Statistical Machine Translation: Language (N-gram) Models

Jakub Waszczuk

Heinrich Heine Universität Düsseldorf

Winter Semester 2018/19

Today

The plan:

- Bayes' theorem
- Parameter estimation
- N-gram models

Outline

Bayes' theorem

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(B)} \cdot \frac{P(B)}{P(A)} = \frac{P(A|B) \cdot P(B)}{P(A)} \tag{1}$$

Bayes' theorem: example

Example

Suppose we know that we have a biased coin, with $p = 0.4$ (probability of getting heads). We throw the coin and get the following sequence:

$$H, T, T, T, H, H, T, H, T, H \tag{2}$$

Thus, instead of getting H the expected 4 times, we got it 5 times.

We can calculate the probability of such an event happening:

$$P(5) = \binom{10}{5} \times 0.4^5 \times 0.6^5 = 0.201$$

Suppose, however, that the coin is not biased. Then we get:

$$P(5) = \binom{10}{5} \times 0.5^5 \times 0.5^5 = 0.236$$

Bayes' theorem: example

Example

Suppose we know that we have a biased coin, with $p = 0.4$ (probability of getting heads). We throw the coin and get the following sequence:

$$H, T, T, T, H, H, T, H, T, H \tag{2}$$

Thus, instead of getting H the expected 4 times, we got it 5 times.

We can calculate the probability of such an event happening:

$$P(5) = \binom{10}{5} \times 0.4^5 \times 0.6^5 = 0.201$$

Suppose, however, that the coin is not biased. Then we get:

$$P(5) = \binom{10}{5} \times 0.5^5 \times 0.5^5 = 0.236$$

Question

Let's assume that we know that the coin is either biased with $p = 0.4$ or not biased at all ($p = 0.5$). What is the probability of the coin being biased if we throw 5 heads out of 10?

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 times in 10 trials

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 times in 10 trials

Result

Let $\alpha := P(B)$. Then:

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 times in 10 trials

Result

Let $\alpha := P(B)$. Then:

$$P(B|E) = 0.201 \cdot \frac{\alpha}{0.236 - 0.035\alpha}$$

Calculations

$$P(B|E) = P(E|B) \cdot \frac{P(B)}{P(E)} = 0.201 \cdot \frac{\alpha}{P(E)}$$

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 times in 10 trials

Result

Let $\alpha := P(B)$. Then:

$$P(B|E) = 0.201 \cdot \frac{\alpha}{0.236 - 0.035\alpha}$$

Calculations

$$P(B|E) = P(E|B) \cdot \frac{P(B)}{P(E)} = 0.201 \cdot \frac{\alpha}{P(E)}$$

$$P(E) = P(E \cap B) + P(E \cap N) = P(E|B) \cdot P(B) + P(E|N) \cdot P(N) =$$

$$0.201 \cdot \alpha + 0.236 \cdot (1 - \alpha) = 0.236 - 0.035\alpha$$

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 times in 10 trials

Result

Let $\alpha := P(B)$. Then:

$$P(B|E) = 0.201 \cdot \frac{\alpha}{0.236 - 0.035\alpha}$$

Prior

$P(B) = \alpha$ can be seen as a parameter representing our **prior** knowlege about the coin.

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 times in 10 trials

Result

Let $\alpha := P(B)$. Then:

$$P(B|E) = 0.201 \cdot \frac{\alpha}{0.236 - 0.035\alpha}$$

Prior

$P(B) = \alpha$ can be seen as a parameter representing our **prior** knowlege about the coin.

- if $\alpha = 0.5$, then $P(B|E) = 0.46$

Bayes' theorem: example

Events

- B – the coin is biased with $h = 0.4$
- N – the coin is not biased ($h = 0.5$)
- E – we get heads 5 times in 10 trials

Result

Let $\alpha := P(B)$. Then:

$$P(B|E) = 0.201 \cdot \frac{\alpha}{0.236 - 0.035\alpha}$$

Prior

$P(B) = \alpha$ can be seen as a parameter representing our **prior** knowlege about the coin.

