Best Language Modeling Result

T.R. Niesler, E.W.D. Whittaker, P.C. Woodland; *Comparison of Part-of-Speech and Automatically Derived Category-Based Language Models for Speech Recognition*

- cluster words into classes according to bigram mutual information (Brown *et al.*, 1990; Ney *et al.*, 1994)
- build variable-length class n-gram model (up to 4-grams and higher)
- linearly interpolate word trigram model with class n-gram model
- 10% decrease in perplexity
- WER 12.5% $\rightarrow$ 11.7% (WSJ)

Most Interesting LM Paper, #1

D. Klakow, *Language-Model Optimization by Mapping of Corpora*

- observation: don't build LM on original text, build on mapped text
  - `at the White House today` $\Rightarrow$ `at the White_House today`
- transformations
  - joining adjacent words: `White House` $\Rightarrow$ `White_House`
  - joining non-adjacent words: `the king of` $\Rightarrow$ `the_?_of king`
  - spelling OOV's: `bugugu` $\Rightarrow$ `b u g u g u`
- improve bigram perplexity by 29%, trigram by 7%

Most Interesting LM Paper, #2

J.R. Bellagarda, *Exploiting Both Local and Global Contraints for Multi-Span Statistical Language Modeling*

- topic adaptation $p(w_0|w_{-1}) \Rightarrow p(w_0|w_{-1}, \text{topic})$
- cluster documents, words
  - build co-occurrence matrix of documents, words
  - map documents/words into lower-dimensional space (LSA/PCA/SVD)
  - cluster document/word vectors (K-means)
- to combine topic info and bigram model
  - find nearest document clusters, combine cluster and bigram probs multiplicatively
- WSJ perplexities
  - bigram: 215, trigram: 142, bigram + topic: 102

Most Interesting Paper, Overall

K.A. Papineni, S. Roukos, R.T. Ward, *Maximum Likelihood and Discriminative Training of Direct Translation Models*

- convert ATIS natural language query to database query
  - *show me all nonstop flights from airport-1 to city-1 leaving airport-1 before time-1 on day-1*
  - `LIST FLIGHTS NONSTOP DEPARTING BEFORE TIME-1 FLYING-ON DAY-1 FROM:AIRPORT AIRPORT-1 TO:CITY CITY-1`
- statistical machine translation!
  - exponential model p(database query $T|$NL query S)
  - training maximum likelihood, maximum discrimination
- 84% queries correct

ROVER Revisited

T.J. Hazen, A.K. Halberstadt, *Using Aggregation to Improve the Performance of Mixture Gaussian Acoustic Models*

- ROVER (Fiscus, 1997): multiple recognizers voting reduces WER
- random starting points for K-means clustering for training Gaussian mixtures
- average output of four different models
- TIMIT phone error-rate: 29.3% $\rightarrow$ 28.4%, WER: 4.5% $\rightarrow$ 4.0%

Discriminative Training

- instead of training parameters to maximum likelihood
  - train parameters to minimize word-error rate, or smoothed version thereof
- training acoustic models
  - R. Shlüter, W. Macherey, *Comparison of Discriminative Training Criteria*
    - continuous digits, 3.0% $\rightarrow$ 2.5% WER
  - C. Chesta, A. Girardi, P. Laface, M. Nigra, *Discriminative Training of Hidden Markov Models Using a Classification Measure Criterion*
    - isolated word, 68 word vocab
    - 98.5% $\rightarrow$ 98.8% WA
  - helps more with worse models, *e.g.*, fewer Gaussians per mixture

Discriminative Training

- P. Beyerlein, *Discriminative Model Combination*
  - find language weight(s)
  - does as well as exhaustive search
  - combine (1 AM + 1 LM) *vs.* (2 AM + 3 LM): 3.5% $\rightarrow$ 3.2% WER (WSJ)
- V. Goel, W. Byrne, S. Khudanpur, *LVCSR Rescoring with Modified Loss Functions: A Decision Theoretic Perspective*
  - search for hypothesis that minimizes WER as opposed to sentence error-rate (sentence likelihood)
  - 38.5% $\rightarrow$ 38.0% WER (SWB)