**Subword language modeling with neural networks**

**Toma ́sˇ Mikolov1, Ilya Sutskever2, Anoop Deoras3, Hai-Son Le4, Stefan Kombrink 1, Jan Cˇ ernocky ́1 12**

Brno University of Technology, University of Toronto,

Johns Hopkins University, Universite ́ Paris Sud and LIMSI/CNRS,

{imikolov,kombrink,cernocky}@fit.vutbr.cz ilya@cs.utoronto.ca, adeoras@jhu.edu, lehaison@limsi.fr

Abstract

We explore the performance of several types of language models on the word-level and the character-level language modeling tasks. This includes two recently proposed recurrent neural network archi- tectures, a feedforward neural network model, a maximum entropy model and the usual smoothed n-gram models. We then propose a simple technique for learning sub-word level units from the data, and show that it combines advantages of both character and word- level models. Finally, we show that neural network based language models can be order of magnitude smaller than compressed n-gram models, at the same level of performance when applied to a Broad- cast news RT04 speech recognition task. By using sub-word units, the size can be reduced even more.

**摘要**

我们探讨了几种语言模型在单词级别和字符级语言建模任务中的表现。其中包括最近提出的两个递归神经网络架构，前馈神经网络模型，最大熵模型和通常的平滑n-gram模型。然后，我们提出了一种从数据中学习子词级单元的简单技术，并表明它结合了字符和单词级模型的优点。最后，我们表明基于神经网络的语言模型可以比压缩的n-gram模型小一个数量级，在应用于广播新闻RT04语音识别任务时处于相同的性能水平。通过使用子字单元，可以进一步减小尺寸。