Language modelling
with probabilities

Natural Language Processing
September 14, 2017
Mark Fishel
“All your base are belong to us”
“All your base are belong to us”

“The girl with the flowers is cute”

“The girl with the flowers are cute”
Language model

Task:
- estimate the quality/fluency/grammaticality of a natural language sentence or segment

Why?
- generate new sentences
- choose between several variants, picking the best sounding one
Language model

Word: \( w \)

Sentence: \( w = w_1 w_2 w_3 \ldots w_n \)
Can we use some grammatical checking rules to determine the fluency of $w$?
Can we use some grammatical checking rules to determine the fluency of \( w \)?

- **in theoretically -- yes,**

- **in practice:**
  - grammatical checking software is unreliable
  - and not available for many languages
  - and its output is often non-continuous
    - cannot be used in optimization
    - cannot lead to a better output out of many viable hypotheses
Can we use some grammatical checking rules to determine the fluency of $w$?

- in theoretically -- yes,
- in practice:
  - grammatical checking software is unreliable
  - and not available for many languages
  - and its output is often non-continuous
    - cannot be used in optimization
    - cannot lead to a better output out of many viable hypotheses

Instead we will try to calculate/model $p(w)$
Why probabilities?
Why probabilities?

To calculate the probability of “heads” or “tails” / predict the weather with a simplistic sunny/cloudy/rainy model/etc?
Why probabilities?

To calculate the probability of “heads” or “tails” / predict the weather with a simplistic sunny/cloudy/rainy model/etc?
Why probabilities?

To find the unknown distribution $P(X)$, given that the distributions $P(X \mid Y, Z)$ and $P(Z)$ are …
To find the unknown distribution $P(X)$, given that the distributions $P(X \mid Y, Z)$ and $P(Z)$ are …
Why probabilities?

To use them! E.g:
Why probabilities?

To use them! E.g:

- disease diagnosis

\[ p(\text{cold} \mid \text{symptoms: cough, runny nose, went outside without a hat and with wet feet.}) \]
Why probabilities?

To use them! E.g:

- **disease diagnosis**
  \[ p(\text{cold} \mid \text{symptoms: cough, runny nose, went outside without a hat and with wet feet}.) \]

- **insurance, assessing risks**
  \[ p(\text{the house will burn within a year} \mid \text{the owner likes playing with matches}) \]
Why probabilities?

To use them! E.g:

- **disease diagnosis**
  \[ \Pr(\text{cold} \mid \text{symptoms: cough, runny nose, went outside without a hat and with wet feet.}) \]

- **insurance, assessing risks**
  \[ \Pr(\text{the house will burn within a year} \mid \text{the owner likes playing with matches}) \]

- **parts of speech**
  \[ \Pr(\text{Det Noun Verb Det Noun} \mid \text{the old man the boat}) \]
Why probabilities?

To use them! E.g:

- **disease diagnosis**
  \( p(\text{cold} \mid \text{symptoms: cough, runny nose, went outside without a hat and with wet feet.}) \)

- **insurance, assessing risks**
  \( p(\text{the house will burn within a year} \mid \text{the owner likes playing with matches}) \)

- **parts of speech**
  \( p(\text{Det Noun Verb Det Noun} \mid \text{the old man the boat}) \)

- **speech recognition**
  \( p(\text{"wreck a nice beach"} \mid \text{)}) \), \( p(\text{"recognize speech"} \mid \text{)}) \)
Why probabilities?

To use them! E.g:

- **disease diagnosis**
  \[ p(\text{cold} \mid \text{symptoms: cough, runny nose, went outside without a hat and with wet feet}) \]

- **insurance, assessing risks**
  \[ p(\text{the house will burn within a year} \mid \text{the owner likes playing with matches}) \]

- **parts of speech**
  \[ p(\text{Det Noun Verb Det Noun} \mid \text{the old man the boat}) \]

- **speech recognition**
  \[ p(\text{"wreck a nice beach"} \mid \text{ }) , \ p(\text{"recognize speech"} \mid \text{ }) \]

- **machine translation**
  \[ p(\text{"All your base are belong to us."} \mid \text{ "君達の基地は、全てCATSがいただいた。"}) \]
Choosing the more probable output:

\[ p("All your base are belong to us." | "君達の...") < p("All of your bases now belong to us." | "君達の...") \]
Applying probabilities:

