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.