KONVENS 2021

Proceedings of the 17th Conference on Natural Language Processing/Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)

6–9 September, 2021
Heinrich Heine University Düsseldorf
Düsseldorf, Germany

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

ISBN 978-1-954085-83-1

Introduction

The papers of these proceedings have been presented at the 17th edition of KONVENS (Konferenz zur Verarbeitung natürlicher Sprache/Conference on Natural Language Processing). KONVENS is a conference series on computational linguistics established in 1992 that was held biennially until 2018 and has been held annually since. KONVENS is organized under the auspices of the German Society for Computational Linguistics and Language Technology, the Special Interest Group on Computational Linguistics of the German Linguistic Society and the Austrian Society for Artificial Intelligence.

The 17th KONVENS took place from September 6 to September 9, 2021 at Heinrich Heine University Düsseldorf. Due to the COVID-19 pandemic situation, KONVENS was held as a hybrid event in order to allow both speakers and regular participants to attend the conference either on-site or online. The special theme of this year's meeting was *Deep Linguistic Modeling*. The KONVENS main conference was accompanied by two workshops, three shared task meetings, and a 'PhD Day'.

Many thanks to all who submitted their work to KONVENS and to our board of reviewers for supporting us greatly with evaluating the submissions. Moreover we would like to thank Heinrich Heine University Düsseldorf for providing the conference rooms and all people from the CL department in Düsseldorf who made the conference possible. Our special thanks go to Tobias Koch from the 'Multimediazentrum' for his generous technical support.

Kilian Evang
Laura Kallmeyer
Rainer Osswald
Jakub Waszczuk
Torsten Zesch

People

Local Organization:

Tatiana Bladier, Rafael Ehren, Kilian Evang, Laura Kallmeyer, Rainer Osswald, Esther Seyffarth

Program Chairs:

Laura Kallmeyer, Rainer Osswald, Jakub Waszczuk, Torsten Zesch

Program Committee:

Adrien Barbaresi, Felix Bildhauer, Marcel Bollmann, Ernst Buchberger, Stephan Busemann, Miriam Butt, Mark Cieliebak, Berthold Crysmann, Stefanie Dipper, Sarah Ebling, Kilian Evang, Stefan Evert, Diego Frassinelli, Abhijeet Gupta, Anke Holler, Laura Kallmeyer, Manfred Klenner, Roman Klinger, Valia Kordoni, Brigitte Krenn, Udo Kruschwitz, Sandra Kübler, Gabriella Lapesa, Ekaterina Lapshinova-Koltunski, Anke Lüdeling, Alexander Mehler, Günter Neumann, Rainer Osswald, Sebastian Pado, Simone Paolo Ponzetto, Simon Petitjean, Hannes Pirker, Ines Rehbein, Georg Rehm, Michael Roth, Josef Ruppenhofer, Younes Samih, Felix Sasaki, Roland Schäfer, Yves Scherrer, Helmut Schmid, Gerold Schneider, Roman Schneider, Sabine Schulte im Walde, Marcin Skowron, Manfred Stede, Ludovic Tanguy, Jakub Waszczuk, Thomas Weskott, Magdalena Wolska, Torsten Zesch, Heike Zinsmeister

Invited Speakers:

Afra Alishahi, Tilburg University
Milica Gašić, Heinrich Heine University Düsseldorf
Mirella Lapta, University of Edinburgh
Johann-Mattis List, Max Planck Institute for the Science of Human History Jena

Satellite Events

1st Workshop on Computational Linguistics for Political Text Analysis

Organizers: Ines Rehbein, Goran Glavaš, Simone Ponzetto, Gabriella Lapesa

Workshop on Multimodal and Multilingual Hate Speech Detection

Organizers: Özge Alaçam, Seid Muhie Yimam

Shared Task on Scene Segmentation (STSS)

