Learning Translation Model to Translate Croatian Dialects to Modern Croatian Language

Blagica Penkova[1], Maja Mitreska[1], Kiril Ristov[1], Kostadin Mishev[1,2] and Monika Simjanoska[1,2]
[1]iReason, LLC, Skopje, N. Macedonia
[2]Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering,
Skopje, N. Macedonia
{blagica.penkova, maja.mitreska, kiril.ristov, kostadin.mishev, monika.simjanoska}@ireason.mk,
{kostadin.mishev, monika.simjanoska}@finki.ukim.mk

***Abstract*—The task of translating dialects into modern language is a challenging task since it requires enormous parallel data. Such data is hard to find especially when it comes to low-resource languages. Among them is the Croatian language which has very few datasets in the standardized version, let alone enough resources for its dialects. In the Croatian language, there are three main groups depending on the geographical position Shtokavian, Kajkavian, and Chakavian which also include other more specific local dialects. For solving these kinds of problems, unsupervised neural machine translation (UNMT) models are considered a good solution since they can be trained on monolingual data. In this paper, we propose an application of a modified version of the state-of-the-art UNMT model for dialect translation on monolingual data of the standardized Croatian language and its dialects. We experimented with several types of cross-lingual embeddings of the input data to determine the best approach that can leverage the similarities and differences between the language and the dialect. All techniques are evaluated on a small parallel dataset using the BLEU metric. Translating these dialects to the modern Croatian language helps in improving communication and access to information for all speakers.**

***Keywords*—Unsupervised Neural Machine Translation, Dialect Translation, Low-resource languages, Cross-lingual embeddings, Croatian Language, Croatian Dialects.**

I. Introduction

Machine translation (MT) is an automatic approach to converting written text from one language to another. The translating task is widely explored in the field of Natural Language Processing (NLP) and many researchers are introducing novel solutions as the development of the technology continues. The translation task can be roughly divided into two groups: statistical MT and neural MT, which took a rise with the introduction of sequence neural networks [1]. The automatic MT replaces the need to use human resources to translate textual forms, hence accelerating the process of their generation. The translation task enables more content to be available to larger groups of audiences and therefore all the information is publicly accessible.

There is extensive research in this field and many different translation architectures have been covered throughout the years. Beginning with more traditional approaches from the statistical MT branch and continuing with state-of-the-art neural networks. The traditional approaches require extensive calculations and are more time and resources-consuming. On the other hand, the unsupervised approach of MT is gaining momentum since the need for parallel sentences cannot be always satisfied. This approach tries to find the underlying hidden meaning and characteristics of an unlabeled dataset. Since the unsupervised approach finds the internal structure of the dataset it solves the problems which emerge from the disadvantages of the supervised approach.

Most of the papers that propose using the unsupervised approach for machine translation are basing their architectures on using monolingual datasets in the two languages and construing the network around them. The establisher of the usage of monolingual data in an unsupervised approach is [2] where they provide several decipherment approaches. In [3], the authors map the two monolingual datasets into the same latent space and learn to reconstruct the sentences from the shared features. The state-of-the-art model [4] is the most widely used and explored model when experimenting with translation between two languages.

With the advancement of unsupervised approaches, not only the translation of two mostly standardized languages is enabled, but a number of possibilities emerge for re-purposing this approach for dialect translation. Dialect translation is an even more challenging task since many dialects can be found in spoken form rather than written form, especially when it comes to dialects spoken by a small number of people. Moreover, the acquisition of sentences in a dialect that comes from a low-resource language is particularly complex. Same to regular translation, dialect translation can be solved using supervised or unsupervised approaches. In the supervised approach, first, a number of parallel sentences are gathered and then phrase-based architectures [5] are used for learning the mappings of the words.

