Findings of the LoResMT 2021 Shared Task on COVID and Sign Language for Low-resource Languages

Atul Kr. Ojha[1,2] atulkumar.ojha@insight-centre.org
Chao-Hong Liu[3] ch.liu@acm.org
Katharina Kann[4] katharina.kann@colorado.edu
John Ortega[5] jortega@cs.nyu.edu
Sheetal Shatam[2] panlingua@outlook.com
Theodorus Fransen[1] theodorus.fransen@insight-centre.org
[1]Data Science Institute, NUIG, Galway
[2]Panlingua Language Processing LLP, New Delhi
[3]Potamu Research Ltd
[4]University of Colorado at Boulder
[5]New York University

Abstract

We present the findings of the LoResMT 2021 shared task which focuses on machine translation (MT) of COVID-19 data for both low-resource spoken and sign languages. The organization of this task was conducted as part of the fourth workshop on technologies for machine translation of low resource languages (LoResMT). Parallel corpora is presented and publicly available which includes the following directions: English↔Irish, English↔Marathi, and Taiwanese Sign language↔Traditional Chinese. Training data consists of 8112, 20933 and 128608 segments, respectively. There are additional monolingual data sets for Marathi and English that consist of 21901 segments. The results presented here are based on entries from a total of eight teams. Three teams submitted systems for English↔Irish while five teams submitted systems for English↔Marathi. Unfortunately, there were no systems submissions for the Taiwanese Sign language↔Traditional Chinese task. Maximum system performance was computed using BLEU and follow as 36.0 for English–Irish, 34.6 for Irish–English, 24.2 for English–Marathi, and 31.3 for Marathi–English.

1 Introduction

The workshop on technologies for machine translation of low resource languages (LoResMT)[1] is a yearly workshop which focuses on scientific research topics and technological resources for machine translation (MT) using low-resource languages. Based on the success of its three predecessors (Liu, 2018; Karakanta et al., 2019, 2020), the fourth LoResMT workshop introduces a shared task section based on COVID-19 and sign language data as part of its research objectives. The hope is to provide assistance with translation for low-resource languages where it could be needed most during the COVID-19 pandemic.

[1]https://sites.google.com/view/loresmt/

To provide a trajectory of the LoResMT shared task success, a summary of the previous tasks follows. The first LoResMT shared task (Karakanta et al., 2019) took place in 2019. There, monolingual and parallel corpora for Bhojpuri, Magahi, Sindhi, and Latvian were provided as training data for two types of machine translation systems: neural and statistical. As an extension to the first shared task, a second shared task (Ojha et al., 2020) was presented in 2020 which focused on zero-shot approaches for MT systems.

This year, the shared task introduces a new objective focused on MT systems for COVID-related texts and sign language. Participants for this shared task were asked to submit novel MT systems for the following language pairs:

- English↔Irish
- English↔Marathi
- Taiwanese Sign Language↔Traditional Chinese

The low-resource languages presented in this shared task were found to be sufficient data for baseline systems to perform translation on the latest COVID-related texts and sign language. Irish, Marathi, and Taiwanese Sign Language can be considered low-resource languages and are translated to either English or traditional Chinese – their high-resource counterpart.

The rest of our work is organized as follows. Section 2 presents the setup and schedule of the shared task. Section 3 presents the data set used for the competition. Section 4 describes the approaches used by participants in the competition and Section 5 presents and analyzes the results obtained by the competitors. Lastly, in Section 6 a conclusion is presented along with potential future work.

2 Shared task setup and schedule

This section describes how the shared task was organized along with the systems. Registered participants were sent links to the training, development, and/or monolingual data (refer to Section 3 for more details). They were allowed to use additional data to train their system with the condition that any additional data used should be made publicly available. Participants were moreover allowed to use pre-trained word embeddings and linguistic models that are publicly available. As a manner of detecting which data sets were used during training, participants were given the following markers for denotation:

- *"-a"* - Only provided development, training and monolingual corpora.
- *"-b"*- Any provided corpora, plus publicly available language's corpora and pre-trained/linguistic model (e.g. systems used pre-trained word2vec, UDPipe, etc. model).
- *"-c"* - Any provided corpora, plus any publicly external monolingual corpora.

Each team was allowed to submit any number of systems for evaluation and their best 3 systems were included in the final ranking presented in this report. Each submitted system was evaluated on standard automatic MT evaluation metrics; BLEU (Papineni et al., 2002), CHRF (Popović, 2015) and TER (Post, 2018).

