MMNLU-22 2022

Massively Multilingual Natural Language Understanding 2022

Proceedings of MMNLU-22

December 7, 2022

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-959429-15-9

Introduction

Let's scale natural language understanding technology to every language on Earth!

By 2023 there will be over 8 billion virtual assistants worldwide, the majority of which will be on smartphones. Additionally, over 100 million smart speakers have been sold, most of which exclusively use a voice interface and require Natural Language Understanding (NLU) during every user interaction in order to function. However, even as we approach the point in which there will be more virtual assistants than people in the world, major virtual assistants still only support a small fraction of the world's languages. This limitation is driven by the lack of labeled data, the expense associated with human-based quality assurance, model maintenance and update costs, and more. Innovation is how we will jump these hurdles. The vision of this workshop is to help propel natural language understanding technology into the 50-language, 100-language, and even the 1,000-language regime, both for production systems and for research endeavors.

For an overview of the workshop and competition, please see the paper entitled "The Massively Multilingual Natural Language Understanding 2022 (MMNLU-22) Workshop and Competition," included in these proceedings.

Organizing Committee

Organizers

Jack FitzGerald, Amazon Alexa, USA
Kay Rottmann, Amazon Alexa, Germany
Julia Hirschberg, Columbia University, USA
Mohit Bansal, University of North Carolina, USA
Anna Rumshisky, University of Massachusetts Lowell, USA
Charith Peris, Amazon Alexa, USA
Christopher Hench, Amazon Alexa, USA

Program Committee

Reviewers

Christopher Church

Yulia Grishina

Thanh-Le Ha, He He, Kuan-Hao Huang

Tushar Jain

Philipp Koehn

Thu Le

Yixin Nie, Jan Niehues

Udita Patel

Sayambhu Sen, Pankaj Kumar Sharma, Shubham Shukla, Veselin Stoyanov

Gokhan Tur

Xiang Zhou

Keynote Talk: Fine-grained Multi-lingual Disentangled Autoencoder for Language-Agnostic Representation Learning

Zhongkai Sun
Amazon

Abstract: Encoding both language-specific and language-agnostic information into a single high-dimensional space is a common practice of pre-trained Multi-lingual Language Models (pMLM). Such encoding has been shown to perform effectively on natural language tasks requiring semantics of the whole sentence (e.g., translation). However, its effectiveness appears to be limited on tasks requiring partial information of the utterance (e.g., multi-lingual entity retrieval, template retrieval, and semantic alignment). In this work, a novel Fine-grained Multilingual Disentangled Autoencoder (FMDA) is proposed to disentangle fine-grained semantic information from language-specific information in a multi-lingual setting. FMDA is capable of successfully extracting the disentangled template semantic and residual semantic representations. Experiments conducted on the MASSIVE dataset demonstrate that the disentangled encoding can boost each other during the training, thus consistently outperforming the original pMLM and the strong language disentanglement baseline on monolingual template retrieval and cross-lingual semantic retrieval tasks across multiple languages.

Bio: TBD

Keynote Talk: Towards Efficient Transfer Learning Across Languages

Mahdi Namazifar
Amazon Alexa AI

Abstract: Unbalanced distribution of text resources necessary for AI research and development across languages is well known to result in biases and unfairness in access to benefits of advances in AI. Multilingual language models have played a big role in transferring learnings across languages, facilitating addressing this imbalance to some extend. This talk focuses on additional approaches that could potentially further enable transfer of learnings across languages.

Bio: Mahdi Namazifar received his PhD in Operations Research with a focus in Optimization from University of Wisconsin-Madison in 2011. After PhD he worked at various companies such as Cisco, Twitter, and Uber on applications of machine learning in different industries. In 2020 he joined Amazon Alexa's Conversation Modeling team where he is working on different problems in NLP and Conversational AI.

