Computational Linguistics

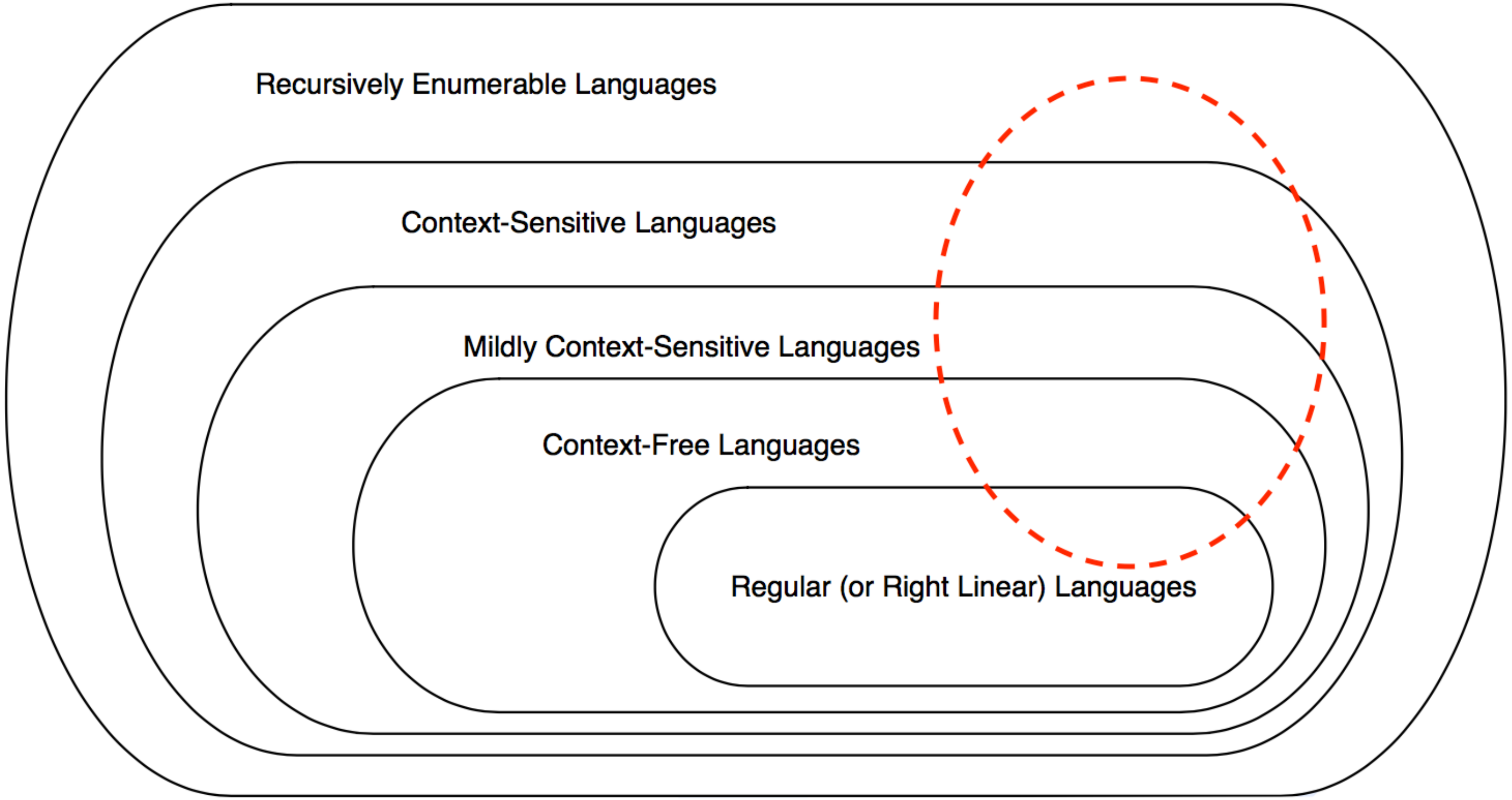
4. N-gram Language Models

Xiaojing Bai

Tsinghua University

https://bxjthu.github.io/CompLing

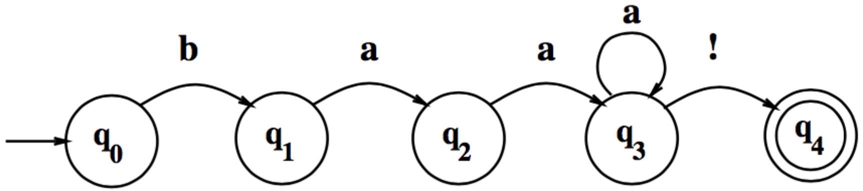
Recap: Chomsky hierarchy


Recursively Enumerable Languages
Context-Sensitive Languages
Mildly Context-Sensitive Languages
Context-Free Languages
Regular (or Right Linear) Languages

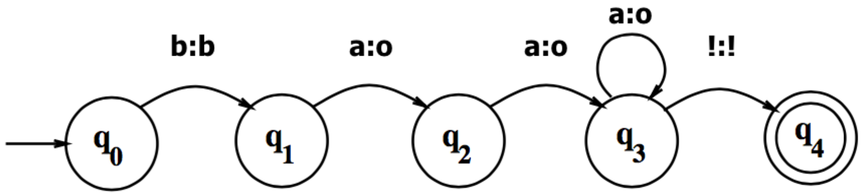
Recap: FSA vs. FST

Recognizer (acceptor) vs. generator

A recognizer takes a string as input and outputs *accept* if the string is in a string of the language, and *reject* if it is not.



