

COLING 2020

**The 28th International Conference
on Computational Linguistics**

Proceedings of the Industry Track

December 12, 2020
Barcelona, Spain (Online)

Copyright of each paper stays with the respective authors (or their employers).

ISBN 978-1-952148-29-3

Preface

We welcome you to the 28th International Conference on Computational Linguistics (COLING 2020) Industry Track. This year marks the first in which COLING has a dedicated track for research related to computational linguistics deployed in real-world settings. In recent years, industrial research has been increasingly influential in the field of computational linguistics—both in the form of research departments that contribute to the advancement of computational linguistics, and also by way of the knowledgeable inventors and developers of innovative language and speech products.

The goals of this session are to foster connections between industry practitioners, share insights from industry research to the broader community, and increase engagement with academia on research questions of high priority in industry. This session will showcase commercially-driven research from diverse angles, including the challenges of doing applied research at scale, production scalability, and a shifting data landscape.

We received 124 submissions (79 long and 45 short) and had an acceptance rate of 23%.¹ Based on the first author’s affiliation, an estimated 76% of submissions came from industry and 24% from academia. Geographically, most submissions were from North America (44%), 27–28% from Europe and Asia (respectively), and 1% from Africa.

We thank the many people who have made this track a success. We are grateful to Donia Scott, general chair, for her support and for providing us with this opportunity. Thank you to program chairs Núria Bel and Chengqing Zong; local chairs Leo Wanner, Horacio Saggion, and Mónica Domínguez; publication chairs Derek Wong, Liang Huang, and Yang Zhao; web chairs Laura Pérez-Mayos and Amita Misra; virtual infrastructure chairs Paul Piwek, Lluís Padró Cirera, and Luis Espinosa Anke; publicity chairs Ghazaleh Kazeminejad, Tiejun Zhao, Ted Pedersen, and Anna Rogers; and all other members of the organizing committee. We were inspired by the success of the NAACL 2018 and 2019 industry tracks, and we thank Anastassia Loukin and the organizers of the NAACL 2019 Industry Track for their guidance. We are grateful to area chairs Juri Ganitkevitch and Greg Hanneman, and the program committee, whose dedication and hard work made this program possible.

Organizers
COLING 2020 Industry Track

¹Not including 17 desk rejects and 13 withdrawals.

Industry Track Chairs

Ann Clifton, Spotify, USA
Courtney Napoles, Grammarly, USA

Area Chairs

Juri Ganitkevitch, Apple, USA
Greg Hanneman, Amazon, USA

Program Committee

Shazia Afzal, IBM Research, India
Sachin Agarwal, Apple, USA
Khaled Ammar, Thomson Reuters Labs, Canada
Arturo Argueta, Apple, USA
Mehdi Bahrami, Fujitsu Laboratories of America, USA
Tamali Banerjee, IIT Bombay, India
Oren Barkan, Microsoft, Israel
Frederic Bechet, Aix Marseille Universite - LIS/CNRS, France
Akash Bharadwaj, Facebook AI Research, USA
Sravan Bodapati, Amazon, USA
Daniele Bonadiman, Amazon AI, USA
Trung Bui, Adobe Research, USA
Vitor Carvalho, Carnegie Mellon University, USA
Alessandra Cervone, Amazon Alexa AI, Italy
Ashwini Challa, Facebook, USA
Nicholas Chamansingh, TTLab, Trinidad and Tobago
Praveen Chandar, University of Delaware, USA
Sourish Chaudhuri, Google Inc, USA
John Chen, Interactions LLC, USA
Li Chen, JD.COM, USA
Yue Chen, Indiana University, USA
Laura Chiticariu, IBM Data & AI, USA
Eunah Cho, Amazon Alexa AI, USA
Ido Cohn, Google, Israel
Deborah Dahl, Conversational Technologies, USA
Tirthankar Dasgupta, Tata Consultancy Services Ltd., India
Heidar Davoudi, Ontario Tech University, Canada
Budhaditya Deb, Microsoft Corporation, USA
Lingjia Deng, Bloomberg L.P., USA
Giuseppe Di Fabbrizio, VUI, Inc., USA
Erika Doggett, Walt Disney Studios, USA
Xin Dong, Rutgers University, USA
Matthew Dunn, LivePerson, USA
Jonathan Engel, Megagon Labs, USA

