

ICNLSP 2022

**Proceedings of the 5th International Conference on
Natural Language and Speech Processing**

16–17 December, 2022 (virtual)



©2022 The Association for Computational Linguistics

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 978-1-959429-36-4

<https://www.icnlsp.org>

Introduction

Welcome to the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022), held online on December 16th, 17th 2022.

ICNLSP is the right choice to select as a forum for researchers, students, and also for industrials to exchange ideas and discuss research and trends in the field of Natural Language Processing, and also to publish their results in the field. As examples of companies present during the conference, we mention here, Mercedes Benz (Germany), and Vail Systems company (USA), and Elm (KSA) and many others.

The program committee accepted 37 papers (long and short ones) which is around 40% of the received submissions (from 31 countries). The accepted papers are of good quality thanks to the high-quality level of the reviews done by the program committee members. All papers have been presented orally, that is why the program was quite long. Various topics of NLP are discussed, as Semantics, language modelling, text classification, speech recognition, information extraction, natural language understanding, etc.

As it is mentioned in the program of the conference, there are three keynotes. The first one was presented by Prof. Eric Laporte from Gustave Eiffel University (France), who exposed his thoughts about hybrid natural language processing in the deep learning era. The second one, dealing with an interesting and challenging topic, was given by Dr. Ahmed Ali from Qatar Computing Research Institute (Qatar), entitled “Multilingual and Code-Switching Speech Recognition”. The third talk was programmed to be presented by Prof. Jan Niehues, from Karlsruhe Institute of Technology (Germany), and entitled “Plug-and-Play Abilities for Neural Machine Translation”. We will be happy to make all the talks and presentations available on the website of the conference.

We hope readers enjoy reading the content of the 5th ICNLSP proceedings. We would like also to invite them to check the proceedings of the past versions of ICNLSP:

Mourad Abbas, Abed Alhakim Freihat, Proceedings of the Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021), 12-13 November 2021, Association for Computational Linguistics, <https://aclanthology.org/2021.icnls-1>

Mourad Abbas, Abed Alhakim Freihat, Proceedings of the 3rd International Conference on Natural Language and Speech Processing (ICNLSP 2019), 12-13 September 2019, Association for Computational Linguistics, <https://aclanthology.org/volumes/W19-74/>

Mourad Abbas, Proceedings of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP 2018), 25-26 April 2018, IEEE, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8374402>

Mourad Abbas, Ahmed Abdelali, Proceedings of the 1st International Conference on Natural Language and Speech Processing, Procedia Computer Science, 128, Elsevier. <https://www.sciencedirect.com/journal/procedia-computer-science/vol/128>

We would like to express our gratitude to the organizing and the program committees for making this event a success.

Mourad Abbas and Abed Alhakim Freihat

Organizers:

General Chair: Dr. Mourad Abbas

Chair: Dr. Abed Alhakim Freihat

Program Chair: Dr. Mourad Abbas

Publicity Chair: Dr. Muhammad Al-Qurishi

Program Committee:

Ahmed Abdelali, QCRI, Qatar.

Hend Al-Khalifa, King Saud University, Saudi Arabia.

Somaya Al-Maadeed, Qatar University, Qatar.

Muhammad Al-Qurishi, Elm, Saudi Arabia.

Yuan An, Drexel University, USA.

Fayssal Bouarourou, University of Strasbourg, France.

Markus Brückl, TU Berlin, Germany.

Hadda Cherroun, Amar Telidji University, Algeria.

G rard Chollet, CNRS, France.

Dirk Van Compernelle, KU Leuven, Belgium.

Kareem Darwish, aiXplain, USA.

Najim Dehak, Johns Hopkins University, USA.

Abed Alhakim Freihat, University of Trento, Italy.

Munir Georges, Technische Hochschule Ingolstadt, Germany.

Fausto Giunchiglia, University of Trento, Italy.

Ahmed Guessoum, USTHB, Algeria.

Fouzi Harrag, Ferhat Abbas University, Algeria.

Valia Kordoni, Humboldt University, Germany.

Eric Laporte, Gustave Eiffel University, France.

Shang-Wen Li, Facebook AI., USA.

Mohamed Lichouri, USTHB, Algeria.

Mhamed Mataoui, EMP, Algeria.

Mohammed Mediani, University of Adrar, Algeria.

Fatiha Merazka, USTHB, Algeria.

Farid Meziane. University of Derby, UK.

Hamdy Mubarak, QCRI, Qatar.

Preslav Nakov, QCRI, Qatar.

Alexis Neme, UPEM, France.

Rasha Obeidat, Jordan university of science and technology, Jordan.

Axel Roebel, IRCAM, France.

Hassan Satori, Sidi Mohammed Ben Abdallah University, Morocco.

Tim Schlippe, Silicon Surfer, Germany.

Nasredine Semmar, CEA, France.

