Part of Speech Tagging

CSCI 1460: Computational Linguistics
Lecture 16

Ellie Pavlick
Fall 2022

Announcements

- Assignment 4 extension
- Final Projects: More info soon!
- Next week: Virtual Lectures
- Final Lectures: Your suggestions welcome!

Topics

- Follow up: Multlingual LMs and Zero-Shot
- Part of Speech Tags
- Hidden Markov Models
- Conditional Random Fields

Topics

- **Follow up: Multlingual LMs and Zero-Shot**
- Part of Speech Tags
- Hidden Markov Models
- Conditional Random Fields

Cross Lingual Transfer

- Goal: Train only on one language (English), but work in all languages
- Intuition: If the model learns a good representation of the languages under the hood (“interlingua”), it should be able to map the training it receives in one language to any other language

Multilingual Language Models

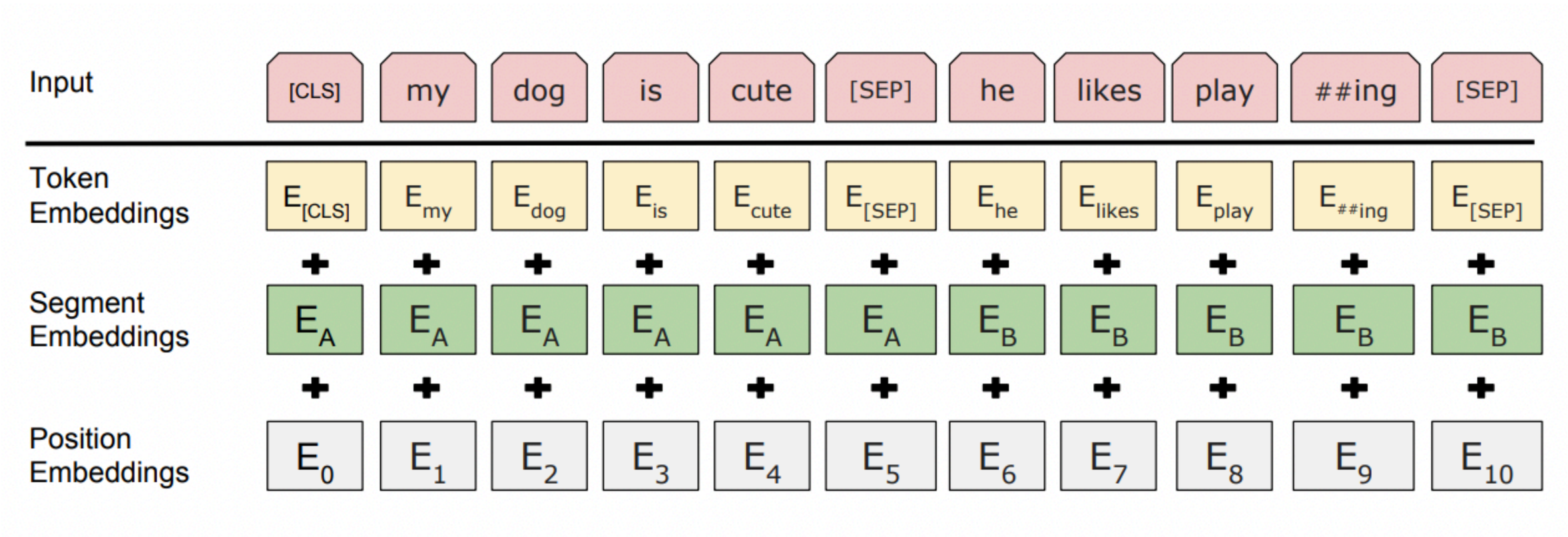
Cross-lingual Language Model Pretraining

Guillaume Lample*
Facebook AI Research
Sorbonne Universités
glample@fb.com

Alexis Conneau*
Facebook AI Research
Université Le Mans
aconneau@fb.com

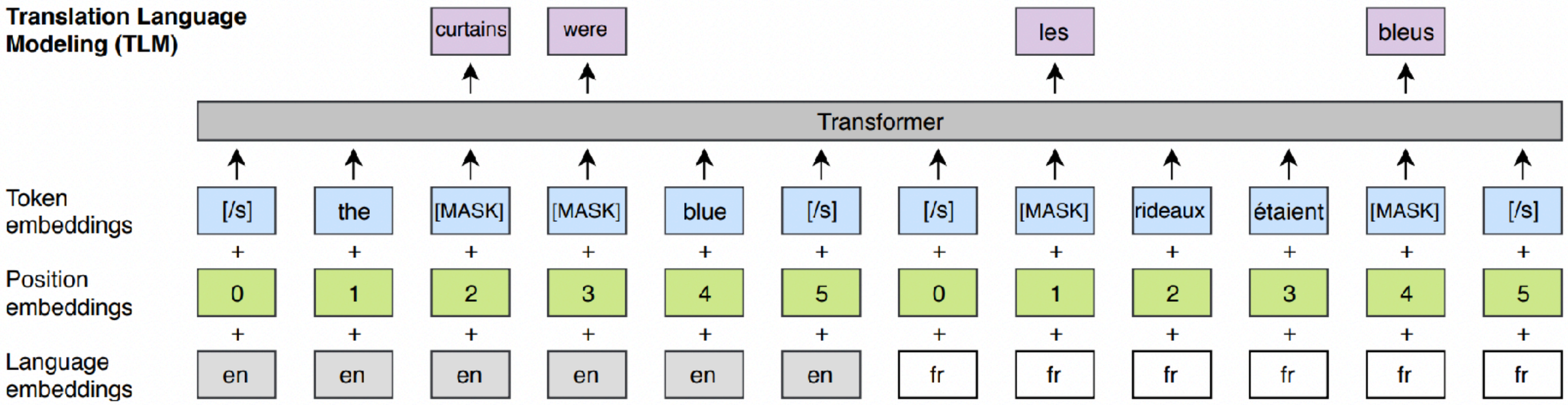
Multilingual Language Models

Recap: BERT



Multilingual Language Models

Multilingual MLM



Multilingual Language Models

- Common multilingual models: mBERT, XLM-RoBERTa
- Just recently: mGPT, BLOOM
- Just monolingual data
- Shared wordpiece vocabulary
- No explicit mechanisms to encourage translation etc.

Unsupervised Cross-lingual Representation Learning at Scale

Alexis Conneau* Kartikay Khandelwal*

Naman Goyal Vishrav Chaudhary Guillaume Wenzek Francisco Guzmán

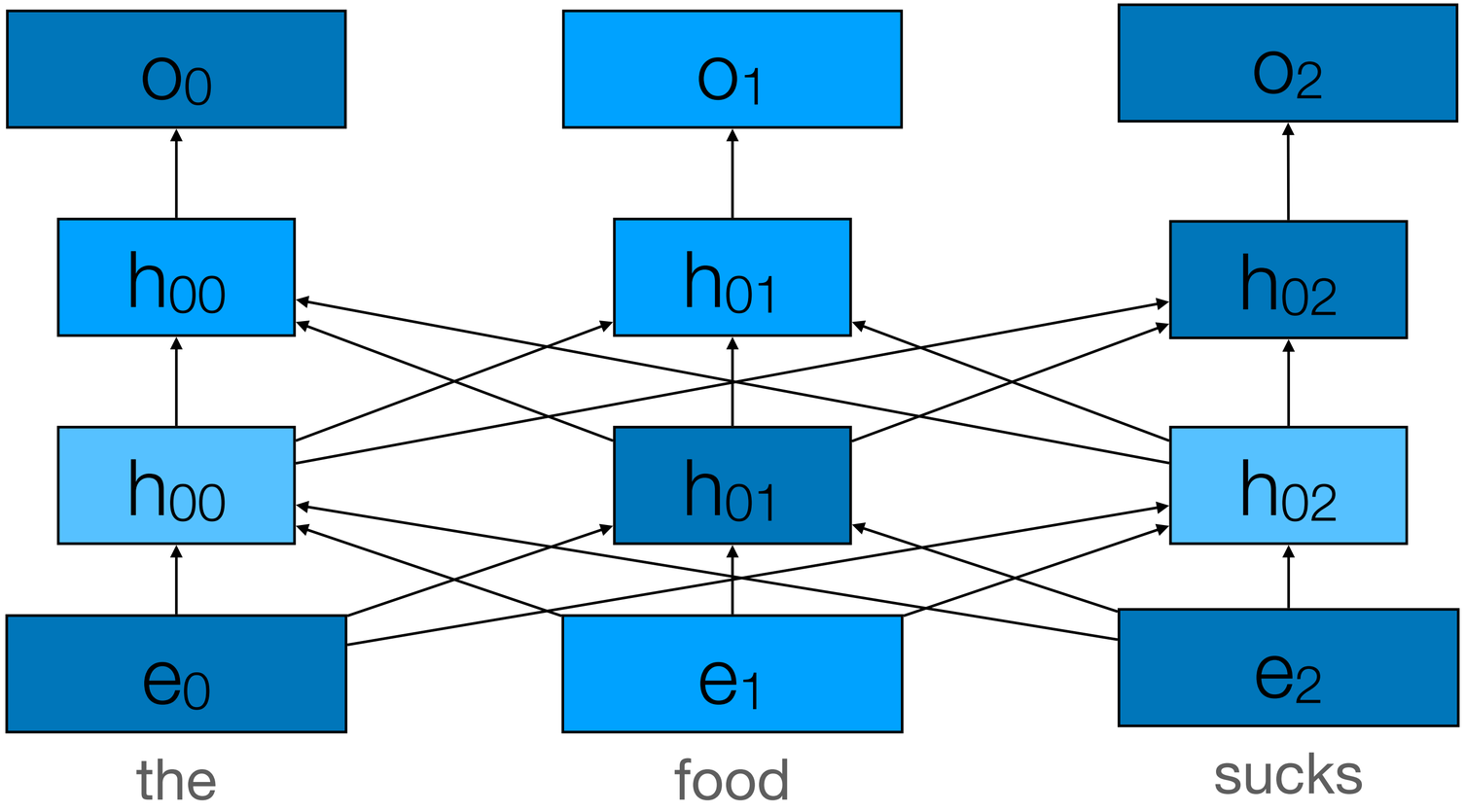
Edouard Grave Myle Ott Luke Zettlemoyer Veselin Stoyanov