Keelan Evanini, Educational Testing Service, USA
Adam Faulkner, Capital One, USA
Michael Flor, Educational Testing Service, USA
Kavita Ganesan, GitHub, USA
Rashmi Gangadharaiah, Amazon AWS AI, USA
Anna Lisa Gentile, IBM Research Almaden, USA
Frédéric Godin, Chatlayer.ai, Belgium
Anuj Goyal, Amazon Alexa, USA
Abhirut Gupta, Google Research, India
Le An Ha, RGCL, RIILP, University of Wolverhampton, UK
Kristian Hammond, Northwestern University, USA
Bo Han, Accenture, Australia
Sanjika Hewavitharana, eBay, USA
Derrick Higgins, Illinois Institute of Technology, USA
Silja Hildebrand, Amazon, USA
Lynette Hirschman, MITRE, USA
Yuta Hitomi, Media Lab, The Asahi Shimbun Company, Japan
Yufang Hou, IBM Research, Ireland
Jinxin Hu, Alibaba, China
David Huggins Daines, Nuance Communications, Canada
Javid Huseynov, IBM, USA
Tomoki Ito, The University of Tokyo, Japan
Michael Johnston, Interactions Corporation, USA
Mahesh Joshi, LinkedIn, USA
Anup Kalia, IBM T. J. Watson Research Center, USA
Saurabh Khanwalkar, Bose Corporation, USA
Doo Soon Kim, Adobe Research, USA
Kunho Kim, Microsoft Corporation, USA
Young-Bum Kim, Amazon Alexa Brain, USA
Sun Kim, Amazon Alexa AI; NCBI/NIH, USA
Jared Kramer, Amazon, USA
Rohit Kumar, VoiceThesis LLC, USA
Sanjeev Kumar, Quark.ai, USA
Varun Kumar, Amazon Alexa, USA
Gakuto Kurata, IBM Research, Japan
Claudia Leacock, Consultant, USA
Jennifer J Liang, IBM Research, USA
Zhengyuan Liu, Institute for Infocomm Research, A*STAR, Singapore
Phoebe Liu, Figure Eight, USA
Honglei Liu, Facebook Conversational AI, USA
Alexander Loeser, Beuth-University of Applied Sciences Berlin, Germany
Anastassia Loukina, Educational Testing Service, USA
Xiaoqiang Luo, Google, USA
Nitin Madnani, Educational Testing Service, USA
Ashraf Mahgoub, Purdue University, USA
Courtney Mansfield, University of Washington, USA
Gabriel Marzinotto, Orange Labs, Aix Marseille Univ, CNRS, LIS, France
Yuji Matsumoto, Riken Center for Advanced Intelligence Project, Japan
Tomoki Matsuno, LAPRAS Inc., Japan
David McClosky, Google, USA
Mahnoosh Mehrabani, Interactions LLC, USA