Organizers: Albin Zehe, Leonard Konle, Lea Dümpelmann, Evelyn Glus, Svenja Guhr, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Annekea Schreiber

GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments

Organizers: Julian Risch, Anke Stoll, Lena Wilms, Michael Wiegand

Shared Task on the Disambiguation of German Verbal Idioms

Organizers: Rafael Ehren, Laura Kallmeyer, Timm Lichte, Jakub Waszczuk

PhD Day

Organizers: Esther Seyffarth, Oliver Deck, Jannis Pagel, Ronja Laarmann-Quante, Stefan Grünewald

Invited Talks

Afra Alishahi: Decoding what deep, grounded neural models learn about language

Humans learn to understand speech from weak and noisy supervision: they extract structure and meaning from speech by simply being exposed to utterances situated and grounded in their daily sensory experience. Emulating this remarkable skill has been the goal of numerous studies; however, researchers have often used severely simplified settings where either the language input or the extralinguistic sensory input, or both, are small-scale and symbolically represented.

Recently, deep neural network models have been successfully used for visually grounded language understanding, where representations of images are mapped to those of their written or spoken descriptions. Despite their high performance, these architectures come at a cost: we know little about the type of linguistic knowledge these models capture from the input signal in order to perform their target task.

I present a series of studies on modelling visually grounded language learning and analyzing the emergent linguistic representations in these models. Using variations of recurrent neural networks to model the temporal nature of spoken language, we examine how form and meaning-based linguistic knowledge emerges from the input signal.

Mirella Lapata: Summarization and Paraphrasing in Quantized Transformer Spaces

Deep generative models with latent variables have become a major focus n of NLP research over the past several years. These models have been used both for generating text and as a way of learning latent representations of text for downstream tasks. While much previous work uses continuous latent variables, discrete variables are attractive because they are more interpretable and typically more space efficient. In this talk we consider learning discrete latent variable models with Quantized Variational Autoencoders, and show how these can be ported to two NLP tasks, namely opinion summarization and paraphrase generation for questions. For the first task, we provide a clustering interpretation of the quantized space and a novel extraction algorithm to discover popular opinions among hundreds of reviews, while for the second task we show that a principled information bottleneck leads to an encoding space that separately represents meaning and surface from, thereby allowing us to generate syntactically varied paraphrases.

Johann-Mattis List: Chances and Challenges for Computational Comparative Linguistics in the 21st Century

The quantitative turn at the beginning of the 21st century has drastically changed the field of comparative linguistics. Had individual genious and expert insights dominated historical linguistics in the past, we now find many studies by interdisciplinary teams who use complex computational techniques to investigate the history of invididual language families based on large amounts of data. Had the identification of linguistic universals in hand-crafted language samples dominated linguistic typology for a long time, scholars now use large cross-linguistic databases to investigate dependencies among linguistic and non-linguistic variables with the help of complex statistical models.

However, despite a period of more than two decades in which quantitative approaches have been increasingly used in comparative linguistics, gaining constantly more popularity even among predominantly qualitatively oriented linguists, we still find many problems, which have only sporadically been addressed. In the talk, I will present three of these so far unsolved problems, which I find particularly important for the future of the field of comparative linguistics. These are: (1) the problem of modeling and comparing

sound change patterns across the languages of the world; (2) the problem of identifying cross-linguistic patterns of semantic change, and (3) the problem of estimating the borrowability of linguistic traits across languages and times.

While none of these problems has been solved so far, I will argue that substantial progress on their solution can be made by improving the integration of cross-linguistic data and by developing dedicated problem-solving strategies in computational linguistics which take the specifics of cross-linguistic data and language evolution into account.

