AMTA 2018
March 17 - 21, 2018
Boston, MA, USA

The 13th Conference of The Association for Machine Translation in the Americas

www.conference.amtaweb.org

WORKSHOP PROCEEDINGS

March 21, 2018

Technologies for MT of Low Resource Languages (LoResMT 2018)

Organizer: Chao-Hong Liu *(ADAPT Centre, Dublin City University)*

Contents

I Introduction

II Organizing Committee

III Program Committee

IV Program

V Invited Talk 1: Research and Development of Information Processing Technologies for Chinese Minority/Cross-border Languages

Bei Wang and Xiaobing Zhao

VI Invited Talk 2: DeepHack.Babel: Translating Data You Cannot See

Valentin Malykh and Varvara Logacheva

1 Using Morphemes from Agglutinative Languages like Quechua and Finnish to Aid in Low-Resource Translation

John E. Ortega and Krishnan Pillaipakkamnatt

12 SMT versus NMT: Preliminary comparisons for Irish

Meghan Dowling, Teresa Lynn, Alberto Poncelas and Andy Way

21 Tibetan-Chinese Neural Machine Translation based on Syllable Segmentation

Wen Lai, Xiaobing Zhao and Wei Bao

30 A Survey of Machine Translation Work in the Philippines: From 1998 to 2018

Nathaniel Oco and Rachel Edita Roxas

37 Semi-Supervised Neural Machine Translation with Language Models

Ivan Skorokhodov, Anton Rykachevskiy, Dmitry Emelyanenko, Sergey Slotin and Anton Ponkratov

81 System Description of Supervised and Unsupervised Neural Machine Translation Approaches from “NL Processing” Team at DeepHack.Babel Task

Ilya Gusev and Artem Oboturov

53 Apertium’s Web Toolchain for Low-Resource Language Technology

Sushain Cherivirala, Shardul Chiplunkar, Jonathan North Washington and Kevin Brubeck Unhammer

Introduction
AMTA 2018 Workshop on
Technologies for MT of Low Resource Languages
(LoResMT 2018)

Recently we have observed the developments of cross-lingual NLP tools, e.g. MLP 2017 Shared Tasks on Cross-lingual Word Segmentation and Morpheme Segmentation and IJCNLP 2017 Shared Task on Customer Feedback Analysis. The results showed clearly now we are able to build one NLP system for multiple languages in a specific task and the system can perform very well compared to its monolingual counterparts. The development of this kind of cross-lingual tools will be very beneficial to the many low resource languages and will definitely improve machine translation (MT) performance for these languages. We would like to see if this idea could be further extended and realized in other NLP tools, e.g. several kinds of word tokenizers/de-tokenizers, morphology analyzers, and what impacts these tools could bring to MT systems.

In this workshop, we solicit work on the NLP tools as well as research on MT systems/methods for low resource languages in general. The scopes of the workshop are not limited to these tools for MT pre-processing and post-processing. We would like to bring together researchers who work on these topics and help review/overview what are the most important tasks we need from these tools for MT in the following years.

Two speeches on organized events dedicated to this line of research in China and Russia will be given. Our speakers will also give the overview of NLP tools developed for and research on minority languages in China and Russia. Seven papers are archived in the proceedings, in which languages involved include Catalan, Finnish, Filipino, Irish, Korean, Latvian, Quechua, Russian, Sámi, Tibetan, Turkic languages, as well as Mandarin Chinese, French and English.