Marie Meter, Pryon, USA
Chad Mills, University of Washington, USA
Prateeti Mohapatra, IBM Research Lab, India
Michelle Morales, CUNY Graduate Center, USA
Isabelle Moulinier, Capital One, USA
Matthew Mulholland, Educational Testing Service, USA
Maria Nadejde, Amazon AWS AI, USA
Nobal B. Niraula, Boeing Research & Technology, USA
Mari Olsen, Lionbridge AI, USA
Girish Palshikar, Tata Consultancy Services Limited, India
Shimei Pan, UMBC, USA
Youngja Park, IBM T. J. Watson Research Center, USA
Ioannis Partalas, Expedia Group, Switzerland
Siddharth Patwardhan, Apple, USA
Stanislav Peshterliev, Facebook, USA
Lahari Poddar, Amazon, Germany
Rashmi Prasad, Interactions Corporation, USA
Nishant Prateek, Amazon Research Cambridge, UK
Long Qin, Singsound Inc, China
Ying Qin, Beijing Foreign Studies University, China
Preethi Raghavan, IBM Research, USA
Vipul Raheja, Grammarly, USA
Anil Ramakrishna, Amazon, USA
Meghana Ravikumar, SigOpt, USA
Anupama Ray, IBM Research, India
Sravana Reddy, Spotify, USA
Ehud Reiter, University of Aberdeen, UK
Brian Riordan, ETS, USA
Salim Roukos, IBM Research AI, USA
Nicholas Ruiz, Interactions, LLC, USA
Alicia Sagae, Consulting Scientist, USA
Avneesh Saluja, Netflix, USA
Rajhans Samdani, Spoke, USA
Stefan Scherer, Embodied, Inc., USA
Frank Schilder, Thomson Reuters, USA
Ethan Selfridge, Interactions LLC, USA
Sapan Shah, Tata Consultancy Services Ltd, India
Arpit Sharma, WalmartLabs, USA
Michal Shmueli-Scheuer, IBM Research, Israel
Sunayana Sitaram, Microsoft Research India, India
Kazoo Sone, Google, USA
Biplav Srivastava, IBM, USA
Fabian Suchanek, Telecom ParisTech University, France
Lichao Sun, University of Illinois at Chicago, USA
Lin Sun, Zhejiang University City College, China
Kyle Swanson, ASAPP, Inc., USA
György Szarvas, Amazon Development Center Germany GmbH, Germany
Aniruddha Tammewar, University Of Trento, Italy
Isabel Trancoso, INESC-ID / IST Univ. Lisbon, Portugal
Keith Trnka, 98point6, Inc., USA
Ling Tsou, SDL, USA

Elena Tutubalina, Insilico Medicine, Russian Federation
Ngoc Phuoc An Vo, IBM Research, USA
Stephen Wan, CSIRO, Australia
Yao Wan, Huazhong University of Science and Technology, China
Varden Wang, Google Inc, USA
Xinhao Wang, Educational Testing Service, USA
Yi-Chia Wang, Facebook AI, USA
Jason D Williams, Apple, USA
Kyle Williams, Microsoft, USA
Joan Xiao, Figure Eight, USA
Wei Yang, Borealis AI, Canada
Jianguo Zhang, University of Illinois at Chicago, USA
Wen Zhang, Zhejiang University, China

Table of Contents

<i>Evaluating Cross-Lingual Transfer Learning Approaches in Multilingual Conversational Agent Models</i> Lizhen Tan and Olga Golovneva	1
<i>Data-Efficient Paraphrase Generation to Bootstrap Intent Classification and Slot Labeling for New Features in Task-Oriented Dialog Systems</i> Shailza Jolly, Tobias Falke, Caglar Tirkaz and Daniil Sorokin	10
<i>Leveraging User Paraphrasing Behavior In Dialog Systems To Automatically Collect Annotations For Long-Tail Utterances</i> Tobias Falke, Markus Boese, Daniil Sorokin, Caglar Tirkaz and Patrick Lehnen	21
<i>Query Distillation: BERT-based Distillation for Ensemble Ranking</i> Wangshu Zhang, Junhong Liu, Zujie Wen, Yafang Wang and Gerard de Melo	33
<i>Semantic Diversity for Natural Language Understanding Evaluation in Dialog Systems</i> Enrico Palumbo, Andrea Mezzalana, Cristina Marco, Alessandro Manzotti and Daniele Amberti	44
<i>An Empirical Study on Multi-Task Learning for Text Style Transfer and Paraphrase Generation</i> Pawel Bujnowski, Kseniia Ryzhova, Hyungtak Choi, Katarzyna Witkowska, Jaroslaw Piersa, Tymoteusz Krumholz and Katarzyna Beksa	50
<i>Best Practices for Data-Efficient Modeling in NLG: How to Train Production-Ready Neural Models with Less Data</i> Ankit Arun, Soumya Batra, Vikas Bhardwaj, Ashwini Challa, Pinar Donmez, Peyman Heidari, Hakan Inan, Shashank Jain, Anuj Kumar, Shawn Mei, Karthik Mohan and Michael White	64
<i>Interactive Question Clarification in Dialogue via Reinforcement Learning</i> Xiang Hu, Zujie Wen, Yafang Wang, Xiaolong Li and Gerard de Melo	78
<i>Towards building a Robust Industry-scale Question Answering System</i> Rishav Chakravarti, Anthony Ferritto, Bhavani Iyer, Lin Pan, Radu Florian, Salim Roukos and Avi Sil	90
<i>Delexicalized Paraphrase Generation</i> Boya Yu, Konstantine Arkoudas and Wael Hamza	102
<i>Multi-task Learning of Spoken Language Understanding by Integrating N-Best Hypotheses with Hierarchical Attention</i> Mingda Li, Xinyue Liu, Weitong Ruan, Luca Soldaini, Wael Hamza and Chengwei Su	113
<i>Misspelling Detection from Noisy Product Images</i> Varun Nagaraj Rao and Mingwei Shen	124
<i>hinglishNorm - A Corpus of Hindi-English Code Mixed Sentences for Text Normalization</i> Piyush Makhija, Ankit Kumar and Anuj Gupta	136
<i>Assessing Social License to Operate from the Public Discourse on Social Media</i> Chang Xu, Cecile Paris, Ross Sparks, Surya Nepal and Keith VanderLinden	146
<i>Extreme Model Compression for On-device Natural Language Understanding</i> Kanthashree Mysore Sathyendra, Samridhi Choudhary and Leah Nicolich-Henkin	160

