

# Assignment 4,5 + Smoothing 1

## (SNLP Tutorial 5)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

25th, 27th May 2021

# Assignment 4

- Exercise 1: Huffman encoding
- Exercise 2: Conditional entropy of DNA
- Bonus: Huffman encoding adaptations

# OOV words

## Corpus

- Train set:



- Test set:



# OOV words

## Corpus

- Train set:



- Test set:



## Accumulate counts

		6		5		3		2				
		4		2		2		2		1		1

## OOV words

### Corpus







- Train set:





- Test set:



### Accumulate counts

•  6	 5	 3	 2		
•  4	 2	 2	 2	 1	 1

## OOV words

- What about  and ?

## OOV words

### Corpus







- Train set:



- Test set:



### Accumulate counts

•  6	 5	 3	 2		
•  4	 2	 2	 2	 1	 1

## OOV words

- What about  and ?
- OOV rate?

## OOV words

### Corpus





- Train set:





- Test set:



### Accumulate counts

• 	6		5		3		2		
• 	4		2		2		2		1
									1

## OOV words

- What about  and ?
- OOV rate?
- $3/12 = 25\%$

## OOV words

### Corpus




- Train set:





- Test set:



### Accumulate counts

• 	6		5		3		2		
• 	4		2		2		2		1
									1

## OOV words

- What about  and ?
- OOV rate?
- $3/12 = 25\%$
- Solutions?



## OOV words

### Corpus








- Train set:





- Test set:



### Accumulate counts

• 	6		5		3		2		
• 	4		2		2		2		1
									1

## OOV words

- What about  and ?
- OOV rate?
- $3/12 = 25\%$
- Solutions?

## OOV words

### Corpus






- Train set:





- Test set:



### Accumulate counts

•  6	 5	 3	 2		
•  4	 2	 2	 2	 1	 1

## OOV words

- What about  and .
- OOV rate?
- $3/12 = 25\%$
- Solutions?

## OOV words

- How do we even know this will be an issue?

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece

## Questions



# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece
- ▶ Start with the alphabet, merge and add frequent character-level n-grams

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece
- ▶ Start with the alphabet, merge and add frequent character-level n-grams

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece
- ▶ Start with the alphabet, merge and add frequent character-level n-grams
- ▶ E.g. bedclothes became white  $\rightarrow$  bed @cloth @es be @came @white

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece
- ▶ Start with the alphabet, merge and add frequent character-level n-grams
- ▶ E.g. bedclothes became white  $\rightarrow$  bed @cloth @es be @came @white

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece
- ▶ Start with the alphabet, merge and add frequent character-level n-grams
- ▶ E.g. bedclothes became white  $\rightarrow$  bed @cloth @es be @came @white
- ▶ Used heavily in any modern NLP (MT, LM, QA, ...)

## Questions

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece
- ▶ Start with the alphabet, merge and add frequent character-level n-grams
- ▶ E.g. bedclothes became white  $\rightarrow$  bed @cloth @es be @came @white
- ▶ Used heavily in any modern NLP (MT, LM, QA, ...)

## Questions

- Can we still get an unknown “word”?

# Subword Units

Solution to OOV words: go lower

- Characters:  $V = \{a, b, c, \dots, \_ \}$
- Syllables:  $V = \{bo, ve, r, how, \dots, \_ \}$
- Data-driven units (subwords):  $V = \{smi, les, es, clo, \dots, \_ \}$
- ▶ Byte Pair Encoding, Word Piece, Sentence Piece
- ▶ Start with the alphabet, merge and add frequent character-level n-grams
- ▶ E.g. bedclothes became white  $\rightarrow$  bed @cloth @es be @came @white
- ▶ Used heavily in any modern NLP (MT, LM, QA, ...)

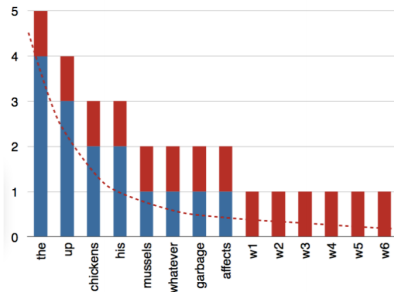
## Questions

- Can we still get an unknown “word”?
- How do we define perplexity for subword language models?