I would like to express my sincere gratitude to the many researchers who helped as advisers, organizers, and reviewers and made the workshop successful. They are Alberto Poncelas, Alex Huynh, Alina Karakanta, Daria Dzendzik, Erlyn Manguilimotan, Francis Tyers, Hamidreza Ghader, Iacer Calixto, Ian Soboroff, Jonathan Washington, Josef van Genabith, Koel Dutta Chowdhury, Majid Latifi, Marzieh Fadaee, Nathaniel Oco, Peyman Passban, Prachya Boonkwan, Qun Liu, Sangjie Duanzhu, Santanu Pal, Sivaji Bandyopadhyay, Sudip Kumar Naskar, Thepchai Supnithi, Tommi Pirinen, Valentin Malykh, Vinit Ravishankar, Wei Bao, Yalemisew Abgaz, as well as colleagues in ADAPT Centre. I am thankful to AMTA organizers Steve Richardson, Priscilla Rasmussen and Olga Beregovaya for their continuous help on the workshop from the very beginning. We are very grateful to the authors who submitted their work to the workshop. Xiaobing Zhao, Bei Wang, Valentin Malykh and Varvara Logacheva, who prepared the two invited speeches are much appreciated. Thank you so much!

Boston, March 2018

Chao-Hong Liu
Workshop Chair
ADAPT Centre
Dublin City University
Glasnevin Dublin 9, Ireland.

Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. The organization of this workshop has partially received funding from the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie Actions (Grant No. 734211).

Organizing Committee

Organizers

Alina Karakanta	Universität des Saarlandes
Chao-Hong Liu	ADAPT Centre, Dublin City University
Daria Dzendzik	ADAPT Centre, Dublin City University
Erlyn Manguilimotan	Weathernews Inc., Japan, formerly with NAIST
Francis Tyers	Higher School of Economics, National Research University
Iacer Calixto	ADAPT Centre, Dublin City University
Ian Soboroff	National Institute of Standards and Technology (NIST)
Jonathan Washington	Swarthmore College
Majid Latifi	Universitat Politècnica de Catalunya - BarcelonaTech
Nathaniel Oco	National University (Philippines)
Peyman Passban	ADAPT Centre, Dublin City University
Prachya Boonkwan	National Electronics and Computer Technology Center
Sangjie Duanzhu	Qinghai Normal University
Santanu Pal	Universität des Saarlandes
Sivaji Bandyopadhyay	Jadavpur University
Sudip Kumar Naskar	Jadavpur University
Thepchai Supnithi	National Electronics and Computer Technology Center
Tommi A Pirinen	Universität Hamburg
Valentin Malykh	Moscow Institute of Physics and Technology
Vinit Ravishankar	Charles University in Prague
Yalemisew Abgaz	ADAPT Centre, Dublin City University

Program Committee

Reviewers

Alberto Poncelas	ADAPT Centre, Dublin City University
Alex Huynh	CLC Center, University of Science, VNU-HCMC-VN
Alina Karakanta	Universität des Saarlandes
Chao-Hong Liu	ADAPT Centre, Dublin City University
Daria Dzendzik	ADAPT Centre, Dublin City University
Erlyn Manguilimotan	Weathernews Inc., Japan, formerly with NAIST
Francis Tyers	Higher School of Economics, National Research University
Hamidreza Ghader	University of Amsterdam
Iacer Calixto	ADAPT Centre, Dublin City University
Jonathan Washington	Swarthmore College
Koel Dutta Chowdhury	ADAPT Centre, Dublin City University
Majid Latifi	Universitat Politècnica de Catalunya - BarcelonaTech
Marzieh Fadaee	University of Amsterdam
Nathaniel Oco	National University (Philippines)
Peyman Passban	ADAPT Centre, Dublin City University
Santanu Pal	Universität des Saarlandes
Sivaji Bandyopadhyay	Jadavpur University
Sudip Kumar Naskar	Jadavpur University
Thepchai Supnithi	National Electronics and Computer Technology Center
Tommi A Pirinen	Universität Hamburg
Valentin Malykh	Moscow Institute of Physics and Technology
Vinit Ravishankar	Charles University in Prague
Yalemisew Abgaz	ADAPT Centre, Dublin City University

LoResMT 2018 Program

Session 1

09:00AM–10:30AM

Introduction to LoResMT 2018 Workshop Chao-Hong Liu

INVITED TALK
Research and Development of Information Processing Technologies for Chinese Minority/Cross-border Languages Bei Wang & Xiaobing Zhao

