Deep Learning Approaches for Low-Resource Natural Language Processing (DeepLo)

Proceedings of the Second Workshop

November 3, 2019
Hong Kong, China
Introduction

The EMNLP-IJCNLP 2019 Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing (DeepLo) takes place on Sunday, November 3rd, in Hong Kong, China, immediately before the main conference.

Natural Language Processing is being revolutionized by deep learning with neural networks. However, deep learning requires large amounts of annotated data, and its advantage over traditional statistical methods typically diminishes when such data is not available; for example, SMT continues to outperform NMT in many bilingually resource-poor scenarios. Large amounts of annotated data do not exist for many low-resource languages, and for high-resource languages it can be difficult to find linguistically annotated data of sufficient size and quality to allow neural methods to excel. Our workshop aimed to bring together researchers from the NLP and ML communities who work on learning with neural methods when there is not enough data for those methods to succeed out-of-the-box. Techniques of interest include self-training, paired training, distant supervision, semi-supervised and transfer learning, and human-in-the-loop algorithms such as active learning.

Our call for papers for this second workshop met with a strong response. We received 85 paper submissions, of which 10 were “extended abstracts” with non-archival status—work that will be presented at the workshop, but will not appear in the proceedings in order to allow it to be published elsewhere. We accepted 32 papers and 7 extended abstracts.

Our program covers a broad spectrum of applications and techniques. It is augmented by invited talks from Heng Ji, Barbara Plank, Dan Roth, Kristina Toutanova, and Luke Zettlemoyer.

We would like to thank the members of our Program Committee for their timely and thoughtful reviews.

Colin Cherry, Greg Durrett, George Foster, Reza Haffari, Shahram Khadivi, Nanyun Peng, Xiang Ren, and Swabha Swayamdipta
Organizers:
Colin Cherry, Google
Greg Durrett, University of Texas, Austin
George Foster, Google
Gholamreza (Reza) Haffari, Monash University
Shahram Khadivi, eBay
Nanyun Peng, University of Southern California
Xiang Ren, University of Southern California
Swabha Swayamdipta, Allen Institute for Artificial Intelligence

Invited Speakers:
Heng Ji, University of Illinois Urbana-Champaign
Barbara Plank, IT University of Copenhagen
Dan Roth, University of Pennsylvania
Kristina Toutanova, Google AI Language
Luke Zettlemoyer, University of Washington / Facebook AI Research

Program Committee:

Afshin Rahimi
Alexander Spangher
Ana Marasovic
Daniel Fried
Ekaterina Vylomova
Fei Liu
Gaurav Kumar
Hainan Xu
Jean-David Ruvini
Jiacheng Xu
Jifan Chen
Johnny Wei
Jonathan Kummerfeld
José G. C. de Souza
Julia Kreutzer
Ke Tran
Kenton Lee
Kevin Duh
Luheng He
Melvin Johnson
Mingyu Derek Ma
Nazneen Fatema Rajani
Orhan Firat
Parminder Bhatia
Parnia Bahar

Patrick Littell
Pavel Petrushkov
Peifeng Wang
Pengxiang Cheng
Poorya Zaremoodi
Robin Jia
Rujun Han
Sameen Maruf
Sanjika Hewavitharana
Sarvnaz Karimi
Selcuk Kopru
Shuoyang Ding
Suchin Gururangan
Te-Lin Wu
Tomer Lancewicki
Thamme Gowda
Vered Shwartz
Weiuye Wang
Xiaolei Huang
Xisen Jin
Yasumasa Onoe
Yi Luan
Yuchen Lin
Yunsu Kim
# Table of Contents

**A Closer Look At Feature Space Data Augmentation For Few-Shot Intent Classification**  
Varun Kumar, Hadrien Glaude, Cyprien de Lichy and William Campbell ................................. 1

**A Comparative Analysis of Unsupervised Language Adaptation Methods**  
Gil Rocha and Henrique Lopes Cardoso....................................................................................... 11

**A logical-based corpus for cross-lingual evaluation**  
Felipe Salvatore, Marcelo Finger and Roberto Hirata Jr. ............................................................... 22

