on Intelligent Systems*, pages 37–43.

Alex Rudnick, Taylor Skidmore, Alberto Samaniego, and Michael Gasser. 2014. Guampa: a toolkit for collaborative translation. In *LREC*, pages 1659–1663.

Secretaría de Políticas Lingüísticas del Paraguay. 2019. Corpus de Referencia del Guaraní Paraguayo Actual – COREGUAPA. `http://www.spl.gov.py`. Accessed: 2021-03-13.

Pamela Shapiro and Kevin Duh. 2018. Morphological word embeddings for arabic neural machine translation in low-resource settings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 1–11.

Guillaume Thomas. 2019. Universal dependencies for mbyá guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. pages 255–264.

Jiajun Zhang and Chengqing Zong. 2020. Neural machine translation: Challenges, progress and future.