The Helsinki-NLP Submissions at NADI 2023 Shared Task: Walking the Baseline

Yves Scherrer[1,2] first.last@ifi.uio.no
Aleksandra Miletić[1] first.last@helsinki.fi
Olli Kuparinen[1,3] first.last@tuni.fi

[1]Department of Digital Humanities, University of Helsinki
[2]Department of Informatics, University of Oslo
[3]Faculty of Information Technology and Commmunication Sciences, Tampere University

Abstract

The Helsinki-NLP team participated in the NADI 2023 shared tasks on Arabic dialect translation with seven submissions. We used statistical (SMT) and neural machine translation (NMT) methods and explored character- and subword-based data preprocessing. Our submissions placed second in both tracks. In the open track, our winning submission is a character-level SMT system with additional Modern Standard Arabic language models. In the closed track, our best BLEU scores were obtained with the leave-as-is baseline, a simple copy of the input, and narrowly followed by SMT systems. In both tracks, fine-tuning existing multilingual models such as AraT5 or ByT5 did not yield superior performance compared to SMT.

1 Introduction

This paper presents the Helsinki-NLP submissions to the NADI 2023 shared tasks. We participated in Subtasks 2 and 3, which consisted in translating dialectal data into Modern Standard Arabic (MSA) (Abdul-Mageed et al., 2023). This was the first time the NADI shared task involved translation, following past tasks on dialect identification and sentiment analysis (Abdul-Mageed et al., 2020, 2021, 2022).

The Arabic dialectal continuum stretches from Morocco in the west to Oman in the east. Various classifications of the dialects have been proposed, ranging from large regions to country-level or even city-level divisions (Bouamor et al., 2018; Habash, 2022). The Arabic language area is also well known for its diglossic situation. While Modern Standard Arabic is used in education, media and culture across the continuum, it is not native to any of the dialectal regions.

The translation subtasks focused on four Arabic dialects: Egyptian, Emirati, Jordanian, and Palestinian. The shared task organizers provided the MADAR corpus (Bouamor et al., 2018) as the training material for the closed track (Subtask 2), which did not allow for the use of additional training data. The Subtask 3 was described as open track where any additional training material was allowed.

Since our initial experiments showed that neural models were particularly affected by the small size of the MADAR training data, a large part of our efforts went into creating additional parallel data for the in Subtask 3 models. In particular, we focused on freely available monolingual MSA corpora, which we then back-translated to three target dialects, grouping Jordanian and Palestinian together. Adding the back-translated data to the original training corpus allowed our neural models to perform on par with less data-hungry statistical models.

We participated in Subtask 2 with three submissions and in Subtask 3 with four submissions. Our submissions can be divided into four different approaches:

- **LAI** – the leave-as-is baseline consisting of a copy of the input text,
- **SMT** – character-level statistical machine translation models;
- **NMT** – Transformer-based neural machine translation models;
- **ByT5** and **AraT5** – pretrained sequence-to-sequence models fine-tuned with task-specific data.

Our best performing translation system was SMT for both subtasks, but it was not able to outperform the LAI baseline in Subtask 2, at least in terms of the BLEU score (Papineni et al. 2002; see Section 5.1 for a critical discussion of evaluation measures). Our submissions placed second on both subtasks.

Section 2 describes the data collection and preparation whereas Section 3 outlines the proposed models in more detail. Our results are presented in Section 4 and further discussed in Section 5.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 670–677

Section 6 offers conclusions of our work.

2 Data Collection and Augmentation

2.1 MADAR3

The training resource provided by the organizers was the MADAR corpus (Bouamor et al., 2018). The dataset contains the same sentences in different Arabic dialects from 25 cities, as well as in English, French and MSA. The corpus was created by translating sentences from the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007) into Arabic dialects.

We found in early experiments that our models achieved better results when excluding the data from Maghrebi and Yemeni dialects, since the development and test sets do not cover these dialect groups. Therefore, for all our submissions, we use the subset of MADAR that covers the Nile Basin, Levant and Gulf regions. We refer to this subcorpus as **MADAR3** throughout the paper.

2.2 MSA data

Considering that parallel resources for the target dialects are hard to come by, we focused our collection efforts on monolingual MSA data, taking inspiration from the AraT5 pretraining setup (Nagoudi et al., 2022). In particular, we used the following resources:

AraNews is a collection of Arabic newspaper texts from 15 Arab nations, the United States of America and the United Kingdom (Nagoudi et al., 2020).