Keynote Talk: Byte-Level Massively Multilingual Semantic Parsing (Zero-Shot Shared Task Winners)

Massimo Nicosia
Google

Abstract: Token free approaches have been successfully applied to a series of word and span level tasks. In this work, we evaluate a byte-level sequence to sequence model (ByT5) on the 51 languages in the MASSIVE multilingual semantic parsing dataset. We examine multiple experimental settings: (i) zero-shot, (ii) full gold data and (iii) zero-shot with synthetic data. By leveraging a state-of-the-art label projection method for machine translated examples, we are able to reduce the gap in exact match to only 5 points with respect to a model trained on gold data from all the languages. We additionally provide insights on the cross-lingual transfer of ByT5 and show how the model compares with respect to mT5 across all parameter sizes

Bio: Massimo Nicosia is a Senior Software Engineer in Research at Google in Zurich working on making natural language understanding models multilingual.

Keynote Talk: Machine Translation for Multilingual Intent Detection and Slots Filling (Organizers' Choice Award)

Maxime De Bruyn
University of Antwerp

Abstract: We expect to interact with home assistants irrespective of our language. However, scaling the Natural Language Understanding pipeline to multiple languages while keeping the same level of accuracy remains a challenge. In this work, we leverage the inherent multilingual aspect of translation models for the task of multilingual intent classification and slot filling. Our experiments reveal that they work equally well with general-purpose multilingual text-to-text models. Furthermore, their accuracy can be further improved by artificially increasing the size of the training set. Unfortunately, increasing the training set also increases the overlap with the test set, leading to overestimating their true capabilities. As a result, we propose two new evaluation methods capable of accounting for an overlap between the training and test set.

Bio: Maxime is a PhD student in computational linguistics at the University of Antwerp (Belgium) under the supervision of Prof. Walter Daelemans. His work mainly focuses on conversational agents and question answering. Prior to starting his PhD, Maxime was a fund manager at a Belgian private bank.

Keynote Talk: Multilingual NLP for Customer Relationship Management

Géraldine Damnati
Orange Labs

Abstract: Natural Language Processing has become a key technology to improve Customer Relationship Management. Extracting key insights from customer feedbacks, mining opinions from surveys or reviews, designing interactive chatbots or voicebots for commercial or technical assistance are examples among several applications where processing language helps managing Customer Relationship. Being able to handle multiple languages is a central feature for companies, whether when operating in a country where several languages are spoken or when operating in several countries. Recent advances in multilingual NLP represent a huge opportunity towards leveraging customer feedbacks expressed in different languages but many challenges remain.
In this talk, I will present some issues encountered in an international company when analyzing its interactions with customers. In the case of Orange, which also operates in Africa and Middle east, low resource languages are also at stakes. I will address the design of multilingual NLP models in a context where using multipurpose Large Language Models or even any model needing GPU computation is not always a realistic scalable solution. I will share experience of data collection in the context of highly regulated domain with European General Data Protection Regulation and of data annotation in a context where micro-tasking is generally not used. I will also discuss how to bridge the gap between academic research on unconstrained benchmark corpora that do not always fit the reality of deployment constraints and how these constraints can fuel new research questions.

Bio: Géraldine Damnati is a Research Engineer at Orange Innovation, DATA&AI, Lannion, France. After an engineering degree from Telecom Bretagne, she obtained in 2000 a PhD in Computer Science from University of Avignon. Her research interests include Natural Language Processing, Spoken Language Understanding, Text and Speech Mining, Semantic Analysis, Question Answering and Information Extraction in general. She has a research activity, contributing to collaborative projects, being co-author of around 80 publications in international conferences. She is also involved in the conception and development of tools in various applicative domains, such as Customer Relationship or Multimedia Content exploration. She is currently involved in several research projects, including the ARCHIVAL pluridisciplinary project (http://archival.msh-paris.fr/) for archive valorisation in the context of Digital Humanities. She is also active in the French NLP community, as a member of the ATALA board (Association pour le Traitement Automatique des Langues) and as the coordinator of the French CNRS GDR-TAL partners club.

Keynote Talk: Towards Massively Multilingual Modular Models

Sebastian Ruder
Google

Abstract: State-of-the-art multilingual models are trained on data of around 100 languages. These models can be adapted to perform better in under-represented languages but such adaptation does not directly benefit the original models. In order to make progress on NLU capabilities for the next 1,000 languages, we need to make it easier for researchers from diverse backgrounds to build upon and share improvements on base models. To this end, I will first discuss the tools currently at our disposal for extending multilingual models, from sparse subnetworks to parameter-efficient adaptation and vocabulary extension. I will then highlight the benefits of modularity compared to current model monoliths. Finally, I will sketch a vision of how we can build, train, and evaluate modular multilingual models that can cover the next 1,000 languages.