- if $\alpha = 0.5$, then $P(B|E) = 0.46$
- if $\alpha = 0.6$, then $P(B|E) = 0.56$

Bayes' theorem

General interpretation

Let α represent model parameters and D the observed event (data!). Then:

$$P(\alpha|D) = \frac{P(D|\alpha) \cdot P(\alpha)}{P(D)} \tag{3}$$

where:

- $P(D|\alpha)$ – the probability of D given parameters α
- $P(D)$ – the probability of D regardless of parameters
- $P(\alpha)$ – the *prior*

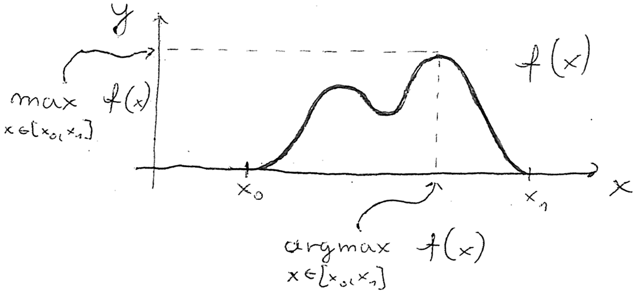
Outline

Argmax

Definition

$$\arg\max_{x \in X} f(x) = \{x : x \in X, \forall_{y \in X} f(x) \geq f(y)\} \qquad (4)$$

Example



Proposition

Let $C > 0$ be a constant. Then, $\arg\max_{x \in X}(Cf(x)) = \arg\max_{x \in X}(f(x))$.

Maximum a-posteriori (MAP) esimation

MAP method

Given an observed event D and the parameter space Θ, the MAP estimates θ_{MAP} are defined as:

$$\theta_{MAP} = \arg\max_{\theta \in \Theta} P(\theta|D) = \arg\max_{\theta \in \Theta} \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \arg\max_{\theta \in \Theta} P(D|\theta) \cdot P(\theta) \qquad (5)$$

Maximum a-posteriori (MAP) esimation

MAP method

Given an observed event D and the parameter space Θ, the MAP estimates θ_{MAP} are defined as:

$$\theta_{MAP} = \arg\max_{\theta \in \Theta} P(\theta|D) = \arg\max_{\theta \in \Theta} \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \arg\max_{\theta \in \Theta} P(D|\theta) \cdot P(\theta) \tag{5}$$

Example

We get back to the example with a coin which is either biased ($p = 0.4$) or not ($p = 0.5$):

- $\Theta = \{p = 0.4, p = 0.5\}$ (somewhat informally)
- D – the event of getting heads 5 times in 10 trials
- Let's assume uniform prior ($P(p = 0.4) = P(p = 0.5) = 0.5$)
- $P(D|p = 0.4) \cdot P(p = 0.4) = 0.201 \times 0.5 = 0.1005$
- $P(D|p = 0.5) \cdot P(p = 0.5) = 0.236 \times 0.5 = 0.118$
- $\arg\max_{\theta \in \Theta} P(D|\theta) \cdot P(\theta) = \{p = 0.5\}$

Likelihood function

Definition

Let θ represent model parameters and D an event. The *likelihood* of θ given D is defined as:

$$L_D(\theta) = P(D|\theta) \tag{6}$$

Warning

The likelihood is *not* a probability. In particular, the following does not necessarily hold:

$$\sum_{\theta \in \Theta} L_D(\theta) = 1 \tag{7}$$

where Θ is the space of possible parameter values.

