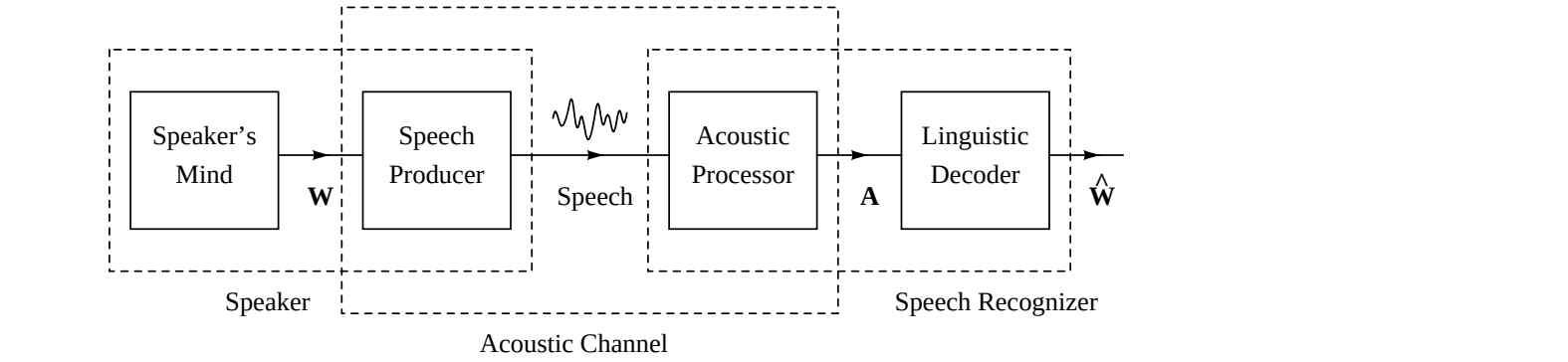
Language Modeling in the Era of Abundant Data

Ciprian Chelba

Statistical Modeling in Automatic Speech Recognition



$$\hat{W} = \text{argmax}_{\text{W}} P(W|A) = \text{argmax}_{\text{W}} P(A|W) \cdot P(W)$$

- $P(A|W)$ *acoustic model* (AM, Hidden Markov Model); varies depending on problem (machine translation, spelling correction, soft keyboard input)
- $P(W)$ *language model* (LM, usually Markov chain)
- *search* for the most likely word string $\hat{W}$

Language Modeling Usual Assumptions

- we have a word level tokenization of the text (not true in all languages, e.g. Chinese)
- some vocabulary is given to us (usually also estimated from data);
- out-of-vocabulary (OoV) words are mapped to <UNK> (“open” vocabulary LM)
- sentences are assumed to be independent and of finite length; LM needs to predict end-of-sentence symbol </S>
-

```
On my second day , I managed the uphill walk
to a waterfall called <UNK> Skok . </S>
```

Language Model Evaluation (1)

Word Error Rate (WER)

```
TRN: UP UPSTATE NEW YORK SOMEWHERE UH      OVER
HYP:    UPSTATE NEW YORK SOMEWHERE UH ALL  ALL
     D  0       0   0    0         0  I    S
   :3 errors/7 words in transcript; WER = 43%
```

Perplexity (PPL) (Jelinek, 1997)

$$PPL(M) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\ln\left[P_M(w_i|w_1\ldots w_{i-1})\right]\right)$$

- good models are “smoothed” ML estimates: $P_M(w_i|w_1\ldots w_{i-1}) > \epsilon$; also guarantees a proper probability model over sentences
- other metrics: out-of-vocabulary rate/n-gram hit ratios

Language Model Smoothing

Markov assumption leads to N-gram model:

$$P_\theta(w_i|w_1 \dots w_{i-1}) = P_\theta(w_i|w_{i-N+1} \dots w_{i-1}), \theta \in \Theta, w_i \in \mathcal{V}$$

Smoothing using Deleted Interpolation:

$$\begin{aligned} P_n(w|h) &= \lambda(h) \cdot P_{n-1}(w|h') + (1 - \lambda(h)) \cdot f_n(w|h) \\ P_{-1}(w) &= uniform(\mathcal{V}) \end{aligned}$$

where:

- $h = (w_{i-n+1} \dots w_{i-1})$ is the n-gram context, and $h' = (w_{i-n+2} \dots w_{i-1})$ is the back-off context
- weights $\lambda(h)$ must be estimated on held-out (cross-validation) data.