**Bad Form: Comparing Context-Based and Form-Based Few-Shot Learning in Distributional Semantic Models**  
Jeroen Van Hautte, Guy Emerson and Marek Rei ....................................................................... 31

**Bag-of-Words Transfer: Non-Contextual Techniques for Multi-Task Learning**  
Seth Ebner, Felicity Wang and Benjamin Van Durme ................................................................. 40

**BERT is Not an Interlingua and the Bias of Tokenization**  
Jasdeep Singh, Bryan McCann, Richard Socher and Caiming Xiong ................................. 47

**Cross-lingual Joint Entity and Word Embedding to Improve Entity Linking and Parallel Sentence Mining**  
Xiaoman Pan, Thamme Gowda, Heng Ji, Jonathan May and Scott Miller ............................... 56

**Deep Bidirectional Transformers for Relation Extraction without Supervision**  
Yannis Papanikolaou, Ian Roberts and Andrea Pierleoni ............................................................. 67

**Domain Adaptation with BERT-based Domain Classification and Data Selection**  
Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati and Bing Xiang ................................... 76

**Empirical Evaluation of Active Learning Techniques for Neural MT**  
Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy and Matthias Paulik 84

**Fast Domain Adaptation of Semantic Parsers via Paraphrase Attention**  
Avik Ray, Yilin Shen and Hongxia Jin ......................................................................................... 94

**Few-Shot and Zero-Shot Learning for Historical Text Normalization**  
Marcel Bollmann, Natalia Korchagina and Anders Søgaard ...................................................... 104

**From Monolingual to Multilingual FAQ Assistant using Multilingual Co-training**  
Mayur Patidar, Surabhi Kumari, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, Lovekesh Vig and Gautam Shroff .......................................................... 115

**Generation-Distillation for Efficient Natural Language Understanding in Low-Data Settings**  
Luke Melas-Kyriazi, George Han and Celine Liang .................................................................... 124

**Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual**  
He He, Sheng Zha and Haohan Wang ......................................................................................... 132

**Metric Learning for Dynamic Text Classification**  
Jeremy Wohlwend, Ethan R. Elenberg, Sam Altschul, Shawn Henry and Tao Lei .................. 143