Leipzig News is a dataset of Arabic news curated by the Leipzig Corpora Collection. The data comes from mostly Saudi Arabian news outlets (Goldhahn et al., 2012).

OSIAN is an Arabic news corpus crawled from the web (Zeroual et al., 2019). It contains articles from 31 international Arabic news broadcasting platforms.

Tatoeba is a project collecting translations of sentences in the web. The data is available in OPUS (Tiedemann, 2012).

TED is a corpus of translated subtitles from over 4000 TED talks (Reimers and Gurevych, 2020). The data is available in OPUS (Tiedemann, 2012).

Wikipedia is a Wikipedia-based corpus we extracted from the Arabic Wikipedia using WikiExtractor (Attardi, 2015)[1]. The extracted data was subsequently sentence-segmented and deduplicated at sentence level (only the exact matches were removed).

Corpus	Sentences	Words
AraNews	59,270	2,643,313
Leipzig News	1,000,000	23,972,851
OSIAN	1,000,000	21,532,389
Tatoeba	47,471	231,507
TED	403,845	5,652,867
Wikipedia	11,368,818	193,912,867

Table 1: Size of additional datasets

An overview of corpus sizes is given in Table 1.[2] Of these resources, AraNews, OSIAN and Wikipedia were used to pretrain AraT5 (V1).

2.3 Backtranslation of MSA data

While monolingual target-side data can easily be included into SMT systems in the form of additional language models, this is more difficult for neural models. The most common approach in this situation is to produce synthetic parallel data using backtranslation (Sennrich et al., 2016).

To this end, we reversed our dialect-specific SMT-mono models from Subtask 2 (see Section 3.2 for details) to produce three dialectal versions of all monolingual MSA data presented in Section 2.2. The backtranslated data was used to train or fine-tune the neural models for Subtask 3 (see Sections 3.3, 3.4 and 3.5).

The quality of the backtranslations is most likely poor, but we nevertheless expect backtranslation to work better than simpler data augmentation methods such as noise injection. Since the authors are not speakers of Arabic, the quality of the backtranslations could not be evaluated.

3 Models

3.1 LAI

As the shared task organizers did not provide an official baseline, we propose the leave-as-is (LAI)

[1]The extraction was done from the Wikimedia data dump `arwiki-20230801-pages-articles-multistream.xml.bz2`

[2]Note that we did not perform full tokenization of the corpora: the word counts in the table are based on whitespace-delimited tokens.

	Training data			Development set BLEU						Test set BLEU				
Model	MADAR3	MSA	BT MSA	Overall	EGY	EMI	JOR	PAL	Subm.	Overall	EGY	EMI	JOR	PAL
LAI	—	—	—	**15.78**	14.87	**26.31**	12.75	**12.90**	2.2	14.28	12.22	**23.13**	11.15	13.41
SMT-multi	✓	—	—	15.62	15.01	25.74	12.52	12.64	2.1	13.60	12.02	21.82	10.46	12.66
SMT-mono	✓	—	—	15.39	**15.91**	17.94	**14.84**	11.78	2.3	12.53	11.91	16.50	9.83	11.42
NMT	✓	—	—	2.61	3.24	2.52	0.00	2.32	—	—	—	—	—	—
ByT5	✓	✓	—	6.63	6.89	4.86	4.94	7.60	—	—	—	—	—	—
AraT5 V2	✓	✓	—	7.41	7.61	5.55	5.98	8.01	—	—	—	—	—	—
Best competitor										**14.76**	**16.04**	14.30	**12.55**	**13.55**
SMT-multi	✓	✓	—	**19.19**	18.88	25.66	17.24	**17.16**	3.1	17.69	16.11	25.81	15.60	15.91
SMT-mono	✓	✓	—	18.61	**19.03**	26.89	13.12	17.14	—	—	—	—	—	—
NMT	✓	—	✓	18.40	16.78	25.34	**19.59**	14.36	3.2	16.88	15.17	24.77	15.41	14.45
ByT5	✓	✓	✓	17.69	16.68	24.90	16.01	14.03	3.3	16.10	15.55	21.79	13.73	13.34
AraT5 V2	✓	✓	✓	19.14	18.49	**28.25**	17.19	14.80	3.4	17.46	15.50	25.06	15.97	15.06
Best competitor										**21.10**	**17.65**	**28.46**	**22.03**	**17.29**

