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`

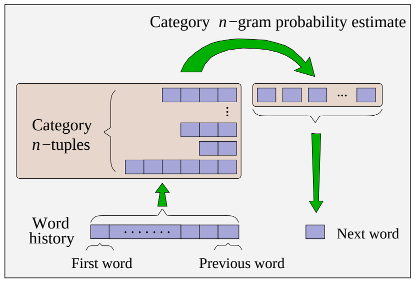
1. Introduction

- Word n−grams suffer from data sparseness
- Category n−grams generalise to unseen word sequences => improved robustness
- Competitive performance for small training sets
- Combining word with category n−grams improves performance, even for large training sets
- Performance depends on category definitions
- Here we compare
  - Part−of−speech based categories, and
  - Automatically determined categories

  in terms of
  - Perplexity, and
  - Word−error rate

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



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}

  "light" ∈ {ADJ, NOUN, VERB}

Automatically determined categories

Initialisation:

- Most frequent words in individual categories
- Remaining words grouped in single category

Algorithm: (Kneser & Ney, Eurospeech 93)

```
•For all words
  •Take word from its category
  •For all categories:
    •Put word in category
    •Calc bigram training set  ΔLL
  • Move word to category for which
    ΔLL  is greatest
```

Example categories:

{ however, meanwhile, indeed, separately moreover, nor, neither, nevertheless, nonetheless, similarly ... }

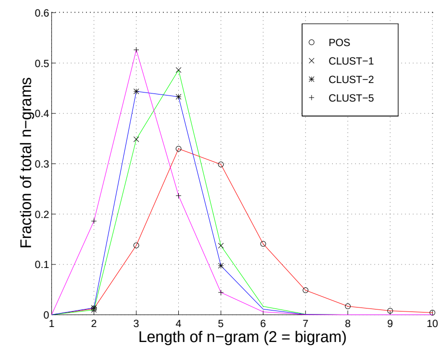
{ iran, dextrel, anyone, brazil, someone, everyone, moscow, israel, iraq, parliament, everybody ... }

{ march, december, midnight, midday, noon, midyear, diligence, midafternoon, midmorning, sept ... }

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



n	n−grams (millions)	Interpolated perplexity	WER (eval)
2	0.21	152.0	12.4
3	2.70	139.2	11.8
4	1.47	132.8	11.7
5	0.27	131.8	11.7
6	0.03	131.7	11.7

Effect of n−gram length for 1994 HUB−1

Category type	Number of categories	n−grams (millions)	Standalone perplexity	Interpolated perplexity	WER (eval)
POS	152	0.91	448.5	139.4	12.3
CLUST−0	150	1.04	301.1	142.2	12.2
CLUST−1	150	2.13	289.5	139.1	11.9
CLUST−2	200	2.70	265.8	136.9	11.9
CLUST−3	500	4.68	212.2	131.7	11.7
CLUST−4	1,000	6.38	184.4	129.7	11.8
CLUST−5	2,000	8.38	167.8	129.4	12.0
Word	65,000	4.88	148.8	148.8	12.5

Performance of various category language models for 1994 HUB−1

DARPA 97 Broadcast News Eval

- Use 1000 categories
- Recognition by lattice rescoring

Model	n−grams (millions)	Perplexity	WER
Word 4g	23.9	147	17.3
Cat 3g	7.9	238	−
Interp	31.8	137	16.8

Language model performance for BN−97

5. Conclusions

- Even with equal number of n−grams, automatically−derived categories perform better
  - Clustering distributes words evenly among categories
  - Uneven distibution 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

Presented at the International Conference on Acoustics, Speech and Signal Processing, Seattle, May 1998.