Some papers exploring this field focus on constructing the parallel dataset and then proposing appropriate architecture for solving the translation problem. In [6] a grapheme-to-phoneme (g2p) model is proposed for learning the Malay dialect translation. The vast majority of the papers that propose parallel datasets for supervised learning implement more methods from the statistical approach [7]–[10], or combine them with a language model [11]. The unsupervised approach also

developed popularity in the field of dialect translation since here even fewer resources are available. The pioneer in this field is [12] where they propose two systems, the first one is a standard attention sequence-to-sequence model with cosine similarity to capture the similarities between the language and the dialect, and a second system based on the Google NMT (GNMT) model. The paper explored the methods of the Modern Standard Arabic (MSA) language. In [13] and [14] Transformer based architectures are implemented like in [14] for Mandarian-Cantonese language translation.

Since most of the papers do not include specific language characteristics and leave it to the model to extract the hidden relationships, can be repurposed for other languages. The usage of monolingual data is immensely helpful when working with low-resource languages where the resources for standardized languages are scarce, and the resources in dialects are even scarcer. Such a language is the Croatian language, the standard language in Croatia. The Croatian language consists of three primary dialects: Shtokavian, Kajkavian and Chakavian. The Shtokavian dialect is the one that the standard Croatian language is based on. The dialects are named after the interrogatory pronouns "što", "kaj" and "ča". Since there is no modern literature written in these languages we combined them to create one model that can learn to distinguish the dialects from the standard Croatian language, and also learn the similarities and the differences between the dialects and the modern version.

In this paper, we propose developing a dialect translation model that can translate a dialect sentence into a sentence from the standard Croatian language. The model uses an unsupervised approach on two monolingual datasets. The first one consists only of Croatian sentences in the modern version and the other only dialect sentences. It is necessary to be mentioned that these sentences are not parallel and are acquired from various resources. So, the first contribution of this paper is constructing three datasets, one consisting only of Shtokavian sentences, one of the Kajkavian sentences and the last one consisting of sentences in Chakavian. All combined form the dialect dataset with 53K sentences. The second contribution of this work is that, to our knowledge, the dialect translation model is the first one in Croatian. There is no prior work in the dialect translation tasks for this kind of low-resource language, especially in Croatian, or any of the Slavic languages.

In the next few paragraphs, we talk about recent advancements in the field of UNMT architectures for dialect translation. In section III, we discuss the introduced dataset and the implemented translation model. In IV we present the results obtained through the experiment as well as their fluency with the help of a human expert in the Croatian language. Our work is concluded in the final section V where we present our findings. The presented work is made publicly available at [1].

II. Related Work

A major problem when dealing with low-resource languages is the lack of available data. In the course of our research, we found related work that allows translation to be performed without a parallel corpus [4]. This paper suggested a way to be able to translate using only monolingual data from the source and the target language. The paper proposes two model variants, a neural and a phrase-based model. Long Short-Term Memory [15] and Transformer cells [16] are the basis for the NMT models. For the Transformer, they use 4 layers both in the encoder and in the decoder. During the first generation, Moses' smoothed n-gram language model is used by default in the PBSMT. In the paper, five language pairs are considered: English-French, English-German, English-Romanian, English-Russian and English-Urdu. They use a dataset of around 2 million sentences of the source and the target language. For tokenization, they use Moses scripts.

The NMT is trained with a 60,000 Byte-Pair Encoding (BPE) tokenizer in the pre-processing phase so the vector embedding is kept in a reasonable dimension while PBSMT is trained with true-casing. Both models required either n-gram embeddings or cross-lingual BPE embeddings. With an embedding dimension of 512 and a context window of size 5 and 10 negative samples, the embeddings are made using fastText [17]. For NMT, fastText is used after the source and target corpora have been concatenated. For PBSMT, n-gram embeddings are first created independently on the source and target corpora, and they are then aligned using the MUSE library. The results in this paper demonstrate that the unsupervised NMT and PBSMT systems are significantly superior to previous unsupervised baselines. For the translation task, the Urdu-English result that the model achieves 12.3 BLEU using only a validation set of 1800 sentences and on the Russian-English translation task the model receives a BLEU score of 16.6.

