# One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling

**Ciprian Chelba, Tomas Mikolov, Mike Schuster,**

**Qi Ge，Thorsten Brants**

Google  1600 Amphitheatre Parkway Mountain View, CA 94043, USA

**Phillipp Koehn**

University of Edinburgh  10 Crichton Street, Room 4.19 Edinburgh, EH8 9AB, UK

**Tony Robinson**

Cantab Research Ltd  St Johns Innovation Centre Cowley Road, Cambridge, CB4 0WS, UK

**Abstract**

We propose a new benchmark corpus to be used for measuring progress in statistical language modeling. With almost one billion words of training data, we hope this benchmark will be useful to quickly evaluate novel language modeling techniques, and to compare their contribution when combined with other advanced techniques. We show performance of several well-known types of language models, with the best results achieved with a recurrent neural network based language model. The baseline unpruned KneserNey 5-gram model achieves perplexity 67.6. A combination of techniques leads to 35% reduction in

perplexity, or 10% reduction in cross-entropy (bits), over that baseline.

The benchmark is available as a code.google.com project; besides the scripts needed to rebuild the training/held-out data, it also makes available log-probability values for each word in each of ten held-out data sets, for each of the baseline n-gram models.

**摘要**

我们提出了一个新的基准语料库，可用于评估统计语言模型的进展。凭借近10亿字的训练数据，我们希望这一基准将有助于快速评估新颖的语言模型技术，并将其与其他先进技术相结合时的贡献进行比较。我们展示了几种众所周知的语言模型的性能，其中使用基于递归神经网络的语言模型获得了最佳结果。未经微调的KneserNey 5-gram模型实现了67.6的困惑。与该基线相比，这些技术的组合使得困惑度减少35％，或交叉熵（比特）减少10％。

该基准测试可作为code.google.com项目使用;除了重建训练/保持数据所需的脚本之外，它还为每个基线n-gram模型中的每个十个保持数据集中的每个单词提供对数概率值。