This folder contains all the experiments including:

|  |  |  |
| --- | --- | --- |
| Folder | Descriptions | Related Papers |
| ‘word’ | the word morphology experiment | [3] |
| ‘word2’  [abandoned] | a new word morphology experiment, where we want to model the joint probabilities of character sequences and corresponding tag sequences |  |
| ‘ptb\_wsj0’ | a typical experiment, where various LMs are trained on PTB training set and evaluated in terms of WERs by rescoring the 1000-best list of WSJ’92 test set. | [2][3][4] |
| ‘ptb\_tagged\_wsj0’ | semi-supervised experiments. We defined TRF LMs to model the joint probability of words and POC tags, and then use these TRF LMs to rescore the 1000-best list of WSJ’92 test set. |  |
| ‘ptb\_fake\_nbest’ | datasets are same to ‘ptb\_wsj0’. We create some fake n-best by randomly adding errors of substitution, insertion or deletion. In these experiments, we propose DNCE training method and compared with NCE. | [6] |
| ‘ptb\_chime4test’ | LMs trained on PTB and used to rescore the n-best list of the development and test sets of CHiME 4 challenge. The n-best lists are generated by Hongyu Xiang. |  |
| ‘pku\_rmrb’  [not finished] | Train LMs on People’s Daily data set, which is manually annotated by Peking University and bought by our Lab. |  |
| ‘hkust’ | Experiments on HKUST Chinese dataset. The 100-bset list is generated by me using Kaldi scripts. (For the scripts to train acoustic models and generate n-best list, see svn:  <https://101.6.68.39/svn/Gsp/user/wangbin/work/kaldi/egs/hkust/s5>). | [6] |
| ‘google1B’ | LMs are trained on google one-billion word benchmark and used to rescoring the 1000-best list of WSJ’92 test set. | [6] |
| ‘CHiME4’ | LM experiments on CHiME 4 Challenge. The the acoustic models are trained and n-best list are generated by Honeyu Xiang. | [5] |

References:

1. **Bin Wang**, Zhijian Ou, Jian Li, Akinori Kawamura. "Joint-Character-POC N-Gram Language Modeling For Chinese Speech Recognition". International Symposium on Chinese Spoken Language Processing (ISCSLP), 2014. (B类会议，EI收录，检索号：20144900274060)
2. **Bin Wang**, Zhijian Ou, Zhiqiang Tan. "Trans-dimensional Random Fields for Language Modeling". Annual meeting of the Association for Computational Linguistics (ACL), 2015. (A类会议，EI收录，检索号：20154201386950)
3. **Bin Wang**, Zhijian Ou, Zhiqiang Tan. "Learning Trans-Dimensional Random Fields with Applications to Language Modeling". IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2017, 40(4): 876-890. (SCI收录, 影响因子:8.329, 论文检索号:FY2ZU)
4. **Bin Wang**, Zhijian Ou. "Language Modeling with Neural Trans-dimensional Random Fields". IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017. (B类会议，已录用)
5. **Bin Wang**, Zhijian Ou. "Learning Neural Trans-dimensional Random Field Language Models with Noise-contrastive Estimation". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018. (顶尖级国际会议，已录用)
6. **Bin Wang**, Zhijian Ou, the latest paper.