DL4NLP 2019

Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing

30 September, 2019

University of Turku

Turku, Finland

Introduction

The Nordic Language Processing Laboratory (NLPL; http://nlpl.eu) is a collaboration of university research groups in Natural Language Processing (NLP) in Northern Europe. Our vision is to implement a virtual laboratory for large-scale NLP research by (a) creating new ways to enable data- and compute-intensive Natural Language Processing research by implementing a common software, data and service stack in multiple Nordic HPC centres, (b) by pooling competencies within the user community and among expert support teams, and (c) by enabling internationally competitive, data-intensive research and experimentation on a scale that would be difficult to sustain on commodity computing resources.

One of the clear strengths of NLPL is its community-building component, with the very successful winter school being the highlight of the year for many of us. As these proceedings demonstrate, the NLPL community-building effort now also includes the first edition of what will hopefully develop into a series of regular workshops organized under the NLPL umbrella.

The purpose of this first NLPL workshop was to bring together researchers with a special interest in the deep learning techniques, specifically inviting also papers on computational and practical aspects of these methods. We are happy to be able to present seven peer-reviewed papers on diverse deep learning topics, together with the invited talks of Barbara Plank and Jussi Karlgren.

On behalf of the organizers,

Joakim Nivre, Filip Ginter, Stephan Oepen, Jörg Tiedemann

September 2019

Organisers:

Chair: Joakim Nivre

Leon Derczynski, Filip Ginter, Bjørn Lindi, Stephan Oepen, Anders Søgaard, Jörg Tiedemann

Program Committee:

Željko Agić, Ali Basirat, Johannes Bjerva, Hande Celikkanat, Mathias Creutz, Leon Derczynski, Filip Ginter, Christian Hardmeier, Mareike Hartmann, Jenna Kanerva, Andrey Kutuzov, Juhani Luotolahti, Joakim Nivre, Stephan Oepen, Alessandro Raganato, Yves Scherrer, Natalie Schluter, Miikka Silfverberg, Sara Stymne, Umut Sulubacak, Aarne Talman, Gongbo Tang, Jörg Tiedemann, Raul Vazquez, Erik Velldal, Aleksi Vesanto, Miryam de Lhoneux, Lilja Øvrelid

Invited Speakers:

Barbara Plank, IT University of Copenhagen, Denmark
Jussi Karlgren, Gavagai and KTH Royal Institute of Technology, Sweden

Invited Talks

Deep Transfer Learning: Learning across Languages, Modalities and Tasks

Barbara Plank, ITU (IT University of Copenhagen)

Transferring knowledge to solve a related problem is an extraordinary human ability. Yet, how can we build Natural Language Processing technology which can transfer to new conditions, such as learning to process a new language or learning to answer new types of questions? In this talk, I will present some recent work to addresses this ubiquitous challenge using neural networks for transfer learning. In particular, I will focus on cross-lingual learning for syntactic processing and work at the interface of language and vision, e.g., multi-task and continual learning for visual question answering.

High-Dimensional Semantic Spaces and the Squinting Linguist

Jussi Karlgren, Gavagai and KTH Royal Institute of Technology

High-dimensional distributed semantic spaces have proven useful and effective for aggregating and processing visual and lexical information for many tasks related to heterogenous data generated by human behaviour or relevant for human information processing. Models to process such high-dimensional spaces have proliferated in recent years with impressive results on quite various tasks.

In general, a representation used in research should hold to some basic qualities.

- It should have descriptive and explanatory power;

- be practical and convenient for further application;

- allow generalisations to be made and analogies to be inferred;

- be reasonably true to human performance, providing defaults to smooth over situations where data are lacking and constraints where the decision space is too broad,

- perform seamlessly and incrementally online in face of novel data, allowing new features and new analyses to be incorporated without recompiling previous understanding.

High-dimensional semantic spaces are usually not designed with (all of) these qualities in mind.

Human language has a large and varying number of features of various regularity, incorporating both lexical items and constructions. The field of linguistics provides a large body of research to understand such regularities. Yet the models used to process human language disregard those regularities, starting from the general principle that they should be discovered rather than given and that learning should be in the form an end-to-end classifier from raw data to relevant categories. This principle is intuitively appealing, in view of the specificity and avowed situation- and task-independence of linguistic rules, but the tools built end up being black boxes and do not guarantee explanatory generality of the results.

This talk will discuss how tasks where human language is the most important source of information might do well to incorporate information other than the most typical lexical features, in effect providing a better input layer to recent neurally inspired models. This talk demonstrates how high-dimensional semantic spaces can accommodate several types of feature concurrently, and that the convenient computing framework can be used for hypothesis-driven research. Such models can represent broad ranges of linguistic features in a common integral framework which is suitable as a bridge between symbolic and continuous representations, as an encoding scheme for symbolic information.

Table of Contents