Even though some languages are widely spoken, there is no known natural language processing (NLP) work on these languages, such as Pidgin English, which is the most widely spoken language in West Africa [18]. The goal of this paper is to translate from Pidgin English to English only by using a monolingual dataset. The model that is used in this paper is based on the previous paper, UNMT [4]. They used a Transformer with 10 attention heads. There are 4 encoder and 4 decoder layers with 3 encoder and decoder layers shared across both languages. They obtained the data from newspapers and collected a corpus of 56695 Pidgin sentences. They trained cross-lingual embedding via monolingual mapping where a linear mapping is learned between already trained monolingual word embedding. They got the English embeddings by using already pre-trained Glove English embeddings. And the Pidgin embeddings are done by training new word embedding on Pidgin (since they share common words with English) meaning first, they got the English embeddings and fine-tuned them with new Pidgin words. And the Unsupervised cross-lingual embeddings are made using MUSE. The training with an Unsupervised Neural Machine Translation model between Pidgin and English achieves BLEU scores of 7.93 from Pidgin to English and 5.18 from English to Pidgin.

Related research on dialect translation has been done in many languages. For instance, the multitask learning model proposed in this article [19]is for converting Arabic dialects

[1]https://github.com/Blagica88/CroatianDialectTranslation.git

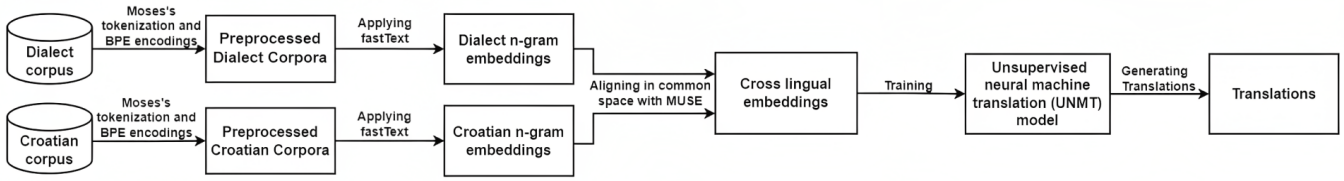

Fig. 1. The structure of our proposed methodology

into Modern Standard Arabic (MSA). In this study, an integrated neural machine translation model was developed and trained with decoders shared across all language pairs and each source language with its own encoder. Consequently, 13805 sentence pairs were trained for Levantine dialects (LD) and 17736 sentence pairs for Maghrebi dialects (MD) that are collected from TV shows, movies and social media. Arabic tokens are space-separated, and modern standard Arabic and English languages are tokenized using the Python tokenizer with default English settings. The findings in this publication demonstrate that, in comparison to the individually trained model, their suggested MTL model may guarantee superior translation quality.

This paper [14] is most similar to our problem because it attempts to exploit the commonality and diversity between dialects to create an unsupervised translation model that accesses only monolingual data. The translation task is between Mandarin and Cantonese, which are official languages and the most widely used dialect in China. Each of the MAN and CAN monolingual training corpora consist of 20M sentences and they also use parallel datasets for evaluation. There are two stages to the training procedure. 1) A commonality model that learns to identify common features in all dialects 2) Diversity modelling that creates connections between various expressions. The commonality modelling is done by training one model for both two dialects and diversity modelling is done with back translation. The model architecture is the same as in the previous paper [4]. They also use UNMT. The main difference is in the embeddings because they propose pivot-private embedding and layer coordination to jointly balance commonality and diversity. The Pivot learns to share a part of the features while the Private captures the word-level characteristics in different dialects. The results are that their model outperforms rule-based simplified and traditional Chinese conversion and conventional unsupervised translation models over 12 BLEU scores.

III. Methodology

In this section, we explain our proposed methodology demonstrated in Figure 1. We present the dataset and its creation, data pre-processing, the cross-lingual embeddings, and the utilization of the UNMT architecture to build our translation model.

A. Dataset

The biggest challenge was gathering enough data for the language model that would be used to translate dialects of Croatian into standard Croatian language. The data used to generate the Croatian dataset was collected from a "24sata" online news portal. The 24sata news portal covers a variety of news articles. More than 650,000 Croatian articles from 2007 to 2019 are included in the collection, along with tags [20]. Our dataset consists of 53K sentences extracted from the news articles from "24sata".