Choosing the more probable output:

\[
p("All your base are belong to us." | "君達の...") < \ p("All of your bases now belong to us." | "君達の...")
\]

\[
\text{translation} = \arg \max_{\text{hypothesis}} p(\text{hypothesis} | \text{source})
\]
Applying probabilities:

Choosing the more probable output:

\[ p("All your base are belong to us." \mid "君達の...") < p("All of your bases now belong to us." \mid "君達の...") \]

\[ \text{translation} = \text{argmax}_{hypothesis} p(hypothesis \mid source) \]

BUT HOW?
What is a probability

● to a programmer?
What is a probability

● to a programmer?
  ○ a number (a numeric variable)
What is a probability

- to a programmer?
  - a number (a numeric variable)
    - $p \in [0, 1]$, e.g:

$$p("psst") = 0.003$$
• does \( p = 0.003 \) mean the event is probable?
does $p = 0.003$ mean the event is probable?
  ○ if there are 3 possible outcomes?
does $p = 0.003$ mean the event is probable?
  ○ if there are 3 possible outcomes?
  ○ or it there is 10 000?
Probabilities

- does $p = 0.003$ mean the event is probable?
  - if there are 3 possible outcomes?
  - or if there is 10,000?
    - if all the other outcomes’ $p < 0.0001$?
● Does $p = 0.003$ mean the event is probable?
  ○ If there are 3 possible outcomes?
  ○ Or if there is 10,000?
    ■ If all the other outcomes’ $p < 0.0001$?
    ■ If some other outcome has a $p = 0.4$?
Probabilities

● does $p = 0.003$ mean the event is probable?
  ○ if there are 3 possible outcomes?
  ○ or it there is 10 000?
    ■ if all the other outcomes’ $p < 0.0001$?
    ■ if some other outcome has a $p = 0.4$?

● pointless question
  ○ instead of single probabilities we will look at probability distributions
Probability distribution

(set of probabilities for each possible outcome of the event:

\[ P(\text{word}) = (p_{\text{the}}, p_{\text{of}}, p, \ldots) = (0.05, 0.03, 0.05, \ldots) \]
Probability distribution

- set of probabilities for each possible outcome of the event:

\[ P(\text{word}) = (p_{\text{the}}, p_{\text{of}}, p, ...) = (0.05, 0.03, 0.05, ...) \]

- Probability: just a number \( \in [0, 1] \)

- Prob. distribution: just a list / map of numbers
  - summing to 1
Probability distribution

- set of probabilities for each possible outcome of the event:

\[ P(\text{word}) = (p_{\text{the}}, p_{\text{of}}, p, \ldots) = (0.05, 0.03, 0.05, \ldots) \]

- Probability: just a number \( \in [0, 1] \)

- Prob. distribution: just a list / map of numbers
  - summing to 1

+ nice framework around them
Uniform distribution: all probabilities are the same:
● event $x$ didn’t happen:

$$p(\text{not } x) = 1 - p(x)$$
Probabilistic framework

- conditional probability for events $x$ and $y$:

$$p(x \mid y) = \frac{p(x \cdot y)}{p(y)}$$

$$= \frac{\text{count}(x \text{ and } y)}{\text{count}(y)}$$
Probabilistic framework

- Conditional probability for events $x$ and $y$:
  \[
p(x \mid y) = \frac{p(x \cdot y)}{p(y)} = \frac{\text{count}(x \text{ and } y)}{\text{count}(y)}\]

- $x$ is independent of $y = \text{the distribution } P(x \mid y) \text{ stays the same regardless of } y$; basically:
  \[
p(x \mid y) = p(x)\]
• \( x \) depends on \( y \):

\[
p(x \cdot y) = p(x \mid y) \times p(y) \\
= p(x) \times p(y \mid x)
\]
Proportional framework

- \( x \) depends on \( y \):

\[
p(x \cdot y) = p(x | y) \times p(y)
= p(x) \times p(y | x)
\]

- \( x \) is independent of \( y \):

\[
p(x \cdot y) = p(x) \times p(y)
\]
Sentence probability

- Sentence = a complex event that consists of simpler events: words

\[ w = w_1 \ w_2 \ w_3 \ \ldots \ \ w_m \]

- Words are dependent/independent?
Sentence probability

- Sentence = a complex event that consists of simpler events: words

\[ w = w_1 \ w_2 \ w_3 \ \ldots \ \ w_m \]