# Smoothing

- Words present in vocabulary, but have  $\sim 0$  probabilities
- Words present in vocabulary, but have unseen context

Solution: Assign probability mass from frequent events to infrequent events  
(Smoothing/Discounting)



- Will cover different smoothing methods over the next few tutorials









# Additive smoothing (add- $\alpha$ -smoothing)

## Distribution






- Add zero counts to frequency table

 6     5     3     2     0     0

- Increase all counts by  $\alpha = 1$

 6+1     5+1     3+1     2+1     0+1     0+1

- Divide by  $N = 22$

 0.32     0.27     0.18     0.13     0.05     0.05

## Perplexity

- Relative frequencies on test corpus:

 0.33     0.17     0.17     0.17     0.08     0.08

- Recall perplexity formula:

$$PP = \sum_{w,h} f(w, h) \cdot \log_2 p(w|h) \quad (1)$$







# Additive smoothing (add- $\alpha$ -smoothing)

## Distribution






- Add zero counts to frequency table

 6     5     3     2     0     0

- Increase all counts by  $\alpha = 1$

 6+1     5+1     3+1     2+1     0+1     0+1

- Divide by  $N = 22$

 0.32     0.27     0.18     0.13     0.05     0.05

## Perplexity

- Relative frequencies on test corpus:

 0.33     0.17     0.17     0.17     0.08     0.08

- PP:  $2^{-(0.33 \cdot (-1.64) + 0.17 \cdot (-1.89) + 0.17 \cdot (-4.32) + 0.17 \cdot (-2.47) + 0.08 \cdot (-2.94) + 0.08 \cdot (-4.32))} = 2^{(-2.6)} \approx 6$
- What would be PP with unsmoothed model?

## Additive smoothing: Bigrams

Recall the additive smoothing formula for unigrams:

$$C^*(w_i) = C(w_i) + \alpha \quad (2)$$

$$N^* = \sum_{w_i \in V} C^*(w_i) = N + \alpha|V| \quad (3)$$

## Additive smoothing: Bigrams

Recall the additive smoothing formula for unigrams:

$$C^*(w_i) = C(w_i) + \alpha \quad (2)$$

$$N^* = \sum_{w_i \in V} C^*(w_i) = N + \alpha|V| \quad (3)$$

$$p^*(w_i) = \frac{C(w_i) + \alpha}{N^*} = \frac{C(w_i) + \alpha}{N + \alpha|V|} \quad (4)$$

## Additive smoothing: Bigrams

Recall the additive smoothing formula for unigrams:

$$C^*(w_i) = C(w_i) + \alpha \quad (2)$$

$$N^* = \sum_{w_i \in V} C^*(w_i) = N + \alpha|V| \quad (3)$$

$$p^*(w_i) = \frac{C(w_i) + \alpha}{N^*} = \frac{C(w_i) + \alpha}{N + \alpha|V|} \quad (4)$$

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} \quad (5)$$

## Additive smoothing: Bigrams

Recall the additive smoothing formula for unigrams:

$$C^*(w_i) = C(w_i) + \alpha \quad (2)$$

$$N^* = \sum_{w_i \in V} C^*(w_i) = N + \alpha|V| \quad (3)$$

$$p^*(w_i) = \frac{C(w_i) + \alpha}{N^*} = \frac{C(w_i) + \alpha}{N + \alpha|V|} \quad (4)$$

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} \quad (5)$$

Smoothen the bigram count:  $C(w_{i-1}, w_i) \rightarrow C(w_{i-1}, w_i) + \alpha$

# Additive smoothing: Bigrams

## Corpus



Bigrams: , , , , ..., ,  ← circular bigram!

Bigrams: AA, AA, AE, EA, ..., AE, EA

## Additive smoothing: Bigrams: bigram counts

- Collect bigram counts & conditional probabilities for history  $A$

Bigram	$C(A, w_i)$	$C(A)$	$\frac{C(A, w_i)}{C(A)}$
AE	3	6	1/2
AA	2	6	1/3
AB	1	6	1/6



## Additive smoothing: Bigrams: add alpha

- We encounter an unknown bigram  $AF$

Bigram	$C(A, w_i)$	$C(A)$	$\frac{C_\alpha(A, w_i)}{C(A)}$
AE	3+1	6	4/6
AA	2+1	6	3/6
AB	1+1	6	2/6
→ AF	0+1	6	1/6

## Additive smoothing: Bigrams: add alpha

- We encounter an unknown bigram  $AF$

Bigram	$C(A, w_i)$	$C(A)$	$\frac{C_\alpha(A, w_i)}{C(A)}$
AE	3+1	6	4/6
AA	2+1	6	3/6
AB	1+1	6	2/6
→ AF	0+1	6	1/6

- Not a probability distribution!