A generator takes a string as input and outputs a new string.



Recap: a formal definition of FST

Q: a finite set of N **states**

$\{q_0, q_1, q_2, \ldots q_{N-1}\}$

Σ: a finite **input alphabet** of symbols

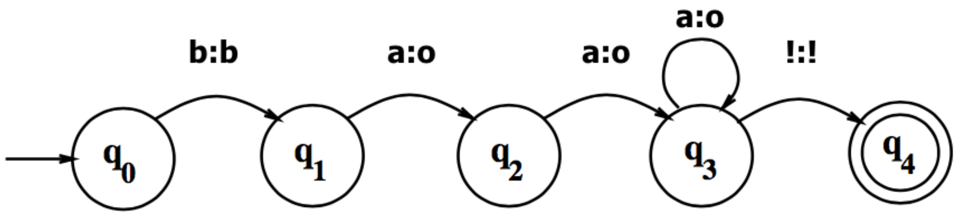
Δ: a finite **output alphabet** of symbols

q_0: the **start state**

F: the set of **final states**, $F \subseteq Q$

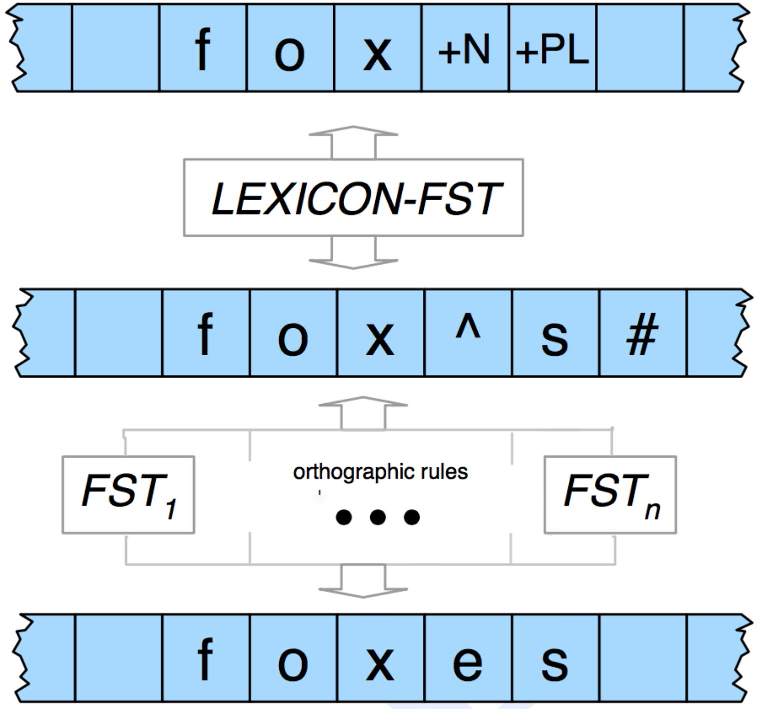
$\delta(q, i)$: the **transition function**. Given a state $q \in Q$ and an input symbol $i \in \Sigma$, $\delta(q, i)$ returns a set of new states, each state $q' \in Q$.