Using Morphemes from Agglutinative Languages like Quechua and Finnish to Aid in Low-Resource Translation John E. Ortega & Krishnan Pillaipakkamnatt

Break **10:30AM–11:00AM**

Session 2

11:00AM–12:30AM

SMT versus NMT: Preliminary comparisons for Irish
Meghan Dowling, Teresa Lynn, Alberto Poncelas & Andy Way

Tibetan-Chinese Neural Machine Translation based on Syllable Segmentation
Wen Lai, Xiaobing Zhao & Wei Bao

A Survey of Machine Translation Work in the Philippines: From 1998 to 2018
Nathaniel Oco & Rachel Edita Roxas

Lunch **12:30PM–02:00PM**

Session 3 **DeepHack.Babel**

02:00PM–03:30PM

INVITED TALK
DeepHack.Babel: Translating Data You Cannot See
Valentin Malykh & Varvara Logacheva

Semi-Supervised Neural Machine Translation with Language Models
Ivan Skorokhodov, Anton Rykachevskiy, Dmitry Emelyanenko, Sergey Slotin & Anton Ponkratov

System Description of Supervised and Unsupervised Neural Machine Translation Approaches from “NL Processing” Team at DeepHack.Babel Task Ilya Gusev & Artem Oboturov

Break **03:30PM–04:00PM**

Session 4 **Tools, Demos, Discussions**

04:00PM–05:30PM

Apertium’s Web Toolchain for Low-Resource Language Technology
Sushain Cherivirala, Shardul Chiplunkar, Jonathan North Washington & Kevin Brubeck Unhammer

Demos and Discussions

Closing.

Invited Talk
Research and Development of Information Processing Technologies for Chinese Minority/Cross-border Languages

Bei Wang bjwangbei@qq.com
National Language Resource Monitoring and Research Center,
Minority Languages Branch, Minzu University of China

Xiaobing Zhao nmzxb_cn@163.com
National Language Resource Monitoring and Research Center,
Minority Languages Branch, Minzu University of China

Natural Language Processing for Chinese minority languages is a challenging and important task. In this talk, we will present the current status of Chinese minority languages, including the situations in general of 56 ethnic groups in China and 33 different groups of cross-border languages used by 30 cross-border ethnic population in China. Secondly, research on minority languages information processing and its difficulties and challenges will be presented. We will also introduce notable projects and scientific publications on minority languages information processing. Specifically, we will give an overview of the minority language word segmentation task we held in 2017, as well as the work we have done on Chinese minority languages natural language processing in our center.

Biography

Prof. Xiaobing Zhao is the chair of National Language Resource Monitoring & Research Center, Minority Languages Branch, Minzu University of China. Xiaobing Zhao obtained her M.S. degree in Artificial Intelligence from Chungwoon University in 2003, and her Ph.D. in Computational Linguistics in Beijing Language & Culture University in 2007. Xiaobing Zhao has published more than 50 papers, authored 2 books and 1 China National invention patent. She has supervised 13 PhD students and 10 master students. Xiaobing Zhao has undertaken a number of small or large scale projects, including Chinese NSFC Key projects, "863" Key projects, provincial and ministerial Key projects, among others. Xiaobing Zhao has won the first prize of Qian Weichang Science and Technology Award, second prize of Science & technology Development/Achievement of the Ministry of Education the ministry of education, first prize in scientific and technological achievements of the State Archives Administration of the People's Republic of China, and other awards.

Invited Talk
DeepHack.Babel:
Translating Data You Cannot See

Valentin Malykh valentin.malykh@phystech.edu
Moscow Institute of Physics and Technology

Varvara Logacheva logacheva.vk@mipt.ru
Moscow Institute of Physics and Technology

Neural networks were introduced in Machine Translation (MT) quite recently and immediately became state of the art in the field. Today their advantage over phrase-based statistical MT systems is unquestionable and neural networks are used in the majority of online MT engines. What is more important, neural MT beats SMT not only in usual data-rich scenario, but also allows accomplishing tasks which are impossible for SMT systems.