Otakar Smrz, Džám-e Džam Language Institute, Czech Republic.
Rudolph Sock, University of Strasbourg, France.
R. V. Swaminathan, Amazon, USA.
Irina Temnikova, Big Data for Smart Society Institute, Bulgaria.
Jan Trmal, Johns Hopkins University, USA.
Rodrigo Wilkens, UCLouvain, Belgium.
Fayçal Ykhlef, CDTA, Algeria.
Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar.
Hasna Zaouali, University of Strasbourg, France.

Organizing Committee:

Hadi Khalilia, University of Trento
Khaled Lounnas, USTHB, Algeria
Nandu C Nair, University of Trento

Invited Speakers:

Prof. Eric Laporte, Gustave Eiffel University, France
Dr. Ahmed Ali, Qatar Computing Research Institute, Qatar
Prof Jan Niehues, Karlsruhe Institute of Technology, Germany

Invited Talks

Hybrid natural language processing in the deep learning era

Prof. Eric Laporte, Gustave Eiffel University, France

In this talk, we examine critically the current wave of interest in pure deep learning for natural language processing. What can symbolic resources do for natural language processing? Among other examples, we take into account the languages with more restricted graphical delimitation than English. Then we discuss the foreseeable future of the synergy between machine learning and symbolic resources: are the goals of formalisation, precision, reliability, adaptability within reach for linguistic data?



Multilingual and Code-Switching Speech Recognition

Dr. Ahmed Ali, Qatar Computing Research Institute, Qatar

The prevalence of code-switching (CS) in spoken content has enforced automatic speech recognition (ASR) systems to handle mixed input. Yet, designing a CS-ASR has many challenges, mainly due to the data scarcity, grammatical structure complexity and mismatch along with unbalanced language usage distribution. Our CS will feature both intersentential (switching between-utterances) and intrasentential (within utterances). The evaluation of the designed system and the analysis of the phenomena will be driven based on real test cases, collected from real meetings and interviews.



We show our results on investigating novel techniques to build practical large vocabulary continuous speech recognition systems capable of dealing with both monolingual and code-switching spoken utterances. We study data augmentation and state of the art modelling techniques to address the lack of balanced transcribed CS data. Moreover, we investigate various challenges of evaluating code-switching ASR output. Finally, we highlight our effort in understanding where/why CS happens in speech analysis for system/human code-switching points.

Plug-and-Play Abilities for Neural Machine Translation

Prof Jan Niehues, Karlsruhe Institute of Technology, Germany

Advances in neural machine translation have led to impressive results and broad areas of application. Using multitask learning, these models have even abilities to process different input and generate a variety of output languages. However, this progress is often backed by millions of training examples. In order to cover the approximately 7000 languages in the words, it is essential to not only generalize to unseen examples, but also to unseen tasks. Therefore, we need to recombine the abilities of NMT systems to process and generate different languages in a plug-and-play fashion.



In this presentation, we will investigate two use cases: translating zero-shot directions in multilingual machine translation and end-to-end speech translation. First, we will dissect the challenges in the zero-shot condition. Motivated by the findings, we will present several methods to promote the possibility to combine the different abilities of an NMT system in order to perform unseen tasks. Finally, we will discuss the effect of the presented ideas on multi-lingual machine translation and speech translation.

Table of Contents

Error correction and extraction in request dialogs	2
<i>Stefan Constantin and Alex Waibel</i>	
Efficient Task-Oriented Dialogue Systems with Response Selection as an Auxiliary Task	12
<i>Radostin Cholakov and Todor Kolev</i>	
TopicRefine: Joint Topic Prediction and Dialogue Response Generation for Multi-turn End-to-End Dialogue System	19
<i>Hongru Wang, Mingyu Cui, Zimo Zhou and Kam-Fai Wong</i>	
Prior Omission of Dissimilar Source Domain(s) for Cost-Effective Few-Shot Learning	30
<i>Zezhong Wang, Hongru Wang, Wai Chung Kwan and Kam-Fai Wong</i>	
Linguistic Knowledge in Data Augmentation for Natural Language Processing: An Example on Chinese Question Matching	40
<i>Zhengxiang Wang</i>	
Detecting Security Patches in Java Projects Using NLP Technology	50
<i>Andrea Stefanoni, Šarūnas Girdzijauskas, Christina Jenkins, Zekarias T. Kefato, Licia Sbattella, Vincenzo Scotti and Emil Wåreus</i>	
Improving NL-to-Query Systems through Re-ranking of Semantic Hypothesis	57
<i>Pius von Däniken, Jan Deriu, Eneko Agirre, Ursin Brunner, Mark Cieliebak and Kurt Stockinger</i>	
Experimenting with ensembles of pre-trained language models for classification of custom legal datasets	68
<i>Tamara Matthews and David Lillis</i>	
Handling Class Imbalance when Detecting Dataset Mentions with Pre-trained Language Models	78
<i>Yousef Younes and Brigitte Mathiak</i>	
Performance of two French BERT models for French language on verbatim transcripts and online posts	88
<i>Emmanuelle Kelodjoue, Jérôme Goulian and Didier Schwab</i>	
Semi-supervised Automated Clinical Coding Using International Classification of Diseases	95
<i>Hlynur Hlynsson, Steindór Ellertsson, Jon Dadason, Emil Sigurdsson and Hrafn Loftsson</i>	
Recent Advances in Long Documents Classification Using Deep-Learning	107
<i>Muhammad Al-Qurishi</i>	
Optimizing singular value based similarity measures for document similarity comparisons	113
<i>Jarkko Lagus and Arto Klami</i>	
Semantic Similarity Based Filtering for Turkish Paraphrase Dataset Creation	119
<i>Besher Alkurdi, Hasan Yunus Sarioglu and Mehmet Fatih Amasyali</i>	
Second-order Document Similarity Metrics for Transformers	128
<i>Jarkko Lagus, Niki Loppi and Arto Klami</i>	
Semantic Similarity-Based Clustering of Findings From Security Testing Tools	134
<i>Phillip Schneider, Markus Voggenreiter, Abdullah Gulraiz and Florian Matthes</i>	
Contextual Embeddings Can Distinguish Homonymy from Polysemy in a Human-Like Way	144