$\sigma(q, i)$: the **output function**. Given a state $q \in Q$ and an input symbol $i \in \Sigma$, $\sigma(q, i)$ returns a set of output symbols, each symbol $o \in \Delta$.



Recap: morphological parsing

- **Lexion**: a list of the stems and affixes of a language
- **morphotactics**: a model to show how the stems and affixes can fit together
- **Orthographic rules**: a model to show the changes that occur in a word



Recap: questions

Difficulties of normalization and morphological parsing in Chinese?

这个门的把手坏了好几天了。
你把手抬高一点儿。

人身上哪怕有一点小病痛，都会影响到工作学习。
这种病痛起来真要人命。

报名选手持本人学生证，于比赛当日指定时段到达赛场。
把“手持”改成“携带”?

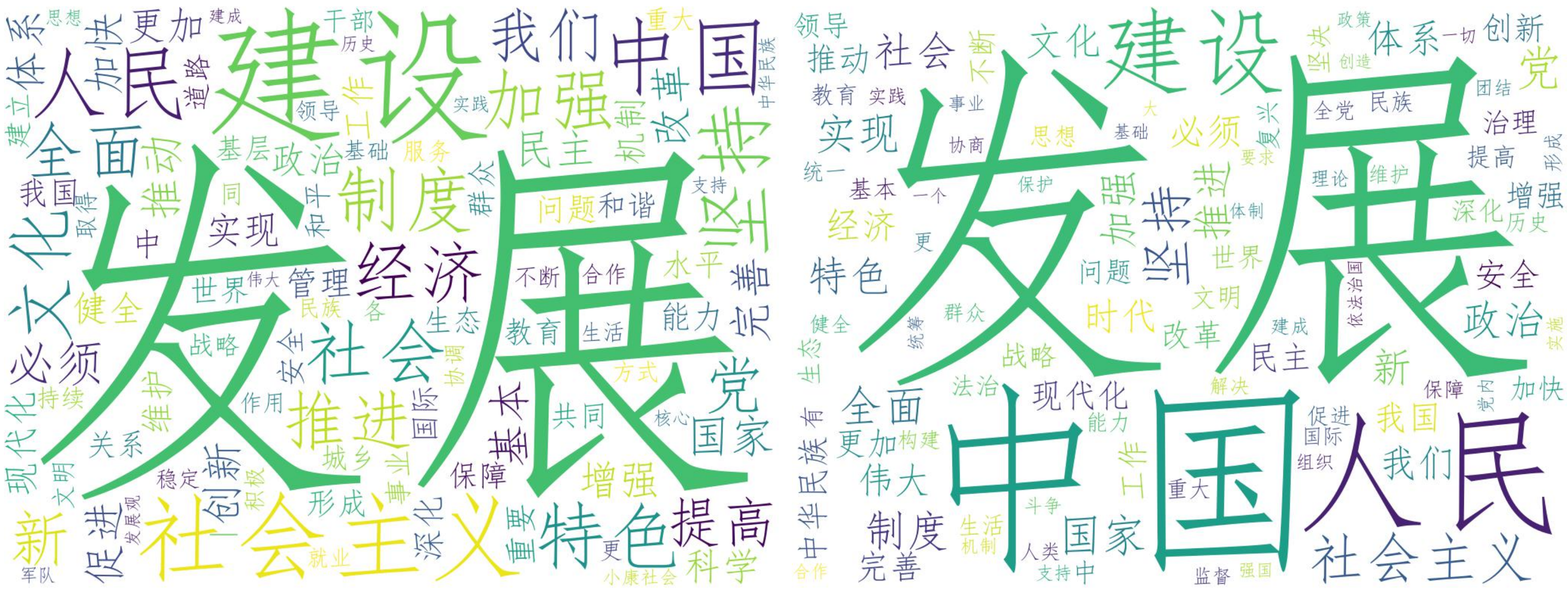
Recap: questions

Difficulties of normalization and morphological parsing in Chinese?