- Words are dependent events:

\[ p(w) = p(w_1) \times p(w_2 \mid w_1) \times p(w_3 \mid w_1, w_2) \times \ldots \]

\[ \ldots p(w_m \mid w_1, w_2, \ldots, w_{m-1}) \]
Sentence probability

- cannot estimate it directly:

\[ p(w_1 \ w_2 \ w_3 \ \ldots \ w_m) = \frac{\text{count}(w_1 \ w_2 \ w_3 \ \ldots \ w_m)}{\#\text{sentences}} \]
Sentence probability

- cannot estimate it directly:

\[ p(w_1 w_2 w_3 \ldots w_m) = \frac{\text{count}(w_1 w_2 w_3 \ldots w_m)}{\# \text{sentences}} \]

- but word probabilities are doable:
  - take a huge text (millions/billions of words)
  - much more running words than word types (different words, so:

\[ p(w_i) = \frac{\text{count}(w_i)}{\text{text length}} \]
Sentence probability

\[ p(w) = p(w_1) \times p(w_2 \mid w_1) \times p(w_3 \mid w_1, w_2) \times \ldots \]
\[ \ldots p(w_m \mid w_1, w_2, \ldots, w_{m-1}) \]

\[ p(w_i) = \frac{\text{count}(w_i)}{\text{text length}} \]

\[ p(w_i \mid w_{i-1}) = \frac{\text{count}(w_{i-1}w_i)}{\text{count}(w_{i-1})} \]
To be, or not to be, that is the question:
Whether 'tis Nobler in the mind to suffer
The Slings and Arrows of outrageous Fortune,
Or to take Arms against a Sea of troubles,
And by opposing end them: to die, to sleep
No more; and by a sleep, to say we end
The Heart-ache, and the thousand Natural shocks
That Flesh is heir to? 'Tis a consummation
Devoutly to be wished. To die, to sleep,
To sleep, perchance to Dream; aye, there's the rub,
For in that sleep of death, what dreams may come,
When we have shuffled off this mortal coil,
Must give us pause. There's the respect
N-grams

To be, or not to be, that is the question:
Whether 'tis Nobler in the mind to suffer
The Slings and Arrows of outrageous Fortune,
Or to take Arms against a Sea of troubles,
And by opposing end them: to die, to sleep
No more; and by a sleep, to say we end
The Heart-ache, and the thousand Natural shocks
That Flesh is heir to? 'Tis a consummation
Devoutly to be wished. To die, to sleep,
To sleep, perchance to Dream; aye, there's the rub,
For in that sleep of death, what dreams may come,
When we have shuffled off this mortal coil,
Must give us pause. There's the respect

\[ p(\text{"be"}) = \frac{\text{count(\text{"be"})}}{\text{count(all words)}} = \frac{3}{136} = 0.022 \]
N-grams

To be, or not to be, that is the question:
Whether 'tis Nobler in the mind to suffer
The Slings and Arrows of outrageous Fortune,
Or to take Arms against a Sea of troubles,
And by opposing end them: to die, to sleep
No more; and by a sleep, to say we end
The Heart-ache, and the thousand Natural shocks
That Flesh is heir to? 'Tis a consummation
Devoutly to be wished. To die, to sleep,
To sleep, perchance to Dream; aye, there's the rub,
For in that sleep of death, what dreams may come,
When we have shuffled off this mortal coil,
Must give us pause. There's the respect

\[ p("be" \mid "to") = \frac{\text{count}("to be")}{\text{count}("to")} = \frac{3}{13} = 0.23 \]
N-grams

To be, or not to be, that is the question:
Whether 'tis Nobler in the mind to suffer
The Slings and Arrows of outrageous Fortune,
Or to take Arms against a Sea of troubles,
And by opposing end them: to die, to sleep
No more; and by a sleep, to say we end
The Heart-ache, and the thousand Natural shocks
That Flesh is heir to? 'Tis a consummation
Devoutly to be wished. To die, to sleep,
To sleep, perchance to Dream; aye, there's the rub,
For in that sleep of death, what dreams may come,
When we have shuffled off this mortal coil,
Must give us pause. There's the respect

\[ p("be" \mid "not to") = \frac{\text{count("not to be")}}{\text{count("not to")}} = \frac{1}{1} = 1 \]
Sentence probability