Table 2: Overview of the tested models and their BLEU scores (↑) on the development and test sets. *MSA*: monolingual MSA data, *BT MSA* monolingual MSA data back-translated to three dialects. ✓: used for pre-training, ✓: used for training or fine-tuning. EGY: Egyptian, EMI: Emirati, JOR: Jordanian, PAL: Palestinian. The horizontal line separates closed (Subtask 2) from open (Subtask 3) submissions according to the organizer-defined criteria.

baseline: an unchanged copy of the input file. We do not suggest that LAI is a potential solution to the task; rather, we introduce it as a way of estimating the task difficulty.

We were unable to beat this baseline with the systems that only use the MADAR corpus for training or fine-tuning in terms of BLEU score. Therefore, we decided to submit LAI as one of our contributions. We think it is interesting to also compare the other participants' systems with this baseline. For example, even the best submitted subtask 2 system scores behind LAI on the Emirati dialect (see Table 2).

3.2 SMT

We use a character-level statistical machine translation model based on the Moses toolkit. We split all sentences into character sequences and treat each character as a separate translation unit.[3]

We provide two variants of the SMT approach. **SMT-multi** is a single model trained on all dialects from MADAR3. **SMT-mono** is a collection of 3 models, each of which is trained on the MADAR texts of one major dialect area (Nile Basin, Levant, Gulf). At prediction time, the relevant model is chosen according to the provided dialect labels.

Furthermore, each of the two models is made available in a closed and an open variant. The closed variant contains a single language model trained on the MSA side of MADAR. The open variant contains a total of 7 language models, corresponding to the different MSA corpora listed in Section 2.2 in addition to MADAR.

3.3 NMT

Our neural machine translation method is based on the Transformer architecture. The model was trained with OpenNMT-py (Klein et al., 2017).[4]

We tokenized the data using the unigram model implemented in the SentencePiece library (Kudo and Richardson, 2018), as it has outperformed BPE-based segmentation when the studied texts include inconsistent writing or non-standard language (Kanjirangat et al., 2023). We experimented with three different vocabulary sizes (300, 500, 1000) and found the smallest (300) to offer the best performance.

Furthermore, we found that the NMT model's performance was enhanced by adding a dialect tag at the beginning of the source sentence. We used the three dialect labels of MADAR3.

The NMT model trained on MADAR3 alone did not produce competitive scores. We only submitted an NMT model trained both on MADAR3 and on the backtranslations.

3.4 ByT5

ByT5 (Xue et al., 2022) is a multilingual pretrained model of the T5 family (Raffel et al., 2020)

[3]Character-level models outperformed SMT models with words and subwords in preliminary experiments. Model parameters are presented in Table 5 in the Appendix.

[4]Experimental details for each model are provided in Table 5 in Appendix A.

that encodes all text as UTF-8 encoded byte sequences. It is pre-trained on the multilingual m4C corpus (Xue et al., 2021), with 1.66% of the data in Arabic. ByT5 was used by the winning team (Samuel and Straka, 2021) in the MultiLexNorm shared task (van der Goot et al., 2021), in which the participants had to normalize social media texts of various languages. We expect that Arabic dialect-to-standard translation consists to a large extent of local changes of individual characters. We therefore find that a byte-based model is a good fit for this task.

We fine-tuned the `byt5-base` model with MADAR3, but found the performance subpar. For our submission, due to computational limitations, we fine-tuned the `byt5-small` model with a random sample of 1M sentences from our backtranslated data and MADAR3.

3.5 AraT5

AraT5 (Nagoudi et al., 2022) is a pre-trained model of the T5 family specifically focused on Arabic, enabling tasks like machine translation into and out of Arabic, summarization, transliteration and other sequence-to-sequence transformation tasks. During the competition, the second version AraT5-V2 was made available. We use the `AraT5v2-base-1024` foundation model for our experiments.

