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.

Topics of Interest:

- tokenization-free models
- character-level machine translation
- character-ngram information retrieval
- transfer learning for character-level models
- models of within-token and cross-token structure
- NL generation (of words not seen in training etc)
- out of vocabulary words
- morphology and segmentation
- relationship b/w morphology and character-level models
- stemming and lemmatization
- inflection generation
- orthographic productivity
- form-meaning representations
- true end-to-end learning
- spelling correction
- efficient and scalable character-level models
Organizers

Organizers:
Manaal Faruqui, Google
Hinrich Schütze, LMU Munich
Isabel Trancoso, INESC-ID/IST
Yulia Tsvetkov, CMU
Yadollah Yaghoobzadeh, MSR Montreal

Program Committee:
Heike Adel, LMU Munich
Rami Al-Rfou, Google
Ehsan Asgari, UC Berkeley
Miguel Ballesteros, IBM New York
Gayatri Bhat, CMU
Marta R. Costa-jussà, UPC Barcelona
Ryan Cotterell, JHU
Vera Demberg, Saarland
Dan Garrette, Google
Kevin Gimpel, TTIC
Katharina Kann, LMU Munich
Sachin Kumar, CMU
Wang Ling, DeepMind
Nanyun Peng, JHU
Christopher Potts, Stanford
Marek Rei, Cambridge
Gaurav Singh Tomar, Google
Eva Schlinger, Google
Cicero dos Santos, IBM
Laura Rimell, Google DeepMind
Ngoc Thang Vu, Stuttgart
Francois Yvon, LIMSI
Alexander Rush, Harvard

Invited Speakers:
Jacob Eisenstein, Georgia Tech
Wang Ling, DeepMind
Graham Neubig, CMU
Barbara Plank, University of Groningen
Brian Roark, Google
# Table of Contents

**Morphological Word Embeddings for Arabic Neural Machine Translation in Low-Resource Settings**  
Pamela Shapiro and Kevin Duh ............................................................... 1

**Entropy-Based Subword Mining with an Application to Word Embeddings**  
Ahmed El-Kishky, Frank Xu, Aston Zhang, Stephen Macke and Jiawei Han ................. 12

**A Comparison of Character Neural Language Model and Bootstrapping for Language Identification in Multilingual Noisy Texts**  
Wafia Adouane, Simon Dobnik, Jean-Philippe Bernardy and Nasredine Semmar ........... 22

**Addressing Low-Resource Scenarios with Character-aware Embeddings**  
Sean Papay, Sebastian Padó and Ngoc Thang Vu ........................................ 32

**Subword-level Composition Functions for Learning Word Embeddings**  
Bofang Li, Aleksandr Drozd, Tao Liu and Xiaoyong Du ................................ 38

**Discovering Phonesthemes with Sparse Regularization**  
Nelson F. Liu, Gina-Anne Levow and Noah A. Smith .................................. 49

**Meaningless yet meaningful: Morphology grounded subword-level NMT**  
Tamali Banerjee and Pushpak Bhattacharyya ............................................. 55

**Fast Query Expansion on an Accounting Corpus using Sub-Word Embeddings**  
Hrishikesh Ganu and Viswa Datha P ......................................................... 61

**Incorporating Subword Information into Matrix Factorization Word Embeddings**  
Alexandre Salle and Aline Villavicencio .................................................. 66

**A Multi-Context Character Prediction Model for a Brain-Computer Interface**  
Shiran Dudy, Shaobin Xu, Steven Bedrick and David Smith .............................. 72
Conference Program

Wednesday, June 6, 2018

09:30–09:45 Opening Remarks
Manaal Faruqui

09:45–10:30 Invited Talk: Orthographic Social Variation in Online Writing
Jacob Eisenstein

10:30–11:00 Coffee Break

11:00–11:45 Invited Talk: Not All that Glitters is Gold
Barbara Plank

11:45–12:00 Best paper talk 1. Sponsor: Microsoft Research

12:00–14:00 Lunch Break

14:00–14:45 Invited Talk: Morphology – When is it Useful in Neural Models?
Graham Neubig

14:45–15:45 Poster Session and Coffee Break

- Morphological Word Embeddings for Arabic Neural Machine Translation in Low-Resource Settings
  Pamela Shapiro and Kevin Duh

- Entropy-Based Subword Mining with an Application to Word Embeddings
  Ahmed El-Kishky, Frank Xu, Aston Zhang, Stephen Macke and Jiawei Han

- A Comparison of Character Neural Language Model and Bootstrapping for Language Identification in Multilingual Noisy Texts
  Wafia Adouane, Simon Dobnik, Jean-Philippe Bernardy and Nasredine Semmar

- Addressing Low-Resource Scenarios with Character-aware Embeddings
  Sean Papay, Sebastian Padó and Ngoc Thang Vu
Wednesday, June 6, 2018 (continued)

*Subword-level Composition Functions for Learning Word Embeddings*
Bofang Li, Aleksandr Drozd, Tao Liu and Xiaoyong Du

*Discovering Phonesthemes with Sparse Regularization*
Nelson F. Liu, Gina-Anne Levow and Noah A. Smith

*Meaningless yet meaningful: Morphology grounded subword-level NMT*
Tamali Banerjee and Pushpak Bhattacharyya

*Fast Query Expansion on an Accounting Corpus using Sub-Word Embeddings*
Hrishikesh Ganu and Viswa Datha P

*Incorporating Subword Information into Matrix Factorization Word Embeddings*
Alexandre Salle and Aline Villavicencio

*A Multi-Context Character Prediction Model for a Brain-Computer Interface*
Shiran Dudy, Shaobin Xu, Steven Bedrick and David Smith

15:45–16:30  *Invited Talk: Romanization, Non-standard Orthography and Text Entry*  
Brian Roark

16:30–16:45  *Best paper talk 2. Sponsor: Microsoft Research*

16:45–17:30  *Invited Talk: What Makes a Character-level Neural Model work?*  
Wang Ling