For instance, in the example with the coin:

$$P(D|p = 0.4) + P(D|p = 0.5) = 0.201 + 0.236 = 0.437 \neq 1.$$

Maximum likelihood estimation (MLE)

MLE method

Given an observed event D and the parameter space Θ, the maximum likelihood estimates θ_{ML} are defined as:

$$\theta_{ML} = \arg\max_{\theta \in \Theta} L_D(\theta) = \arg\max_{\theta \in \Theta} P(D|\theta) \tag{8}$$

Maximum likelihood estimation (MLE)

MLE method

Given an observed event D and the parameter space Θ, the maximum likelihood estimates θ_{ML} are defined as:

$$\theta_{ML} = \arg\max_{\theta \in \Theta} L_D(\theta) = \arg\max_{\theta \in \Theta} P(D|\theta) \tag{8}$$

Example

We get back to the example with a coin which is either biased ($p = 0.4$) or not ($p = 0.5$):

- $\Theta = \{p = 0.4, p = 0.5\}$ (somewhat informally)
- D – the event of getting heads 5 times in 10 trials
- $P(D|p = 0.4) = 0.201$
- $P(D|p = 0.5) = 0.236$
- $\arg\max_{\theta \in \Theta} P(D|\theta) = \{p = 0.5\}$

MAP vs ML estimation

Proposition

Determining the ML estimates is equivalent with finding the MAP estimates assuming uniform prior (meaning $P(\theta_1) = P(\theta_2)$ for any two $\theta_1, \theta_2 \in \Theta$).

Note: uniform prior is not always the best choice, but it makes sense if we don't know anything about the parameters in the first place.

Proof

Using Bayes' theorem:

$$\arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \arg\max_{\theta} P(D|\theta) \cdot P(\theta) = \arg\max_{\theta} L_D(\theta) \cdot P(\theta)$$

Since we assume uniform prior, $P(\theta)$ is effectively a constant. Therefore:

$$\arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} L_D(\theta) \cdot P(\theta) = \arg\max_{\theta} L_D(\theta)$$

Maximum likelihood estimation in language modeling

Example

Suppose we have a corpus of 10^6 words in which the word *rabbit* occurs 60 times. What is the probability of *rabbit* occuring in a text?

Assumption

The number of occurrences of *rabbit* follows a binomial distribution with $p = P(\textit{rabbit})$.

MLE solution

- Let D be the observation made – in a text of 10^6 words *rabbit* occurs 60 times
- The likelihood of a particular value of the parameter p is:

$$L_D(p) = P(D|p) = \binom{10^6}{60} \times p^{60} \cdot (1-p)^{10^6-60}$$

- When maximizing $L_D(p)$, we can ignore the constant $\binom{10^6}{60}$:

$$\hat{p}_{ML} = \arg\max_p L_D(p) = \arg\max_p (p^{60} \cdot (1-p)^{10^6-60}) = \frac{60}{10^6}$$

Maximum likelihood estimation in language modeling

In general

Let's say that:

- We have a corpus of n words
- A word w occurs in this corpus k times
- We assume binomial distributions

Then, the ML estimates are:

$$\hat{P}_{ML}(w) = \frac{k}{n} \tag{9}$$

Good exercise – the proof (sketch below)

- Consider the binomial distribution for each word w separately (see the previous slide)
- Determine the value of the parameter $p = P(w)$ such that:

$$\frac{\partial L_D(p)}{\partial p} = 0 \tag{10}$$

By the way: we have just discovered the so-called *unigram* language model!

Outline

Language models in SMT

Motivation

We want an SMT system to:

- output words that are true to the original in meaning – **translation model**
- string the words together in fluent English sentences – **language model**

$$P(e|f) = P_{TM}(e|f) \cdot P_{LM}(e) \tag{11}$$

Examples

The language model supports difficult decisions about word order and grammaticality:

$$P_{LM}(\text{the house is small}) > P_{LM}(\text{small the is house})$$

and appropriate word translation (*Haus* → *house*, *home*, *building*?) in the given context:

$$P_{LM}(\text{I am going home}) > P_{LM}(\text{I am going house})$$

Language modeling: naive approach

Question

Let $w = w_1, w_2, \ldots, w_n$ be a sentence of length n. How can we estimate $P(w)$?

