Machine Translation of Omani Arabic Dialect from Social Media

Khoula Kahlan Al Kharusi
Sultan Qaboos University
khoula.alkharusi@gmail.com

Abdurahman Khalifa AAlAbdulsalam
Sultan Qaboos University
a.aalabdulsalam@squ.edu.om

Abstract

Research studies on Machine Translation (MT) between Modern Standard Arabic (MSA) and English are abundant. However, studies on MT between Omani Arabic (OA) dialects and English are very scarce. This research study focuses on the lack of availability of an Omani dialect parallel dataset, as well as MT of OA to English. The study uses social media data from X (formerly Twitter) to build an authentic parallel text of the Omani dialects[1]. The research presents baseline results on this dataset using Google Translate, Microsoft Translation, and Marian NMT. A taxonomy of the most common linguistic errors is used to analyze the translations made by the NMT systems to provide insights on future improvements. Finally, transfer learning is used to adapt Marian NMT to the Omani dialect, with significant improvement of 9.88 points in the BLEU score.

1 Introduction

In the era of social media and worldwide communication, Machine Translation (MT) has become essential in lowering or eliminating the language barrier between people (Franceschini et al., 2020). Using artificial intelligence, users can translate any post from any language without human involvement. Recently MT underwent a remarkable evolution thanks to deep learning and artificial neural network models (Baniata et al., 2021). Although MT research attempted to produce high-quality translations of the most widely used languages, which are well documented with abundant sources, it still has a long way to go in terms of languages that are not as well documented, such as Arabic dialects.

Over the past decade, the Arabic language has drawn much interest from the MT community. However, most MT contributions focus on Modern Standard Arabic (MSA), while the translation of Arabic dialects is still in its early stages. Arabic is the world's fifth most widely used language, with almost 450 million speakers in 22 countries. Classical Arabic (CA) and MSA are the standard Arabic varieties recognized by Western linguists. The Quran, classical texts, and old Arabic literature are written in CA. MSA is a modern form that is based on the syntactic, morphological, and phonological structures of CA. MSA is the primary form of official communication in the Arab world that is used in education, business, news, and legislation (Al-Qaraghuli et al., 2021). Arabic dialects are used informally in day-to-day conversations throughout the Arab world. Arabic dialects are primarily spoken-only languages; however, in the last decade, these dialects have become increasingly prevalent in social media, text messages, TV shows, and other forms of informal communication. Nowadays, Arabic dialects are being used increasingly in written format for informal communication online (Harrat et al., 2019).

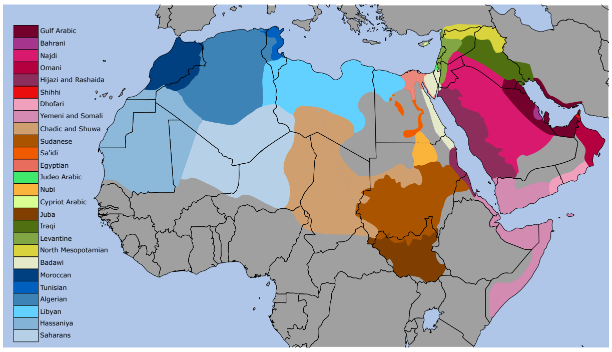

Figure 1: Geographic spread of Arabic dialects (Schmitt, 2020)

Despite the extensive use of Arabic dialects, they are considered low-resource language which hinder MT development. Arabic dialects vary from MSA in terms of phonology, semantics, morphology, and syntax (Harrat et al., 2019). They simplify many standard Arabic rules while simultaneously

[1] Dataset availabel in Github https://github.com/khoula-k/OmaniArabicTranslation

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 302–309
December 7, 2023 ©2023 Association for Computational Linguistics

introducing new sets of rules that add additional complications. Therefore, most MSA resources and tools cannot be easily adapted to translate Arabic dialects (Harrat et al., 2019). The lack of standard orthography is one of the fundamental challenges associated with Arabic dialects. Arabic dialects have diglossia, a linguistic phenomenon in which the speakers mix two or more varieties of the same language (e.g., standard official language and local dialect) within the same context (Farghaly and Shaalan, 2009; Harrat et al., 2019). It is worth noting that Arabic has a diverse range of colloquial varieties, with over 27 variations existing worldwide. These varieties exhibit varying degrees of mutual understanding, highlighting Arabic's nuanced and diverse nature (Elgabou and Kazakov, 2017). Figure 1 provides a basic overview of the geographic distribution of these dialects. There are two primary ways to approach colloquial Arabic MT. The first involves translation between MSA and colloquial Arabic dialects then to foreign language therefore MSA acting as an intermediate language. The second approach involves translation of Arabic dialects into foreign languages directly (Harrat et al., 2019). It is worth noting that all contributions in this field are primarily related to the English language.