https://github.com/google-research/bert/blob/master/multilingual.md

Multilingual Language Models

Recap: Pretraining and Finetuning

Task: Sentiment Analysis
Input: “the food sucks”
Expected: Negative

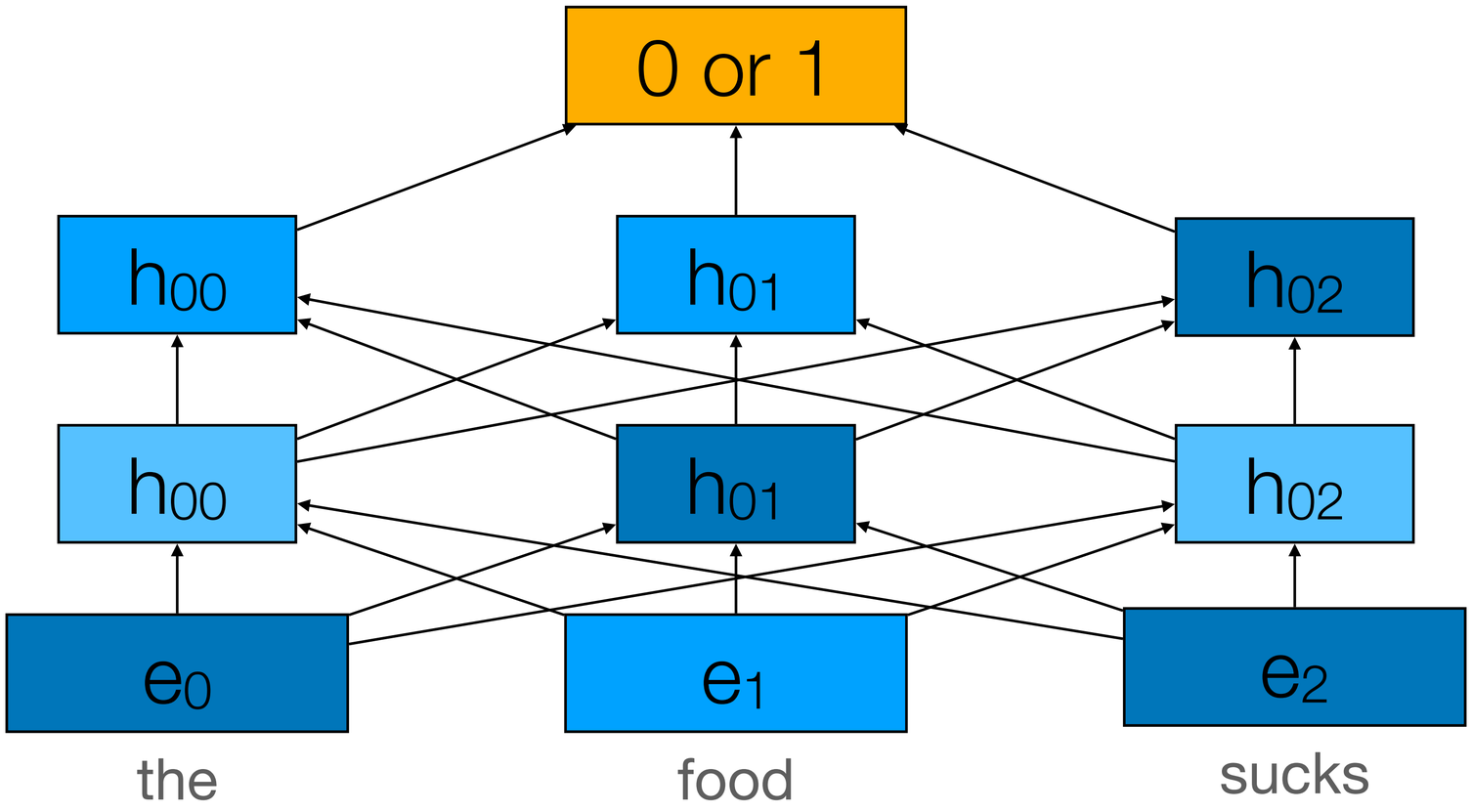


Multilingual Language Models

Recap: Pretraining and Finetuning

Task: Sentiment Analysis
Input: “the food sucks”
Expected: Negative

Adjust output layers to reflect target task

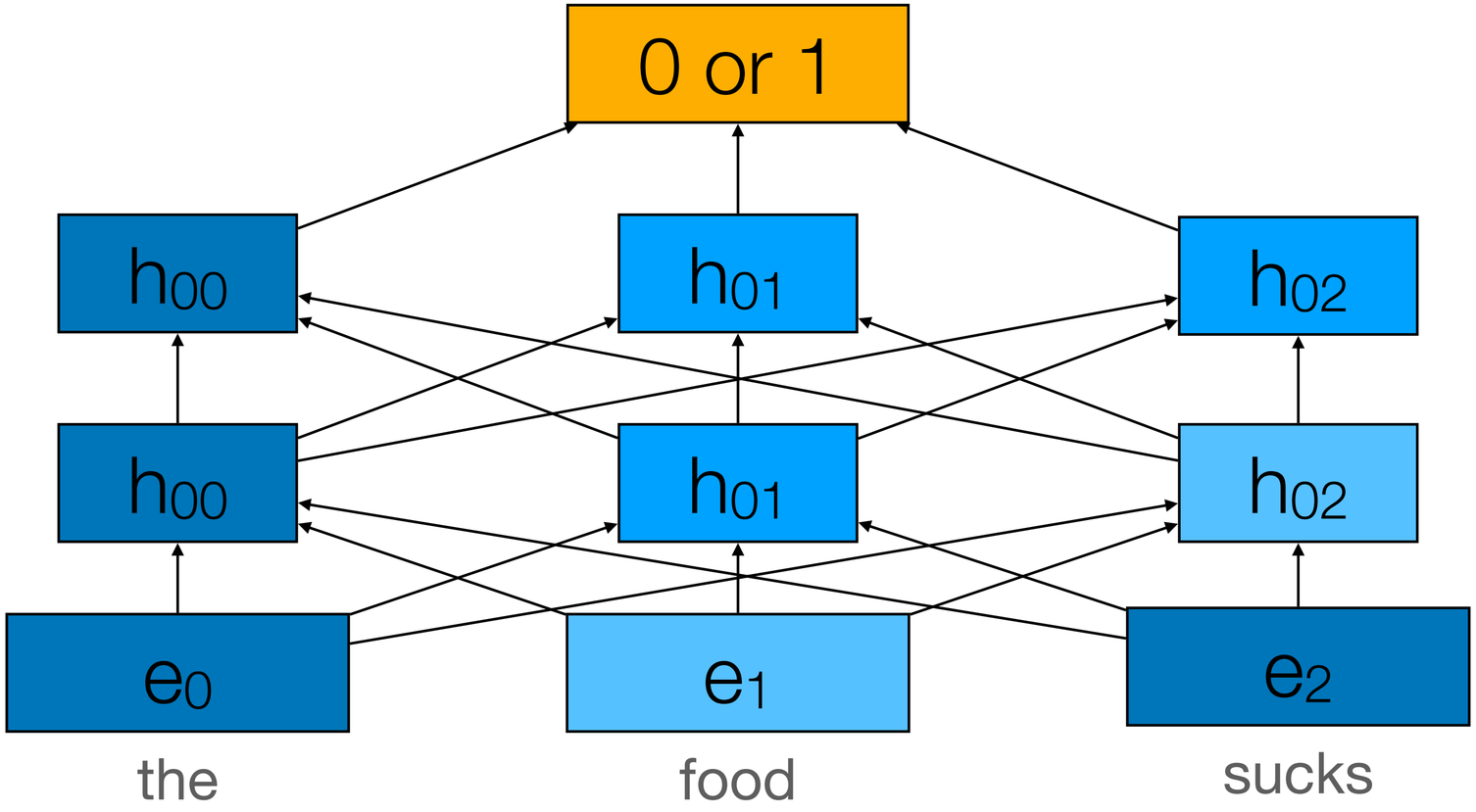


Multilingual Language Models

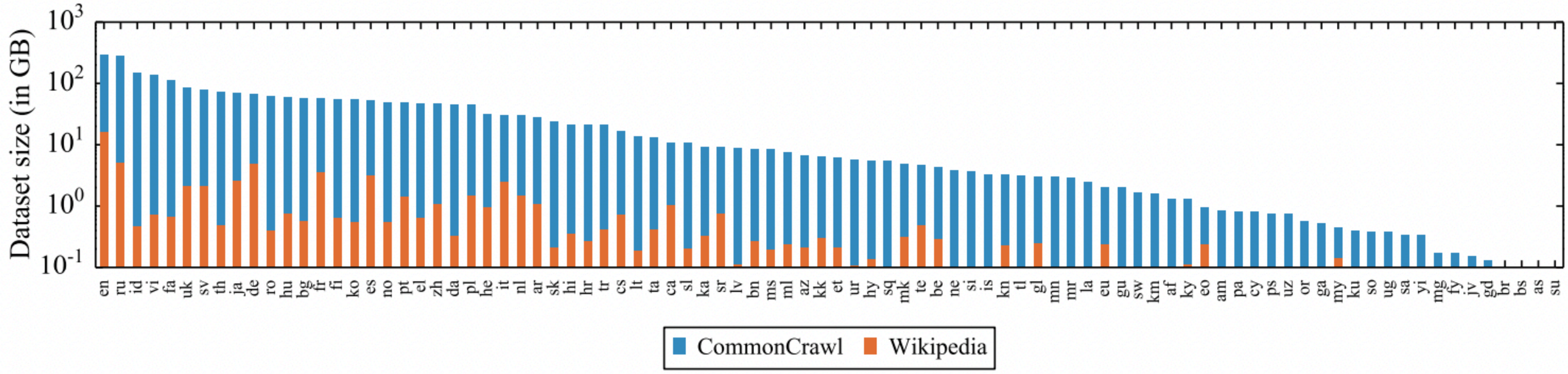
Recap: Pretraining and Finetuning

Task: Sentiment Analysis
Input: “the food sucks”
Expected: Negative

Keep training network, backproping through everything as needed.



Multilingual Language Models



Multilingual Language Models

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{\text{Base}}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

Multilingual Language Models

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{\text{Base}}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

Multilingual Language Models

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{\text{Base}}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

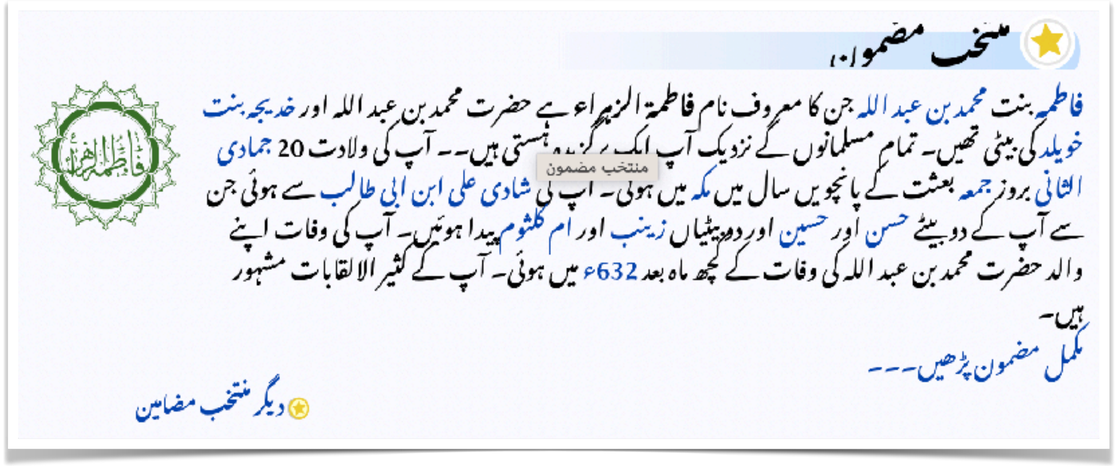
Cross Lingual Transfer

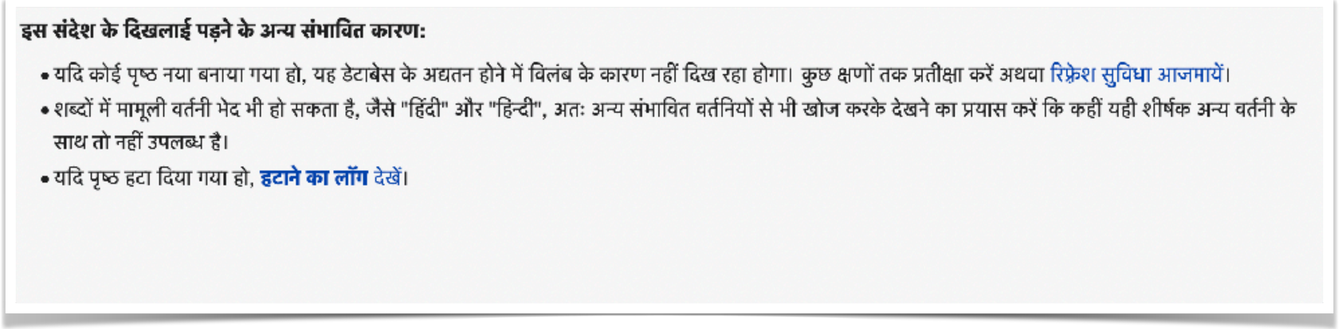
How multilingual is Multilingual BERT?

Telmo Pires* **Eva Schlinger** **Dan Garrette**
Google Research
{telmop,eschling,dhgarrette}@google.com

Cross Lingual Transfer

- What about transferring across languages with different scripts?
- I.e., zero overlap in vocabulary?
- E.g., train on Hindi (written in Devanagri), test on Urdu (written in Arabic)





