**Character-Aware Neural Language Models**

**Yoon Kim† Yacine Jernite∗**

†School of Engineering and Applied Sciences Harvard University {yoonkim,srush}@seas.harvard.edu

**David Sontag Alexander M.**

**Rush†**

∗Courant Institute of Mathematical Sciences New York University {jernite,dsontag}@cs.nyu.edu

**Abstract**

We describe a simple neural language model that re- lies only on character-level inputs. Predictions are still made at the word-level. Our model employs a con- volutional neural network (CNN) and a highway network over characters, whose output is given to a long short-term memory (LSTM) recurrent neural net- work language model (RNN-LM). On the English Penn Treebank the model is on par with the existing state-of-the-art despite having 60% fewer parameters. On languages with rich morphology (Arabic, Czech, French, German, Spanish, Russian), the model outperforms word-level/morpheme-level LSTM baselines, again with fewer parameters. The results suggest that on many languages, character inputs are sufficient for language modeling. Analysis of word representations obtained from the character composition part of the model reveals that the model is able to encode, from characters only, both semantic and orthographic information.

**摘要**

本文描述了一个简单的神经语言模型，仅仅依赖于字符级的输入，其预测过程仍然是在单词级别进行的。我们的模型采用了卷积神经网络（CNN）和字符快速网络，其输出被赋予长期短期记忆（LSTM）递归神经网络语言模型（RNN-LM）。在英语Penn Treebank数据集上，尽管参数减少了60％，但该模型与现有技术水平相当。在具有丰富形态学的语言（如阿拉伯语，捷克语，法语，德语，西班牙语，俄语）上，该模型同样使用更少的参数，且优于单词级/语素级的LSTM基线。结果表明，在许多语言中，利用字符输入足以进行语言建模。利用模型的字符组成部分获得的单词表示揭示了该模型能够仅从字符编码获得语义和正交信息。