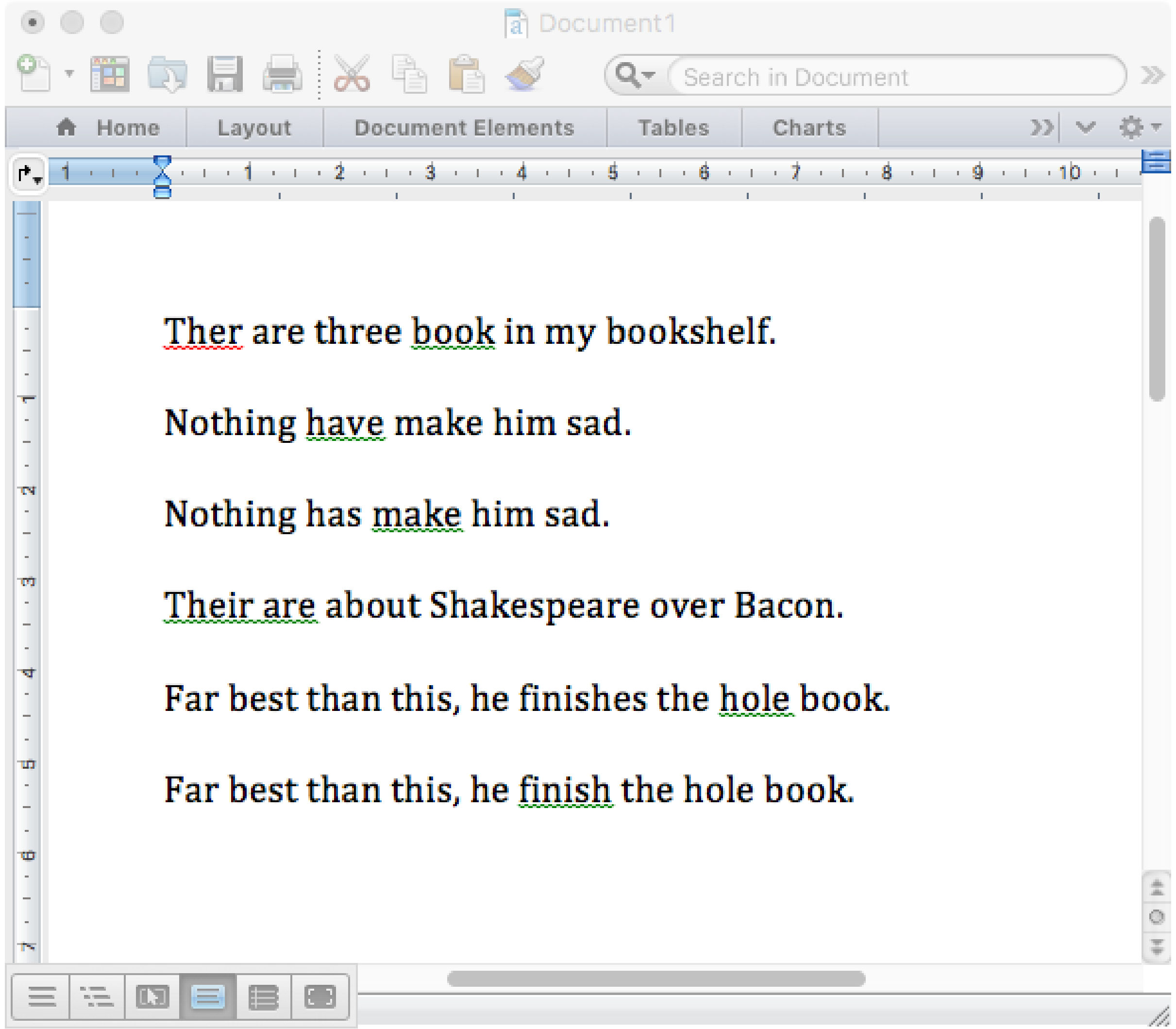
- Ambiguities
- Unknown words
- What is a WORD?