The schedule for deliver of training data and release of test data along with notification and submission can be found in Table 1.

Date	Event
May 10, 2021	Release of training data
July 01, 2021	Release of test data
July 13, 2021	Submission of the systems
July 20, 2021	Notification of results
July 27, 2021	Submission of shared task papers
August 01, 2021	Camera-ready

Table 1: LoResMT 2021 Shared Task programming

3 Languages and data sets

In this section, we present background information about the languages and data sets featured in the shared task along with a itemized view of the linguistic families and number of segments in Table 2.

3.1 Training data set

- **English↔Irish** Irish (also known as Gaeilge) has around 170,000 L1 speakers and "1.85 million (37%) people across the island (of Ireland) claim to be at least somewhat proficient with the language". In the Republic of Ireland, it is the national and first official language. It is also one of the official languages of the European Union and a recognized minority language in Northern Ireland with the ISO *ga* code.[2]

  English-Irish bilingual COVID sentences/documents were extracted and aligned from the following sources: (a) Gov.ie[3] - Search for services or information , (b) Ireland's Health Services[4] - HSE.ie , (c) Revenue Irish Taxes and Customs[5] and (d) Europe Union[6]. In addition, the Irish bilingual training data was built from monolingual data using back translation (Sennrich et al., 2016). English and Irish monolingual data was compiled from Wikipedia pages and newspapers such as The Irish Times[7], RTE[8] and COVID-19 pandemic in the Republic of Ireland[9]. Back-translated and crawled data were cross-validated for accuracy by language experts leaving approximately 8,112 Irish parallel sentences for the training data set.

- **English↔Marathi** Marathi, which has the ISO code *mr*, is dominantly spoken in India's Maharashtra state. It has around 83,026,680 speakers.[10] It belongs to the Indo-Aryan language family.

  English–Marathi parallel COVID sentences were extracted from the Government of India website and online newspapers such as PMIndia[11], myGOV[12], Lokasatta[13], BBC

[2] https://cloud.dfki.de/owncloud/index.php/s/sAs23JKXRwEEacn
[3] www.gov.ie
[4] https://www.hse.ie/
[5] https://www.revenue.ie/
[6] https://europa.eu
[7] https://www.irishtimes.com/
[8] https://www.rte.ie/news/ & https://www.rte.ie/gaeilge/
[9] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_Republic_of_Ireland
[10] https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf
[11] https://www.pmindia.gov.in/
[12] https://www.mygov.in/
[13] https://www.loksatta.com/

Marathi and English[14]. After pre-processing and manual validation, approximately 20,993 parallel training sentences were left. Additionally, English and Marathi monolingual sentences were crawled from the online newspapers and Wikipedia (see Table 2).

- **Taiwanese Sign Language ↔ Traditional Chinese** According to UN, there are "72 million deaf people worldwide... they use more than 300 different sign languages."[15] In Taiwan, Taiwanese Sign Language is a recognized national language, with a population of less than thirty thousand "speakers". Taiwanese Sign Language (and Korean Sign Language) evolved from Japanese Sign Language and share about 60% of "words" between them.
  The sign language data set is prepared from press conferences for COVID-19 response, which were held daily or weekly depending on the pandemic situation in Taiwan. Fig. 1 shows a sample video of sign language and its translations in Traditional Chinese (excerpted from the corpus) and English.