All the resources that we managed to find in Kaykavian, Shtokavian and Chakavian are before the standardization of the Croatian language hence all of them are old poems and/or dramas. We found the following books on the Shtokavian language: "Pesme Meneti I Ora Dria" collected by Vatroslav Jagic [21], "Novela od Stanca" by Marin Držić [22], "Pjesni ljubene" by Ore Dria [23],and "Razgovor ugodni naroda slovinskoga" by Andrije Kaia Mioia [24], "Satir Iliti Divji Čovik" by Matija Antun Relković [25], "Priče iz davnine" by Ivana Brlić Mažuranić [26], "Kvas bez kruha" by Antun Nemčić [27] and "Izabrane štokavske pjesme" by Fran Galović [28]. The following books are in Chakavian: "Skladanja izvarsnih pisan razliih" by Hanibala Lucia [29], "Ribanje i ribarsko prigovaranje" by Petra Hektorovia [30], "Jeđupka" and "Pelegrin Sabu Mietiu" by Mike Pelegrinovia [31], [32] and "Vazetje Sigeta Grada" by Brne Kar [33]. Additionally, we were only able to find one book, "Matija, Grabancija" by Tituš Brezovačk [34] on the Kajkavian dialect. The discovery of TV series on Croatian dialects had the largest influence on our dataset. We were able to find two written books for the series "Projsaci i Sinovi" by Ivan Raos [35] and "Velo Misto" by Miljenko Smoje [36] which are written in the Shtokavian and Kajkavian dialects, respectively. From all of the collected data, we extracted each sentence as a separate input sample. After cleaning, we were able to gather 53K sentences.

Even though we acquired monolingual data for training, parallel phrases were necessary for testing and validating our model. Since parallel sentences were required for the translation of Croatian dialects into the standardized form of the language, we collaborated with a human Croatian specialist to create them. As a result, our validation and test datasets include 30 sentences each in a dialect as well as 30 sentences in standard Croatian.

Moreover, we were able to find parallel word dictionaries for both Shtokavian and Chakavian and utilized them in the cross-lingual phase.

B. Cross-lingual Embedding

For the purposes of the paper and the model, before training the model on the monolingual data, we first need to embed the

words of the sentences to capture their context as well as the context of the words. The architecture described in III-C requires monolingual embeddings for both of the languages and shared cross-lingual embeddings learnt over the concatenated datasets.

1) Dialect Monolingual Embedding: The dialect dataset is first tokenized using Moses tokenizer and encoded using Byte-Pair Encoding (BPE). Furthermore, the dataset is used to train a fastText model to learn the representations of the word.

2) Croatian Monolingual Embedding: Similarly to the dialect monolingual embeddings, the Croatian monolingual embeddings are learnt using the fastText skip-gram model. The learnt embeddings are further used in the creation of cross-lingual embeddings.

3) Cross-lingual Embedding: For the cross-lingual embeddings, we experimented with two approaches. The first approach is based on [14] where besides the monolingual embeddings, we train another fastText model on the concatenated and shuffled dataset created from the two monolingual datasets. The second approach is based on MUSE [37] where the monolingual fastText embeddings are aligned in a common space to obtain multilingual word embeddings. The alignment is done using the unsupervised approach and only utilizes a bilingual dictionary of pairs of dialect words and standard words for the evaluation. The MUSE-aligned centred embeddings between the two languages are learnt using adversarial training and (iterative) Procrustes refinement. For the purposes of our work, we experiment with both approaches and obtained better results with the MUSE-aligned approach.