Language Model Smoothing: Katz

Katz Smoothing (Katz, 1987) uses Good-Turing discounting:

$$P_n(w|h) = \begin{cases} f_n(w|h), C(h,w) > K \\ (r+1)\frac{t_{r+1}}{t_r} \cdot f_n(w|h), 0 < C(h,w) \leq K \\ \beta(h) P_{n-1}(w|h') \end{cases}$$

where:

- t_r represents the number of n-grams (types) that occur r times: $t_r = |(w_{i-n+1} \dots w_i), C(w_{i-n+1} \dots w_i) = r|$
- $\beta(h)$ is the back-off weight ensuring proper normalization

Language Model Smoothing: Kneser-Ney

Kneser-Ney Smoothing (Kneser & Ney, 1995):

$$P_n(w|h) = \begin{cases} \frac{C(h,w)-D_1}{C(h)} + \lambda(h)P_{n-1}(w|h'), n = N \\ \frac{LeftDivC(h,w)-D_2}{\sum_w LeftDivC(h,w)} + \lambda(h)P_{n-1}(w|h'), 0 \leq n < N \end{cases}$$

where:

- $LeftDivC(h,w) = |v, C(v,h,w) > 0|$ is the "left diversity" count for an n-gram (h,w)

See (Goodman, 2001) for a detailed presentation on LM smoothing.

Language Model Representation: ARPA Back-off

```
p(wd3|wd1,wd2)=
 if(trigram exists)     p_3(wd1,wd2,wd3)
 else if(w1,w2 exists) bo_2(w1,w2)*p(wd3|wd2)
 else                   p(wd3|w2)
p(wd2|wd1)=
 if(w1,w2 exists)       p_2(wd1,wd2)
 else                   bo_1(wd1)*p_1(wd2)

\1-grams:
p_1     wd       bo_1
\2-grams:
p_2     wd1 wd2 bo_2
\3-grams:
p_3     wd1 wd2 wd3
```

Language Model Size Control: Entropy Pruning

Entropy pruning (Stolcke, 1998) is required for use in 1st pass:

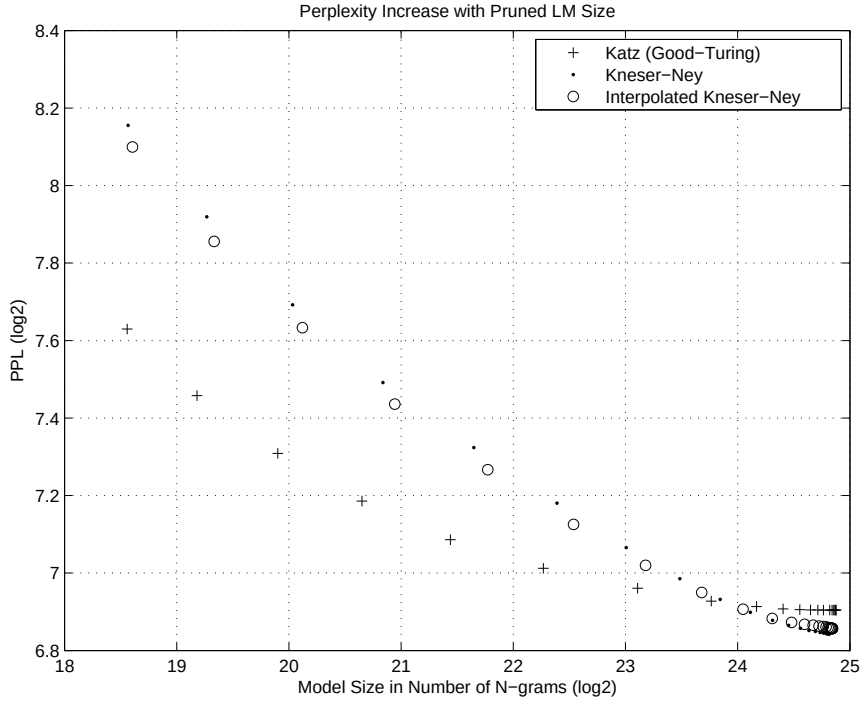
- should one remove n-gram (h, w)?