Bio: Sebastian is a research scientist at Google based in Berlin, Germany working on natural language processing (NLP) for under-represented languages. Before that he was a research scientist at DeepMind. He completed his PhD in Natural Language Processing and Deep Learning at the Insight Research Centre for Data Analytics, while working as a research scientist at Dublin-based text analytics startup AYLIEN. Previously, he studied Computational Linguistics at the University of Heidelberg, Germany and at Trinity College, Dublin. He's interested in cross-lingual and transfer learning for NLP and making ML and NLP more accessible.

Keynote Talk: HIT-SCIR at MMNLU-22: Consistency Regularization for Multilingual Spoken Language Understanding (Best Paper and Full-Data Shared Task Winner)

Bo Zheng
Research Center for Social Computing and Information Retrieval (SCIR) of Harbin Institute of Technology

Abstract: Multilingual spoken language understanding (SLU) consists of two sub-tasks, namely intent detection and slot filling. To improve the performance of these two sub-tasks, we propose to use consistency regularization based on a hybrid data augmentation strategy. The consistency regularization enforces the predicted distributions for an example and its semantically equivalent augmentation to be consistent. We conduct experiments on the MASSIVE dataset under both full-dataset and zero-shot settings. Experimental results demonstrate that our proposed method improves the performance on both intent detection and slot filling tasks. Our system ranked 1st in the MMNLU-22 competition under the full-dataset setting.

Bio: Bo Zheng is a final-year Ph.D. student at the Research Center for Social Computing and Information Retrieval (SCIR) of Harbin Institute of Technology, advised by Prof. Wanxiang Che. His research interests include cross-lingual NLP, machine reading comprehension, and language analysis. He has published many papers in international conferences and journals such as ACL, EMNLP, CoNLL, etc. He was ranked first on multiple official leaderboards, including the leaderboard of Google's XTREME benchmark, Google's Natural Questions dataset, and Amazon's MASSIVE dataset. He was also ranked first in multiple international competitions, including Amazon's MMNLU-2022 competition, CoNLL 2018 shared task, and NLP-TEA 2016 shared task.

Keynote Talk: Massively Multilingual NLP in 1600+ Languages

David Yarowsky
Johns Hopkins University

Abstract: The talk will cover a range of topics in massively multilingual and very low-resource NLP and speech recognition, in core functionalities, at a nearly unprecedented language-universal scale.

Bio: David Yarowsky is a Professor of Computer Science at Johns Hopkins University, and a member of its Center for Language and Speech Processing. He received his PhD from the University of Pennsylvania in 1996. He is an ACL Fellow, NSF CAREER award winner, Rockefeller Fellow, summa-cum-laude graduate from Harvard, ACL Test-of-time award winner, ACL Treasurer, co-founder of the EMNLP conference series and longtime ACL/SIGDAT executive committee member. He has pioneered the field of cross-lingual information projection via bilingual word alignments and done extensive work in low-resource and massively multilingual NLP, and is also known for an eponymous influential algorithm used for co-training, multi-view machine learning and low-resource bootstrapping.

Keynote Talk: Learning in the Wild: Modeling Language in Real-World Scenarios

Anna Rumshisky
University of Massachusetts Lowell

Abstract: Scientific progress in NLP is often measured by model performance on standardized benchmarks. But in many cases, existing benchmarks fail to reflect the settings in which algorithmic solutions are applied in practice. The challenges of modeling language in real-world scenarios often go beyond covariate shift and related well-studied phenomena. In this talk, I will discuss some of these challenges, using user interactions with digital assistants as a case study. I will describe some recent work aimed at addressing such challenges, including (a) learning from a combination of positive and negative noisy user feedback in a federated setting, and (b) learning from frequency-enriched data in a setting where a different treatment is required for the head and tail of the distribution.