C. Architecture

We used [14] approach as a reference and built our model on top of their source code. The used architecture is a Transformer based. The number of layers in the encoder and decoder and the number of layers to share between the encoder and decoder were all assigned to 4. Adam is the optimizer utilized for the encoder and decoder, and the learning rate is 0.0001. The pre-trained language embeddings described above are employed in the training process. A larger embedding dimension can potentially capture more complex relationships between words and their context, and as a result, the dimensionality of our embeddings is set to 1024 and on shared embedding layers to 512. The number of training examples per epoch is 200 and the number of examples in each training batch is 32. Unlike [14] which uses cross-lingual embeddings consisting of pivot-private embeddings, we employ MUSE-aligned embedding on both of the datasets. That way we obtained higher results when evaluating the model.

D. Evaluation Metrics

The evaluation of the model is done using the BLEU metric which is standard in machine learning problems, especially in translating tasks. It measures the quality of the translated text to the original text. The standard definition of the BLEU metric is that it compares the machine-generated sentence to a set of reference sentences. The obtained score indicated how similar the candidate text is to the reference texts [38].

IV. Results

In this section, a comparison is made of the experiments done over the architecture with the two different embeddings. The cross-lingual shared embeddings learned with MUSE are in fact aligned embeddings from the two separate fastText models trained on separate language-specific data. For the second experiment, the fastText embeddings are trained on the concatenated dataset of the sentences in Croatian and in dialect. Moreover, the embeddings utilized in the training process of the model are in the form of shared-private, meaning that the final word embedding consists of shared embeddings learned from the concatenated dataset, and private embeddings learned from the language-specific datasets.

TABLE I
Results of the models during evaluation

Metric	Valid	Test
Dialects-Croatian	9.83	12.8
Croatian-Dialects	7.86	9.71
Dialects-Croatian-Dialects	42.27	29.07

Table I presents the best-obtained results from the first experiment with MUSE embeddings. We evaluated the experiments on a manually gathered and translated dataset with the help of a human expert. The models were trained to translate the sentences in both directions, dialect-to-standard language translation and standard-to-dialect language translation. Moreover, we included the dialect-standard-dialect back translation to check the quality of the translated sentences.

In Figure 2, four different BLEU scores are reported where each BLEU score indicates the degree of overlap between the predicted translations and reference translations in terms of n-gram precision. The notation is in the form of score metric, source language, target language, and type of dataset, where the source language and the target language can receive a 'di' value for the dialect sentences or a 'cr' value for the sentences in the standard Croatian language and the type of dataset can be validation or test set.

- **bleu_di_cr_valid:** Model's performance in translating from Croatian dialect to standard Croatian on the validation set.
- **bleu_cr_di_valid:** Model's performance in translating from standard Croatian to Croatian dialect on the validation set.
- **bleu_di_cr_test:** Model's performance in translating from Croatian dialect to standard Croatian on the test set.
- **bleu_cr_di_test:** Model's performance in translating from standard Croatian to Croatian dialect on the test set.

In Figure 3, two different BLEU scores are reported that measure the quality of the back-to-back translations:

- **bleu_di_cr_di_valid:** Model's performance in translating from dialect to standard Croatian and back to dialect on the validation set.
- **bleu_di_cr_di_test:** Model's performance in translating from dialect to standard Croatian and back to dialect on the test set.