Cross Lingual Transfer

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

Cross Lingual Transfer

Only ~6 points worse when trained on Urdu rather than Hindi

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

Cross Lingual Transfer

Larger performance drops for less similar language pairs

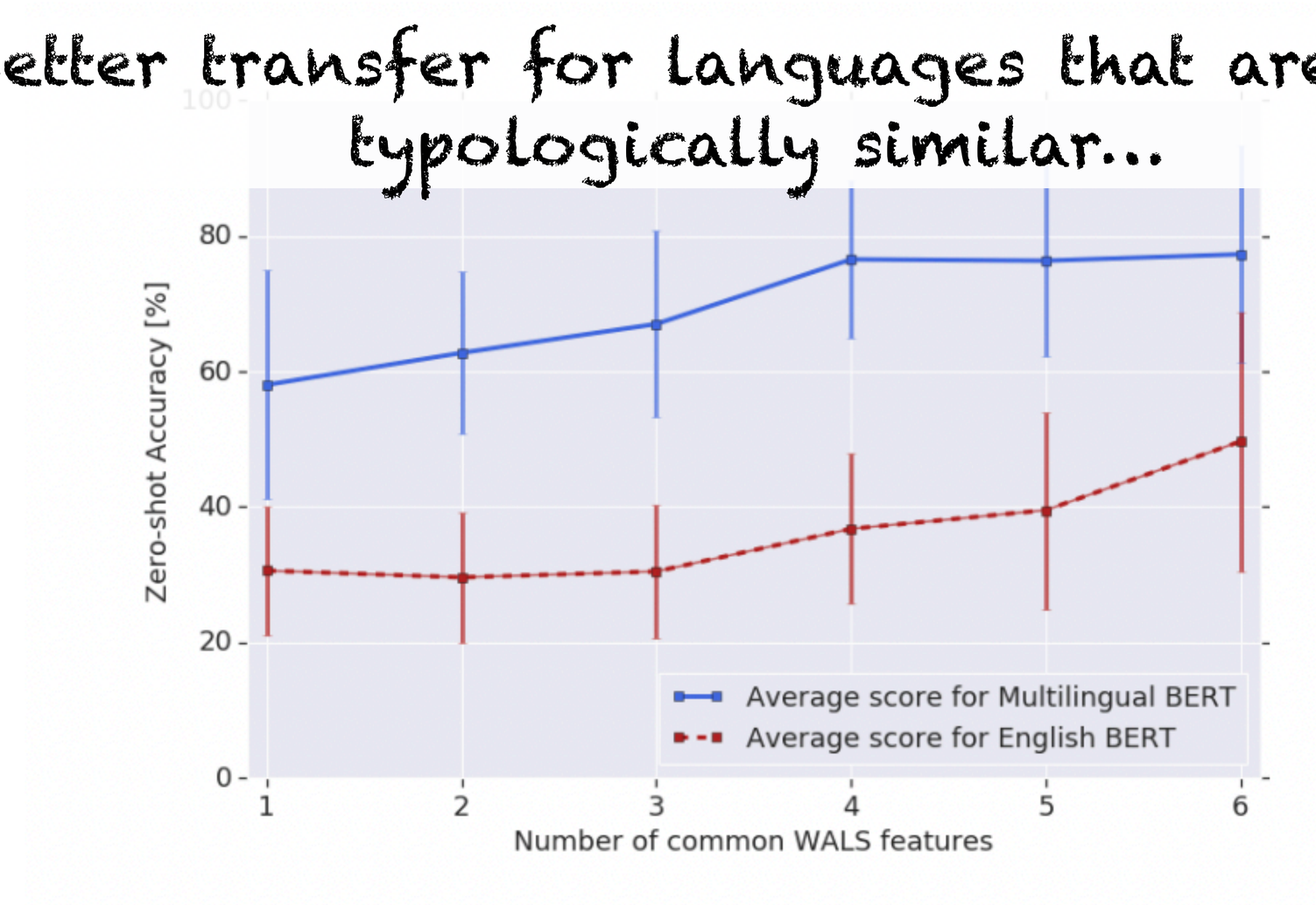
	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

Cross Lingual Transfer

Better transfer for languages that are more typologically similar...



Cross Lingual Transfer

...e.g., more syntactically similar

	SVO	SOV
SVO	**81.55**	66.52
SOV	63.98	**64.22**

(a) Subj./verb/obj. order.

	AN	NA
AN	**73.29**	70.94
NA	75.10	**79.64**

(b) Adjective/noun order.

Table 5: Macro-average POS accuracies when transferring between SVO/SOV languages or AN/NA languages. Row = fine-tuning, column = evaluation.

Cross Lingual Transfer

- Training on the language you care about is still better, in general
- Why would we do cross-lingual transfer then?
- A few reasons:
  - You might not have training data in the language you care about
  - You might not even have enough data to train an MT system for the language you care about
  - Cross-lingual models require only unlabeled, monolingual data, so are more feasible to train

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

BigScience Workshop

Major contributors

Teven Le Scao, Albert Villanova del Moral, Lucile Saulnier, Stas Bekman, Ellie Pavlick, Hugo Laurençon, Roman Castagné, Benoît Sagot, Jonathan Tow, Rachel Bawden, Niklas Muennighoff, Yacine Jernite, Alexandra Sasha Luccioni, Colin Raffel, François Yvon, Suzana Ilić, Margaret Mitchell, Albert Webson, Julien Launay, Victor Sanh, Thomas Wang, Stella Biderman, Daniel Hesslow, Iz Beltagy, Angelina McMillan-Major, Christopher Akiki, Olatunji Ruwase, Alexander M. Rush, Thomas Wolf, Pedro Ortiz Suarez, Samson Tan, Huu Nguyen

Dataset

Aaron Gokaslan, Aitor Soroa, Albert Villanova del Moral, Alexandra Sasha Luccioni, Alham Fikri Aji, Angelina McMillan-Major, Anna Rogers, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Akiki, Christopher Klamm, Colin Leong, Colin Raffel, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Ethan Kim, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hiêu Trần, Hugo Laurençon, Huu Nguyen, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Lucile Saulnier, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Margaret Mitchell, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Pawan Sasanka Ammanamanchi, Pedro Ortiz Suarez, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Roman Castagné, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Samson Tan, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stella Biderman, Suhas Pai, Suzana Ilić, Sydney Zink, Teven Le Scao, Thomas Wang, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Yacine Jernite, Zaid Alyafeai, Zeerak Talat, Amit Alfassy, Adi Simhi, Efrat Levkovizh, Ariel Kreisberg Nitzav, Eyal Bar Natan, Stanislav Silberberg

Tokenization

Arun Raja, Benjamin Heinzerling, Benoît Sagot, Chenglei Si, Colin Raffel, Elizabeth Salesky, Lucile Saulnier, Manan Dey, Matthias Gallé, Pedro Ortiz Suarez, Roman Castagné, Sabrina J. Mielke, Samson Tan, Teven Le Scao, Thomas Wang, Wilson Y. Lee, Zaid Alyafeai

Prompt engineering

Abheesht Sharma, Albert Webson, Alexander M. Rush, Alham Fikri Aji, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Canwen Xu, Colin Raffel, Debajyoti Datta, Dragomir Radev, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jonathan Chang, Jos Rozen, Khalid Almubarak, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Manan Dey, Matteo Manica, Mike Tian-Jian Jiang, Nihal Nayak, Niklas Muennighoff, Rachel Bawden, Ryan Teehan, Samuel Albanie, Shanya Sharma, Sheng Shen, Srulik Ben-David, Stella Biderman, Stephen H. Bach, Taewoon Kim, Tali Bers, Teven Le Scao, Thibault Fevry, Thomas Wang, Thomas Wolf, Trishala Neeraj, Urmish Thakker, Victor Sanh, Vikas Raunak, Xiangru Tang, Zaid Alyafeai, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh

Architecture and objective

Adam Roberts, Colin Raffel, Daniel Hesslow, Hady Elsahar, Hyung Won Chung, Iz Beltagy, Jaesung Tae, Jason Phang, Julien Launay, Lintang Sutawika, Lucile Saulnier, M Saiful Bari, Niklas Muennighoff, Ofir Press, Sheng Shen, Stas Bekman, Stella Biderman, Teven Le Scao, Thomas Wang, Victor Sanh, Zheng Xin Yong

Engineering

Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Niklas Muennighoff, Nouamane Tazi, Olatunji Ruwase, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stas Bekman, Stéphane Requena, Suraj Patil, Teven Le Scao, Thomas Wang, Tim Dettmers

Evaluation and interpretability

Ahmed Baruwa, Albert Webson, Alexandra Sasha Luccioni, Alham Fikri Aji, Amanpreet Singh, Anastasia Cheveleva, Anne Louie Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Dragomir Radev, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Ellie Pavlick, François Yvon, Gerta Indra Winata, Hailey Schoelkopf, Jaesung Tae, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Khalid Almubarak, Lintang Sutawika, Manan Dey, Maraim Masoud, Margaret Mitchell, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Niklas Muennighoff, Oleg Serikov, Oskar van der Wal, Pawan Sasanka Ammanamanchi, Pierre Colombo, Rachel Bawden, Rui Zhang, Ruochen Zhang, Samson Tan, Sebastian Gehrmann, Shanya Sharma, Shayne Longpre, Stella Biderman, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Urmish Thakker, Verena Rieser, Vikas Raunak, Vitaly Protasov, Vladislav Mikhailov, Wilson Y. Lee, Yada Pruksachatkun, Zdeněk Kasner, Zeerak Talat, Zheng-Xin Yong, Shani Pais, Zachary Bamberger, Liam Hazan, Omer Antverg, Eli Bogdanov, Yonatan Belinkov

Broader impacts

Aaron Gokaslan, Alexandra Sasha Luccioni, Alham Fikri Aji, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Angelina McMillan-Major, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Chenghao Mou, Minh Chien Vu, Christopher Akiki, Danish Contractor, David Ifeoluwa Adelani, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Gérard Dupont, Giada Pistilli, Habib Rezanejad, Hessie Jones, Huu Nguyen, Ian Yu, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jaesung Tae, Jenny Chim, Jesse Passmore, Josh Seltzer, Julien Launay, Julio Bonis Sanz, Karen Fort, Khalid Almubarak, Livia Dutra, Long Phan, Mairon Samagaio, Manan Dey, Maraim Elbadri, Maraim Masoud, Margaret Mitchell, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Niklas Muennighoff, Nishant Subramani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Olivier Nguyen, Paulo Villegas, Pawan Sasanka Ammanamanchi, Priscilla Amuok, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Shanya Sharma, Shayne Longpre, Silas Wang, Somaieh Nikpoor, Sourav Roy, Stas Bekman, Stella Biderman, Suhas Pai, Suzana Ilić, Sylvain Viguier, Teven Le Scao, Thanh Le, Tobi Oyebade, Trieu Le, Tristan Thrush, Yacine Jernite, Yoyo Yang, Zach Nguyen, Zeerak Talat, Zheng Xin Yong

Applications

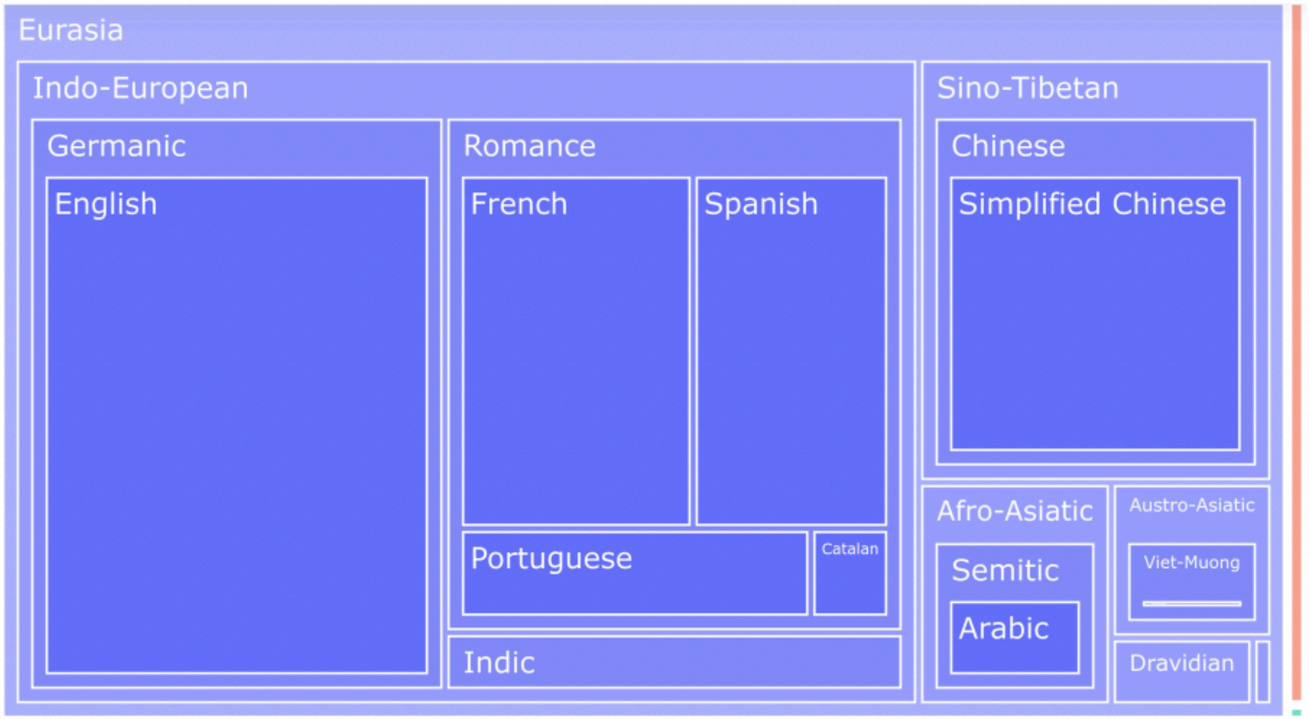
Abhinav Ramesh Kashyap, Albert Villanova del Moral, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Carlos Muñoz Ferrandis, Chenxi Zhou, Chirag Jain, Christopher Akiki, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Daniel van Strien, Danish Contractor, David Lansky, Debajyoti Datta, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Francesco De Toni, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jason Alan Fries, Javier de la Rosa, Jenny Chim, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Leon Weber, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minh Chien Vu, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shamik Bose, Shilok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Stella Biderman, Stephen H. Bach, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Trishala Neeraj, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye

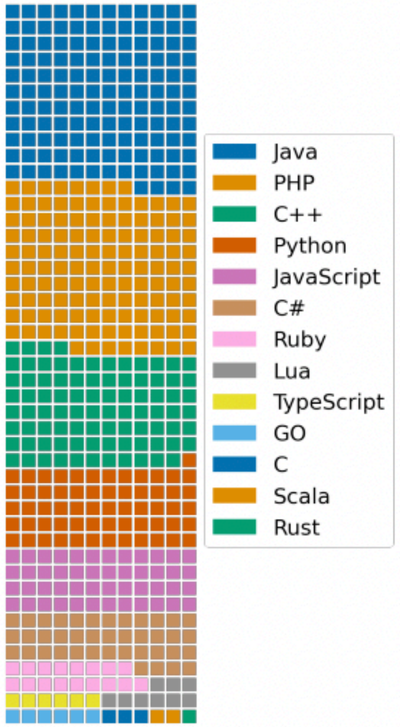
Organization

Angela Fan, Christopher Akiki, Douwe Kiela, Giada Pistilli, Margot Mieskes, Mathilde Bras, Matthias Gallé, Suzana Ilić, Yacine Jernite, Younes Belkada, Thomas Wolf

: A ... ultilingual

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model





Pretraining

In-Context Learning

Task: Sentiment Analysis
Input: “the food sucks”
Expected: Negative

P(“positive”) vs. P(“negative”)

Frozen, Pretrained Model

Is the following review positive or negative?
Awesome place -> positive
Would not recommend -> negative
My absolute fav -> positive
This food sucks. ->

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

Src↓	Trg→	eng	ben	hin	swh	yor
eng	M2M	–	23.04	28.15	29.65	2.17
	BLOOM	–	25.52	27.57	21.7	2.8
ben	M2M	22.86	–	21.76	14.88	0.54
	BLOOM	30.23	–	16.4	–	–
hin	M2M	27.89	21.77	–	16.8	0.61
	BLOOM	35.40	23.0	–	–	–
swh	M2M	30.43	16.43	19.19	–	1.29
	BLOOOM	37.9	–	–	–	1.43
yor	M2M	4.18	1.27	1.94	1.93	–
	BLOOM	3.8	–	–	0.84	–

(a) Low-resource languages

Src↓	Trg→	cat	spa	fre	por
cat	M2M	–	25.17	35.08	35.15
	BLOOM	–	29.12	34.89	36.11
spa	M2M	23.12	–	29.33	28.1
	BLOOM	31.82	–	24.48	28.0
glg	M2M	30.07	27.65	37.06	34.81
	BLOOM	38.21	27.24	36.21	34.59
fre	M2M	28.74	25.6	–	37.84
	BLOOM	38.13	27.40	–	39.60
por	M2M	30.68	25.88	40.17	–
	BLOOM	40.02	28.1	40.55	–

(b) Romance languages

Topics

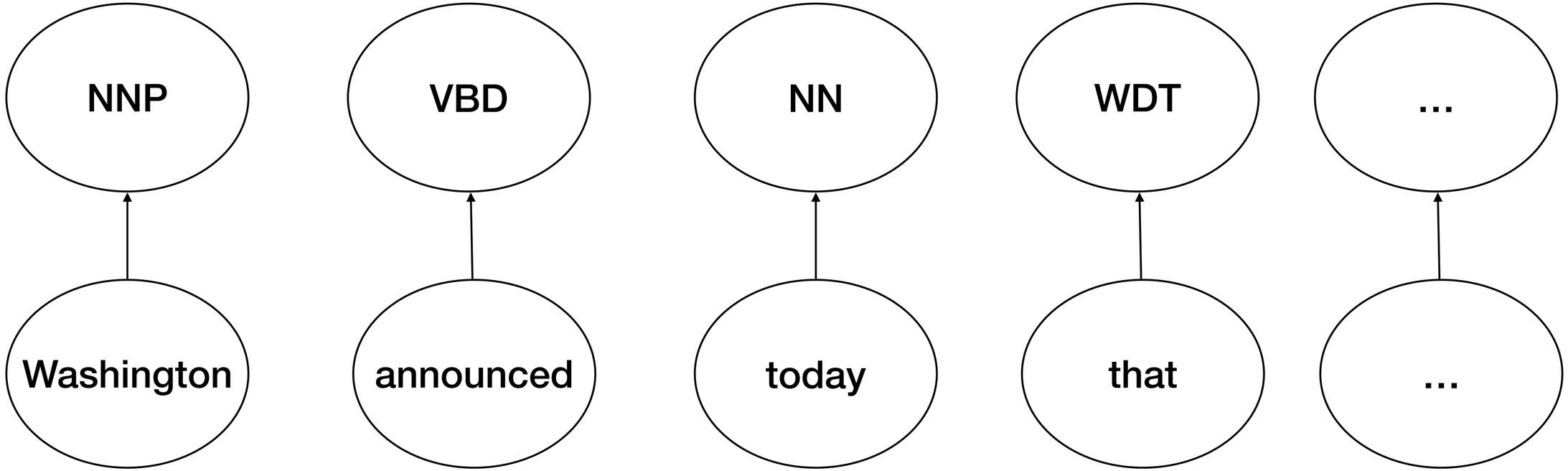
- Follow up: Multlingual LMs and Zero-Shot
- **Part of Speech Tags**
- Hidden Markov Models
- Conditional Random Fields

Sequence Labeling Tasks

- POS Tagging: NOUN, VERB, ADJ, etc
- Named Entity Recognition (NER): PERSON, LOCATION, ORG, etc
- Goal: Given words, predict tags

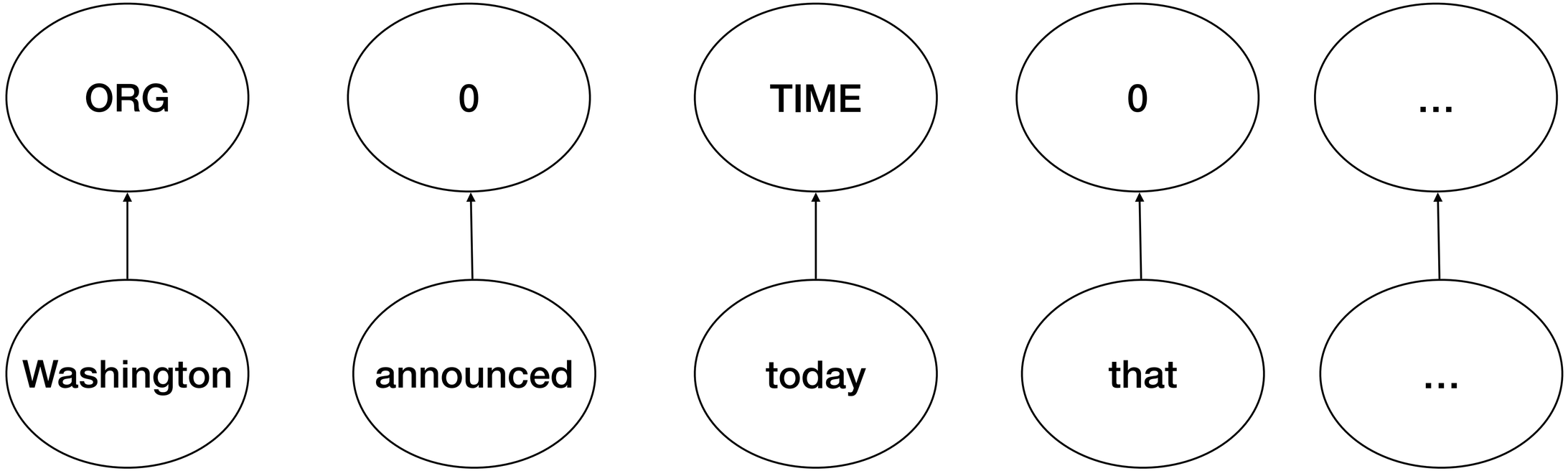
Sequence Labeling Tasks

- POS Tagging: NOUN, VERB, ADJ, etc
- Named Entity Recognition (NER): PERSON, LOCATION, ORG, etc
- Goal: Given words, predict tags



Sequence Labeling Tasks

- POS Tagging: NOUN, VERB, ADJ, etc
- Named Entity Recognition (NER): PERSON, LOCATION, ORG, etc
- Goal: Given words, predict tags



Sequence Labeling Tasks

- In linguistics, tags are *descriptive*—they help us understand constraints on grammar and meaning
  - E.g., “the NOUN” is grammatical, “the VERB” is not
- In NLP, tags are used as *features*—they provide information that is important for understanding sentence, anticipating next words, etc
  - Knowing if a document uses “cat” as a noun or “cat” as a verb is important for knowing the topic (animals vs. linux…)

Part of Speech Tags

Penn Treebank Tags

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	"to"	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential 'there'	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*'s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past partici-ple	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your, one's*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	+,%, &	WRB	wh-adverb	*how, where*

Figure 8.2 Penn Treebank part-of-speech tags.

Part of Speech Tags

Penn Treebank Tags

```
parsed = spacy_processor(text)
print(word, word.tag_)

Will MD
it PRP
just RB
cease VB
to TO
function VB
and CC
orbit VB
the DT
earth NN
forever RB
```

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*			
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*			
DT	determiner	*a, the*	NNS	noun, plural	*llamas*			
EX	existential 'there'	*there*	PDT	predeterminer	*all, both*			
FW	foreign word	*mea culpa*	POS	possessive ending	*'s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past partici-ple	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your, one's*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	*+,%, &*	WRB	wh-adverb	*how, where*

Figure 8.2 Penn Treebank part-of-speech tags.

	Tag	Description	Example
Open Class	**ADJ**	Adjective: noun modifiers describing properties	*red, young, awesome*
	ADV	Adverb: verb modifiers of time, place, manner	*very, slowly, home, yesterday*
	NOUN	words for persons, places, things, etc.	*algorithm, cat, mango, beauty*
	VERB	words for actions and processes	*draw, provide, go*
	PROPN	Proper noun: name of a person, organization, place, etc..	*Regina, IBM, Colorado*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh, um, yes, hello*
Closed Class Words	**ADP**	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by, under*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and, or, but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two, first, second*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that, which*
Other	**PUNCT**	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

Figure 8.1 The 17 parts of speech in the Universal Dependencies tagset (Nivre et al., 2016a). Features can be added to make finer-grained distinctions (with properties like number, case, definiteness, and so on).

	Tag	Description	Exa
Open Class	ADJ	Adjective: noun modifiers describing properties	*red,*
	ADV	Adverb: verb modifiers of time, place, manner	*very,*
	NOUN	words for persons, places, things, etc.	*algo*
	VERB	words for actions and processes	*draw*
	PROPN	Proper noun: name of a person, organization, place, etc..	*Regi*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh, u*
Closed Class Words	ADP	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, o*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and, or, but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two, first, second*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that, which*
Other	PUNCT	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

```
parsed = spacy_processor(text)
print(word, word.pos_)

Will AUX
it PRON
just ADV
cease VERB
to PART
function VERB
and CCONJ
orbit VERB
the DET
earth NOUN
forever ADV
```

Figure 8.1 The 17 parts of speech in the Universal Dependencies tagset (Nivre et al., 2016a). Features can be added to make finer-grained distinctions (with properties like number, case, definiteness, and so on).