This research focuses on the MT of Omani Arabic (OA). Oman's location, surrounded by the Indian Subcontinent, Persia, Arabia, and East African coasts, played a significant role in shaping its history and the languages spoken by its people. Despite Oman's small population, its linguistic context is diverse. Some Omanis speak multiple indigenous languages, such as Jibbali, Shahri, and Mehri, each with thousands of speakers, in addition to Bathari, Harsusi, and Hobyot, with a few hundred speakers each (Al-Balushi, 2017). Additionally, some Omanis speak non-indigenous languages, including Persian, Aajmi, Kumzari from Iran, Baluchi from Baluchistan, Zidjali from Pakistan, Kojki/Luwati from India, and Swahili from East Africa (Al-Balushi, 2017). The impact of various languages on OA is particularly evident in its vocabulary, featuring words borrowed from Hindi, such as guniyyah (meaning sack) and bigli (referring to an electric torch), as well as Persian words like drishah (window) and saman (stuff). English has also contributed words such as sekal (bicycle), batri/betri (battery), swik (switch), and beb (pipe), while Portuguese brings in banderah (flag) and mez (table).

The prevalent dialect in Oman differs from that dominant in the rest of the Arabian Gulf. It is mostly in the form of the Hadari (Sedentary) dialect rather than a Bedouin one (Nabhani, 2011). The Hadari dialect is prevalent in the northern part of Oman, including the capital Muscat, and is also used in most TV shows.

Limited research is available on translating colloquial Arabic dialects, particularly Omani dialects. While most prior works group OA dialects with other Gulf dialects, more research is necessary. It is important to note that while the Gulf region may share cultural similarities, it cannot be assumed that they share linguistic homogeneity. Moreover, OA datasets used in prior works are not publicly available. This research aims to close this gap by creating an authentic Omani Arabic-English parallel corpus that is available for public use. The dataset will be used to adapt an existing Arabic Neural Machine Translation (NMT) system to the Omani dialect.

2 Related Works

In this section, we will explore the literature on dialectical datasets and the MT of Arabic dialects.

2.1 Dialectical Arabic Datasets

In the literature, various dialectical parallel Arabic datasets have been mentioned. Nonetheless, this subsection will focus on the datasets that are publicly available. The **MADAR** corpus (Multi-Arabic Dialect Applications and Resources) (Bouamor et al., 2018) is a collection that comprises parallel sentences encompassing the dialects of 25 cities in the Arab world, along with MSA, English, and French. The corpus is created by translating select sentences from the Basic Traveling Expression Corpus (BTEC), which was in Japanese, English, and Chinese (Takezawa et al., 2007) to the different dialects.

The **MPCA** (Multidialectal Parallel Corpus of Arabic), as documented in (Bouamor et al., 2014), is comprised of 2,000 sentences that represent five Arabic dialects, as well as English and MSA. The corpus was developed by tasking four translators who are native speakers of Palestinian, Syrian, Jordanian, and Tunisian colloquial Arabic varieties to translate 2,000 sentences originally written in Egyptian Arabic into their respective dialects.

The **PADIC** (Parallel Arabic DIalect Corpus) (Meftouh et al., 2015) multi-dialectal corpus contains six dialects in addition to MSA. Two Algerian dialect corpora were created: Annaba's dialect (a city in Algeria) from daily conversations and the dialect from movies/TV shows in the Algiers dialect. Both were transcribed and translated manually. They were later used to obtain other MSA and dialectal corpora.

Currently, the MADAR dataset is the only source we found for Omani Arabic, with the dialect of the capital city, Muscat. Out of the 25 dialects in the MADAR corpus, the dialect of Muscat is the most similar to MSA with an overlap score of 37.5% (Bouamor et al., 2018; Salameh et al., 2018). It has been stated that the translators were native speakers of the dialects, and they got access to English and French versions of the corpus without the MSA to avoid biased translation. Upon analyzing the OA in the MADAR dataset by a native speaker of the Omani dialect, it was observed that it predominantly reflects a dialect that is more oriented toward MSA with some Bedouin influence rather than the sedentary Muscat-Omani dialect.