Fig. 2. Comparison of BLEU CR-DI and DI-CR scores across epochs

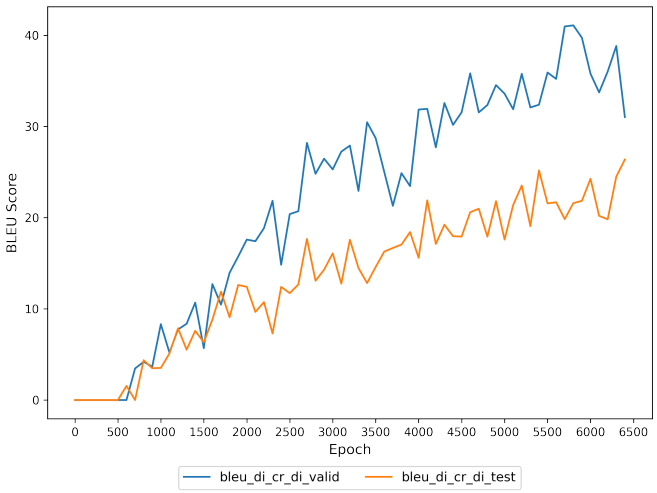

Fig. 3. Comparison of BLEU scores from the back translation across epochs

The flow of how the number of epochs affects our model's ability to learn more efficiently is depicted in Figure 2. For instance, when we tested the model to translate from dialect to standard Croatian language, the BLEU score in the 600 epoch is only 6, but as the number of epochs increases, so does the BLEU score. We currently have the best BLEU score in the 5000 epoch, which is 12.8. The lowest score was received in the validation phase when translating from Croatian to dialect languages. The comparison of BLEU scores from the back translation across epochs is shown in Figure 3. As we can see, we received the highest score of 42 BLEU in the validation phase when translating a back-to-back dialect-Croatian-dialect.

An effort was made to create a language model for translating Croatian dialects to the country's official language, and initial success was achieved. The model did in fact learn some interesting cases and captured the differences between the dialects and the standard language. In the examples below, we present the reference sentences written in the standard language and in dialect, and also the outputs of the model when the different versions of translations are done over the input sentences.

Example 1:

Reference sentence in standard language: E, moje dijete, tako je onda bilo.

Reference sentence in dialect language: E, moje dite, tako je unda bilo.

Dialect-Standard Translation: E, moje, dijete ako je unda bilo.

Standard-Dialect Translation: E moje dite drugačije je unda bilo

Dialect-Standard-Dialect Translation: E moje dite drugačije je bilo unda ništa.

Example 2:

Reference sentence in standard language: Ustao samo rano jer samo morao ići na tržnicu po rajčice.

Reference sentence in dialect language: Usto sam rano jer sam moro ić na tržnicu po rajčice.

Dialect-Standard Translation: Usto sam jer sam rano moro na tržnicu po suzama .

Standard-Dialect Translation: Ustao samo rano jer morao na tržnicu po kući.

Dialect-Standard-Dialect Translation: Usto sam jer sam ja moro na tržnicu po zdravlje rajčice.

From the outputs, it can be concluded that the model is able to translate some different words and make changes. As we can see from Example 1, the model was able to change the word "dite" which is in the Shtokavian dialect to "dijete" in the standard Croatian language. That is a sign that the model indeed learned the difference and has the ability to evolve and upgrade itself. However, the lack of sufficient data in the dialect dataset prevents the model from learning more effectively.

V. Conclusion

This paper represents the first-ever work done in the field of Croatian dialect translation to the standardized version of the language. Since it is the first one of its kind it has its deficiencies due to the lack of available digital data in all of the dialects. The unsupervised approach is proven to work in various types of settings and languages, but it requires a vast amount of monolingual data. Despite the need for no parallel data, it needs a lot of sentences to learn the underlying characteristics, similarities and differences between the languages so it can learn and align the words in a shared latent space. This sort of model can be used in speech-to-text systems in the Croatian language to serve as many customers as possible from different backgrounds and from different locations all over Croatia. At the moment we are trying to acquire as many dialect resources as possible to improve the results and create a system that can translate from Shtokavian, Kajkavian and Chakvian dialects to the standardized version of the Croatian language.

REFERENCES

[1] P. Koehn, "Neural machine translation," *arXiv preprint arXiv:1709.07809*, 2017.

[2] S. Ravi and K. Knight, "Deciphering foreign language," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 12–21. [Online]. Available: https://aclanthology.org/P11-1002

[3] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," *arXiv preprint arXiv:1711.00043*, 2017.

[4] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," *arXiv preprint arXiv:1804.07755*, 2018.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://aclanthology.org/P07-2045

[6] T.-P. Tan, S.-S. Goh, and Y.-M. Khaw, "A malay dialect translation and synthesis system: Proposal and preliminary system," in *2012 International Conference on Asian Language Processing*. IEEE, 2012, pp. 109–112.

[7] S. Kchaou, R. Boujelbane, and L. H. Belguith, "Parallel resources for tunisian arabic dialect translation," in *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, 2020, pp. 200–206.