Part of Speech Tags

Differences between tag sets

	There	**are**	**70**	**children**	**there**	**.**
Penn Treebank	EX	VBP	CD	NNS	RB	.
Universal Dependency	PRO	VERB	NUM	NOUN	ADV	PUNC

Part of Speech Tagging

Ambiguity

- Ambiguity: about 55% of tokens are ambiguous, i.e., have more than one possible tag
  - e.g.,
    - earnings growth took a **back/JJ** seat
    - a small building in the **back/NN**
    - a clear majority of senators **back/VBP** the bill
    - Dave began to **back/VB** toward the door
    - enable the country to buy **back/RP** debt
    - I was twenty-one **back/RB** then
- Usually easy to tell from context
- Most frequent class baseline: when ambiguous, guess the most likely tag given the training data
  - Achieves 92%; human performance is 97%

Part of Speech Tagging

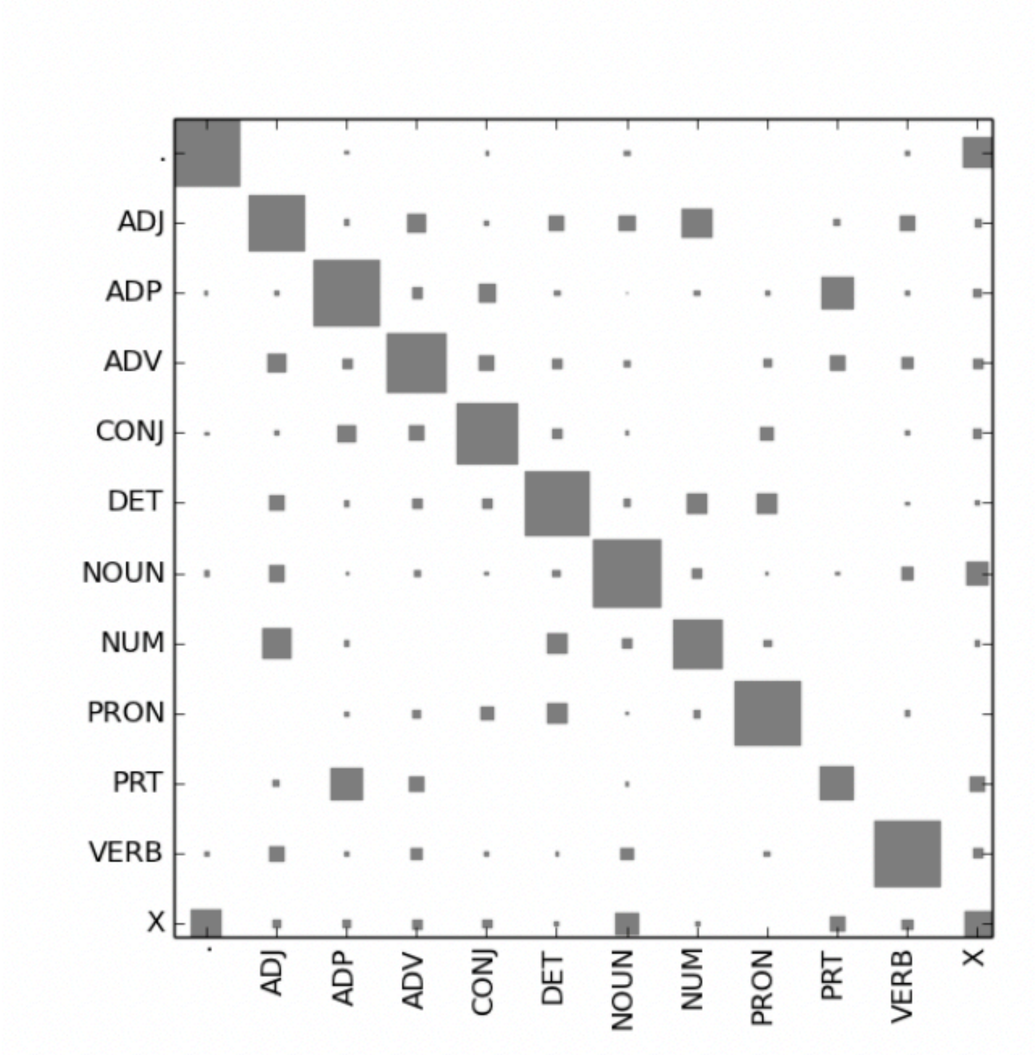
Ambiguity

Learning part-of-speech taggers with inter-annotator agreement loss

Barbara Plank, Dirk Hovy, Anders Søgaard
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140, DK-2300 Copenhagen S
bplank@cst.dk,dirk@cst.dk,soegaard@hum.ku.dk

Part of Speech Tagging

Ambiguity



RITTER

...	you/PRON	**come/VERB**	**out/PRT**	to/PRT		...
	it/PRON	**comes/VERB**	**out/ADP**	nov/NOUN		

GIMPEL

	Advances/NOUN	and/CONJ	**Social/NOUN**	**Media/NOUN**	.../X	
...	Journalists/NOUN	and/CONJ	**Social/ADJ**	**Media/NOUN**	experts/NOUN	...

Topics

- Follow up: Multlingual LMs and Zero-Shot
- Part of Speech Tags
- **Hidden Markov Models**
- Conditional Random Fields

Markov Chains

- Markov Assumption: state t+1 depends *only* on state t (nothing prior)
  - $P(t+1|t0\ldots.t) = P(t+1|t)$
  - Like a bigram language model!
- Markov Chain: Simple model for assigning probabilities to sequences using Markov assumption

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
i	5	827		9		
want	2		608	1	6	5
to	2		4	686		6
eat			2		2	42
food	15		15		4	
lunch	2				1	

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	

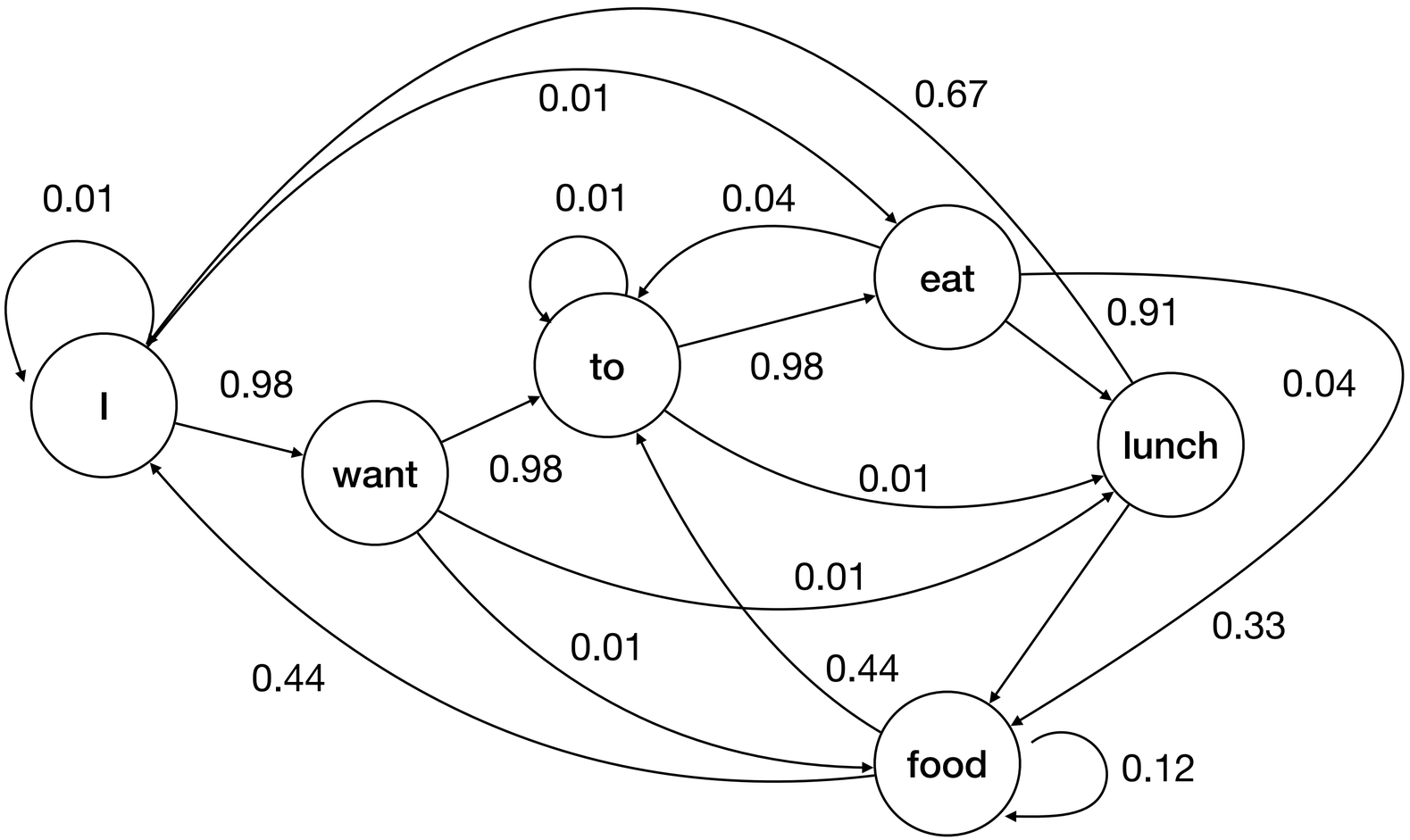
$$P(w_n \mid w_{n-1}) = \frac{\#(w_{n-1} w_n)}{\# w_{n-1}}$$

$$P(w_0 \ldots w_n) \approx P(w_0 \mid \text{<s>}) \times P(w_1 \mid w_0) \times P(w_2 \mid w_1) \times \ldots \times P(w_n \mid w_{n-1})$$

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	



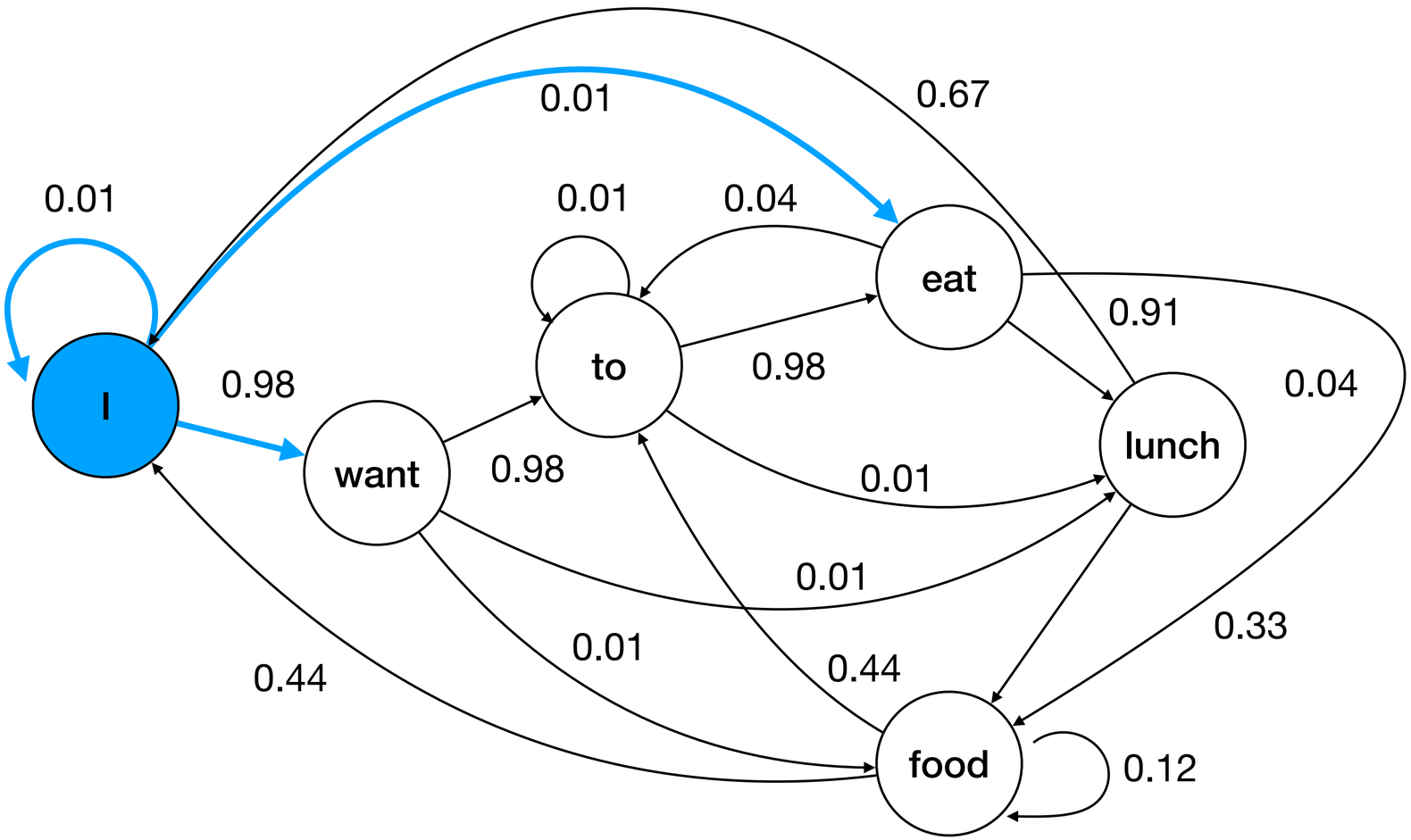
$$P(w_n \mid w_{n-1}) = \frac{\#(w_{n-1}w_n)}{\#w_{n-1}}$$

$$P(w_0 \ldots w_n) \approx P(w_0 \mid \text{<s>}) \times P(w_1 \mid w_0) \times P(w_2 \mid w_1) \times \ldots \times P(w_n \mid w_{n-1})$$