Naive approach

- Take a large collection of sentences T
- Assume the binomial distribution
- $\hat{P}_{ML}(w) = \frac{C(w)}{|T|}$, where $C(w)$ is the *count* – number of occurrences of w in T

Issue

- There are infinitely many sentences one can produce
- Most long sequences of words will not occur in T at all

Language modeling: scaling down

Idea

Break down the calculation of $P(w)$ into smaller steps:

- Assume a sequence of random variables W_1, W_2, W_3, ...
- Variable W_i represents the word on position i
- We introduce a special symbol $\rtimes$, which represents the end of sentence
- $P(w) = P(W_1 = w_1, W_2 = w_2, \ldots, W_n = w_n, W_{n+1} = \rtimes)$

Example

$$P(\text{I am going home}) = P(W_1 = \text{I}, W_2 = \text{am}, W_3 = \text{going}, W_4 = \text{home}, W_5 = \rtimes)$$

Language modeling: chain rule

Chain rule (extension of the product rule)

$$
\begin{aligned}
P(W_1 = w_1, W_2 = w_2, W_3 = w_3, \ldots, W_n = w_n) = \\
& P(W_1 = w_1) \\
& \times P(W_2 = w_2 | W_1 = w_1) \\
& \times P(W_3 = w_3 | W_1 = w_1, W_2 = w_2) \\
& \times \cdots \\
& \times P(W_n = w_n | W_1 = w_1, \ldots, W_{n-1} = w_{n-1})
\end{aligned}
$$

Language modeling: Markov assumptions

Markov property of order 0

Formally:

$$P(W_k = w_k | W_1 = w_1, \ldots, W_{k-1} = w_{k-1}) = P(W_k = w_k) \tag{12}$$

Alternatively, using $A \perp\!\!\!\perp B$ to denote independence of A and B:

$$W_k \perp\!\!\!\perp W_1, W_2, \ldots, W_{k-1} \tag{13}$$

In words:

- The probability of $W_k = w_k$ does *not* depend on the preceding words at all

Language modeling: Markov assumptions

Markov property of order 1

Formally:

$$P(W_k = w_k | W_1 = w_1, \ldots, W_{k-1} = w_{k-1}) = P(W_k = w_k | W_{k-1} = w_{k-1}) \quad (14)$$

Alternatively, using $A \perp\!\!\!\perp B \mid C$ to denote conditional independence of A and B given C:

$$W_k \perp\!\!\!\perp W_1, W_2, \ldots, W_{k-2} \mid W_{k-1} \quad (15)$$

In words:

- The probability of $W_k = w_k$ does *not* depend on the preceding words w_1, w_2, ..., provided that we know w_{k-1}

Language modeling: Markov assumptions

Markov property of order n

Formally:

$$P(W_k = w_k | W_1 = w_1, \ldots, W_{k-n} = w_{k-n}, \ldots, W_{k-1} = w_{k-1}) = P(W_k = w_k | W_{k-n} = w_{k-n}, \ldots, W_{k-1} = w_{k-1}) \tag{16}$$

Alternatively, using $A \perp\!\!\!\perp B \mid C$ to denote conditional independence of A and B given C:

$$W_k \perp\!\!\!\perp W_1, W_2, \ldots, W_{k-n-1} \mid W_{k-n}, W_{k-n+1}, \ldots, W_{k-1} \tag{17}$$

In words:

- The probability of $W_k = w_k$ does *not* depend on the preceding words w_1, w_2, ..., provided that we know w_{k-n}, w_{k-n+1}, ..., w_{k-1}

Language modeling: Markov chain

Markov chain

Let $W_1, W_2, \ldots$ be a sequence of random variables. We call it a *Markov chain* of order n if it satisfies the Markov property of order n.