2.2 Machine Translation of Dialectical Arabic

When it comes to Arabic dialect MT, there are two main approaches. The first approach focuses on translating between MSA and its corresponding dialects, while the other approach aims to translate Arabic dialects into foreign languages. It is important to note that most research in this field is related to translating into English.

In the field of colloquial Arabic MT, one of the earliest studies was conducted by Sawaf in 2010. The study focused on dialect normalization and used a hybrid RBMT and SMT to translate into MSA (Sawaf, 2010).

Wael Salloum and Nizar Habash have contributed several papers to the field of colloquial Arabic translation. One of their approaches, as described in (Salloum and Habash, 2011), involved a rule-based method for producing MSA paraphrases of dialectal Arabic OOV (out of vocabulary words) in the Levantine and Egyptian dialects. They then combined this with the results generated by ADAM (Salloum and Habash, 2014), to create Elissa (Salloum and Habash, 2012), which can handle Levantine, Egyptian, Iraqi, and to a lesser extent Gulf Arabic. (Salloum and Habash, 2013) published an advanced version of their translation system, which translates dialectal Arabic to English by pivoting through MSA.

(Zbib et al., 2012) proposed a massive SMT-based system for Levantine and Egyptian dialects. They created parallel corpora of Levantine-English and Egyptian-English and then trained their statistical translation model using direct translation and pivoting through MSA. In contrast to the previously discussed approach that utilized a statistical model in (Sghaier and Zrigui, 2020), a rule-based system was developed to translate from the Tunisian dialect to MSA without relying on statistical models.

The following works utilized a modern technique of deep neural networks to translate Arabic dialects. AraBench (Sajjad et al., 2020) presented evaluation benchmarks for dialectal Arabic to English. The paper details several experiments conducted in this regard. They used the OpenNMT model (Klein et al., 2017) and trained it in extensive heterogeneous MSA and dialectical Arabic data. This base model is then fine-tuned towards in-domain dialectical training data. Lastly, they used back-translation to increase the dialectal Arabic-English training data size. (Baniata et al., 2021) is using the state-of-the-art Transformer models to translate DA to MSA using subword units for tokenization, effectively solving the issue of out-of-vocabulary words. The subword segmentation algorithm operates under the premise that a word comprises a combination of subwords.

All of the studies that focus on Omani dialect utilized the MADAR corpus. In the research conducted by (Baniata et al., 2021), Omani dialect is grouped with other Gulf dialects, making it difficult to assess the system's performance for the Omani Arabic specifically. On the other hand, AraBench (Sajjad et al., 2020) has tested OA independently and achieved a BLEU score of 39.5% for the translation model trained for MSA-EN translation. However, it is worth noting that the Muscat-MADAR dataset used AraBench may not be representative of OA with a lot of influence from MSA.

2.3 Machine Translation for Low-resource Languages

MT has significantly improved with the use of deep neural networks. However, the downside is that it demands extensive training data and takes up a lot of computing power and time. Fortunately, transfer learning offers a practical solution by utilizing prior knowledge of a trained model to improve performance on related tasks. This approach reduces the need for extensive training data, saving time and resources. (Zoph et al., 2016) used a French-English model as the parent model for low-resource language pairs such as Hausa, Turkish, Uzbek, and Urdu into English. On average, NMT shows 5.6 BLUE points score improvement from transfer learning. The researchers also explored the similarity between the parent and child languages. They conducted a transfer learning method using French and German as parent languages for the Spanish language. The results showed that French was a better parent language for Spanish, which could be the result of its greater similarity to the child's language. (Zoph et al., 2016) employs a transfer learning approach with a single parent and one child, whereas (Goyal et al., 2020) utilizes transfer learning by leveraging related languages. Two simple and effective methods are introduced: Multilingual Transfer Learning, which helps improve low-resource languages by utilizing parallel data from related languages, regardless of their resource levels, and Unified Transliteration and Subword Segmentation, which takes advantage of the similarities between related language pairs.

3 Omani Parallel Corpus

This study aims to translate Omani Arabic, utilizing original data of language in use on social media posts. X (formerly Twitter) was utilized to collect text representatives of the Omani dialect. X is a leading social media platform that contains trending news and topics and has a very large user base. Therefore, it offers a valuable resource for conducting large-scale text analysis. Furthermore, API allows users to execute complex queries, such as retrieving all text related to a particular topic or extracting a specific user's posts.