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	



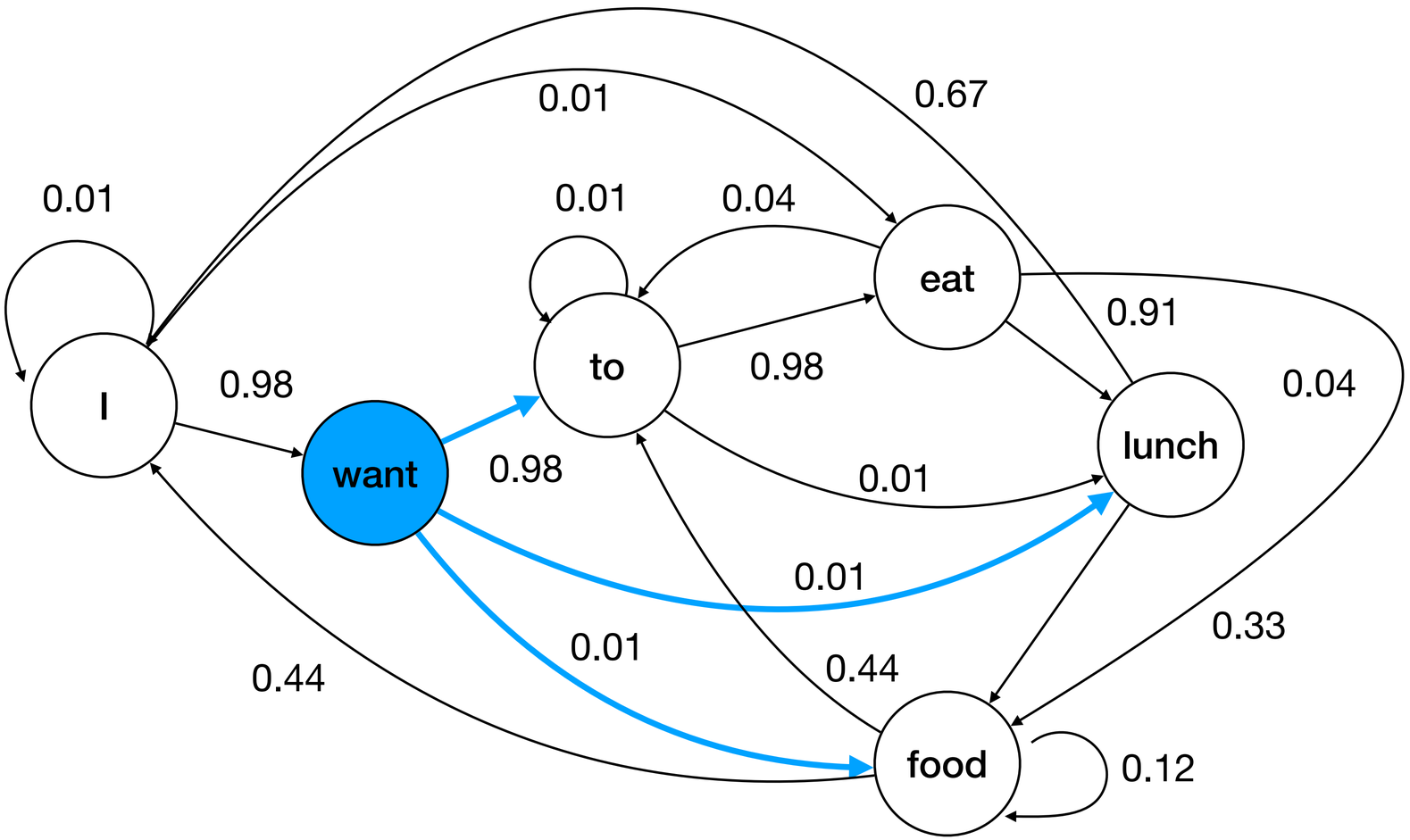
$$P(w_n \mid w_{n-1}) = \frac{\#(w_{n-1}w_n)}{\#w_{n-1}}$$

$$P(w_0 \ldots w_n) \approx P(w_0 \mid \text{<s>}) \times P(w_1 \mid w_0) \times P(w_2 \mid w_1) \times \ldots \times P(w_n \mid w_{n-1})$$

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	



$$P(w_n \mid w_{n-1}) = \frac{\#(w_{n-1}w_n)}{\#w_{n-1}}$$

$$P(w_0 \ldots w_n) \approx P(w_0 \mid \text{<s>}) \times P(w_1 \mid w_0) \times P(w_2 \mid w_1) \times \ldots \times P(w_n \mid w_{n-1})$$

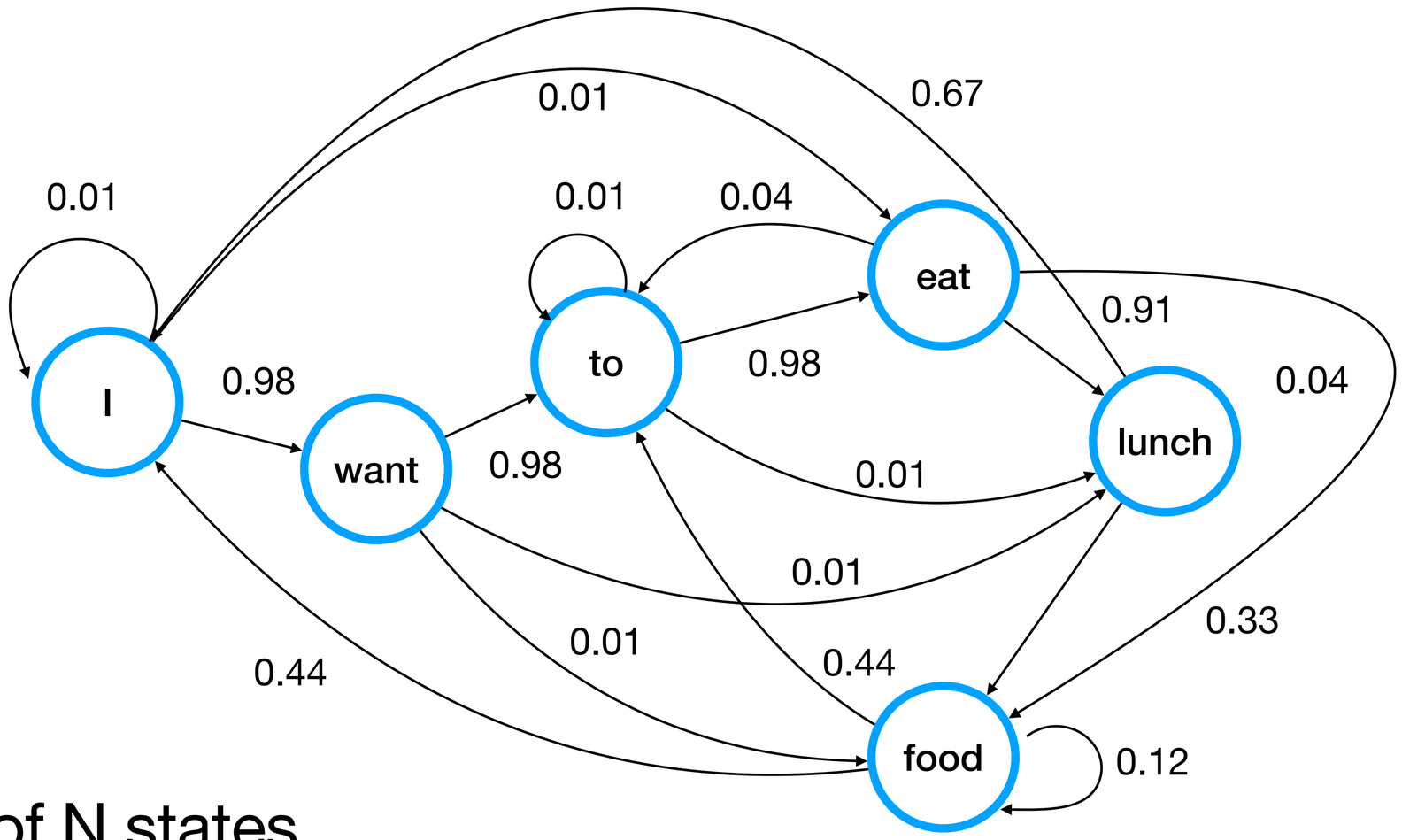
Markov Chains

- Formally:
  - $Q = \{q_1, q_2, \ldots q_N\}$: a set of N states
  - $A = \{a_{ij}\}$ for all i,j from 1 to N: a transition probability matrix
  - $\pi = \{\pi_1, \pi_2, \ldots \pi_N\}$: an initial probability distribution over states

