IWCLUL 2019

The fifth International Workshop on Computational Linguistics for Uralic Languages

Proceedings of the Workshop

January 7—January 8, 2019
Tartu, Estonia

ISBN 978-1-948087-92-6

Introduction

These are the proceedings for the fifth international workshop on computational linguistics of Uralic languages, with a special focus on neural methods. Uralic languages are interesting from the computational-linguistic perspective, and also less-researched in the up-and-coming methodology of neural networks in natural language processing. The Uralic languages share large parts of morphological and morphophonological complexity that is not present in the Indo-European language family, which has traditionally dominated computational-linguistic research. This can be seen for example in number of morphologically complex forms belonging to one word, which in Indo-European languages is in range of ones or tens whereas for Uralic languages, it can be in the range of hundreds and thousands. Furthermore, Uralic language situations share a lot of geo-political aspects: the three national languages—Finnish, Estonian and Hungarian—are comparably small languages and only moderately resourced in terms of computational-linguistics while being stable and not in threat of extinction. The recognised minority languages of western-European states, on the other hand—such as North Sámi, Kven and Võro—do clearly fall in the category of lesser resourced and more threatened languages, whereas the majority of Uralic languages in the east of Europe and Siberia are closer to extinction. Common to all rapid development of more advanced computational-linguistic methods is required for continued vitality of the languages in everyday life, to enable archiving and use of the languages with computers and other devices such as mobile applications.

Computational linguistic research inside Uralistics is being carried out only in a handful of universities, research institutes and other sites and only by relatively few researchers. Our intention with organising this conference is to gather these researchers from scattered institutions together in order to share ideas and resources, and avoid duplicating efforts in gathering and enriching these scarce resources. We want to initiate more concentrated effort in collecting and improving language resources and technologies for the survival of the Uralic languages and hope that our effort today will become an ongoing tradition in the future.

For the current proceedings of The Fifth International Workshop on Computational Linguistics for Uralic Languages, we accepted 15 high-quality submissions about topics including semantic parsing, neural models, language documentation, tokenisation, corpora and lexicons, optical character recognition, morphological analysis and disambiguation. The range of languages covered this year again is wide, stretching over large part of Uralic languages including Finnish, Hungarian and Estonian as well as North Sámi, Livonian and Votic, and some multilingual papers touching methods and corpora for also Udmurt, Komi, Erzya, Moksha, Mari, Kildin Sámi, Khanty, Mansi and many others.

The conference was organised in collaboration with Tartu Ülikool, Estonia, on January 7–8 2019. The program consisted of an invited speech by Måns Huldén, a poster session, and six talks during the first day and an open discussion and individual project workshops during the second day. The current proceedings include the written versions all oral and poster presentations.

—Tommi A. Pirinen, Heiki-Jaan Kaalep, Francis M. Tyers
Conference organisers,
January 8, 2019, Tartu

Organizers:

ACL SIGUR, the special interest group for Uralic Languages, and the local organisers:
Anneli Vainumäe, University of Tartu
Heiki-Jaan Kaalep, University of Tartu

Program Committee:

Tommi Pirinen, University of Hamburg
Francis Tyers, Indiana University and Higher School of Economics
Eszter Simon, Research Institute for Linguistics, Hungarian Academy of Sciences
Anna Volkova, School of Linguistics, National Research University, Higher School of Economics, Moscow
Heiki-Jaan Kaalep, University of Tartu
Lene Antonsen, University of Tromsø
Trond Trosterud, University of Tromsø
Thierry Poibeau, LaTTiCe-CNRS
Veronika Vincze, Hungarian Academy of Sciences, Research Group on Articial Intelligence
Kadri Muischnek, University of Tartu
Csilla Horvath, Research Institute for Linguistics, Hungarian Academy of Sciences
Filip Ginter, University of Turku
Mark Fišel, University of Tartu
Kaili Müürisep, University of Tartu
Michael Rießler, Albert-Ludwigs-Universität Freiburg
Jeremy Bradley, Ludwig Maximilian University of Munich

Invited Speaker:

Mans Hulden, University of Colorado in Boulder

Table of Contents

Conference Program

Monday, January 7, 2019

8:30–9:00 ***Registration***

9:00–9:15 ***Opening Remarks***

9:15–10:15 *Invited Talk by Måns Huldén*

10:15–11:15 Presentations 1

10:15–10:45 *Data-Driven Morphological Analysis for Uralic Languages*
Miikka Silfverberg and Francis Tyers

10:45–11:15 *North Sámi morphological segmentation with low-resource semi-supervised sequence labeling*
Stig-Arne Grönroos, Sami Virpioja and Mikko Kurimo

11:15–11:45 Poster Boasters

11:15–11:18 *What does the Nom say? An algorithm for case disambiguation in Hungarian*
Noémi Ligeti-Nagy, Andrea Dömötör and Noémi Vadász

11:18–11:21 *A Contrastive Evaluation of Word Sense Disambiguation Systems for Finnish*
Frankie Robertson

11:21–11:24 *Elliptical Constructions in Estonian UD Treebank*
Kadri Muischnek and Liisi Torga

11:24–11:27 *FiST –towards a free Semantic Tagger of modern standard Finnish*
Kimmo Kettunen

11:27–11:30 *An OCR system for the Unified Northern Alphabet*
Niko Partanen and Michael Rießler

11:30–11:33 *ELAN as a search engine for hierarchically structured, tagged corpora*
Joshua Wilbur

Monday, January 7, 2019 (continued)

11:33–11:36 *Neural and rule-based Finnish NLP models—expectations, experiments and experiences*
Tommi A Pirinen

11:36–11:39 *Uralic multimedia corpora: ISO/TEI corpus data in the project INEL*
Timofey Arkhangelskiy, Anne Ferger and Hanna Hedeland

11:39–11:42 *Corpora of social media in minority Uralic languages*
Timofey Arkhangelskiy

11:45–13:00 ***Lunch***

Presentations 2

13:00–13:30 *Is this the end? Two-step tokenization of sentence boundaries*
Linda Wiechetek, Sjur Nørstebø Moshagen and Thomas Omma

13:30–14:00 *Learning multilingual topics through aspect extraction from monolingual texts*
Johannes Huber and Myra Spiliopoulou

14:00–14:30 *Electronical resources for Livonian*
Valts Ernštreits

14:30–15:00 *The use of Extract Morphology for Automatic Generation of Language Technology for Votic*
Kristian Kankainen

Monday, January 7, 2019 (continued)

15:00	**Posters and Demos**
20:00	***Social Dinner***

Tuesday, January 8, 2019

10:00–11:00	***ACL SIGUR Business Meeting***
11:00–	**Tutorials and Hands-On Session**