Fine-tuning AraT5-V2 on MADAR3 only did not yield competitive results. Instead, we submitted a model fine-tuned on MADAR3 and a random sample of the backtranslations, with a total of 1.4M sentence pairs (15% of the full dataset).[5]

4 Results

4.1 Results on the development set

Our results on the development set are shown in the middle panel of Table 2 with the official evaluation metric BLEU. Our best submission in Subtask 2 is the leave-as-is baseline (LAI; Section 3.1). The fact that unmodified input achieves better results than machine translation approaches can be taken as an indicator of the difficulty of the closed track task. Note, however, that our best-performing machine translation approach (SMT-multi) is in general less than one BLEU point below LAI.

The inclusion of additional training material in Subtask 3 led to a significant improvement for neural methods, as illustrated by the results of NMT, ByT5 and AraT5 in the lower part of Table 2. Nevertheless, our best performing approach remains SMT-multi, which scores first overall, and for all individual dialects except for Jordanian. AraT5 is the second best model overall, but note that SMT-mono outperforms it on Egyptian and Palestinian. The scores across different models are the most stable for Egyptian, and they vary the most on Emirati, where the difference between the best (SMT-multi) and worst model (ByT5) is around 4 BLEU points.

4.2 Official results

The right-hand panel of Table 2 shows the official results on the test set. For comparison, we added the results of the top-performing system of each subtask.

For Subtask 2, the LAI baseline outperformed both of our SMT systems, and got close to the best submission. It can be noted that our LAI model outperformed the best competitor on Emirati by a large margin, suggesting that models tend to over-normalize this dialect.

For Subtask 3, SMT and AraT5 were our best submissions, as could be expected from the development set scores. However, there is a significant gap to the best competitor, especially for Emirati and Jordanian. We would like to note however that our Subtask 3 submissions rely on similar training data as was used for AraT5 pretraining, but in a smaller volume. In that sense, it may be more relevant to compare our systems with Subtask 2 submissions that are based on AraT5.

Note that in all our experiments we systematically use sentence-level contexts. However, our previous work has shown that contexts of sliding windows of three words can bring significant improvements, especially for the TF-based systems (Kuparinen et al., 2023). This approach requires word-level data alignments which are not trivial to produce. Therefore we defer this to future work.

5 Discussion

5.1 Evaluation metrics

The BLEU score (Papineni et al., 2002), which was used as the official metric in this shared task, treats each word as an atomic unit and considers a word as wrong even if only one character is incorrect. However, in dialect-to-standard translation tasks, an large amount of differences is expected to concern changes of individual characters. It

[5]The samples used for byT5 and AraT5 differ due to computational time constraints.

Model		Overall	EGY	EMI	JOR	PAL
		BLEU				
2.2	LAI	**15.78**	14.87	**26.31**	12.75	**12.90**
2.3	SMT-mono	15.39	**15.91**	17.94	**14.84**	11.78
2.1	SMT-multi	15.62	15.01	25.74	12.52	12.64
		chrF				
2.2	LAI	45.02	46.56	49.47	40.85	44.10
2.3	SMT-mono	**46.96**	**49.11**	**51.11**	**43.81**	**44.69**
2.1	SMT-multi	44.96	46.60	49.37	40.81	43.94

Table 3: BLEU and chrF scores on the development set.

might therefore be interesting to consider metrics that reflects this better, for example the chrF score (Popović, 2015), which is based on the precision and recall of character n-grams.

Table 3 compares the development set BLEU scores with the chrF scores of our Subtask 2 submissions. According to BLEU, LAI is the best performing system, mostly thanks to its good performance on Emirati. SMT-mono is the worst of the three despite winning on two individual dialects. In contrast, according to chrF, SMT-mono outperforms all other systems on all four dialects. The large variation on Emirati has also disappeared.

This suggests that our SMT-mono system could in fact be perceived as better than the higher-ranked LAI baseline. It would be instructive to see which of the two evaluation metrics correlates better with human assessment on this particular task.

5.2 Test data domains

While the MADAR corpus contains relatively short and simple sentences from the travel domain, the development and test data provided for the NADI shared task comes from a different source and text domain. It can be interesting to see how the proposed translation models fare on both domains.

To this end, we extracted the test instances from the MADAR3 corpus (which were held out from model training) and evaluated some of our submissions on them. Table 4 provides a comparison of the results on the NADI test data and the MADAR3 test data.

