
Empirical Methods in Natural Language Processing

Lecture 6

Language Models: More smoothing

(most slides from Sharon Goldwater; some adapted from Philipp Koehn)

31 January 2017



Recap: Smoothing for language models

- N -gram LMs reduce sparsity by assuming each word only depends on a fixed-length history.
- But even this assumption isn't enough: we still encounter lots of unseen N -grams in a test set or new corpus.
- If we use MLE, we'll assign 0 probability to unseen items: useless as an LM.
- **Smoothing** solves this problem: move probability mass from seen items to unseen items.
- Add- α smoothing: ($\alpha = 1$ or < 1) very simple, but no good when vocabulary size is large.

Good-Turing smoothing: intuition

- Good-Turing smoothing:
 - estimate the probability of seeing (any) item with N_c counts (e.g., 0 count) as the proportion of items already seen with N_{c+1} counts (e.g., 1 count).
 - Divide that probability evenly between all possible items with N_c counts.

Good-Turing smoothing

- If n is count of history, then $P_{GT} = \frac{c^*}{n}$ where

$$c^* = (c + 1) \frac{N_{c+1}}{N_c}$$

- N_c number of N -grams that occur exactly c times in corpus
- N_0 total number of unseen N -grams
- Ex. for trigram probability $P_{GT}(\text{three}|\text{I spent})$, then n is count of **I spent** and c is count of **I spent three**.

Problems with Good-Turing

- Assumes we know the vocabulary size (no unseen words)
[but again, use UNK: see J&M 4.3.2]
- Doesn't allow “holes” in the counts (if $N_i > 0$, $N_{i-1} > 0$)
[can estimate using linear regression: see J&M 4.5.3]
- Applies discounts even to high-frequency items
[but see J&M 4.5.3]
- But there's a more fundamental problem...

Remaining problem

- In training corpus, suppose we see *Scottish beer* but neither of
 - *Scottish beer drinkers*
 - *Scottish beer eaters*
- If we build a trigram model smoothed with Add- α or G-T, which example has higher probability?

Remaining problem

- Previous smoothing methods assign equal probability to all unseen events.
- Better: use information from lower order N -grams (shorter histories).
 - beer drinkers
 - beer eaters
- Two ways: **interpolation** and **backoff**.

Interpolation

- **Combine** higher and lower order N -gram models, since they have different strengths and weaknesses:
 - high-order N -grams are sensitive to more context, but have sparse counts
 - low-order N -grams have limited context, but robust counts
- If P_N is N -gram estimate (from MLE, GT, etc.; $N \in \{1, 2, 3\}$), use:

$$P_{\text{INT}}(w_3|w_1, w_2) = \lambda_1 P_1(w_3) + \lambda_2 P_2(w_3|w_2) + \lambda_3 P_3(w_3|w_1, w_2)$$

$$P_{\text{INT}}(\text{three}|\text{I, spent}) = \lambda_1 P_1(\text{three}) + \lambda_2 P_2(\text{three}|\text{spent}) \\ + \lambda_3 P_3(\text{three}|\text{I, spent})$$

Interpolation

- Note that λ_i s must sum to 1:

$$\begin{aligned} 1 &= \sum_{w_3} P_{\text{INT}}(w_3|w_1, w_2) \\ &= \sum_{w_3} [\lambda_1 P_1(w_3) + \lambda_2 P_2(w_3|w_2) + \lambda_3 P_3(w_3|w_1, w_2)] \\ &= \lambda_1 \sum_{w_3} P_1(w_3) + \lambda_2 \sum_{w_3} P_2(w_3|w_2) + \lambda_3 \sum_{w_3} P_3(w_3|w_1, w_2) \\ &= \lambda_1 + \lambda_2 + \lambda_3 \end{aligned}$$

Fitting the interpolation parameters

- In general, any weighted combination of distributions is called a **mixture model**.
- So λ_i s are **interpolation parameters** or **mixture weights**.
- The values of the λ_i s are chosen to optimize perplexity on a held-out data set.

Back-Off

- Trust the highest order language model that contains N -gram

$$P_{BO}(w_i|w_{i-N+1}, \dots, w_{i-1}) = \begin{cases} P^*(w_i|w_{i-N+1}, \dots, w_{i-1}) & \text{if } \text{count}(w_{i-N+1}, \dots, w_i) > 0 \\ \alpha(w_{i-N+1}, \dots, w_{i-1}) P_{BO}(w_i|w_{i-N+2}, \dots, w_{i-1}) & \text{else} \end{cases}$$

Back-Off

- Requires
 - adjusted prediction model $P^*(w_i | w_{i-N+1}, \dots, w_{i-1})$
 - backoff weights $\alpha(w_1, \dots, w_{N-1})$
- Exact equations get complicated to make probabilities sum to 1.
- See textbook for details if interested.

Do our smoothing methods work here?

Example from MacKay and Peto (1995):

Imagine, you see, that the language, you see, has, you see, a frequently occurring couplet, ‘you see’, you see, in which the second word of the couplet, see, follows the first word, you, with very high probability, you see. Then the marginal statistics, you see, are going to become hugely dominated, you see, by the words you and see, with equal frequency, you see.

- $P(\text{see})$ and $P(\text{you})$ both high, but *see* nearly always follows *you*.
- So $P(\text{see}|\text{novel})$ should be much lower than $P(\text{you}|\text{novel})$.

Diversity of histories matters!

- A real example: the word *York*
 - fairly frequent word in Europarl corpus, occurs 477 times
 - as frequent as *foods*, *indicates* and *providers*
 - in unigram language model: a respectable probability
- However, it almost always directly follows *New* (473 times)
- So, in unseen bigram contexts, *York* should have low probability
 - lower than predicted by unigram model as used in interpolation/backoff.