This chapter will explain the full process of creating the Omani Parallel Corpus. After the completion of the corpus, we will present a translation baseline results on this corpus obtained using Google Translate, Microsoft Translation, and Marian NMT systems.

3.1 Data Collection

Each post by users on 'X' comes with metadata fields and values containing information such as the author, creation timestamp, message, location, etc. The data has been retrieved in JSON format using the platform API. Using `conversation_id`, each post and its related replies are collected in one file, which we consider conversations surrounding a particular topic initiated by the main post. We collected posts from a prominent news account in Oman (@oman1_news). Each post and its related replies were treated as a single document for a specific topic. As a result, we obtained a corpus consisting of 905 topics in the form of conversation and containing a total of 87,220 posts.

Real-world social media data typically comprises texts, images, and videos, often accompanied by offensive language and hate speech. The text is often noisy with hashtags, URLs, and foreign characters, and there may be instances of spelling errors. Additionally, individuals may use slang, which is an informal language unique to specific groups or geographic regions that carry cultural connotations with different meanings.

3.2 Corpus Linguistics

Each document in the resulting corpus is converted to a CSV file where the first row is the source post, and the following are replies. Table 1 provides a corpus summary in numbers. The most frequent tokens are the linking words, while the least frequent tokens are words with foreign characters.

Table 1: X Omani corpus

Number of topics/conversations	905
Total number of posts/messages	87,220
Total number of tokens	1,102,952
Unique tokens (vocabulary size)	118,821

3.3 Translation

We have asked volunteers to translate each document into English. Nine participants who are native speakers of the dialect have worked on the translations. We assigned unique sets of topics to each translator and asked them to produce a translation that precisely reflects the source sentence without making any assumptions. The translators were provided with the following guidelines:

- The English translation should retain the punctuation marks from the source sentence (like periods, commas, and question marks).
- When translating idioms and slang, it is important to convey the intended meaning rather than translating them literally.
- Disregard any posts containing offensive language, hate speech, or advertisements.

Table 2: Omani Arabic-English parallel corpus

OA	EN
ولي خطف علينا اليومين الماضيات شو؟!!	And what was it that just happened the past two days?!!
ما شوي الي خطف والي جالس يخطف عليهم.	What they faced and are facing isn't easy
يعني بكم جونية العيش دكتور اتّستوي	so how much will be the sack of rice doctor
اول مره نهزمهن	First time we beat them

- Avoid translating posts with Quran verses.

A total of **2906** posts have been translated, covering various topics. The Omani Parallel Corpus was created by combining all the translated posts. Table 2 below shows some examples from the parallel corpus.

4 Error Analysis in Omani Arabic MT

The most used measures for translation accuracy are automated. However, it can be difficult to establish a direct correlation between these measures and the actual errors present in the translations. A comprehensive analysis of errors is crucial for any natural language processing task, as it can reveal valuable insights into what went wrong and guide future research directions (Ângela Costa et al., 2015; Vilar et al., 2006). For the error classification, we adopted a simplified version of the taxonomy (Vilar et al., 2006) shown in figure 2. Error analysis can be a time-consuming task that requires linguistic expertise. Therefore, we conducted an analysis of a sample of sixty source sentences to identify errors generated by Google Translate and Marian NMT.

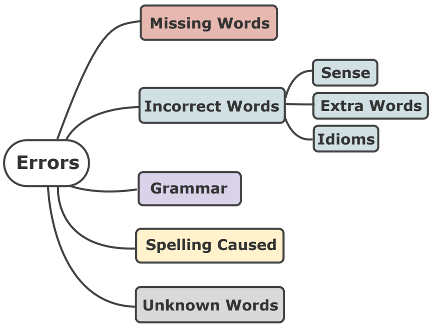

Figure 2: Adapted taxonomy for translation errors

Table 3: Error analysis results on Google Translate and Marian NMT

Error Type		Google	Marian
Missing Words		11	47
Incorrect Words	Sense	64	67
	Extra Words	5	5
	Idioms	8	4
Unknown Words		12	16
Grammar		6	3
Spelling Caused		6	6

Out of the 60 sentences translated by both MT engines, Google had incorrect translations in 54 cases, while Marian had errors in 57. Google's total number of translation errors was 112, while Marian had 150 errors. In Table 3, the number of translation errors produced by each translation system can be observed, categorized by the type of translation error. Both translation systems produce a similar number of translation errors across all categories except for missing words. Marian NMT dropped 36 more words than Google, which only dropped 11. The majority of errors were related to choosing the wrong word sense during translation.