## Additive smoothing: Bigrams: add alpha

- We encounter an unknown bigram  $AF$

Bigram	$C(A, w_i)$	$C(A)$	$\frac{C_\alpha(A, w_i)}{C(A)}$
AE	3+1	6	4/6
AA	2+1	6	3/6
AB	1+1	6	2/6
→ AF	0+1	6	1/6

- Not a probability distribution!
- Solution: We need to adjust the divisor a tiny bit. But how tiny?

## Additive smoothing: Bigrams: normalization

- Add  $\alpha \cdot 4$  to history count
- Pretend that we have seen the history  $|V| = 4$  times more.

## Additive smoothing: Bigrams: normalization

- Add  $\alpha \cdot 4$  to history count
- Pretend that we have seen the history  $|V| = 4$  times more.

Bigram	$C(A) + \alpha V $	$\frac{C_\alpha(A, w_i)}{C(A) + \alpha V }$
AE	$6 + 4$	$4/10$
AA	$6 + 4$	$3/10$
AB	$6 + 4$	$2/10$
$\rightarrow$ AF	$6 + 4$	$1/10$

## Additive smoothing: Bigrams: normalization

- Add  $\alpha \cdot 4$  to history count
- Pretend that we have seen the history  $|V| = 4$  times more.

Bigram	$C(A) + \alpha V $	$\frac{C_\alpha(A, w_i)}{C(A) + \alpha V }$
AE	$6 + 4$	$4/10$
AA	$6 + 4$	$3/10$
AB	$6 + 4$	$2/10$
$\rightarrow$ AF	$6 + 4$	$1/10$

- Now the probabilities sum up to 1:  $4/10 + 3/10 + 2/10 + 1/10 = 1$

## Additive smoothing: Bigrams: normalization

- We encounter another n-gram  $AD$
- What is  $|V|$  now?

## Additive smoothing: Bigrams: normalization

- We encounter another n-gram  $AD$
- What is  $|V|$  now?

Bigram	$C(A) + \alpha V $	$\frac{C_\alpha(A, w_i)}{C(A) + \alpha V }$
AE	$6 + 5$	$4/11$
AA	$6 + 5$	$3/11$
AB	$6 + 5$	$2/11$
$\rightarrow$ AF	$6 + 5$	$1/11$
$\rightarrow$ AD	$6 + 5$	$1/11$



## Additive smoothing: Bigrams: normalization

- We encounter another n-gram  $AD$
- What is  $|V|$  now?

Bigram	$C(A) + \alpha V $	$\frac{C_\alpha(A, w_i)}{C(A) + \alpha V }$
AE	$6 + 5$	$4/11$
AA	$6 + 5$	$3/11$
AB	$6 + 5$	$2/11$
$\rightarrow$ AF	$6 + 5$	$1/11$
$\rightarrow$ AD	$6 + 5$	$1/11$

- $C(A)$  is constant, unsmoothed count
- Probabilities sum up to 1:  $4/11 + 3/11 + 2/11 + 1/11 + 1/11 = 1$

## Additive smoothing: Bigrams: general case

- General formula for smoothed bigram Probabilities:

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i) + \alpha}{C(w_{i-1}) + \alpha|V|} \quad (6)$$

## Additive smoothing: Bigrams: general case

- General formula for smoothed bigram Probabilities:

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i) + \alpha}{C(w_{i-1}) + \alpha|V|} \quad (6)$$

- What is  $V$ ?

## Additive smoothing: Bigrams: general case

- General formula for smoothed bigram Probabilities:

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i) + \alpha}{C(w_{i-1}) + \alpha|V|} \quad (6)$$

- What is  $V$ ?
- $|V|$  = Number of bigram **types** starting with  $w_{i-1}$

## Additive smoothing: Bigrams: general case

- General formula for smoothed bigram Probabilities:

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i) + \alpha}{C(w_{i-1}) + \alpha|V|} \quad (6)$$

- What is  $V$ ?
- $|V|$  = Number of bigram **types** starting with  $w_{i-1}$

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i) + \alpha}{C(w_{i-1}) + \alpha|V_{(w_{i-1}, \bullet)}|} \quad (7)$$

## Additive smoothing: Bigrams: general case

- General formula for smoothed bigram Probabilities:

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i) + \alpha}{C(w_{i-1}) + \alpha|V|} \quad (6)$$

- What is  $V$ ?
- $|V|$  = Number of bigram **types** starting with  $w_{i-1}$

$$p(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i) + \alpha}{C(w_{i-1}) + \alpha|V_{(w_{i-1}, \bullet)}|} \quad (7)$$