Table of Contents

Long Papers

The Impact of Word Embeddings on Neural Dependency Parsing 1
Benedikt Adelmann, Wolfgang Menzel and Heike Zinsmeister

Benchmarking down-scaled (not so large) pre-trained language models 14
Matthias Aßenmacher, Patrick Schulze and Christian Heumann

ArgueBERT: How To Improve BERT Embeddings for Measuring the Similarity of Arguments . . 28
Maike Behrendt and Stefan Harmeling

How Hateful are Movies? A Study and Prediction on Movie Subtitles 37
Niklas von Boguszewski, Sana Moin, Anirban Bhowmick, Seid Muhie Yimam and Chris Biemann

Emotion Recognition under Consideration of the Emotion Component Process Model 49
Felix Casel, Amelie Heindl and Roman Klinger

Identifikation von Vorkommensformen der Lemmata in Quellenzitaten frühneuhochdeutscher Lexikoneinträge 62
Stefanie Dipper and Jan Christian Schaffert

Emotion Stimulus Detection in German News Headlines 73
Bao Minh Doan Dang, Laura Oberländer and Roman Klinger

Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques 86
Jakob Fehle, Thomas Schmidt and Christian Wolff

Definition Extraction from Mathematical Texts on Graph Theory in German and English 104
Theresa Kruse and Fritz Kliche

Extraction and Normalization of Vague Time Expressions in German 114
Ulrike May, Karolina Zaczynska, Julián Moreno-Schneider and Georg Rehm

Automatic Phrase Recognition in Historical German 127
Katrin Ortmann

Automatically Identifying Online Grooming Chats Using CNN-based Feature Extraction 137
Svenja Preuß, Tabea Bayha, Luna Pia Bley, Vivien Dehne, Alessa Jordan, Sophie Reimann, Fina Roberto, Josephine Romy Zahm, Hanna Siewerts, Dirk Labudde and Michael Spranger

Who is we? Disambiguating the referents of first person plural pronouns in parliamentary debates 147
Ines Rehbein, Josef Ruppenhofer and Julian Bernauer

Examining the Effects of Preprocessing on the Detection of Offensive Language in German Tweets 159
Sebastian Reimann and Daniel Dakota

Neural End-to-end Coreference Resolution for German in Different Domains 170
Fynn Schröder, Hans Ole Hatzel and Chris Biemann

How to Estimate Continuous Sentiments From Texts Using Binary Training Data 182
Sandra Wankmüller and Christian Heumann

forumBERT: Topic Adaptation and Classification of Contextualized Forum Comments in German 193
Ayush Yadav and Benjamin Milde

Short Papers

Robustness of end-to-end Automatic Speech Recognition Models – A Case Study using Mozilla DeepSpeech 203
Aashish Agarwal and Torsten Zesch

Effects of Layer Freezing on Transferring a Speech Recognition System to Under-resourced Languages 208
Onno Eberhard and Torsten Zesch

DeInStance: Creating and Evaluating a German Corpus for Fine-Grained Inferred Stance Detection 213
Anne Göhring, Manfred Klenner and Sophia Conrad

Combining text and vision in compound semantics: Towards a cognitively plausible multimodal model 218
Abhijeet Gupta, Fritz Günther, Ingo Plag, Laura Kallmeyer and Stefan Conrad

MobIE: A German Dataset for Named Entity Recognition, Entity Linking and Relation Extraction in the Mobility Domain 223
Leonhard Hennig, Phuc Tran Truong and Aleksandra Gabryszak

Automatically evaluating the conceptual complexity of German texts 228
Freya Hewett and Manfred Stede

WordGuess: Using Associations for Guessing, Learning and Exploring Related Words 235
Cennet Oguz, André Blessing, Jonas Kuhn and Sabine Schulte Im Walde

Towards a balanced annotated Low Saxon dataset for diachronic investigation of dialectal variation 242
Janine Siewert, Yves Scherrer and Jörg Tiedemann

German Abusive Language Dataset with Focus on COVID-19 247
Maximilian Wich, Svenja Räther and Georg Groh

Comparing Contextual and Static Word Embeddings with Small Philosophical Data 253
Wei Zhou and Jelke Bloem