$$D[q(h)p(\cdot|h) \parallel q(h) \cdot p'(\cdot|h)] = q(h) \sum_w p(w|h) \log \frac{p(w|h)}{p'(w|h)}$$

$$| D[q(h)p(\cdot|h) \parallel q(h) \cdot p'(\cdot|h)] | < pruning\ threshold$$

- lower order estimates: $q(h) = p(h_1) \dots p(h_n|h_1 \dots h_{n-1})$ or relative frequency: $q(h) = f(h)$
- greedily reduces LM size at min cost in PPL

On Smoothing and Pruning



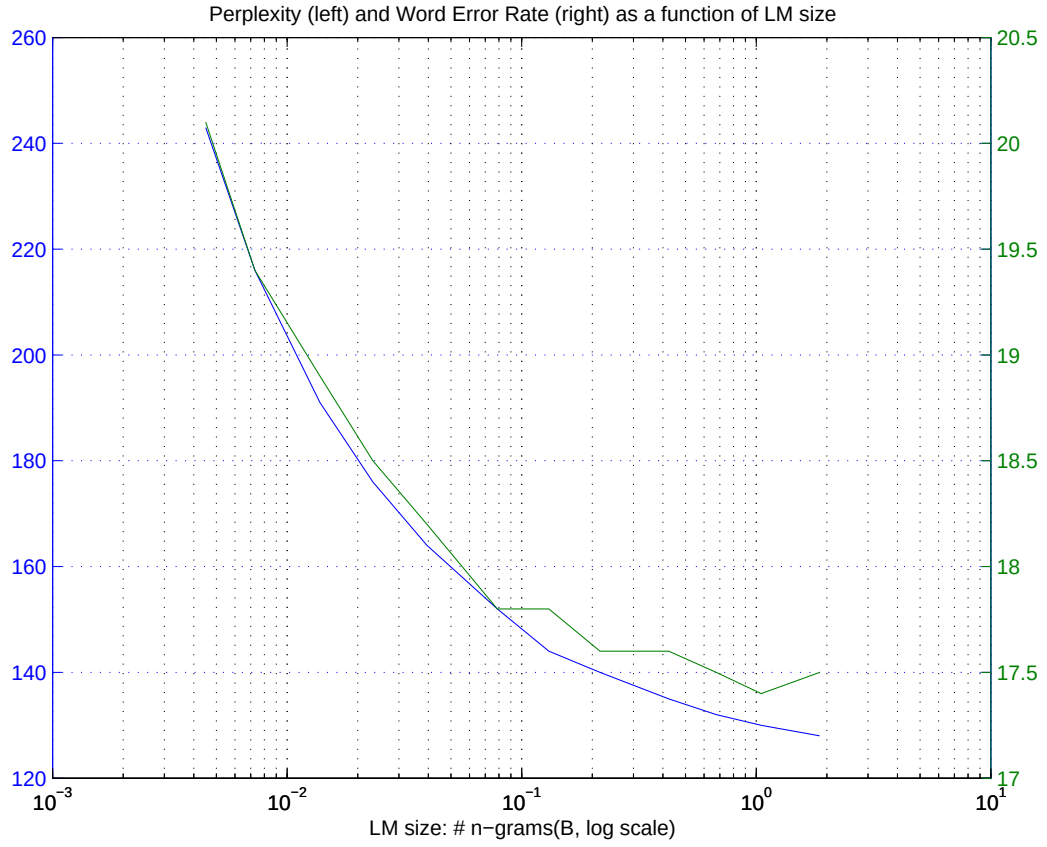
- KN degrades very fast with aggressive pruning (< 10% of original size) (Ciprian Chelba, 2010)
- switch from KN to Katz smoothing: 10% WER gain for voice-search

Voice Search LM Training Setup (Chelba & Schalkwyk, 2013)

- spelling corrected google.com queries, normalized for ASR, e.g. `5th -> fifth`
- vocabulary size: 1M words, OoV rate 0.57% (!), excellent n-gram hit ratios
- training data: 230B words

Order	no. n-grams	pruning	PPL	n-gram hit-ratios
3	15M	entropy	190	47/93/100
3	7.7B	none	132	97/99/100
5	12.7B	1-1-2-2-2	108	77/88/97/99/100

Is Bigger Better? YES!



- PPL is really well correlated with WER when controlling for vocabulary and training set.