Kneser-Ney Smoothing

- Kneser-Ney smoothing takes diversity of histories into account
- Count of distinct histories for a word:

$$N_{1+}(\bullet w_i) = |\{w_{i-1} : c(w_{i-1}, w_i) > 0\}|$$

- Recall: maximum likelihood est. of unigram language model:

$$P_{ML}(w_i) = \frac{C(w_i)}{\sum_w C(w)}$$

- In KN smoothing, replace raw counts with count of histories:

$$P_{KN}(w_i) = \frac{N_{1+}(\bullet w_i)}{\sum_w N_{1+}(\bullet w)}$$

Kneser-Ney in practice

- Original version used backoff, later “modified Kneser-Ney” introduced using interpolation.
- Fairly complex equations, but until recently the best smoothing method for word n -grams.
- See Chen and Goodman (1999) for extensive comparisons of KN and other smoothing methods.
- KN (and other methods) implemented in language modelling toolkits like SRILM (classic), KenLM (good for really big models), OpenGrm Ngram library (uses finite state transducers), etc.

Ken_{neth} HeafieldLM

KenLM is a highly efficient language modeling toolkit.



<https://kheafield.com/code/kenlm/>

Are we done with smoothing yet?

We've considered methods that predict rare/unseen words using

- Uniform probabilities (add- α , Good-Turing)
- Probabilities from lower-order n-grams (interpolation, backoff)
- Probability of appearing in new contexts (Kneser-Ney)

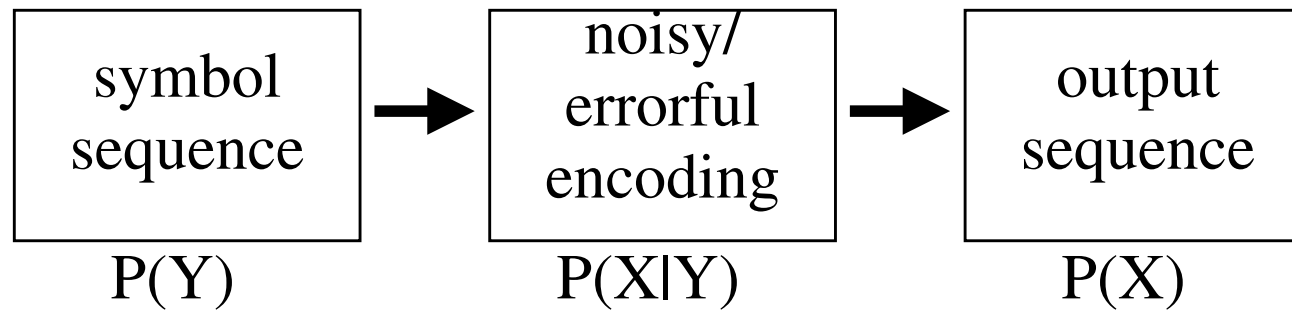
What's left?

Back to the big picture

- However we train our LM, we will want to use it in some application.
- Now, a bit more detail about how that can work.
- We need another concept from information theory: the **Noisy Channel Model**.

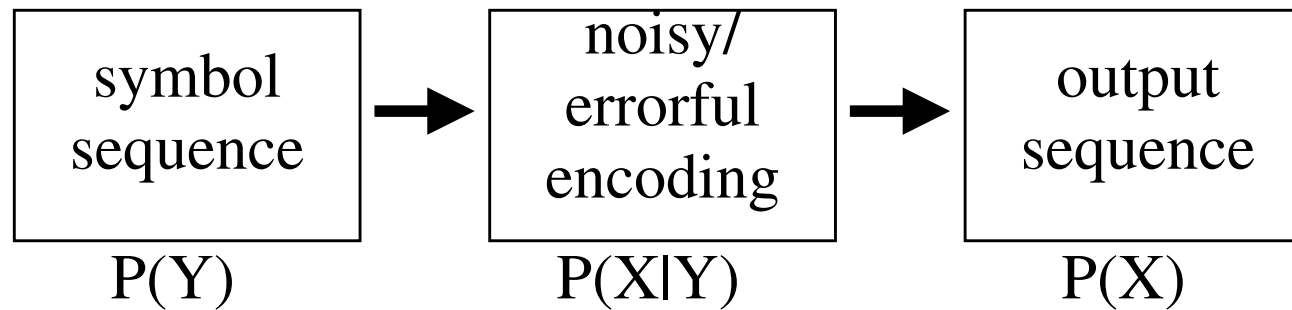
Noisy channel model

- We imagine that someone tries to communicate a sequence to us, but noise is introduced. We only see the output sequence.



Noisy channel model

- We imagine that someone tries to communicate a sequence to us, but noise is introduced. We only see the output sequence.



Application	Y	X
Speech recognition	spoken words	acoustic signal
Machine translation	words in L_1	words in L_2
Spelling correction	intended words	typed words

Example: spelling correction

- $P(Y)$: Distribution over the words the user intended to type. A language model!
- $P(X|Y)$: Distribution describing what user is **likely** to type, given what they **meant**. Could incorporate information about common spelling errors, key positions, etc. Call it a **noise model**.
- $P(X)$: Resulting distribution over what we actually see.
- Given some particular observation x (say, `effert`), we want to recover the most probable y that was intended.

Summary

- Different smoothing methods account for different aspects of sparse data and word behaviour.
 - Interpolation/backoff: leverage advantages of both higher and lower order N -grams.
 - Kneser-Ney smoothing: accounts for diversity of history.
 - Distributed representations: account for word similarity.
- Noisy channel model combines LM with noise model to define the “best” solution for many applications.

References

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Mackay, D. J. C. and Peto, L. C. B. (1995). A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):289–308.