\[ p(w) = p(w_1) \times p(w_2 \mid w_1) \times p(w_3 \mid w_1, w_2) \times \ldots \]
\[ \ldots p(w_m \mid w_1, w_2, \ldots, w_{m-1}) \]

\[ p(w_i) = \frac{\text{count}(w_i)}{\text{text length}} \]

\[ p(w_i \mid w_{i-1}) = \frac{\text{count}(w_{i-1}w_i)}{\text{count}(w_{i-1})} \]
Lause tõenäosus

\[ p(w) = p(w_1) \times p(w_2 | w_1) \times p(w_3 | w_1, w_2) \times \ldots \times p(w_m | w_1, w_2, \ldots, w_{m-1}) \]

\[ p(w_i) = \frac{\text{count}(w_i)}{\text{teksti pikkus}} \]

\[ p(w_i | w_{i-1}) = \frac{\text{count}(w_{i-1}w_i)}{\text{count}(w_{i-1})} \]

\[ p(w_m | w_1, w_2, \ldots, w_{m-1}) = \frac{\text{count}(w_1w_2 \ldots w_m)}{\text{count}(w_1w_2 \ldots w_{m-1})} \]
Lause tõenäosus

\[
p(w) = p(w_1) \times p(w_2 \mid w_1) \times p(w_3 \mid w_1, w_2) \times \ldots \times p(w_m \mid w_1, w_2, \ldots, w_{m-1})
\]

\[
p(w_i) = \frac{\text{count}(w_i)}{\text{teksti pikkus}}
\]

\[
p(w_i \mid w_{i-1}) = \frac{\text{count}(w_{i-1}w_i)}{\text{count}(w_{i-1})}
\]

\[
p(w_m \mid w_1, w_2, \ldots, w_{m-1}) = \frac{\text{count}(w_1w_2\ldots w_m)}{\text{count}(w_1w_2\ldots w_{m-1})}
\]

AS NASTY AS THE PROBABILITY FOR THE WHOLE SENTENCE
Independence assumptions

- let’s assume that:
  - each word does not depend on all previous words, but on a fixed-length context:

\[ p(w_i \mid w_1, w_2, \ldots, w_{i-1}) \approx p(w_i \mid w_{i-2}, w_{i-1}) \]
N-grams

- sequence of N words = “n-gram”
  - trigram: \( w_{i-2} w_{i-1} w_i \)
  - bigram: \( w_{i-1} w_i \)
  - unigram (word): \( w_i \)

- N-gram probability:
  \[
p(w_i \mid w_{i-2} w_{i-1}) = \frac{\text{count}(w_{i-2} w_{i-1} w_i)}{\text{count}(w_{i-2} w_{i-1})}
\]
N-grams

• if
  \[ p(w_i \mid w_1, w_2, \ldots, w_{i-1}) \approx p(w_i \mid w_{i-2}, w_{i-1}) \]

• then
  \[ p(w) = \prod_i p(w_i \mid w_1, \ldots, w_{i-1}) \]
  \[ \approx \prod_i p(w_i \mid w_{i-2}, w_{i-1}) \]

• so we make an independence assumption between word \( i \) and words \( 1 \) to \( i-k \)
... for the securities
-1.19731 found </s>
-0.0002559 tab . </s>
-0.00122599 of sr _tab_
-1.44299 organisation )
-0.693387 perfent '
-2.16691 send alert
-3.44701 set yesterday
-2.52041 the ' execute
-1.26578 timeout and
-0.898618 verkrw field
-0.0507751 was extended ; for
...
Limitations of the approach:

- sentence $\neq$ sequence of words, sentence = hierarchy of words, phrases, etc:
  - “the girl with the flowers is pretty”
  - $\text{count}_{\text{OpenSubtitles}}$ (“the flowers is”) = 1
  - $\text{count}_{\text{OpenSubtitles}}$ (“the flowers are”) = 26
what if a particular n-gram is unseen?

- count = 0
- probability = ...
- sentence probability = ...
More language models

what if a particular n-gram is unseen?
• count = 0
• probability = 0
• sentence probability = 0
what if a particular n-gram is unseen?