**Evaluating Lottery Tickets Under Distributional Shifts**  
Shrey Desai, Hongyuan Zhan and Ahmed Aly ........................................................................... 153
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-lingual Parsing with Polyglot Training and Multi-treebank Learning: A Faroese Case Study</td>
<td>James Barry, Joachim Wagner and Jennifer Foster</td>
<td>163</td>
</tr>
<tr>
<td>Inject Rubrics into Short Answer Grading System</td>
<td>Tianqi Wang, Naoya Inoue, Hiroki Ouchi, Tomoya Mizumoto and Kentaro Inui</td>
<td>175</td>
</tr>
<tr>
<td>Instance-based Inductive Deep Transfer Learning by Cross-Dataset Querying with Locality Sensitive Hashing</td>
<td>Somnath Basu Roy Chowdhury, Annervaz M and Ambedkar Dukkipati</td>
<td>183</td>
</tr>
<tr>
<td>Multimodal, Multilingual Grapheme-to-Phoneme Conversion for Low-Resource Languages</td>
<td>James Route, Steven Hillis, Isak Czeresnia Etinger, Han Zhang and Alan W Black</td>
<td>192</td>
</tr>
<tr>
<td>Natural Language Generation for Effective Knowledge Distillation</td>
<td>Raphael Tang, Yao Lu and Jimmy Lin</td>
<td>202</td>
</tr>
<tr>
<td>Neural Unsupervised Parsing Beyond English</td>
<td>Katharina Kann, Anhad Mohananey, Samuel R. Bowman and Kyunghyun Cho</td>
<td>209</td>
</tr>
<tr>
<td>Reevaluating Argument Component Extraction in Low Resource Settings</td>
<td>Anirudh Joshi, Timothy Baldwin, Richard Sinnott and Cecile Paris</td>
<td>219</td>
</tr>
<tr>
<td>Reinforcement-based denoising of distantly supervised NER with partial annotation</td>
<td>Farhad Nooralahzadeh, Jan Tore Lønning and Lilja Øvrelid</td>
<td>225</td>
</tr>
<tr>
<td>Samvaadhana: A Telugu Dialogue System in Hospital Domain</td>
<td>Suma Reddy Duggenpudi, Kusampudi Siva Subrahmanyam Varma and Radhika Mamidi</td>
<td>234</td>
</tr>
<tr>
<td>Towards Zero-resource Cross-lingual Entity Linking</td>
<td>Shuyan Zhou, Shruti Rijhwani and Graham Neubig</td>
<td>243</td>
</tr>
<tr>
<td>Transductive Auxiliary Task Self-Training for Neural Multi-Task Models</td>
<td>Johannes Bjerva, Katharina Kann and Isabelle Augenstein</td>
<td>253</td>
</tr>
<tr>
<td>Weakly Supervised Attentional Model for Low Resource Ad-hoc Cross-lingual Information Retrieval</td>
<td>Lingjun Zhao, Rabih Zbib, Zhuolin Jiang, Damianos Karakos and Zhongqiang Huang</td>
<td>259</td>
</tr>
<tr>
<td>X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension</td>
<td>Mostafa Abdou, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein and Anders Søgaard</td>
<td>265</td>
</tr>
<tr>
<td>Zero-Shot Cross-lingual Name Retrieval for Low-Resource Languages</td>
<td>Kevin Blissett and Heng Ji</td>
<td>275</td>
</tr>
<tr>
<td>Zero-shot Dependency Parsing with Pre-trained Multilingual Sentence Representations</td>
<td>Ke Tran and Arianna Bisazza</td>
<td>281</td>
</tr>
</tbody>
</table>
Conference Program

Sunday, November 3, 2019

7:30–8:50   Breakfast

8:50–9:00   Opening Remarks

9:00–9:45   Invited Talk 1: Heng Ji

9:45–10:30  Poster Session 1

A Closer Look At Feature Space Data Augmentation For Few-Shot Intent Classification
Varun Kumar, Hadrien Glaude, Cyprien de Lichy and William Campbell

A Comparative Analysis of Unsupervised Language Adaptation Methods
Gil Rocha and Henrique Lopes Cardoso

A logical-based corpus for cross-lingual evaluation
Felipe Salvatore, Marcelo Finger and Roberto Hirata Jr

Adaptively Scheduled Multitask Learning: The Case of Low-Resource Neural Machine Translation
Poorya Zaremoodi and Gholamreza Haffari

Bad Form: Comparing Context-Based and Form-Based Few-Shot Learning in Distributional Semantic Models
Jeroen Van Hautte, Guy Emerson and Marek Rei

Bag-of-Words Transfer: Non-Contextual Techniques for Multi-Task Learning
Seth Ebner, Felicity Wang and Benjamin Van Durme

BERT is Not an Interlingua and the Bias of Tokenization
Jasdeep Singh, Bryan McCann, Richard Socher and Caiming Xiong

Building Bilingual Pre-trained Models In A Day
Ke Tran
Byte-Pair encoding for text-to-SQL generation
Samuel Müller and Andreas Vlachos

Cross-lingual Joint Entity and Word Embedding to Improve Entity Linking and Parallel Sentence Mining
Xiaoman Pan, Thamme Gowda, Heng Ji, Jonathan May and Scott Miller

Deep Bidirectional Transformers for Relation Extraction without Supervision
Yannis Papanikolaou, Ian Roberts and Andrea Pierleoni

Domain Adaptation with BERT-based Domain Classification and Data Selection
Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati and Bing Xiang

Empirical Evaluation of Active Learning Techniques for Neural MT
Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy and Matthias Paulik

Fast Domain Adaptation of Semantic Parsers via Paraphrase Attention
Avik Ray, Yilin Shen and Hongxia Jin

Few-Shot and Zero-Shot Learning for Historical Text Normalization
Marcel Bollmann, Natalia Korchagina and Anders Søgaard