Stationary Markov chain

We say that a Markov chain is *stationary* if the distributions of its variables do not depend on their position in the sequence. For instance, in the 1-order case:

$$P(W_i = x | W_{i-1} = y) = P(W_j = x | W_{j-1} = y) \tag{18}$$

for any two $i, j > 1$.

Language modeling: n-grams

Naming convention

In NLP/CL, a stationary Markov chain of order $n-1$ is also called an *n-gram* model.

- Markov chain of order 0 – **unigram** model
- Markov chain of order 1 – **bigram** model
- Markov chain of order 2 – **trigram** model

In general, the *n*-gram model captures relations between *n* adjacent words at a time.

Language modeling: n-gram parameters

Notation

Let V be a *vocabulary* (a set of words). Let also $y \in V$ and $x_1, \dots, x_n \in V^n$, where n is the order of a Markov chain. Thanks to the stationary property, we can simplify:

$$P(W_i = y | W_{i-n} = x_1, W_{i-n+1} = x_2, \dots, W_{i-1} = x_n) \tag{19}$$

as:

$$P(y | x_1, \dots, x_n) \tag{20}$$

because, regardless of the position i, $P(W_i = y | W_{i-n} = x_1, \dots, W_{i-1} = x_n)$ is the same.

Parameters

The parameter set of a stationary Markov chain of order n takes the following form:

$$\{P(y|x) : y \in V, \vec{x} \in V^n\} \tag{21}$$

where for each $\vec{x} \in V^n$:

$$\sum_{y \in V} P(y|\vec{x}) = 1 \tag{22}$$

Language modeling: example

Example

The following table represents the probabilities in a bigram model. The special symbol $\ltimes$ represents the only word that can be at the beginning of a sentence.

	$P(\cdot)$	$P(\cdot\|\ltimes)$	$P(\cdot\|\text{the})$	$P(\cdot\|\text{house})$	$P(\cdot\|\text{is})$	$P(\cdot\|\text{small})$	$P(\cdot\|\rtimes)$
$\ltimes$	1	0	0	0	0	0	0
the	0	0.4	0	0.1	0.4	0	0
house	0	0.1	0.5	0	0.2	0.5	0
is	0	0.3	0	0.5	0	0	0
small	0	0.2	0.5	0.1	0.4	0	0
$\rtimes$	0	0	0	0.3	0	0.5	1

What are the probabilities of the following sentences in this model?

- „*the house is small*"
- „*is the house small*"
- „*small the is house*"

Maximum likelihood estimation in n-grams

Number of occurrences

Let T be a training corpus. We define $C(w_1, \ldots, w_k)$ as the number of occurrences (*count*) of the sequence $w_1, \ldots, w_k$ in T.

Bigram ($n = 1$)

$$P_{ML}(y|x) = \frac{C(x, y)}{C(x)} \tag{23}$$

Example

Let $T = (\ltimes, a, a, a, b, b, b, b, a, a, a, a, \rtimes)$. Then:

- $P_{ML}(a|a) =$
- $P_{ML}(\rtimes|a) =$
- $P_{ML}(a|b) =$

Maximum likelihood estimation in n-grams

Number of occurrences

Let T be a training corpus. We define $C(w_1, \ldots, w_k)$ as the number of occurrences (*count*) of the sequence $w_1, \ldots, w_k$ in T.

Bigram ($n = 1$)

$$P_{ML}(y|x) = \frac{C(x, y)}{C(x)} \tag{23}$$

Example

Let $T = (\ltimes, a, a, a, b, b, b, b, a, a, a, a, \rtimes)$. Then:

- $P_{ML}(a|a) = \frac{5}{7}$
- $P_{ML}(\rtimes|a) =$
- $P_{ML}(a|b) =$

Maximum likelihood estimation in n-grams

Number of occurrences

Let T be a training corpus. We define $C(w_1, \ldots, w_k)$ as the number of occurrences (*count*) of the sequence $w_1, \ldots, w_k$ in T.