Recap: questions

How might morphological parsing work for us?



[A Computational Linguistic Analysis of Party Congress Reports](#) by Li Yimeng



Ther are three book in my bookshelf.

Nothing have make him sad.

Nothing has make him sad.

Their are about Shakespeare over Bacon.

Far best than this, he finishes the hole book.

Far best than this, he finish the hole book.

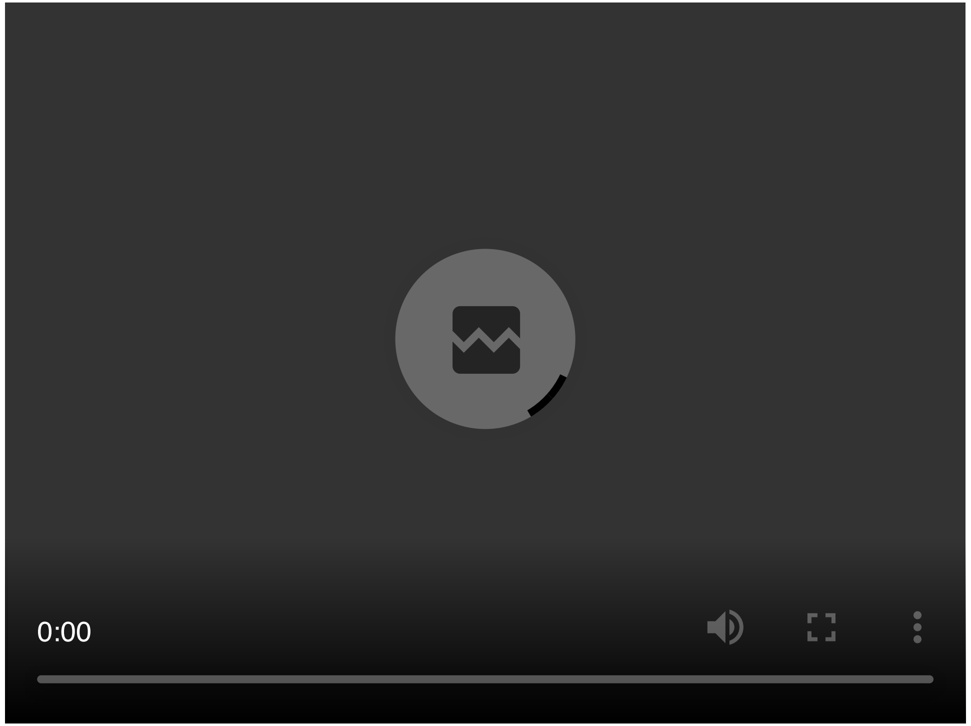
We will have a quiz next ...

At the end of this session you will

- understand how n-grams can model a language;
- learn how to use corpus data to compute the probabilities of n-grams;
- understand how n-grams may help to develop NLP applications;
- learn how to build n-gram models with Python.

Handwriting recognition

I have the gub!



Bank Teller #1:
Does this look like "gub" or "gun"?

