**CHARAGRAM: Embedding Words and Sentences**

**via Character n-grams**

**John Wieting , Mohit Bansal , Kevin Gimpel , Karen Livescu**

Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA

{jwieting,mbansal,kgimpel,klivescu}@ttic.edu

**Abstract**

We present CHARAGRAM embeddings, a simple approach for learning character-based compositional models to embed textual sequences. A word or sentence is represented using a character n-gram count vector, followed by a single nonlinear transformation to yield a low-dimensional embedding. We use three tasks for evaluation: word similarity, sentence similarity, and part-of-speech tagging. We demonstrate that CHARAGRAM embeddings outperform more complex architectures based on character-level recurrent and convolutional neural networks, achieving new state-of-the-art performance on several similarity tasks.

**摘要**

本文提出了CHARAGRAM嵌入方法，这是一种学习基于字符的组合模型以嵌入文本序列的简单方法。使用字符n-gram计数向量表示单词或句子，然后进行单个非线性变换以产生低维的向量。我们使用三个任务来进行评估：单词相似度，句子相似度和词性标注。可以证明，基于字符级递归和卷积神经网络，CHARAGRAM嵌入优于更复杂的架构，并且在几个相似性任务上实现了新的最先进性能。