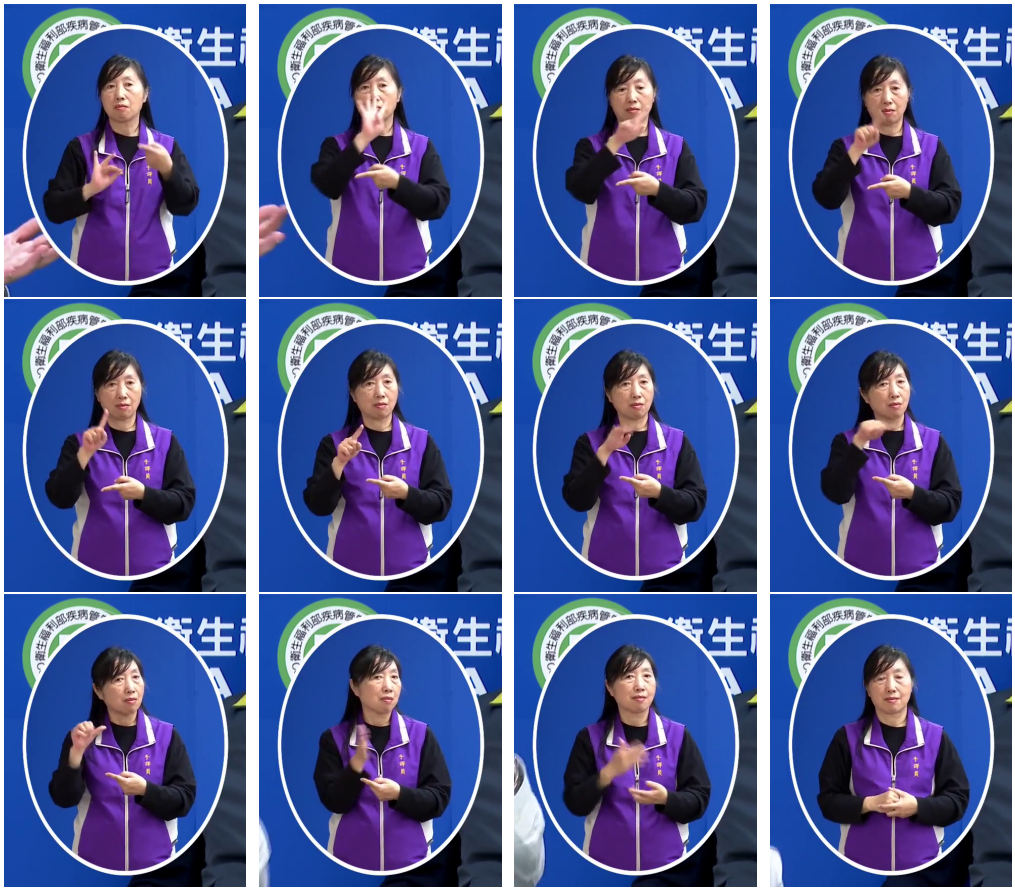

Figure 1: Sample of a sign language video in frames (excerpted from C00207_711.mp4 in the corpus; Translations in Traditional Chinese: "4.1吧或4.5 大概是這樣的一個比例", and English: "The ratio is approximately 4.1 or 4.5")

[14] https://www.bbc.com/marathi & https://www.bbc.com/
[15] https://www.un.org/en/observances/sign-languages-day

3.2 Development and test data sets

Similar to the training data, English-Irish and English-Marathi language pair's dev and test data sets were crawled from bilingual and/or monolingual websites. Additionally, some parallel segments and terminology were taken from the Translation Initiative for COVID-19 (Anastasopoulos et al., 2020), a manually translated and validated data set created by professional translators and native speakers of the target languages. The participants of the shared task were provided with the manual translations of which 502 Irish and 500 Marathi development segments were used while 250 (Irish-English), 500 (English-Irish), 500 (English-Marathi) and 500 (Marathi-English) manually translated segments were used for testing. Taiwanese Sign Language ↔ Traditional Chinese language pair's participants were provided with 3071 segments and videos for development and 7,053 videos for sign language testing.

The detailed statistics of the data set in each language is provided in Table2. The complete shared task data sets are available publicly[16].

Language	Code	Family	Train	Dev	Monolingual	Test
English	en	Indo-Germanic	-	-	8,826	-
Irish	ga	Celtic	8112	502	-	750
Marathi	mr	Indo-Aryan	20,933	500	21,902	1,000
TSign	sgTW	Japanese Sign Language	128,608	3,071	-	7,053
TChinese	zhTW	Mandarin Chinese	128,608	3,071	-	7,053

Table 2: Statistics of the Shared task data (TSign refers to Taiwanese Sign Language and TChinese refers to Traditional Chinese)

4 Participants and methodology

A total of 12 teams registered for the shared task: 5 teams registered to participate for all language pairs, 5 teams registered to participate only for English↔Marathi, one team registered for Taiwanese↔Mandarin (Traditional Chinese) sign language and one team registered for English↔Irish. Out of these, a total of 6 teams submitted their systems on COVID while none of them submitted a system for sign language. Out of the submitted systems, two teams participated for the English↔Irish and English↔Marathi tasks, one team participated for English-Irish and three teams participated for English↔Marathi (see Table 3). All the teams who submitted their systems were invited to submit system description papers describing their experiments. Table 3 identifies the participating teams and their language choices.