- For n-grams of length  $n$ :

$$p(w_i|w_{i-1} : w_{i-n+1}) = \frac{C(w_{i-n+1} : w_i) + \alpha}{C(w_{i-n+1} : w_{i-1}) + \alpha|V_{(w_{i-n+1}:w_{i-1}, \bullet)}|} \quad (8)$$

## Additive smoothing: Bigrams: general case

- For n-grams of length  $n$ :

$$p(w_i | w_{i-1} : w_{i-n+1}) = \frac{C(w_{i-n+1} : w_i) + \alpha}{C(w_{i-n+1} : w_{i-1}) + \alpha |V_{(w_{i-n+1}:w_{i-1}, \bullet)}|} \quad (9)$$

- We already know the shared (train + test) vocabulary  $V$

## Additive smoothing: Bigrams: general case

- For n-grams of length  $n$ :

$$p(w_i | w_{i-1} : w_{i-n+1}) = \frac{C(w_{i-n+1} : w_i) + \alpha}{C(w_{i-n+1} : w_{i-1}) + \alpha |V_{(w_{i-n+1}:w_{i-1}, \bullet)}|} \quad (9)$$

- We already know the shared (train + test) vocabulary  $V$
- $V_{(A, \bullet)}$  is then  $AA, AB, AC, AD, AE, AF \Rightarrow |V_{(A, \bullet)}| = 6 = |V|$



## Additive smoothing: Bigrams: general case

- For n-grams of length  $n$ :

$$p(w_i | w_{i-1} : w_{i-n+1}) = \frac{C(w_{i-n+1} : w_i) + \alpha}{C(w_{i-n+1} : w_{i-1}) + \alpha |V_{(w_{i-n+1}:w_{i-1}, \bullet)}|} \quad (9)$$

- We already know the shared (train + test) vocabulary  $V$
- $V_{(A, \bullet)}$  is then  $AA, AB, AC, AD, AE, AF \Rightarrow |V_{(A, \bullet)}| = 6 = |V|$
- We find that the formula we found is identical to the one on the lecture slides!

$$p(w_i | w_{i-1} : w_{i-n+1}) = \frac{C(w_{i-n+1} : w_i) + \alpha}{C(w_{i-n+1} : w_{i-1}) + \alpha |V|} \quad (10)$$

## MARY HAD A LITTLE LAMB

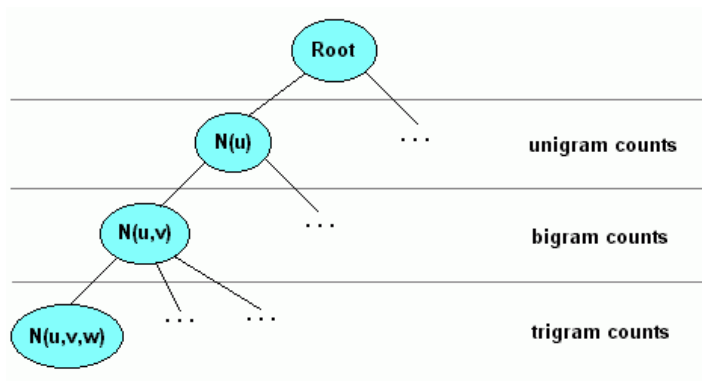
- Consider the bigram (LITTLE MARY)
- Consider the trigram (HAD A LAMB)

For a trigram  $p(w_3|w_2, w_1)$ , use probability of bigram  $P(w_3|w_2)$ , else back-off to unigram probability  $P(w_3)$ .

$$0.5 \cdot p(w_3|w_2, w_1) + 0.25 \cdot p(w_3|w_2) + 0.25 \cdot p(w_3)$$
$$0.5 \cdot p(\text{lamb}|\text{a, had}) + 0.25 \cdot p(\text{lamb}|\text{a}) + 0.25 \cdot p(\text{lamb})$$

Will be covered in more detail in further tutorials.

# Count Trees



# Assignment 5

- Exercise 1: OOV Words
- Exercise 2: Additive smoothing
- Exercise 3: Perplexity, infinite smoothing, interpolation
- Bonus: Other language models

# Resources

- ① UdS SNLP Class: <https://teaching.lsv.uni-saarland.de/snlp/>
- ② Additive smoothing: [https://en.wikipedia.org/wiki/Additive\\_smoothing](https://en.wikipedia.org/wiki/Additive_smoothing)
- ③ n-gram count trees: <http://ssli.ee.washington.edu/WS07/notes/ngrams.pdf>
- ④ n-gram models: <https://web.stanford.edu/~jurafsky/slp3/3.pdf>
- ⑤ Count-trees figure: <https://www.w3.org/TR/ngram-spec/>