- count = 0
- probability = 0
- sentence probability = 0

Counter-measures:
- back-off
- interpolation
if the n-gram \( y_{i-k} \ldots y_i \) is unseen:

- originally its probability would be 0:
  \[
p(y_i \mid y_{i-k} \ldots y_{i-1}) = 0
\]
if the n-gram "y_{i-k}...y_i" is unseen:

- originally its probability would be 0:
  \[ p(y_i \mid y_{i-k}...y_{i-1}) = 0 \]

- instead of 0 we will use the probability of a shorter n-gram:
  \[ p(y_i \mid y_{i-k+1}...y_{i-1}) \]
if the n-gram “$y_{i-k} \ldots y_i$” is unseen:

● originally its probability would be 0:

$$p(y_i \mid y_{i-k} \ldots y_{i-1}) = 0$$

● instead of 0 we will use the probability of a shorter n-gram:

$$p(y_i \mid y_{i-k+1} \ldots y_{i-1})$$

● with a smaller weight like 0.1
  ○ to express our smaller confidence in a shorter context n-gram
if the n-gram “$y_{i-k}...y_i$” is unseen:

● originally its probability would be 0:
  \[ p(y_i \mid y_{i-k}...y_{i-1}) = 0 \]

● instead of 0 we will use the probability of a shorter n-gram:
  \[ p(y_i \mid y_{i-k+1}...y_{i-1}) \]

● with a smaller weight like 0.1
  ○ to express our smaller confidence in a shorter context n-gram

● this can be done recursively
Interpolation

• back-off doesn’t get us a well defined probability distribution
Interpolation

- back-off doesn’t get us a well defined probability distribution
- a prettier alternative that works even better in practice -- interpolation: instead of

\[ p(y_i \mid y_{i-k} \ldots y_{i-1}) \]
Interpolation

- back-off doesn’t get us a well defined probability distribution
- a prettier alternative that works even better in practice -- interpolation: instead of
  \[ p(y_i \mid y_{i-k} \ldots y_{i-1}) \]
- we will use
  \[ \lambda_1 p(y_i \mid y_{i-k} \ldots y_{i-1}) + \lambda_2 p(y_i \mid y_{i-k+1} \ldots y_{i-1}) + \ldots \]
Interpolation

- back-off doesn’t get us a well defined probability distribution
- a prettier alternative that works even better in practice -- interpolation: instead of
  \[ p(y_i | y_{i-k} \ldots y_{i-1}) \]
  - we will use
    \[ \lambda_1 p(y_i | y_{i-k} \ldots y_{i-1}) + \lambda_2 p(y_i | y_{i-k+1} \ldots y_{i-1}) + \ldots \]
  - where \( 0 >> \lambda_1 >> \lambda_2 >> \ldots >> \lambda_k \) and \( \sum_i \lambda_i = 1 \)
N-gram size

- shorter n-grams in LMs:
  - the model will generalize well
  - less data is needed
  - but if the context is not long enough, it will be too weak (think of “the girl with the flowers”)
N-gram size

- **shorter n-grams in LMs:**
  - the model will generalize well
  - less data is needed
  - but if the context is not long enough, it will be too weak (think of “the girl with the flowers”)

- **longer n-grams in LMs:**
  - stronger model
  - much more data needed to avoid data sparsity
Advanced statistical LM

Syntactic n-grams

- $p(\text{sword} \mid \text{P-OBJ, with/PREP, killed/ROOT})$
  instead of $p(\text{sword} \mid \text{with a})$: 

  The viking killed the pirate with a sword
Software: language modelling

- KenLM
  - comes with Moses
- SRI LM
  - old “default”
- IRST LM
  - can handle huuuuuge texts and models without running out of RAM
Applications

- **Statistical machine translation**
  - includes a translation model and a *language model*
Applications

- Statistical machine translation
  - includes a translation model and a language model

- Speech recognition
  - includes an acoustic model and a language model!
Applications

● Statistical machine translation
  ○ includes a translation model and a language model

● Speech recognition
  ○ includes an acoustic model and a language model

● Text domain similarity
  ○ train a language model on news, another one on legal texts, another one on technical manuals
  ○ with a new text of unknown origin see which model scores higher on that text -- likely it belongs to that kind of texts = document classification
Summary

- N-gram = N words in a row
- N-gram probability = weird but useful thing
- “Probability of a sentence” actually makes sense! (sorry, Noam Chomsky)
- Next weeks: how to do it better / learn to predict probability distributions with a regression model (Kairit)
Join the dark side
Questions?