Team	English–Irish	English–Marathi	TSign–TChinese	System Description Paper
IIITT	en2ga & ga2en	en2mr & mr2en	—	(Puranik et al., 2021)
oneNLP-IIITH	—	en2mr & mr2en	—	(Mujadia and Sharma, 2021)
A3108	—	en2mr & mr2en	—	(Yadav and Shrivastava, 2021)
CFILT-IITBombay	—	en2mr & mr2en	—	(Jain et al., 2021)
UCF	en2ga & ga2en	en2mr & mr2en	—	(Chen and Fazio, 2021)
adapt_dcu	en2ga	—	—	(Lankford et al., 2021)
Total	**3**	**5**	**0**	**6**

Table 3: Details of the teams and submitted systems for the LoResMT 2021 Shared Task.

Next, we give a short description of the approaches used by each team to build their systems. More details about the approaches can be found in the papers by respective teams in the accompanying proceeding.

[16] https://github.com/loresmt/loresmt-2021

- **IIITT** (Puranik et al., 2021) used a fairseq pre-trained model Indictrans for English-Marathi. It consists of two models that can translate from Indic to English and vice-versa. The model can perform 11 languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu pre-trained on the Samanantar data set, the largest data set for Indic languages during the time of submission. The model is fine-tuned on the training data set provided by the organizers and a parallel bible corpus for Marathi. The team used the parallel bible parallel corpus from a previous task (MultiIndicMT task in WAT 2020). After conducting various experiments, the best checkpoint was recorded and predicted upon. For Irish, the team fine-tuned an Opus MT model from Helsinki NLP on the training data set, and then predicted results after recording. After careful experimentation, the team observed that the Opus MT model outperformed the other models giving it the highest scoring model award.

- **oneNLP-IIITH** (Mujadia and Sharma, 2021) used a sequence to sequence neural model with a transformer network (4 to 8 layers) with label smoothing and dropouts to reduce overfitting with English-Marathi and Marathi-English. The team explored the use of different linguistic features like part-of-speech and morphology on sub-word units for both directions. In addition, the team explored forward and backward translation using web-crawled monolingual data.

- **A3108** (Yadav and Shrivastava, 2021) built a statistical machine translation (smt) system in both directions for English↔Marathi language pair. Its initial baseline experiments used various tokenization schemes to train models. By using optimal tokenization schemes, the team was able to create synthetic data and train an augmented dat aset to create more statistical models. Also, the team reordered English syntax to match Marathi syntax and further trained another set of baseline and data augmented models using various tokenization schemes.

- **CFILT-IITBombay** (Jain et al., 2021) buildt three different neural machine translation systems; a baseline English–Marathi system, a Baseline Marathi-English system, and a English–Marathi system that was based on back translation. The team explored the performance of the NMT systems between English and Marathi languages. Also, they explored the performance of back-translation using data obtained from NMT systems trained on a very small amount of data. From their experiments, the team observed that back-translation helped improve the MT quality over the baseline for English-Marathi.

- **UCF** (Chen and Fazio, 2021) used transfer learning, uni-gram and sub-word segmentation methods for English–Irish, Irish–English, English–Marathi and Marathi–English. The team conducted their experiment using an OpenNMT LSTM system. Efforts were constrained by using transfer learning and sub-word segmentation on small amounts of training data. Their models achieved the following BLEU scores when constraining on tracks of English–Irish, Irish–English, and Marathi–English: 13.5, 21.3, and 17.9, respectively.

- **adapt_dcu** (Lankford et al., 2021) used a transformer training approach carried out using OpenNMT-py and sub-word models for English–Irish. The team also explored domain adaptation techniques while using a Covid-adapted generic 55k corpus, fine-tuning, mixed fine-tuning and combined data set approaches were compared with models trained on an extended in-domain data set.

5 Results

As discussed, participants were allowed to use data sets other than those provided. The best three results for English-Irish, Irish-English, English-Marathi and Marathi-English language

pairs are presented in Tables 4 and 5. The complete submitted systems results are available publicly[17]. Table 4 depicts how the UCF team were able to gain the highest and lowest results for Irish-English and English-Marathi with shared data. The highest scores were 21.3 BLEU, 0.45 CHRF and 0.711 TER, while the lowest scores were 5.1 BLEU, 0.22 CHRF and 0.872 TER. However, with the additional data and by using pre-trained models (see Table 5), adapt_dcu achieved the best results for English-Irish where scores were 36 BLEU, 0.6 CHRF and 0.531 TER. Contrastingly, UCF scored the lowest for English-Marathi. The lowest scores were 4.8 BLEU, 0.29 CHRF and 1.063 TER.