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	

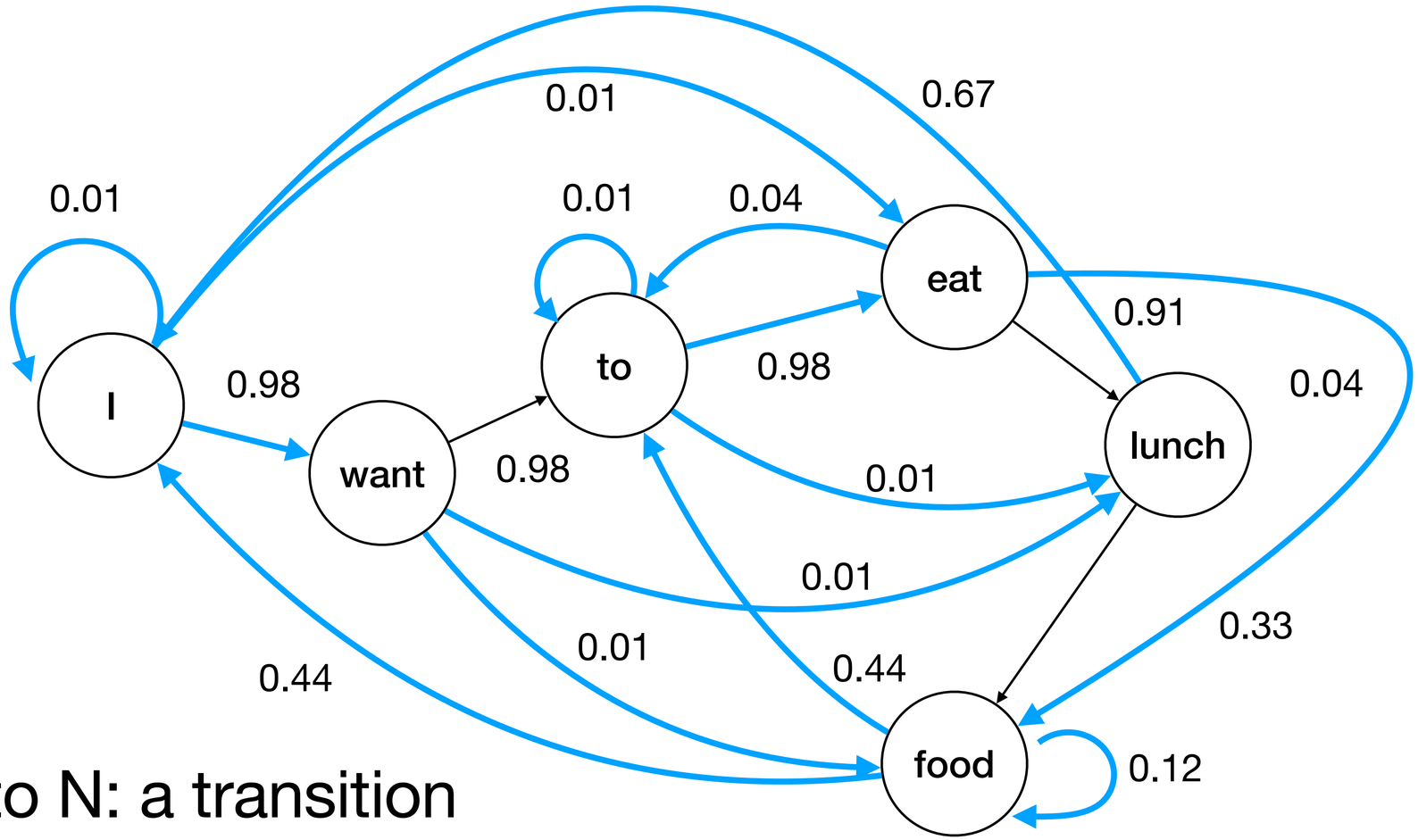


$Q = \{q_1, q_2, \ldots q_N\}$: a set of N states

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	



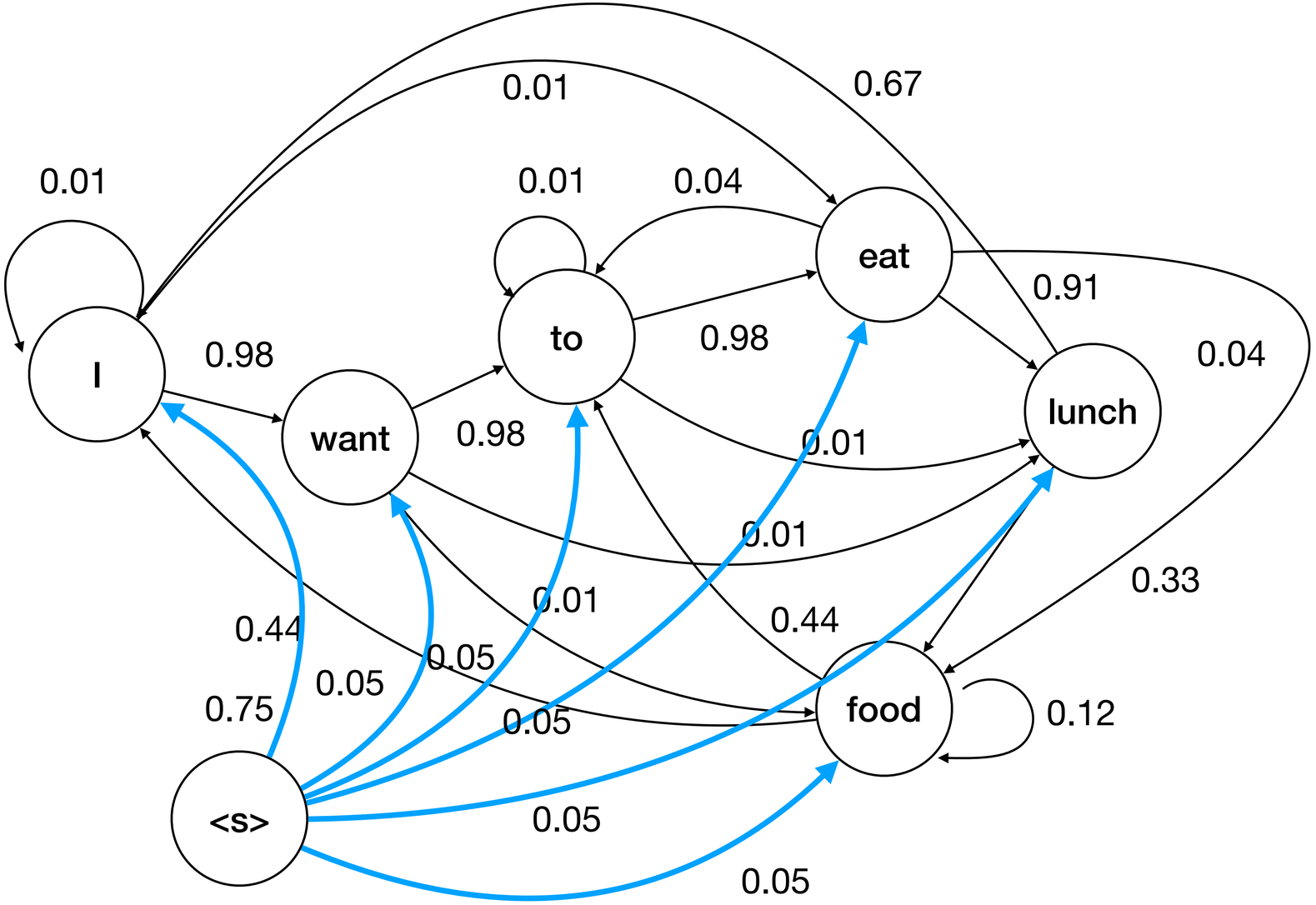
A = {aij} for all i,j from 1 to N: a transition probability matrix

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
π	0.75	0.05	0.05	0.05	0.05	0.05

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	



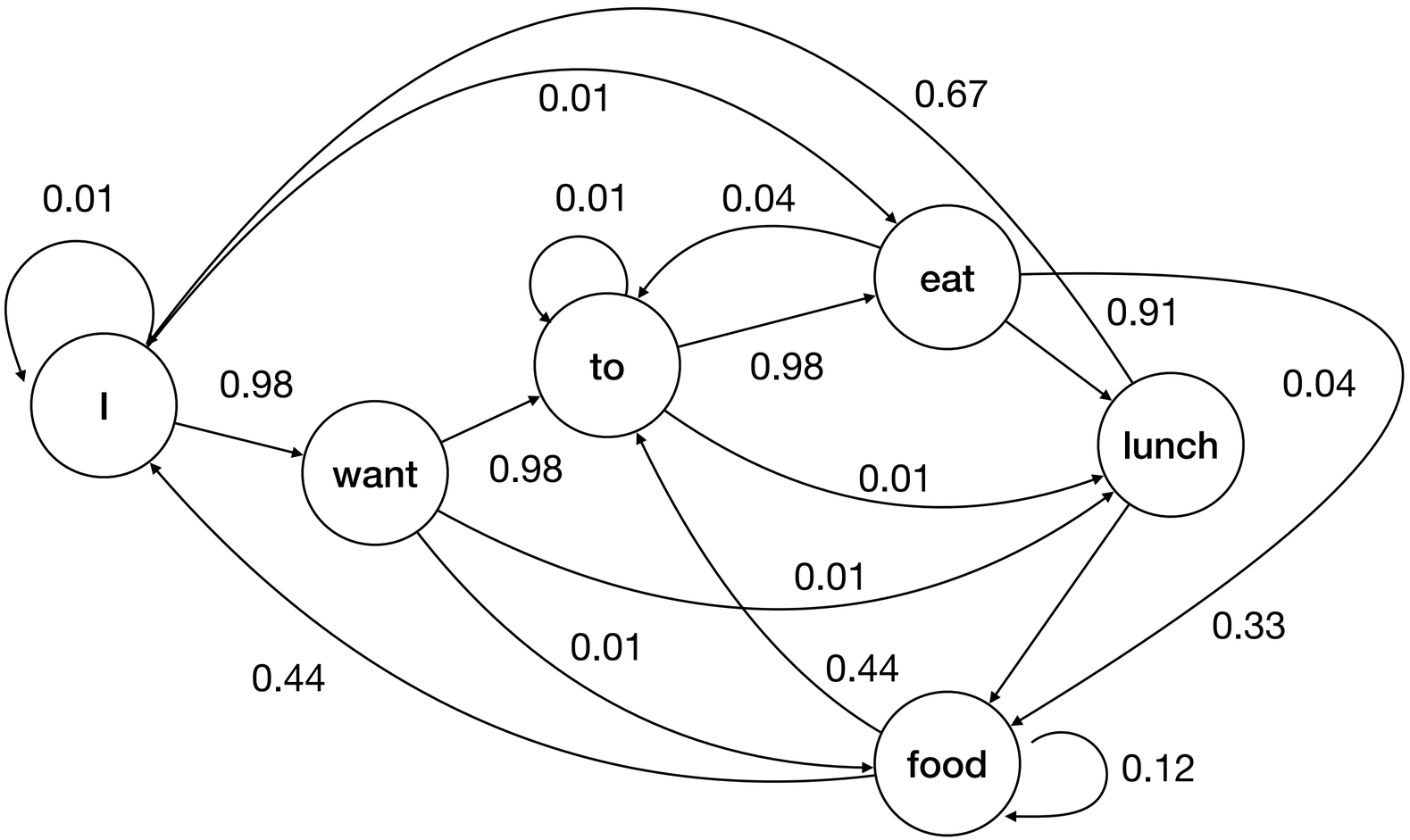
π = {π1, π2, …πN}: an initial probability distribution over states

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
π	0.75	0.05	0.05	0.05	0.05	0.05

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	



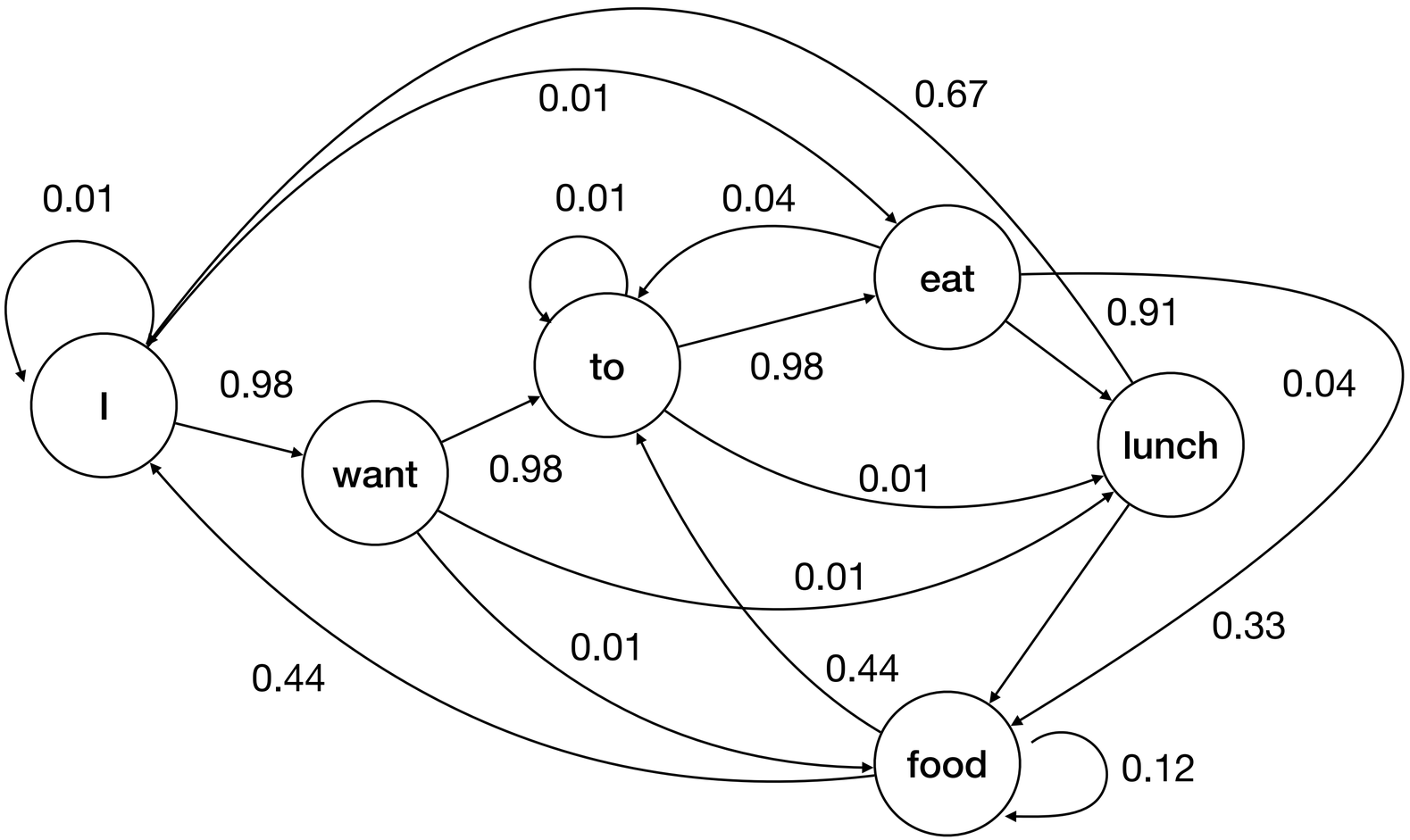
P(I want food to eat)?

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
π	0.75	0.05	0.05	0.05	0.05	0.05

	i	want	to	eat	food	lunch
i	0.01	0.98		0.01		
want			0.98		0.01	0.01
to			0.01	0.98		0.01
eat			0.04		0.04	0.91
food	0.44		0.44		0.12	
lunch	0.67				0.33	



P(I want food to eat)?