[8] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaili, "Machine translation experiments on padic: A parallel arabic dialect corpus," in *Proceedings of the 29th Pacific Asia conference on language, information and computation*, 2015, pp. 26–34.

[9] F. Huang, "Improved arabic dialect classification with social media data," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2118–2126.

[10] S. Chakraborty, A. Sinha, and S. Nath, "A bengali-sylheti rule-based dialect translation system: Proposal and preliminary system," in *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2016, NEHU, Shillong, India*. Springer, 2018, pp. 451–460.

[11] W. Salloum and N. Habash, "Elissa: A dialectal to standard arabic machine translation system," in *Proceedings of COLING 2012: Demonstration Papers*, 2012, pp. 385–392.

[12] W. Farhan, B. Talafha, A. Abuammar, R. Jaikat, M. Al-Ayyoub, A. B. Tarakji, and A. Toma, "Unsupervised dialectal neural machine translation," *Information Processing & Management*, vol. 57, no. 3, p. 102181, 2020.

[13] M. Dare, V. F. Diaz, A. H. Z. So, Y. Wang, and S. Zhang, "Unsupervised mandarin-cantonese machine translation," *arXiv preprint arXiv:2301.03971*, 2023.

[14] Y. Wan, B. Yang, D. F. Wong, L. S. Chao, H. Du, and B. C. Ao, "Unsupervised neural dialect translation with commonality and diversity modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9130–9137.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[18] K. Ogueji and O. Ahia, "Pidginunmt: Unsupervised neural machine translation from west african pidgin to english," *arXiv preprint arXiv:1912.03444*, 2019.

[19] L. H. Baniata, S. Park, S.-B. Park *et al.*, "A neural machine translation model for arabic dialects that utilizes multitask learning (mtl)," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[20] M. Purver, R. Shekhar, M. Pranjić, S. Pollak, and M. Martinc, "24sata news article archive 1.0," 2021, slovenian language resource repository CLARIN.SI. [Online]. Available: http://hdl.handle.net/11356/1410

[21] V. Jagić, *Pjesme Šiška Menčetića Vlahovića i Gjore Držića*, ser. Stari pisci hrvatski. Zagreb: JAZU, 1870, vol. 2.

[22] M. Držić, *Novela od Stanca*. Zagreb: Školska knjiga, 2006.

[23] D. Držić, *Pjesni ljuvene*, ser. Stari pisci hrvatski. Zagreb: JAZU, 1965.

[24] A. Kačić Miošić, *Razgovor ugodni naroda slovinskoga*. pò A. Czesaru, 1801.

[25] M. A. Relković, *Satir ili Divji čovik*. Zagreb, Croatia: Tiskom Josipa Kossicha, 1783.

[26] I. Brlić-Mažuranić, *Priče iz davnine*. Zagreb: Nakladni zavod Hrvatske, 1916.

[27] A. Nemčić, *Kvas bez kruha*. Zagreb: Profil knjiga, 2017.

[28] F. Galović, *Izabrane štokavske pjesme*. Zagreb: Mladost, 1984.

[29] H. Lucić, *Skladanja izvrsnih pisan različitih*. Venice: Tisak Franje pl. Kesterčaneka, 1556.

[30] P. Hektorović, *Ribanje i ribarsko prigovaranje*. Zagreb: Nakladni zavod Matice hrvatske, 1968.

[31] M. Pelegrinović, *Jeđupka*. Zagreb: Matica Hrvatska, 1584.

[32] ——, *Pelegrin Sabu Mišetiću*. Zagreb: Matica Hrvatska, 1578.

[33] B. Karnarutića, *Vazetje Sigeta Grada*. Zagreb: Mletci, 1584.

[34] T. Brezovački, *Matijaš Grabancijaš dijak*. Zagreb, Croatia: Nakladom Stjepana Kugli, 1899.

[35] I. Raos, *Prosjaci i sinovi*. Mladost, 1971.

[36] M. Smoje, *Velo misto*. Zagreb: Školska knjiga, 1981.

[37] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.