Team	System/task description	BLEU	CHRF	TER
adapt_dcu	en2ga-a	9.8	0.34	0.880
UCF	ga2en-TransferLearning-a	21.3	0.45	0.711
CFILT-IITBombay	en2mr-Backtranslation-a	12.2	0.38	0.979
CFILT-IITBombay	en2mr-Baseline_200-a	11	0.38	0.961
CFILT-IITBombay	en2mr-Baseline_1600-a	10.8	0.38	0.935
oneNLP-IIITH	en2mr-Method1-a	10.4	0.32	0.907
A3108	en2mr-Method29transliterate-a	11.8	0.45	0.95
A3108	en2mr-Method29unk-a	11.8	0.45	0.95
A3108	en2mr-Method10unk-a	11.4	0.43	0.934
UCF	en2mr-UnigramSegmentation-a	5.1	0.22	0.872
CFILT-IITBombay	mr2en-Baseline_1000-a	16.6	0.41	0.870
CFILT-IITBombay	mr2en-Baseline_1200-a	16.3	0.40	0.867
CFILT-IITBombay	mr2en-Baseline_1400-a	16.2	0.41	0.879
oneNLP-IIITH	mr2en-Method1-a	16.7	0.40	0.835
oneNLP-IIITH	mr2en-Method2-a	16.2	0.41	0.831
A3108	mr2en-Method7transliterate-a	14.6	0.47	0.945
A3108	mr2en-Method7unk-a	14.6	0.47	0.945
A3108	mr2en-Method20transliterate-a	14.5	0.42	0.866
UCF	mr2en-UnigramSegmentation-a	17.9	0.40	0.744

Table 4: Results of submitted systems at English↔Irish & English↔Marathi in the "-a" method

[17] https://github.com/loresmt/loresmt-2021

Team	System/task description	BLEU	CHRF	TER
adapt_dcu	en2ga-b	36.0	0.60	0.531
IIITT	en2ga-helsnikiopus-b	25.8	0.53	0.629
IIITT	ga2en-helsinkiopus-b	34.6	0.61	0.586
IIITT	en2mr-IndicTrans-b	24.2	0.59	0.597
oneNLP-IIITH	en2mr-Method2-c	22.2	0.56	0.746
oneNLP-IIITH	en2mr-Method3-c	22.0	0.56	0.753
oneNLP-IIITH	en2mr-Method1-c	21.5	0.56	0.746
UCF	en2mr-UnigramSegmentation-b	4.8	0.29	1.063
oneNLP-IIITH	mr2en-Method3-c	31.3	0.58	0.646
oneNLP-IIITH	mr2en-Method2-c	30.6	0.57	0.659
oneNLP-IIITH	mr2en-Method1-c	20.7	0.48	0.735
UCF	mr2en-UnigramSegmentation-b	7.7	0.24	0.833
IIITT	mr2en-IndicTrans-b	5.1	0.22	1.002

Table 5: Results of submitted systems at English↔Irish & English↔Marathi in the "-b" and "-c" method

6 Conclusion

We have reported the findings of the LoResMT 2021 Shared Task on COVID and sign language translation for low-resource languages as part of the fourth LoResMT workshop. All submissions used neural machine translation except for the one from oneNLP-IIITH. We conclude that in our shared tasks the use of transfer learning, domain adaptation, and back translation achieve optimal results when the data sets are domain specific as well as small-sized. Our findings show that uni-gram segmentation transfer learning methods provide comparatively low results for the following metrics: BLEU, CHRF and TER. The highest BLEU scores achieved are 36.0 for English-to-Irish, 34.6 for Irish-to-English, 24.2 for English-to-Marathi, and 31.3 for Marathi-to-English.

In future iterations of the LoResMT shared tasks, extended corpora of the three language pairs will be provided for training and evaluation. Human evaluation on system results will also be conducted. For sign language MT, the tasks will be fine-grained and evaluated separately.

7 Acknowledgements

This publication has emanated from research in part supported by Cardamom-Comparative Deep Models of Language for Minority and Historical Languages (funded by the Irish Research Council under the Consolidator Laureate Award scheme (grant number IRCLA/2017/129)) and we are grateful to them for providing English↔Irish parallel and monolingual COVID-related texts. We would like to thank Panlingua Language Processing LLP and Potamu Research Ltd for providing English↔Marathi parallel and monolingual COVID data and Taiwanese Sign Language↔Traditional Chinese linguistic data, respectively.