<i>Scalable Cross-lingual Treebank Synthesis for Improved Production Dependency Parsers</i> Yousef El-Kurdi, Hiroshi Kanayama, Efsun Sarioglu Kayi, Vittorio Castelli, Todd Ward and Radu Florian	172
<i>An Industry Evaluation of Embedding-based Entity Alignment</i> Ziheng Zhang, Hualuo Liu, Jiaoyan Chen, Xi Chen, Bo Liu, YueJia Xiang and Yefeng Zheng .	179
<i>Learning Domain Terms - Empirical Methods to Enhance Enterprise Text Analytics Performance</i> Gargi Roy, Lipika Dey, Mohammad Shakir and Tirthankar Dasgupta	190
<i>Model-agnostic Methods for Text Classification with Inherent Noise</i> Kshitij Tayal, Rahul Ghosh and Vipin Kumar	202
<i>ScopeIt: Scoping Task Relevant Sentences in Documents</i> Barun Patra, Vishwas Suryanarayanan, Chala Fufa, Pamela Bhattacharya and Charles Lee	214
<i>Uncertainty Modeling for Machine Comprehension Systems using Efficient Bayesian Neural Networks</i> Zhengyuan Liu, Pavitra Krishnaswamy, Ai Ti Aw and Nancy Chen	228
<i>Regularized Graph Convolutional Networks for Short Text Classification</i> Kshitij Tayal, Nikhil Rao, Saurabh Agarwal, Xiaowei Jia, Karthik Subbian and Vipin Kumar . .	236

Conference Program

Tuesday, December 8, 2020 (UTC+1)

16:00–16:30 Session Industry 1: Dialogue

- 16:00–16:06 *Evaluating Cross-Lingual Transfer Learning Approaches in Multilingual Conversational Agent Models*
Lizhen Tan and Olga Golovneva
- 16:06–16:12 *Data-Efficient Paraphrase Generation to Bootstrap Intent Classification and Slot Labeling for New Features in Task-Oriented Dialog Systems*
Shailza Jolly, Tobias Falke, Caglar Tirkaz and Daniil Sorokin
- 16:12–16:18 *Leveraging User Paraphrasing Behavior In Dialog Systems To Automatically Collect Annotations For Long-Tail Utterances*
Tobias Falke, Markus Boese, Daniil Sorokin, Caglar Tirkaz and Patrick Lehnen
- 16:18–16:24 *Query Distillation: BERT-based Distillation for Ensemble Ranking*
Wangshu Zhang, Junhong Liu, Zujie Wen, Yafang Wang and Gerard de Melo
- 16:24–16:30 *Semantic Diversity for Natural Language Understanding Evaluation in Dialog Systems*
Enrico Palumbo, Andrea Mezzalana, Cristina Marco, Alessandro Manzotti and Daniele Amberti

Wednesday, December 9, 2020 (UTC+1)