Bio: Anna Rumshisky is an Associate Professor of Computer Science at the University of Massachusetts Lowell, where she heads the Text Machine Lab for NLP. Her primary research area is machine learning for natural language processing, with a focus on deep learning techniques. She has made contributions in a number of application areas, including computational lexical semantics, temporal reasoning and argument mining, as well as clinical informatics and computational social science. She received her PhD from Brandeis University and completed postdoctoral training at MIT CSAIL, where she is currently a Research Affiliate. She is a recipient of the NSF CAREER award in 2017, and her work won the best thematic paper award at NAACL-HLT 2019.

Keynote Talk: Multilingual Information Extraction for Thousands of Types

Heng Ji
University of Illinois at Urbana-Champaign

Abstract: Supervised information extraction models require a substantial amount of training data to perform well. However, information annotation requires a lot of human effort and costs much time, especially for low-resource languages, which limits the application of existing supervised approaches to new knowledge types. In order to reduce manual labor and shorten the time to build an information extraction system for an arbitrary ontology, we present a new framework to train such systems much more efficiently without large annotations. Our weak supervision approach only requires a set of keywords, a small number of examples and an unlabeled corpus in any language, and takes advantage of naturally existing "hubs" (such as linking to WikiData, Multilingual embedding and universal semantic parsers) for cross-lingual transfer.

Bio: Heng Ji is a professor at Computer Science Department, and an affiliated faculty member at Electrical and Computer Engineering Department of University of Illinois at Urbana-Champaign. She is an Amazon Scholar. She received her B.A. and M. A. in Computational Linguistics from Tsinghua University, and her M.S. and Ph.D. in Computer Science from New York University. Her research interests focus on Natural Language Processing, especially on Multimedia Multilingual Information Extraction, Knowledge Base Population and Knowledge-driven Generation. She was selected as Young Scientist and a member of the Global Future Council on the Future of Computing by the World Economic Forum in 2016 and 2017. The awards she received include AI's 10 to Watch Award by IEEE Intelligent Systems in 2013, NSF CAREER award in 2009, Google Research Award in 2009 and 2014, IBM Watson Faculty Award in 2012 and 2014, Bosch Research Award in 2014-2018, Best-of-ICDM2013 Paper, Best-of-SDM2013 Paper, ACL2020 Best Demo Paper Award, and NAACL2021 Best Demo Paper Award. She is elected as the North American Chapter of the Association for Computational Linguistics (NAACL) secretary 2020-2021. She has served as the Program Committee Co-Chair of many conferences including NAACL-HLT2018, and she has been the coordinator for the NIST TAC Knowledge Base Population track since 2010.

Table of Contents

Program

Wednesday, December 7, 2022

09:00 - 09:30 *Introduction and Shared Task Overview*

09:30 - 10:00 *Fine-grained Multi-lingual Disentangled Autoencoder for Language-Agnostic Representation Learning*

10:00 - 10:30 *Invited Talk by Mahdi Namazifar: Towards Efficient Transfer Learning Across Languages*

10:30 - 11:00 *Break*

11:00 - 11:30 *Zero-Shot Shared Task Winners: Massimo Nicosia and Francesco Piccinno, Google*

11:30 - 12:00 *Invited Talk by Sebastian Ruder, Google: Towards Massively Multilingual Modular Models*

12:00 - 12:30 *Invited Talk by Géraldine Damnati, Orange Labs: Multilingual NLP for Customer Relationship Management*

12:30 - 13:30 *Lunch*

13:30 - 14:00 *Organizers' Choice Award: Maxime De Bruyn and the bolleke team*

14:00 - 14:30 *Best Paper Award and Full-Data Shared Task Winner: Bo Zheng and the HIT-SCIR team*

14:30 - 15:30 *Poster Session*

15:30 - 16:00 *Break*

16:00 - 16:30 *Invited Talk by David Yarowsky, JHU: Massively Multilingual NLP in 1600+ Languages*

16:30 - 17:00 *Invited Talk by Anna Rumshisky, UMass Lowell: Learning in the Wild: Modeling Language in Real-World Scenarios*

17:00 - 17:30 *Invited Talk by Heng Ji, U of Illinois Urbana-Champaign: Multilingual Information Extraction for Thousands of Types*

17:30 - 18:30 *Networking*