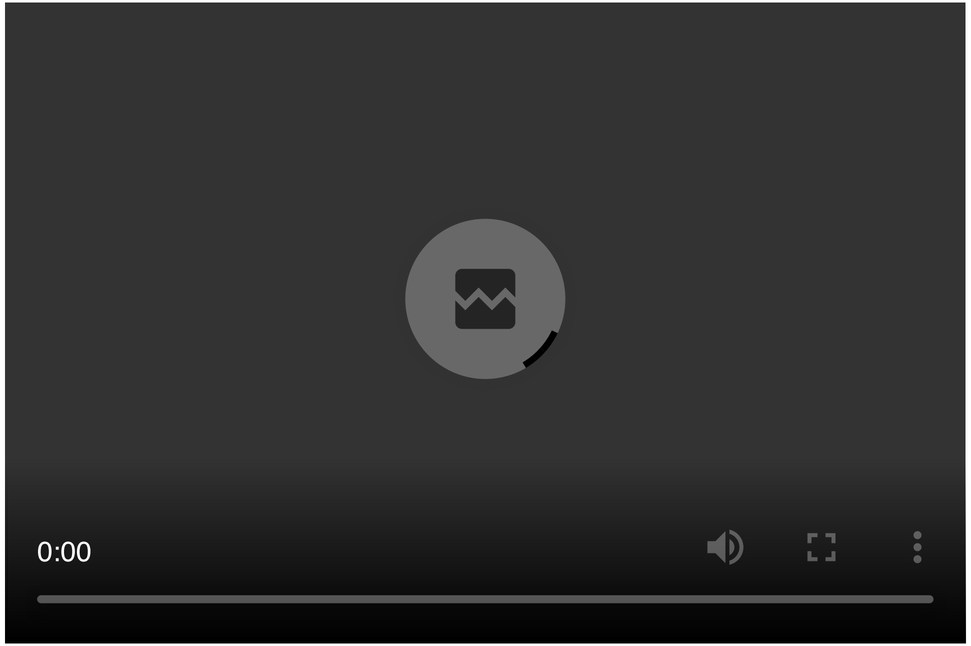
Bank Teller #2:
Gun. See? But what does "abt" mean?

Virgil:
It's "act". A-C-T. Act natural. Please put fifty thousand dollars into this bag and act natural.

Bank Teller #1:
Oh, I see. This is a holdup?

Take the Money and Run **(1969)**

Speech recognition



Augmentative communication

Language generation

"You are uniformly charming!" cried he, with a smile of associating and now and then I bowed and they perceived a chaise and four to wish for.

A random sentence generated from
a **Jane Austen trigram model**

这菜不错！
……

这菜不错！
不咸！

N-grams for 这菜不错！不咸！

- unigram: 这 菜 不 错 ！ 不 咸 ！
- bigram: 这菜 菜不 不错 错！ ！不 不咸 咸！
- trigram: 这菜不 菜不错 不错！ 错！不 ！不咸 不咸！
- 4-gram: 这菜不错 菜不错！ 不错！不 错！不咸 ！不咸！
- 5-gram: 这菜不错！ 菜不错！不 不错！不咸 错！不咸！
- ...
- n-gram

Conditional probability

Conditional probability is the probability of event A given that the occurrence of event B, written as $P(A|B)$.

Joint probability

Joint probability is the probability of two events in conjunction, i.e. the probability of both events together, written as $P(A \cap B)$ or $P(A, B)$.

If A and B are independent, i.e. knowing the outcome of A does not change the probability of B, or $P(B|A) = P(B)$, then $P(A \cap B) = P(A)P(B)$.

If A and B are not independent, e.g. knowing the outcome of A does change the probability of B, or $P(B|A) \neq P(B)$, then $P(A \cap B) = P(A)P(B|A)$.

N-grams as a model of language

Basic problem:
Is this a probable sequence of words in the language and how probable is it?

N-grams as a model of language

Basic problem:
Is this a probable sequence of words in the language and how probable is it?

Using corpus data for probabilities

```
<s>welcome home</s>

<s>welcome back</s>

<s>welcome home</s>

<s>you are a welcome sight</s>

<s>what a welcome</s>
```

A toy corpus

w_{n-1}	w_n	count	probability
<s>	welcome	3	0.60
<s>	what	1	0.20
<s>	you	1	0.20
a	welcome	2	1.00
are	a	1	1.00
back	</s>	1	1.00
home	</s>	2	1.00
sight	</s>	1	1.00
welcome	home	2	0.40
welcome	back	1	0.20
welcome	sight	1	0.20
welcome	</s>	1	0.20
what	a	1	1.00
you	are	1	1.00