16:00–16:30 Session Industry 2: Generation and Question Answering

- 16:00–16:06 *An Empirical Study on Multi-Task Learning for Text Style Transfer and Paraphrase Generation*
Pawel Bujnowski, Kseniia Ryzhova, Hyungtak Choi, Katarzyna Witkowska, Jaroslaw Piersa, Tymoteusz Krumholz and Katarzyna Beksa
- 16:06–16:12 *Best Practices for Data-Efficient Modeling in NLG: How to Train Production-Ready Neural Models with Less Data*
Ankit Arun, Soumya Batra, Vikas Bhardwaj, Ashwini Challa, Pinar Donmez, Peyman Heidari, Hakan Inan, Shashank Jain, Anuj Kumar, Shawn Mei, Karthik Mohan and Michael White
- 16:12–16:18 *Interactive Question Clarification in Dialogue via Reinforcement Learning*
Xiang Hu, Zujie Wen, Yafang Wang, Xiaolong Li and Gerard de Melo
- 16:18–16:24 *Towards building a Robust Industry-scale Question Answering System*
Rishav Chakravarti, Anthony Ferritto, Bhavani Iyer, Lin Pan, Radu Florian, Salim Roukos and Avi Sil

Wednesday, December 9, 2020 (UTC+1) (continued)

- 16:24–16:30 *Delexicalized Paraphrase Generation*
Boya Yu, Konstantine Arkoudas and Wael Hamza

Thursday, December 10, 2020 (UTC+1)

16:00–16:33 Session Industry 3: Applications

- 16:00–16:06 *Multi-task Learning of Spoken Language Understanding by Integrating N-Best Hypotheses with Hierarchical Attention*
Mingda Li, Xinyue Liu, Weitong Ruan, Luca Soldaini, Wael Hamza and Chengwei Su
- 16:06–16:12 *Misspelling Detection from Noisy Product Images*
Varun Nagaraj Rao and Mingwei Shen
- 16:12–16:18 *hinglishNorm - A Corpus of Hindi-English Code Mixed Sentences for Text Normalization*
Piyush Makhija, Ankit Kumar and Anuj Gupta
- 16:18–16:24 *Assessing Social License to Operate from the Public Discourse on Social Media*
Chang Xu, Cecile Paris, Ross Sparks, Surya Nepal and Keith VanderLinden
- 16:24–16:30 *Extreme Model Compression for On-device Natural Language Understanding*
Kanthashree Mysore Sathyendra, Samridhi Choudhary and Leah Nicolich-Henkin
- 16:30–16:33 *Scalable Cross-lingual Treebank Synthesis for Improved Production Dependency Parsers*
Yousef El-Kurdi, Hiroshi Kanayama, Efsun Sarioglu Kayi, Vittorio Castelli, Todd Ward and Radu Florian

Friday, December 11, 2020 (UTC+1)

16:00–16:30 Session Industry 4: Machine Learning Applications

- 16:00–16:06 *An Industry Evaluation of Embedding-based Entity Alignment*
Ziheng Zhang, Hualuo Liu, Jiaoyan Chen, Xi Chen, Bo Liu, YueJia Xiang and Yefeng Zheng
- 16:06–16:12 *Learning Domain Terms - Empirical Methods to Enhance Enterprise Text Analytics Performance*
Gargi Roy, Lipika Dey, Mohammad Shakir and Tirthankar Dasgupta
- 16:12–16:18 *Model-agnostic Methods for Text Classification with Inherent Noise*
Kshitij Tayal, Rahul Ghosh and Vipin Kumar
- 16:18–16:24 *ScopeIt: Scoping Task Relevant Sentences in Documents*
Barun Patra, Vishwas Suryanarayanan, Chala Fufa, Pamela Bhattacharya and Charles Lee
- 16:24–16:27 *Uncertainty Modeling for Machine Comprehension Systems using Efficient Bayesian Neural Networks*
Zhengyuan Liu, Pavitra Krishnaswamy, Ai Ti Aw and Nancy Chen
- 16:27–16:30 *Regularized Graph Convolutional Networks for Short Text Classification*
Kshitij Tayal, Nikhil Rao, Saurabh Agarwal, Xiaowei Jia, Karthik Subbian and Vipin Kumar