There is a striking difference in terms of LAI BLEU between the two datasets: NADI seems to be much "easier" than MADAR, in the sense that fewer replacements are required. For both datasets, the closed-track SMT model does not do any better than the baseline. The two selected open-track models have very similar performances on the NADI test set, but differ greatly on their performance on MADAR. AraT5, presumably thanks to the large amount of pretraining data, generalizes much better to the more difficult MADAR test set.

Model		NADI	MADAR3
2.2	LAI	14.28	3.48
2.1	SMT-multi	13.60	3.53
3.1	SMT-multi	17.69	10.22
3.4	AraT5 V2	17.46	17.85

Table 4: Overall BLEU scores for the NADI and MADAR3 test sets.

6 Conclusions

In this paper, we described our participation in the NADI shared task, where we submitted seven systems to two tracks. Our submissions placed second on both tracks. Our strongest translation method was SMT in both tracks, but given the difficulty of the task, it was outperformed by an LAI baseline in the closed track. Neural models closed the gap to the SMT models only with large amounts of additional parallel data obtained through backtranslation.

We would like to note again that our open track submissions do not use any human-translated parallel training data besides MADAR, and that the total amount of training data is smaller than what was used for AraT5 pre-training. This makes our models, in particular the SMT ones, more data efficient than large pretrained models such as AraT5 or ByT5.

We also showed that the participating systems could have been ranked differently with a character-based evaluation metric, which underlines the importance of the selected metrics.

Acknowledgements

This work has been supported by the Academy of Finland through project No. 342859 "CorCoDial – Corpus-based computational dialectology".

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, ElMoatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification

Shared Task. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Giuseppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

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. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *International Conference on Language Resources and Evaluation*.

Nizar Y Habash. 2022. *Introduction to Arabic natural language processing*. Springer Nature.

Vani Kanjirangat, Tanja Samardžić, Ljiljana Dolamic, and Fabio Rinaldi. 2023. Optimizing the size of subword vocabularies in dialect classification. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 14–30, Dubrovnik, Croatia. 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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-Standard Normalization: A Large-Scale Multilingual Evaluation. In *Findings of EMNLP2023*. (accepted).

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Tariq Alhindi. 2020. Machine generation and detection of Arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 69–84, Barcelona, Spain (Online). Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

David Samuel and Milan Straka. 2021. ÚFAL at MultiLexNorm 2021: Improving multilingual lexical normalization by fine-tuning ByT5. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 483–492, Online. Association for Computational Linguistics.

Yves Scherrer. 2023. Character alignment methods for dialect-to-standard normalization. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 110–116, Toronto, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.

A Experimental Details

We trained all neural models on a single NVIDIA V100 GPU. The SMT models were trained on a Xeon Gold 6230 CPU. We will make the training scripts and the additional data publicly available for the final submission.

Model	Parameter	Selected values	Considered alternatives
SMT	Subword tokenization	characters	words, unigram subwords
	Alignment tool	eflomal	— (Scherrer, 2023)
	Alignment symmetrization	grow-diag-final-and	—
	Language model n-gram size	10	—
	Maximum phrase length	10	—
	Distortion	disabled	—
	Tuning method	MERT	—
NMT	Subword tokenization	unigram subwords	characters
	Encoder + decoder layers	6 + 6	—
	Attention heads	8	—
	Embedding dimensions	512	—
	Hidden layer dimensions	512	—
	Position representation clipping	4	—
	Dropout	0.1	—
	Label smoothing	0.1	—
	Optimizer	Adam	—
	Adam $\beta 2$	0.98	0.998
	Batch size / accumulate gradient	2 * 5000 tokens	—
	Initial learning rate	0.1	0.01, 2.0
	Decay	Noam, 10000 warmup steps	—
	Max. training sequence length	1000	—
	Max. prediction sequence length	1000	—
	Training time	100000 steps	—
ByT5	Foundation model	google/byt5-small	google/byt5-base
	Max. sequence length	512	—
	Batch size	8 sentences	—
	Early stopping	disabled	—
	Training time	5 epochs	—
	Model selection criterion	validation loss	—
AraT5	Foundation model	UBC-NLP/AraT5v2-base-1024	UBC-NLP/AraT5-base
	Max. sequence length	256	—
	Batch size	12 sentences	—
	Early stopping	5 epochs	—
	Max. training time	20 epochs	—
	Model selection criterion	validation loss	—

Table 5: Hyperparameter settings.