From Monolingual to Multilingual FAQ Assistant using Multilingual Co-training
Mayur Patidar, Surabhi Kumari, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, Lovekesh Vig and Gautam Shroff

Generation-Distillation for Efficient Natural Language Understanding in Low-Data Settings
Luke Melas-Kyriazi, George Han and Celine Liang

H-FND: Hierarchical False-Negative Denoising for Robust Distantly-Supervised Relation Extraction
Tsu-Jui Fu and Wei-Yun Ma

10:30–11:00  Break

11:00–11:45  Invited Talk 2: Barbara Plank
Sunday, November 3, 2019 (continued)

11:45–12:30 Contributed Talks

11:45–12:00 Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual He He, Sheng Zha and Haohan Wang

12:00–12:15 Metric Learning for Dynamic Text Classification Jeremy Wohlwend, Ethan R. Elenberg, Sam Altschul, Shawn Henry and Tao Lei


12:30–14:00 Lunch Break

14:00–14:45 Invited Talk 3: Dan Roth

14:45–15:30 Poster Session 2

Cross-lingual Parsing with Polyglot Training and Multi-treebank Learning: A Faroese Case Study James Barry, Joachim Wagner and Jennifer Foster

Inject Rubrics into Short Answer Grading System Tianqi Wang, Naoya Inoue, Hiroki Ouchi, Tomoya Mizumoto and Kentaro Inui

Instance-based Inductive Deep Transfer Learning by Cross-Dataset Querying with Locality Sensitive Hashing Somnath Basu Roy Chowdhury, Annervaz M and Ambedkar Dukkipati

Multimodal, Multilingual Grapheme-to-Phoneme Conversion for Low-Resource Languages James Route, Steven Hillis, Isak Czeresnia Ettinger, Han Zhang and Alan W Black

Natural Language Generation for Effective Knowledge Distillation Raphael Tang, Yao Lu and Jimmy Lin
Neural Rule Grounding for Low-Resource Relation Extraction
Wenxuan Zhou, Hongtao Lin, Ziqi Wang, Leonardo Neves and Xiang Ren

Neural Unsupervised Parsing Beyond English
Katharina Kann, Anhad Mohananey, Samuel R. Bowman and Kyunghyun Cho

Pseudolikelihood Reranking with Masked Language Models
Julian Salazar, Davis Liang, Toan Q. Nguyen and Katrin Kirchhoff

Reevaluating Argument Component Extraction in Low Resource Settings
Anirudh Joshi, Timothy Baldwin, Richard Sinnott and Cecile Paris

Reinforcement-based denoising of distantly supervised NER with partial annotation
Farhad Nooralahzadeh, Jan Tore Lønning and Lilja Øvrelid

Samvaadhana: A Telugu Dialogue System in Hospital Domain
Suma Reddy Duggenpudi, Kusampudi Siva Subrahmananyam Varma and Radhika Mamidi

Towards Zero-resource Cross-lingual Entity Linking
Shuyan Zhou, Shruti Rijhwani and Graham Neubig

Transductive Auxiliary Task Self-Training for Neural Multi-Task Models
Johannes Bjerva, Katharina Kann and Isabelle Augenstein

Weakly Supervised Attentional Model for Low Resource Ad-hoc Cross-lingual Information Retrieval
Lingjun Zhao, Rabih Zbib, Zhuolin Jiang, Damianos Karakos and Zhongqiang Huang

X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension
Mostafa Abdou, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein and Anders Søgaard

XLDA: Cross-Lingual Data Augmentation for Natural Language Inference and Question Answering
Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Richard Socher and Caiming Xiong

Zero-Shot Cross-lingual Name Retrieval for Low-Resource Languages
Kevin Blissett and Heng Ji
Sunday, November 3, 2019 (continued)

Zero-shot Dependency Parsing with Pre-trained Multilingual Sentence Representations
Ke Tran and Arianna Bisazza

15:30–16:00  Break

16:00–16:45  Invited Talk 4: Kristina Toutanova


17:30–17:40  Closing Remarks