EMNLP 2017

First Workshop on Subword and Character Level Models in NLP

Proceedings of the Workshop

September 7, 2017
Copenhagen, Denmark

We thank our sponsor Google Inc. for a generous support.

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 978-1-945626-91-3

Introduction

Traditional NLP starts with a hand-engineered layer of representation, the level of tokens or words. A tokenization component first breaks up the text into units using manually designed rules. Tokens are then processed by components such as word segmentation, morphological analysis and multiword recognition. The heterogeneity of these components makes it hard to create integrated models of both structure within tokens (e.g., morphology) and structure across multiple tokens (e.g., multi-word expressions). This approach can perform poorly (i) for morphologically rich languages, (ii) for noisy text, (iii) for languages in which the recognition of words is difficult and (iv) for adaptation to new domains; and (v) it can impede the optimization of preprocessing in end-to-end learning.

The workshop provides a forum for discussing recent advances as well as future directions on sub-word and character-level natural language processing and representation learning that address these problems.

We received 37 submissions, out of which we accepted 24 as papers and 4 as extended abstracts.

Organizers:

Manaal Faruqui, Google Research, USA
Hinrich Schütze, LMU Munich, Germany
Isabel Trancoso, INESC-ID/IST, Portugal
Yadollah Yaghoobzadeh, LMU Munich, Germany

Program Committee:

Heike Adel, LMU Munich
Ehsaneddin Asgari, UC Berkeley
Miguel Ballesteros, IBM
Kris Cao, Cambridge
Grzegorz Chrupala, Tilburg
Junyoung Chung, Montreal
Trevor Cohn, Melbourne
Marta R. Costa-jussa, UPC
Ryan Cotterell, Johns Hopkins
Chris Dyer, DeepMind
Alex Fraser, LMU Munich
Kevin Gimpel, TTI Chicago
Angeliki Lazaridou, Trento
Wang Ling, DeepMind
Andrew Mass, Stanford
Chris Potts, Stanford
Marek Rei, Cambridge
Rami Al-Rfou, Google
Laura Rimell, Cambridge
Cicero Nogueira dos Santos, IBM
Helmut Schmid, LMU Munich
Jörg Tiedemann, Helsinki
Thang Vu, IMS Stuttgart
Francois Yvon, LIMSI

Invited Speakers:

Kyunghyun Cho, NYU
Karen Livescu, TTIC
Tomas Mikolov, Facebook
Noah Smith, University of Washington

Panel Discussion:

Kyunghyun Cho, NYU
Sharon Goldwater, University of Edinburgh
Karen Livescu, TTIC
Tomas Mikolov, Facebook
Noah Smith, University of Washington

Table of Contents

Conference Program

Thursday, September 7, 2017

09:00–09:10 *Opening Remarks*
Manaal Faruqui

09:10–09:50 *Invited Talk: Subword-level Information in NLP using Neural Networks*
Tomas Mikolov

09:50–10:30 *Invited Talk: Chewing the Fat about Mincing Words*
Noah Smith

10:30–11:00 ***Coffee break***

11:00–11:40 *Invited Tutorial Talk: Neural WFSTs*
Ryan Cotterell

11:40–12:10 **Best paper presentations**

11:40–11:55 *Character and Subword-Based Word Representation for Neural Language Modeling Prediction*
Matthieu Labeau and Alexandre Allauzen

11:55–12:10 *Learning variable length units for SMT between related languages via Byte Pair Encoding*
Anoop Kunchukuttan and Pushpak Bhattacharyya

Thursday, September 7, 2017 (continued)

12:10–14:00 Poster session and Lunch break

Character Based Pattern Mining for Neology Detection
Gaël Lejeune and Emmanuel Cartier

(EXTENDED ABSTRACT) Patterns versus Characters in Subword-aware Neural Language Modeling
Zhenisbek Assylbekov and Rustem Takhanov

Automated Word Stress Detection in Russian
Maria Ponomareva, Kirill Milintsevich, Ekaterina Chernyak and Anatoly Starostin

A Syllable-based Technique for Word Embeddings of Korean Words
Sanghyuk Choi, Taeuk Kim, Jinseok Seol and Sang-goo Lee

Supersense Tagging with a Combination of Character, Subword, and Word-level Representations
Youhyun Shin and Sang-goo Lee

Weakly supervised learning of allomorphy
Miikka Silfverberg and Mans Hulden

Character-based recurrent neural networks for morphological relational reasoning
Olof Mogren and Richard Johansson

(EXTENDED ABSTRACT) Align and Copy: Hard Attention Models for Morphological Inflection Generation
Tatyana Ruzsics, Peter Makarov and Simon Clematide

Glyph-aware Embedding of Chinese Characters
Falcon Dai and Zheng Cai

Exploring Cross-Lingual Transfer of Morphological Knowledge In Sequence-to-Sequence Models
Huiming Jin and Katharina Kann

(EXTENDED ABSTRACT) Language Generation with Recurrent Generative Adversarial Networks without Pre-training
Ofir Press, Amir Bar, Ben Bogin, Jonathan Berant and Lior Wolf

Thursday, September 7, 2017 (continued)

14:00–14:40 *Invited Talk: Fully Character Level Neural Machine Translation*
Kyunghyun Cho

14:40–15:50 Poster session and Coffee break

Unlabeled Data for Morphological Generation With Character-Based Sequence-to-Sequence Models
Katharina Kann and Hinrich Schütze

Vowel and Consonant Classification through Spectral Decomposition
Patricia Thaine and Gerald Penn

Syllable-level Neural Language Model for Agglutinative Language
Seunghak Yu, Nilesh Kulkarni, Haejun Lee and Jihie Kim

Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition
Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura and Tomoko Ohkuma

Word Representation Models for Morphologically Rich Languages in Neural Machine Translation
Ekaterina Vylomova, Trevor Cohn, Xuanli He and Gholamreza Haffari

Spell-Checking based on Syllabification and Character-level Graphs for a Peruvian Agglutinative Language
Carlo Alva and Arturo Oncevay

What do we need to know about an unknown word when parsing German
Bich-Ngoc Do, Ines Rehbein and Anette Frank

A General-Purpose Tagger with Convolutional Neural Networks
Xiang Yu, Agnieszka Falenska and Ngoc Thang Vu

Reconstruction of Word Embeddings from Sub-Word Parameters
Karl Stratos

Inflection Generation for Spanish Verbs using Supervised Learning
Cristina Barros, Dimitra Gkatzia and Elena Lloret

Neural Paraphrase Identification of Questions with Noisy Pretraining
Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit and Dipanjan Das

Sub-character Neural Language Modelling in Japanese
Viet Nguyen, Julian Brooke and Timothy Baldwin

Byte-based Neural Machine Translation
Marta R. Costa-jussà, Carlos Escolano and José A. R. Fonollosa

(EXTENDED ABSTRACT) Natural Language Generation through Character-Based RNNs with Finite-State Prior Knowledge
Raghav Goyal, Marc Dymetman and Eric Gaussier

Improving Opinion-Target Extraction with Character-Level Word Embeddings
Soufian Jebbara and Philipp Cimiano

15:50–16:30 *Invited Talk: Acoustic Word Embeddings*
Karen Livescu

16:30–17:30 *Panel discussion*
Kyunghyun Cho, Sharon Goldwater, Karen Livescu, Tomas Mikolov, Hinrich Schütze and Noah Smith

17:30–17:45 *Closing remarks*
Hinrich Schütze