Bigram ($n = 1$)

$$P_{ML}(y|x) = \frac{C(x, y)}{C(x)} \tag{23}$$

Example

Let $T = (\ltimes, a, a, a, b, b, b, b, a, a, a, a, \rtimes)$. Then:

- $P_{ML}(a|a) = \frac{5}{7}$
- $P_{ML}(\rtimes|a) = \frac{1}{7}$
- $P_{ML}(a|b) =$

Maximum likelihood estimation in n-grams

Number of occurrences

Let T be a training corpus. We define $C(w_1, \ldots, w_k)$ as the number of occurrences (*count*) of the sequence $w_1, \ldots, w_k$ in T.

Bigram ($n = 1$)

$$P_{ML}(y|x) = \frac{C(x, y)}{C(x)} \tag{23}$$

Example

Let $T = (\ltimes, a, a, a, b, b, b, b, a, a, a, a, \rtimes)$. Then:

- $P_{ML}(a|a) = \frac{5}{7}$
- $P_{ML}(\rtimes|a) = \frac{1}{7}$
- $P_{ML}(a|b) = \frac{1}{4}$

Maximum likelihood estimation in n-grams

In general

$$P_{ML}(y|x_1, \ldots, x_n) = \frac{C(x_1, \ldots, x_n, y)}{C(x_1, \ldots, x_n)} \tag{24}$$

Example with $n = 2$ (trigram model)

Let $C = (\ltimes, \ltimes, a, a, a, b, b, b, b, a, a, a, a, \rtimes, \rtimes)$. Then:

- $P_{ML}(a|a, a) =$
- $P_{ML}(b|a, a) =$

Issue

The higher the value of n:

- The smaller the number of occurrences of $(x_1, \ldots, x_n, y)$ in training data
- The higher the number of parameters of the model (training data size stays the same)

Result: the estimates are not reliable.

Maximum likelihood estimation in n-grams

In general

$$P_{ML}(y|x_1,\ldots,x_n) = \frac{C(x_1,\ldots,x_n,y)}{C(x_1,\ldots,x_n)} \qquad (24)$$

Example with $n = 2$ (trigram model)

Let $C = (\ltimes, \ltimes, a, a, a, b, b, b, b, a, a, a, a, \rtimes, \rtimes)$. Then:

- $P_{ML}(a|a,a) = \frac{3}{5}$
- $P_{ML}(b|a,a) =$

Issue

The higher the value of n:

- The smaller the number of occurrences of $(x_1,\ldots,x_n,y)$ in training data
- The higher the number of parameters of the model (training data size stays the same)

Result: the estimates are not reliable.

Maximum likelihood estimation in n-grams

In general

$$P_{ML}(y|x_1, \ldots, x_n) = \frac{C(x_1, \ldots, x_n, y)}{C(x_1, \ldots, x_n)} \tag{24}$$

Example with $n = 2$ (trigram model)

Let $C = (\ltimes, \ltimes, a, a, a, b, b, b, b, a, a, a, a, \rtimes, \rtimes)$. Then:

- $P_{ML}(a|a, a) = \frac{3}{5}$
- $P_{ML}(b|a, a) = \frac{1}{5}$

Issue

The higher the value of n:

- The smaller the number of occurrences of $(x_1, \ldots, x_n, y)$ in training data
- The higher the number of parameters of the model (training data size stays the same)

Result: the estimates are not reliable.

Language modeling: Markov chain

How to choose the order?

It's a trade-off:

- n-th order Markov property *implies* $(n+1)$-th order Markov property
- Too large n leads to sparseness issues (not enough data to estimate reliable statistics)
- Too small n is not realistic („***is the house*** *you've been renting for the last two years* ***small****?*")
- Interpolation can be used to combine several models of different orders

Why not syntax-based models?

It's a question of complexity (both practical and conceptual):

- syntax-based models can be more accurate, but
- are more difficult to integrate with translation models