5 Transfer Learning

To implement this approach, we start with a pre-trained NMT model that has been trained on a large parallel corpus (MSA-EN). We then use this model to initialize a new NMT model called the child model. This model is then trained on the domain dataset with a limited parallel data. Using the pre-trained parent model, the child model commences with established weights inherited from the parent model rather than starting with random weights. This method is particularly useful since the Omani parallel corpus is limited, and the MSA corpus provides a strong prior distribution over language vocabularies.

Marian NMT (Junczys-Dowmunt et al., 2018) is used as a neural translation system for the parent model. Marian is a highly efficient NMT framework that is built on pure C++, requiring minimal dependencies. The framework was mainly developed by the Adam Mickiewicz University and the University of Edinburgh. It is currently utilized in var-

Table 4: Arabic-English Opus corpus

Training	Arabic Tokens	English Tokens
126.6M	2.3G	3.9G

ious European projects and is the primary engine for translation and training behind the NMT launch at the World Intellectual Property Organization. Marian has found its niche in the growing world of open-source NMT toolkits due to two key aspects: it is built entirely on C++ which makes it very efficient. Additionally, it is self-contained with its own back-end that enables reverse-mode automatic differentiation using dynamic graphs.

Marian NMT model follows the original transformer architecture with six encoders and six decoders with eight attention heads in each layer (Junczys-Dowmunt et al., 2018). Language Technology Research Group at the University of Helsinki, Finland, trained Marian NMT on many language pairs from the OPUS-MT datasets. These models have been converted to `PyTorch`[2] using the transformers library by Hugging Face[3]. The Arabic-English[4] translation model was trained with a parallel dataset of 126.6M Arabic-English sentence pairs (see Table 4 below (Tiedemann, 2020; Tiedemann and Thottingal, 2020).

The Omani parallel corpus was split into a train set, a validation set, and a test set. The training set contains 70% of the whole corpus, and the remaining 30% is divided equally between the validation and test sets. The training was done using `Seq2SeqTraining` script from Hugging Face and activation function is `AdamW`. Before jumping to the model results, having a look at the validation and training loss is a good practice to ensure models are generalized and there is no over-fitting. Both validation and training loss decreased up until the fifth epoch. However, after the fifth epoch, the validation loss showed a slight increase. Hence, we have decided to proceed with five epochs for training.

Table 5: Omani corpus split

Dataset	Percentage	Parallel posts
Training Set	70%	2,034
Validation Set	15%	436
Test Set	15%	437

Table 6: Results of Google, Microsoft, Marian NMT, and the transfer learning model on a test set of OA corpus

	Google	Microsoft	Marian	Tuned Model
BLEU	34.98	34.26	24.22	34.11
chrF	60.55	61.28	49.02	59.81

In order to compare the fine-tuned model with various translation systems, we calculated the BLEU score for the validation set translated by Google, Microsoft, Marian NMT, and the transfer learning model. In Figure 6, the results indicate that Google achieved the highest BLEU score of 34.98, but it is noteworthy that Microsoft and our model were not far behind, scoring 34.26 and 34.11, respectively. On the other hand, Marian NMT scored the lowest with 24.22. Fine-tuning Marian NMT closed the performance gap between Marian and other translation engines (Google and Microsoft). Marian's initial BLEU score was 24.22, but after completing transfer learning in the OA training set, it increased significantly to 34.11, representing an improvement of 9.88 points. Our results outperformed (Zoph et al., 2016), which achieved a maximum improvement of 7.5 points. The training process is significantly influenced by the closeness of the parent model language to the child's model language.

Although the transfer learning model has displayed positive outcomes in Marian NMT, it has yet to surpass the MSA-English systems of Google and Microsoft. It would be advantageous to implement transfer learning in these systems, but they do not offer open-source models.

6 Conclusion

Using Machine Translation (MT) is an effective way to overcome language barriers in communication. While there are numerous research studies on MT from Modern Standard Arabic (MSA) to English, there is a significant lack of studies on translating Omani dialects. In this study, we aim to establish a first baseline targeted at the automatic translation of the written text of Omani dialects from social media.