One of the recent notable advances in the field of MT is training of a translation model without parallel data. This task could not be fulfilled by SMT systems which need sentence-aligned datasets to extract the translation variants. One approach is to use a model called auto-encoder to train neural Language Models (LMs) for source and target languages. Such LMs create a representation of a text in a multi-dimensional space. These spaces can be merged so that both source and target sentences can be represented in the same space. With such a pair of LMs a source text can be converted to a vector in this space and then to a textual representation in the target language — which gives us an MT system trained on two unrelated monolingual corpora.

Our hackathon was inspired by these works. In order to make the task easier we relaxed the task. Instead of unsupervised MT we formulated the task of the hackathon as semi-supervised MT — MT which is trained on a very small parallel dataset and larger monolingual corpora for both source and target languages. This scenario is also more realistic. While pairs of languages with no parallel data are quite rare, there exist many pairs where parallel texts exist, but are scarce. In such cases it is desirable to improve MT systems with monolingual corpora.

The hackathon lasted for five days, which of course did not allow to train a state-of-the-art MT model. However, our corpora were relatively small (50,000 parallel sentences and 0.5–1 million monolingual sentences). We also imposed restrictions on the size of models and on training time, so that participants have time to train their models, get results and make changes several times within these five days. In order to spare participants from re-implementing existing models we allowed use of open-source implementations of MT systems. On the other hand, the task was complicated by domain discrepancy between the parallel and monolingual data.

One of the features of our hackathon was hidden training data. We had three sets of corpora, each consisting of a small parallel training corpus, a pair of relatively fair in size monolingual corpora and small parallel dataset for testing. One of these sets was given to participants to make them familiar with the format and let them tune their models. That was Ru-En set, since most of the participants are proficient in both languages. Another set for a different language pair was used for the evaluation of models during hackathon. Neither the dataset itself nor the language pair were not disclosed to participants. They submitted their models, they were trained and tested in a blind mode. The third (hidden) dataset was used to produce the final scores and define the winner in the same manner. The first and the third sets share a feature of domain skew between parallel corpus and monolingual corpora, while the second does not.

The main aim of hiding the data was to prevent participants from cheating — e.g. collecting extra data for a language pair instead of training on the provided datasets. In addition to that, it served for making the scenario more language-independent. The majority of algorithms used for the MT task are in principle language-independent in the sense that they use the same principles for most of languages. Despite that, many common techniques are more effective on languages with poor morphology and "standard" (Subject-Verb-Object) predominant word order. The success of MT systems for other languages often depends not on algorithms, but on additional tweaks which deal with word order errors and rich morphology. By ruling out the information about properties of source and target languages we check the efficiency of MT algorithms themselves and have a possibility to see to what extent they can handle the challenges of a particular language pair.

The majority of participants used the following scenario. They used parallel corpora to train an initial translation model which was initialised with word embeddings trained on monolingual corpora. After that they translated monolingual datasets with this model and retrained it using this synthetic parallel data. Nevertheless, some teams used unsupervised models as a part of ensemble models.

Biographies

Valentin Malykh is a researcher and a PhD student at Moscow Institute of Physics and Technology. His research interests are dialogue systems, in particular robustness to noise in input and style transfer in dialogues. Valentin has been co-organising DeepHack events since 2016, and he also co-organised ConvAI — a competition of chatbots which took part at NIPS-2017.

Varvara Logacheva is a researcher in Moscow Institute of Physics and Technology. She got a PhD in Computer Science from the University of Sheffield, where she was a member of Natural Language Processing group. The main topic of her thesis was Quality Estimation for Machine Translation and its integration into MT systems. Her present research interests are dialogue systems, in particular non-goal-oriented dialogue systems (chatbots) and their automatic and manual evaluation. Varvara has been co-organising track on Quality Estimation for MT at WMT since 2015. She also co-organised ConvAI in 2017.