Better Language Models: More Smarts

1-billion word benchmark (Chelba et al., 2013) results

Model	Num. Params	PPL
Katz 5-gram	1.74 B	79.9
Kneser-Ney 5-gram	1.76 B	67.6
SNM skip-gram	33 B	52.9
RNN	20 B	51.3
ALL, linear interpolation		41.0

- there are LMs that handily beat the N-gram by leveraging longer context (when available)
- how about increasing the amount of data, when we have it?

Better Language Models: More Smarts, More Data? Ideally Both

10/100 billion word query data benchmark results[a]

Model	Data Amount	Num. Params	PPL
Katz 6-gram	10B	3.2 B	123.9
Kneser-Ney 6-gram	10B	4.1 B	114.5
SNM skip-gram	10B	25 B	111.0
RNN	10B	4.1 B	111.1
Katz 6-gram	100B	19.6 B	92.7
Kneser-Ney 6-gram	100B	24.5 B	87.9
RNN	100B	4.1 B	101.0

- more data and model is an easy way to get solid gains
- complex models better scale up gracefully
- KN smoothing loses its edge over Katz

[a] Thanks Babak Damavandi for the RNN experimental results

More Data Is Not Always a Winner: Query Stream Non-stationarity (1)

- USA training data:
  - XX months
  - X months
- test data: 10k, Sept-Dec 2008
- very little impact in OoV rate for 1M wds vocabulary: 0.77% (X months vocabulary) vs. 0.73% (XX months vocabulary)

More Data Is Not Always a Winner: Query Stream Non-stationarity (2)

3-gram LM	Training Set	Test Set PPL
unpruned	X months	121
unpruned	XX months	132
entropy pruned	X months	205
entropy pruned	XX months	209

- bigger is not always better[a]
- 10% rel reduction in PPL when using the most recent X months instead of XX months
- no significant difference after pruning, in either PPL or WER

[a] The vocabularies are mismatched, so the PPL comparison is troublesome. The difference would be higher if we used a fixed vocabulary.

More Locales

- training data across 3 locales: USA, GBR, AUS, spanning same amount of time ending in Aug 2008
- test data: 10k/locale, Sept-Dec 2008

Out of Vocabulary Rate:

Training Locale	Test Locale USA	GBR	AUS
USA	0.7	1.3	1.6
GBR	1.3	0.7	1.3
AUS	1.3	1.1	0.7

- locale specific vocabulary halves the OoV rate

Locale Matters (2)

Perplexity of unpruned LM:

Training Locale	Test Locale USA	GBR	AUS
USA	132	234	251
GBR	260	110	224
AUS	276	210	124

- locale specific LM halves the PPL of the unpruned LM

Open Problems

- Entropy of text from a given source:
  how much are we leaving on the table?
- How much data/model is enough for a given source:
  does such a bound exist for N-gram models?
- More data, relevance, transfer learning:
  not all data is created equal.
- Conditional ML estimation:
  LM estimation should take into account the channel model.

Entropy of English

High variance, depending on estimate, source of data; 0.1-0.2 bits/char is a significant difference in PPL at word level!

- (Cover & King, 1978): 1.3 bits/char
- (Brown, Pietra, Mercer, Pietra, & Lai, 1992): 1.75 bits/char
- 1-billion corpus: $\approx^a$ 1.17 bits/char for KN, $\approx$ 1.03 bits/char for the best reported LM mixing skip-gram SNM with RNN
- 10, 100 -billion query corpus: $\approx$ 1.43, 1.35 bits/char for KN, respectively.

[a] Modulo OoV word modeling

Abundant Data: How Much is Enough for Modeling a Given Source?

A couple of observations:

- one can prune an LM to about 10% of unpruned size without significant impact on PPL
- increasing the amount of data and model size becomes unproductive after a while

