**1. Summary**

This paper focuses on the progress of language modeling. It first introduces several techniques in detail with variations and limits, then describes experiments combining techniques in various ways (with different training data sizes), and finally discusses the results as conclusion. Note that evaluation metrics contains perplexity, entropy and word-error rate.

Techniques including smoothing, high-order n-grams, skipping, sentence mixture models, caching and clustering are systematically explored. (1) For smoothing, Interpolated Kneser-Ney smoothing (with modified backoff distribution and simple discounting scheme) outperforms almost all other techniques. (2) For high-order n-grams, the best performance improvement is achieved with n = 5 and later increment on n does not bring obvious gain. (3) For skipping, it shows that using pairs through to 5-gram level could catch almost all benefit. (4) For sentence mixture models, increasing the number of mixtures brings out-of-expectation performance improvement. (5) For caching, it works excellently for reducing perplexity when the training data size is small/medium. (6) For clustering, a new technique called “fullibmpredict” (which utilizes intuition behind predictive clustering and factors the problem into cluster prediction) is the best performing technique already found without combination.

With combination of various techniques (sentence mixture model working as highest level and predictive clustering combined with other techniques as sentence-specific model), 38%-50% (depending on training data sizes) perplexity reduction is achieved compared with a fair baseline (Katz smoothed trigram model without count cutoffs). And it is also good to obtain 8.9% word-error rate reduction (caching is not included in combination). Note that smoothing might plays the most important role in combination, while other techniques are also necessary.

The results are a bit discouraging since word-error rate reduction is just modest with the proposed complex, large and slow combination architecture. Note that the traditional trigrams model still remains the standard greatly because of its efficiency. However, there are still some encouraging points like young techniques and novel combinations. Also, many other techniques are being explored: maximum entropy models, neural networks, latent semantic analysis… Finally, though language modeling is indeed very difficult, the authors points out that 4 areas of hope: language model compression, language model adaptation, applications of language model to other domains and basic research (e.g., soft-clustering).