The bigram counts and probabilities for the toy corpus

Probabilities of bigrams

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Probabilities of sequences

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

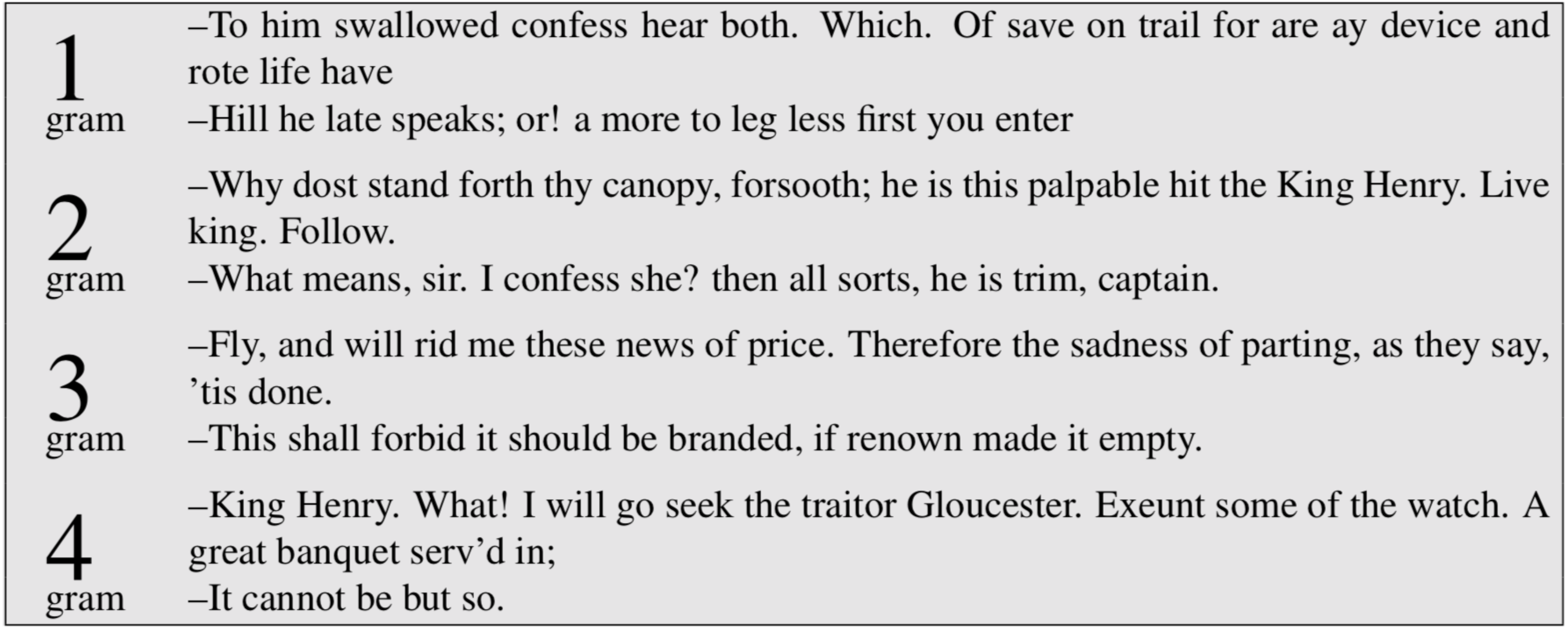
Why approximately equal to?

$$\begin{aligned} P(w_1^n) &= P(w_1)P(w_2|w_1)P(w_3|w_1^2)\cdots P(w_n|w_1^{n-1}) \\ &= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \end{aligned}$$

w_{n-1}	w_n	count	probability
<s>	welcome	3	0.60
<s>	what	1	0.20
<s>	you	1	0.20
a	welcome	2	1.00
are	a	1	1.00
back	</s>	1	1.00
home	</s>	2	1.00
sight	</s>	1	1.00
welcome	home	2	0.40
welcome	back	1	0.20
welcome	sight	1	0.20
welcome	</s>	1	0.20
what	a	1	1.00
you	are	1	1.00