References

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P., Lazar, R., Lewis, W., Neubig, G., Niu, M., Öktem, A., Paquin, E., Tang, G., and Tur, S. (2020). TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Chen, W. and Fazio, B. (2021). The UCF Systems for the LoResMT 2021 Machine Translation Shared Task. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

Jain, A., Mhaskar, S., and Bhattacharyya, P. (2021). Evaluating the Performance of Back-translation for Low Resource English-Marathi Language Pair: CFILT-IITBombay @ LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

Karakanta, A., Ojha, A. K., Liu, C.-H., Abbott, J., Ortega, J., Washington, J., Oco, N., Lakew, S. M., Pirinen, T. A., Malykh, V., Logacheva, V., and Zhao, X., editors (2020). *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, Suzhou, China. Association for Computational Linguistics.

Karakanta, A., Ojha, A. K., Liu, C.-H., Washington, J., Oco, N., Lakew, S. M., Malykh, V., and Zhao, X. (2019). Proceedings of the 2nd workshop on technologies for mt of low resource languages. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*.

Lankford, S., Afli, H., and Way, A. (2021). Machine Translation in the Covid domain: an English-Irish case study for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

Liu, C.-H., editor (2018). *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, Boston, MA. Association for Machine Translation in the Americas.

Mujadia, V. and Sharma, D. M. (2021). English-Marathi Neural Machine Translation for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

Ojha, A. K., Malykh, V., Karakanta, A., and Liu, C.-H. (2020). Findings of the LoResMT 2020 shared task on zero-shot for low-resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 33–37, Suzhou, China. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Popović, M. (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.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Puranik, K., Hande, A., Priyadharshini, R., D, T., Sampath, A., Thamburaj, K. P., and Chakravarthi, B. R. (2021). Attentive fine-tuning of Transformers for Translation of low-resourced languages @LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

Sennrich, R., Haddow, B., and Birch, A. (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.

Yadav, S. and Shrivastava, M. (2021). A3-108 Machine Translation System for LoResMT Shared Task @MT Summit 2021 Conference. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*.

A3-108 Machine Translation System for LoResMT Shared Task @MT Summit 2021 Conference

Saumitra Yadav saumitra.yadav@research.iiit.ac.in
Manish Shrivastava m.shrivastava@iiit.ac.in
Machine Translation - Natural Language Processing Lab, Language Technologies Research Centre, Kohli Center on Intelligent Systems, International Institute of Information Technology - Hyderabad

Abstract

In this paper, we describe our submissions for LoResMT Shared Task @MT Summit 2021 Conference. We built statistical translation systems in each direction for English $\Longleftrightarrow$ Marathi language pair. This paper outlines initial baseline experiments with various tokenization schemes to train models. Using optimal tokenization scheme we create synthetic data and further train augmented dataset to create more statistical models. Also, we reorder English to match Marathi syntax to further train another set of baseline and data augmented models using various tokenization schemes. We report configuration of the submitted systems and results produced by them.

1 Introduction

Machine Translation systems are systems which translate from source language to target. There are multiple ways of creating such a system - rule based, data driven, hybrid etc. We are using data driven methods to create translation system. In data driven methods - statistical (Koehn et al., 2003) and neural methods (Bahdanau et al., 2014) have been employed to build decent MT systems in resource setting like English $\Longleftrightarrow$ French. In LoResMT shared task (Ojha et al., 2021) we are dealing with low resource setting for English, Marathi pair. According to Koehn and Knowles (2017), compared to statistical methods neural methods have a drawback when used in low resource setting. Hence, for this shared task we are using only phrase based statistical models to build translation models using Moses[1] (Koehn et al., 2007).

Marathi is morphologically richer, agglutinative language when compared to English. Also, former follows SOV as canonical syntactic structure while latter follows SVO. Level of difference in morphological richness and syntactic divergence between the two languages suggests to look for methods which can help to address them to certain extent in phrase based statistical models. Since we are in low resource setting, to address data sparsity problem, we use various tokenization schemes, e.g. BPE (Sennrich et al., 2016b), morfessor (Virpioja et al., 2013). Combinations of these tokenization schemes are used with SMT based method to create a baseline systems. After checking the optimal tokenization scheme, we use that scheme to augment training data with synthetic dataset using back translation (Sennrich et al., 2016a). As was the case in baseline systems, augmented dataset goes through prepossessing with various tokenization schemes and SMT method to build more systems. We elevate the amount of learning, the reordering model of SMT has to do, by making use of rule based reordering system

[1] http://statmt.org/moses/