<i>Kyra Wilson and Alec Marantz</i>	
Modeling the Ordering of English Adjectives using Collaborative Filtering	156
<i>Sagar Indurkha</i>	
Comparison of Token- and Character-Level Approaches to Restoration of Spaces, Punctuation, and Capitalization in Various Languages	168
<i>Laurence Dyer, Anthony Hughes, Dhvani Shah and Burcu Can</i>	
New Features for Discriminative Keyword Spotting	179
<i>Punnoose Kuriakose</i>	
Hierarchical Multi-Task Transformers for Crosslingual Low Resource Phoneme Recognition	187
<i>Kevin Glocker and Munir Georges</i>	
Concatenative Phonetic Synthesis for the Proto-Indo-European Language	193
<i>Patrick Donnelly</i>	
A low latency technique for speaker detection from a large negative list	202
<i>Yu Zhou, B. Chandra Mouli and Vijay Gurbani</i>	
Supervised Acoustic Embeddings And Their Transferability Across Languages	212
<i>Sreepratha Ram and Hanan Aldarmaki</i>	
A Dataset for Detecting Humor in Arabic Text	219
<i>Hend Alkhalifa, Fetoun Alzahrani, Hala Qawara, Reema Alrowais, Sawsan Alowa and Luluh Aldhubayi</i>	
A deep sentiment analysis of Tunisian dialect comments on multi-domain posts in different social media platforms	226
<i>Emna Fsih, Rahma Boujelbane and Lamia Hadrich Belguith</i>	
TuniSER: Toward a Tunisian Speech Emotion Recognition System	234
<i>Abir Messaoudi, Hatem Haddad, Moez Benhaj Hmida and Mohamed Graiet</i>	
Evaluating Large-Language Models for Dimensional Music Emotion Prediction from Social Media Discourse	242
<i>Patrick Donnelly and Aidan Beery</i>	
Customer Sentiments Toward Saudi Banks During the Covid-19 Pandemic	251
<i>Dhuha Alqahtani, Lama Alzahrani, Maram Bahareth, Nora Alshameri, Hend Al-Khalifa and Luluh Aldhubayi</i>	
Towards an Automatic Dialect Identification System using Algerian Youtube Videos	258
<i>Khaled Lounnas, Mohamed Lichouri, Mourad Abbas, Thissas Chahboub and Samir Salmi</i>	
Constructing the Corpus of Chinese Textual ‘Run-on’ Sentences (CCTRS): Discourse Corpus Benchmark with Multi-layer Annotations	265
<i>Kun Sun and Rong Wang</i>	
Utilizing BERT Intermediate Layers for Unsupervised Keyphrase Extraction	277
<i>Mingyang Song, Yi Feng and Liping Jing</i>	
"Der Frank Sinatra der Wettervorhersage": Cross-Lingual Vossian Antonomasia Extraction	282
<i>Michel Schwab, Robert Jäschke and Frank Fischer</i>	
The Elementary Scenario Component Metric for Summarization Evaluation	288
<i>Martin Kirilov, Daan Kolkman and Bert-Jan Butijn</i>	

Scaling Native Language Identification with Transformer Adapters	298
<i>Ahmet Yavuz Uluslu and Gerold Schneider</i>	
Arguments to Key Points Mapping with Prompt-based Learning	303
<i>Ahnaf Mozib Samin, Behrooz Nikandish and Jingyan Chen</i>	
Pre-training Language Models for Surface Realization	312
<i>Farhood Farahnak and Leila Kosseim</i>	