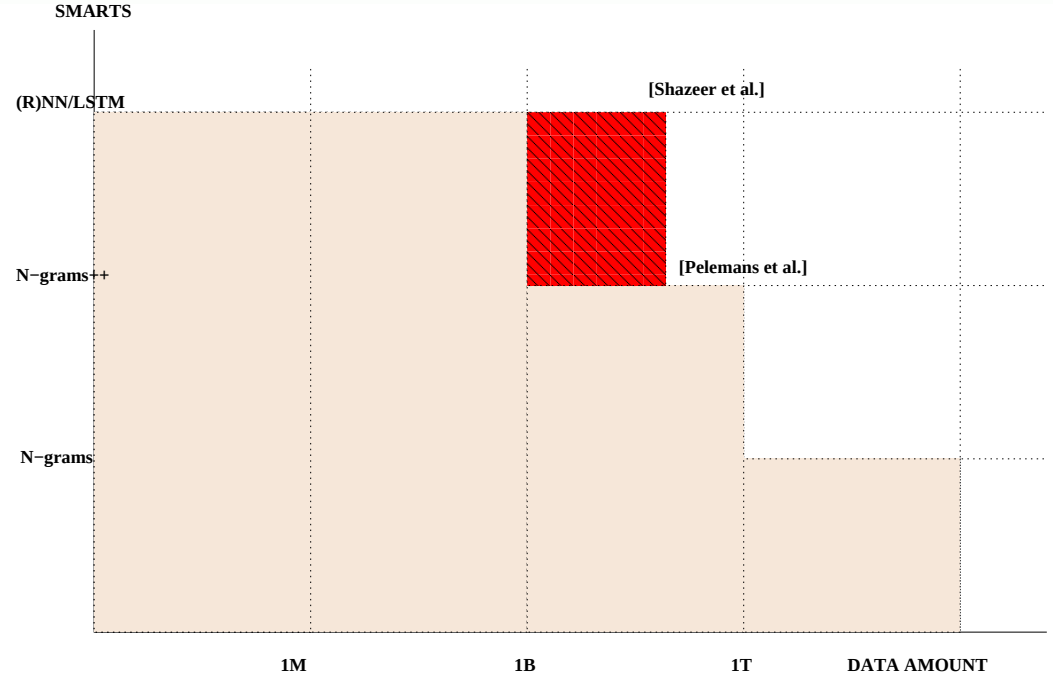
For a given source, and N-gram order, is there a data size beyond which there is no benefit to the model quality?

Abundant Data: Not All Data is Created Equal

- It is not always possible to find very large amounts of data that is well matched to a given application/test set
- E.g. when building an LM for SMS text we may have very little such data, quite a bit more from posts on social networks, and a lot of text from a web crawl.
- LM adaptation: leveraging data in different amounts, and of various degrees of relevance[a] to a given test set.

[a] Relevance of data to a given test set is hard to describe, but you know it when you see it.

More Smarts with Abundant Data



References

Brown, P. F., Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D., & Lai, J. C. (1992, March). An estimate of an upper bound for the entropy of english. *Comput. Linguist.*, *18*(1), 31–40. Available from `http://dl.acm.org/citation.cfm?id=146680.146685`

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., et al. (2013). *One billion word benchmark for measuring progress in statistical language modeling.*

Chelba, C., & Schalkwyk, J. (2013). Empirical exploration of language modeling for the google.com query stream as applied to mobile voice search. In *Mobile speech and advanced natural language solutions* (pp. 197–229). New York: Springer. Available from `http://www.springer.com/engineering/signals/book/978-1-4614-6017-6`

References

Ciprian Chelba, Will Neveitt, Peng Xu, Thorsten Brants. (2010). Study on Interaction between Entropy Pruning and Kneser-Ney Smoothing. In *Proc. interspeech* (pp. 2242–2245). Makuhari, Japan.

Cover, T., & King, R. (1978, September). A convergent gambling estimate of the entropy of english. *IEEE Trans. Inf. Theor.*, *24*(4), 413–421. Available from `http://dx.doi.org/10.1109/TIT.1978.1055912`

Goodman, J. (2001). *A bit of progress in language modeling, extended version* (Tech. Rep.). Microsoft Research.

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, MA, USA: MIT Press.

References

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *Ieee transactions on acoustics, speech and signal processing* (Vol. 35, p. 400-01).

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the ieee international conference on acoustics, speech and signal processing* (Vol. 1, pp. 181–184).

Stolcke, A. (1998). Entropy-based pruning of back-off language models. In *Proceedings of news transcription and understanding workshop* (pp. 270–274). Lansdowne, VA: DARPA.

Pelemans et al. (2016). Sparse Non-negative Matrix Language Modeling. In *Transactions of the Association for Computational Linguistics* (pp. 329–342). TACL: ACL.

Shazeer et al. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *Submitted to ICLR* (pp. –). CoRR: ArXiv.