Our initial step was to thoroughly analyze the literature on the translation of colloquial Arabic dialects and identify the datasets that contained an

[2] `PyTorch` is a machine learning framework based on the Torch library `https://pytorch.org/`

[3] Hugging Face develops tools for building applications using machine learning `https://huggingface.co/`

[4] `https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-ar-en`

OA corpus. Only one source was available for reseacrh use and it may not be representative of the Omani dialect. We collected messages from social media to create an authentic Omani dialect corpus. Then we translated a total of 2906 messages. This corpus has been used to conduct a baseline study on existing MT Models' performance in Omani dialect translation, where we found that Google and Microsoft translation engines got higher BLEU scores reaching 33%, compared to Marian NMT, which scored 22.3%. We conducted a manual evaluation to identify Google and Marian NMT errors. After linguistically classifying the errors, we discovered that the most common error made by both NMTs was choosing the wrong word sense. We enhanced the translation of OA by utilizing transfer learning with Marian NMT. This resulted in a significant improvement of 9.88% in the BLEU score.

The main contribution of this research can be summarized as follows:

- Collecting and creating a parallel corpus of Omani dialects and English.
- Analyzing MT errors to inform future research direction.
- Applying transfer learning for OA using an existing MSA-English model.

We faced challenges because we had limited resources and time constraints. We didn't have the funds or time to hire professional translators for the corpus, and we couldn't review every sentence to select them for translation. Additionally, we utilized transfer learning with the available open-source model.

In the future, we hope to enhance the translation of the OA corpus by collaborating with linguistic experts, increasing the quantity of translated sentences, and providing multiple translations for each sentence to ensure an accurate evaluation.

References

Rashid Al-Balushi. 2017. Omani arabic: More than a dialect. *Macrolinguistics*, 4:80–125.

Mohammed Al-Qaraghuli, Gheith Abandah, and Ashraf Suyyagh. 2021. Correcting arabic soft spelling mistakes using transformers. pages 146–151. IEEE.

Laith H. Baniata, Isaac K.E. Ampomah, and Seyoung Park. 2021. A transformer-based neural machine translation model for arabic dialects that utilizes subword units. *Sensors*, 21.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The madar arabic dialect corpus and lexicon. *European Language Resources Association (ELRA)*.

Hani Elgabou and Dimitar Kazakov. 2017. Building dialectal Arabic corpora. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 52–57, Varna, Bulgaria. Association for Computational Linguistics, Shoumen, Bulgaria.

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing. *ACM Transactions on Asian Language Information Processing*, 8:1–22.

Dario Franceschini, Chiara Canton, Ivan Simonini, Armin Schweinfurth, Adelheid Glott, Sebastian Stüker, Thai-Son Nguyen, Felix Schneider, Thanh-Le Ha, Alex Waibel, Barry Haddow, Philip Williams, Rico Sennrich, Ondřej Bojar, Sangeet Sagar, Dominik Macháček, and Otakar Smrz. 2020. Removing european language barriers with innovative machine translation technology.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing and Management*, 56:262–273.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Karima Meftouh, Salma Jamoussi, Mourad Abbas, Crstdla ‡ Algiers, and Algeria Kamel Smaili. 2015. Machine translation experiments on padic: A parallel arabic dialect corpus.

Hala Al Nabhani. 2011. Language and identity in oman through the voice of local radio broadcasters.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. pages 5094–5107. International Committee on Computational Linguistics.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. pages 10–21. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2012. Elissa: A dialectal to standard arabic machine translation system.

Wael Salloum and Nizar Habash. 2013. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. *NAACL-HLT*.

Wael Salloum and Nizar Habash. 2014. Adam: Analyzer for dialectal arabic morphology. *Journal of King Saud University - Computer and Information Sciences*, 26:372–378. Special Issue on Arabic NLP.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Genevieve A. Schmitt. 2020. *Relevance of Arabic Dialects: A Brief Discussion*, pages 1383–1398. Springer International Publishing, Cham.

Mohamed Ali Sghaier and Mounir Zrigui. 2020. Rule-based machine translation from tunisian dialect to modern standard arabic. *Procedia Computer Science*, 176:310–319. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt – building open translation services for the world.

David Vilar, Jia Xu, Luis Fernando D'haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects.

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

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29:127–161.