1. Introduction

- Word n-grams suffer from data sparseness
- Category n-grams generalise to unseen words better
- Competitive performance for small training sets
- Combining word with category n-grams improves performance, even for large training sets
- Performance depends on category definitions

2. Category definitions

**Part-of-speech categories**
- 152 categories from tagged LOB corpus
- Words may belong to several categories
- Example categories:
  - ADJ = (able, abnormal, ...), light, ... yellow, young)

** Automatically determined categories**
- Most frequent words in individual categories
- Remaining words grouped in single category

Algorithm: (Kneser & Ney, Eurospeech 93)

Example categories:
- however, meanwhile, indeed, separately, moreover, nor, neither, nevertheless, nonetheless, similarly ...
- Iran, dextrel, anyone, Brazil, someone, everyone, Moscow, Israel, Iraq, parliament, everybody ...
- March, December, midnight, midday, noon, midnight, diligence, midafternoon, midmorning, sept ...

3. Variable-length n-grams

- Selectively increase length of individual n-grams according to expected performance benefit
- Leaving one-out cross validation
- Optimise performance while minimising model size

4. Experiments

**ARPA CSR 94 HUB–1 Evaluation**

- Built several category-based models
- One using part-of-speech classes
- Various using automatically-derived categories
- Combined with baseline trigram by linear interpolation
- Training: 37 M words 1987–89 Wall Street Journal
- N-best rescoring (N = 100) with 65K HTK recogniser
- Interpolation weight minimises dev word error rate

**DARPA 97 Broadcast News Eval**

- Use 1000 categories
- Recognition by lattice rescoring

**5. Conclusions**

- Even with equal number of n-grams, automatically-derived categories perform better
- Clustering distributes words evenly among categories
- Uneven distribution in part-of-speech categories
- As number of categories increase, performance reaches optimum
  - Ability to generalise deteriorates with too many categories
  - Generalisation allows word n-gram performance to be improved
- Performance improvement is negligible for n > 4

---

A Comparison of Part-of-Speech and Automatically Derived Category-Based Language Models for Speech Recognition

T.R. Niesler, E.W.D. Whittaker and P.C. Woodland

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England

http://svr-www.eng.cam.ac.uk