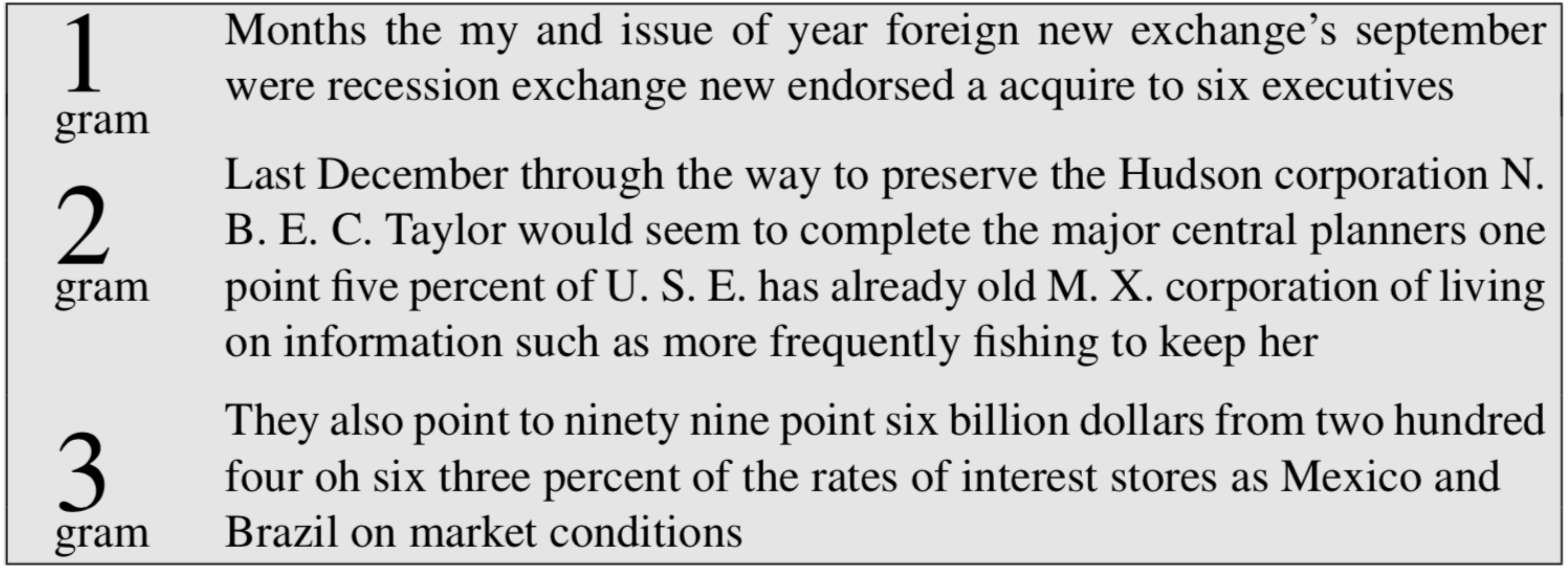
The bigram counts and probabilities for the toy corpus

The increasing power of higher-order n-grams



1 gram	–To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have –Hill he late speaks; or! a more to leg less first you enter
2 gram	–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. –What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, ’tis done. –This shall forbid it should be branded, if renown made it empty.
4 gram	–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv’d in; –It cannot be but so.

N-grams and their dependence on their training sets



1 gram	Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
2 gram	Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
3 gram	They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

N-grams for Machine translation

他	向	记者	介绍了	主要	内容
He	to	reporters	introduced	main	content

1. he introduced reporters to the main contents of the statement
2. he briefed to reporters the main contents of the statement
3. he briefed reporters on the main contents of the statement

Advanced issues and further readings

- J+M_3.2: Evaluating Language Models (required)
- J+M_3.3: Generalizations and Zeros (required)
- J+M_3.3: Smoothing (optional)

At the end of this session you will

- understand how n-grams can model a language;
- learn how to use corpus data to compute the probabilities of n-grams;
- understand how n-grams may help to develop NLP applications;
- learn how to build n-gram models with Python.

Homework

- Review (Quiz 4 on Oct. 24, 2018)
  - J+M_3 (3.1-3.3)
  - J+M_2
  - J+M_second_edition_3 (3.1)]
- Read
  - Mathematical foundations
  - J+M_8 (8.1-8.4)
- Read and practice
  - n-gram.py by Qing Lyu

Next session

Hidden Markov Models and Part-Of-Speech Tagging