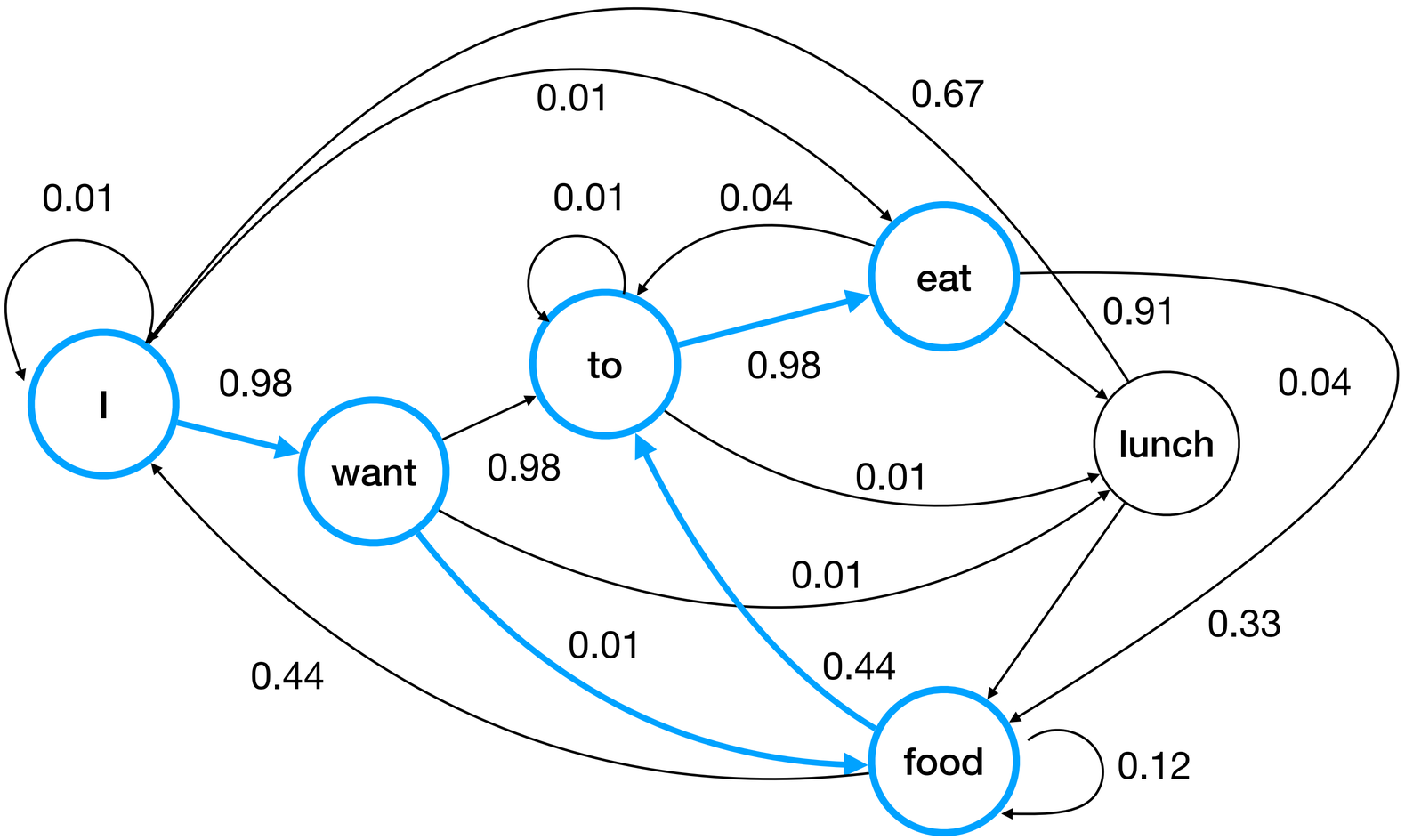
π(I) x P(want|I) x P(food|want) x P(to|food) x P(eat|to) x P(</s>|eat)

Markov Chains

Example: Bigram Language Model

	i	want	to	eat	food	lunch
π	**0.75**	0.05	0.05	0.05	0.05	0.05

	i	want	to	eat	food	lunch
i	0.01	**0.98**		0.01		
want			0.98		**0.01**	0.01
to			0.01	**0.98**		0.01
eat			0.04		0.04	0.91
food	0.44		**0.44**		0.12	
lunch	0.67				0.33	



P(I want food to eat)?

0.75 x 0.98 x 0.01 x 0.44 x 0.98 x P(</s>|eat)

Hidden Markov Models

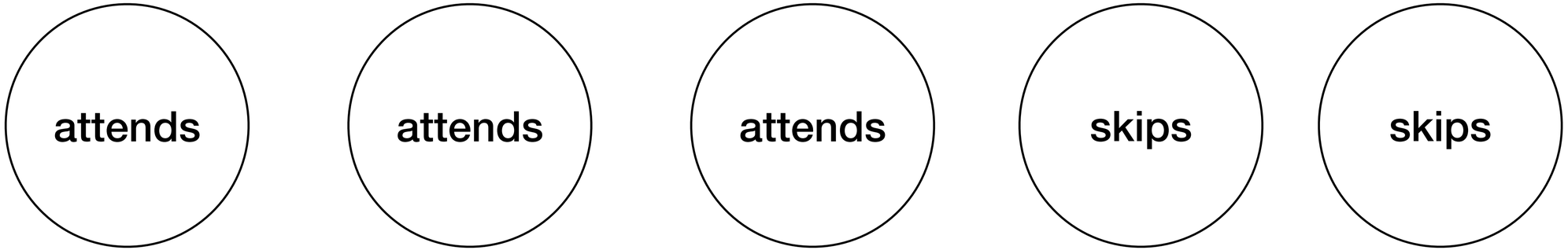
Latent States

- *Hidden* = Latent = some events are not directly observed, but they can nonetheless influence events which *are* observed
  - Observed: whether someone attends lecture, how much time they spend on the quizzes, how much time they spend on the readings
  - Latent: a person's intrinsic "interest" in the material
  - Observed = f(Latent)
- Two assumptions in an HMM:
  - Markov: $P(q_{t+1}|q_0 \ldots q_t) = P(q_{t+1}|q_t)$
  - Output Independence: $P(o_t|o_0 \ldots o_n, q_0 \ldots q_n) = P(o_t|q_t)$

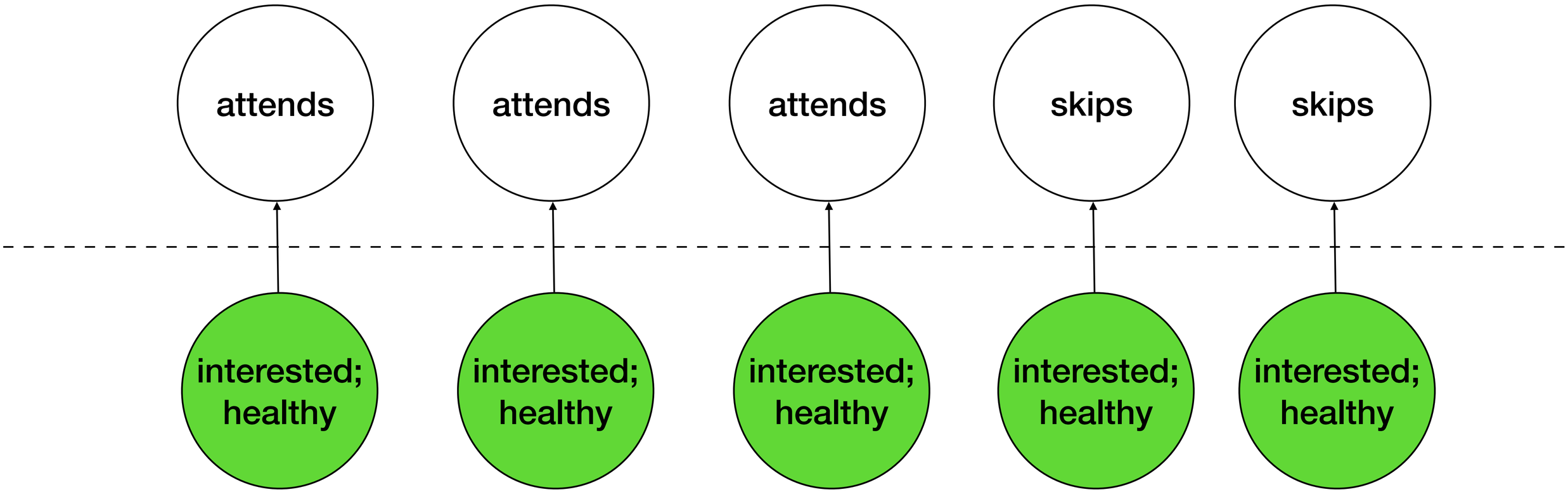
Hidden Markov Models

- Formally:
  - $Q = \{q_1, q_2, \ldots q_N\}$: a set of N states
  - $A = \{a_{ij}\}$ for all i,j from 1 to N: a transition probability matrix
  - $\pi = \{\pi_1, \pi_2, \ldots \pi_N\}$: an initial probability distribution over states
  - $O = \{o_1, o_2, \ldots o_T\}$: a set of T observations
  - $B = \{b_{ij}\}$ for all i from 1 to N; j from 1 to T: an emission probability matrix specifying probability of observation o_j given state q_i

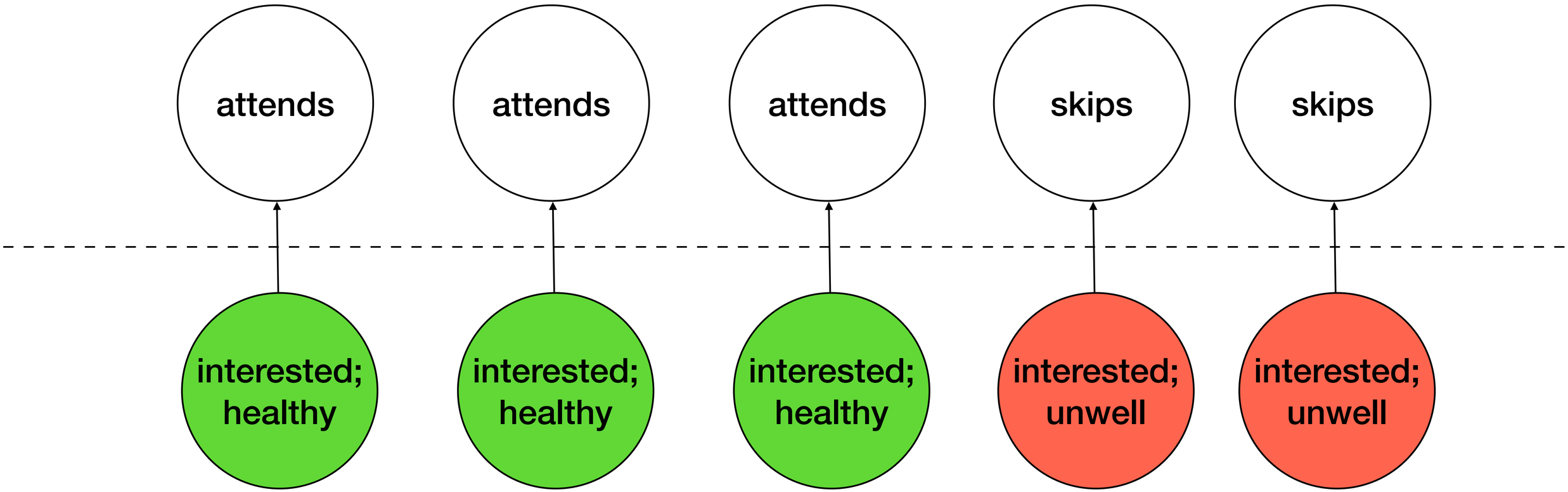
Hidden Markov Models



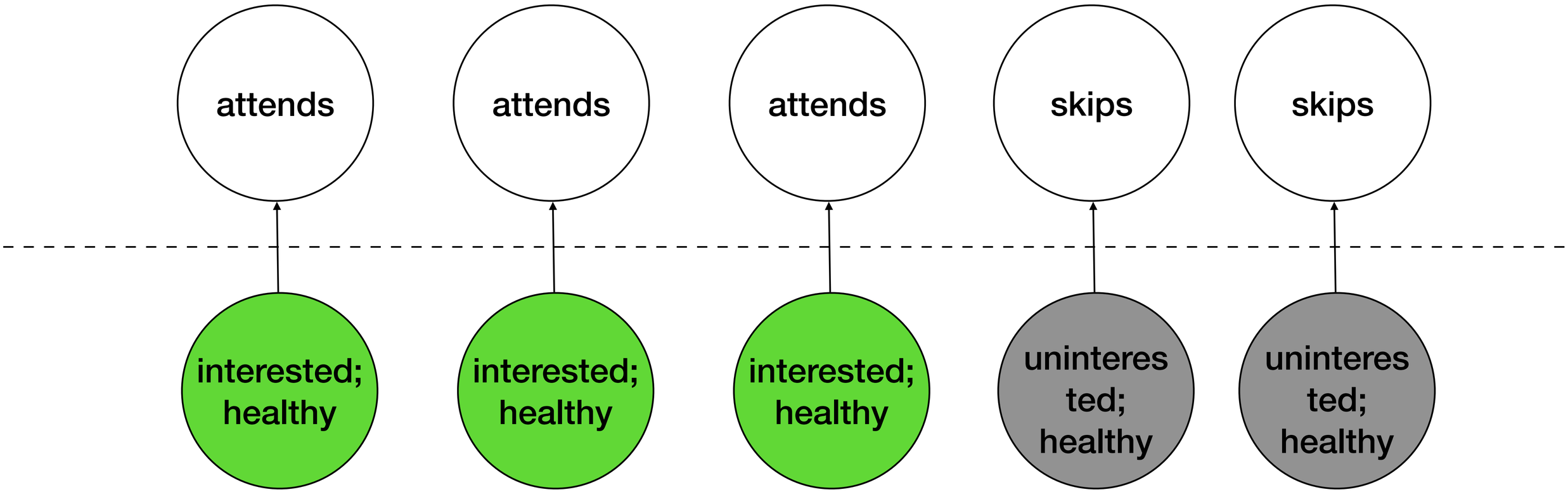
Hidden Markov Models



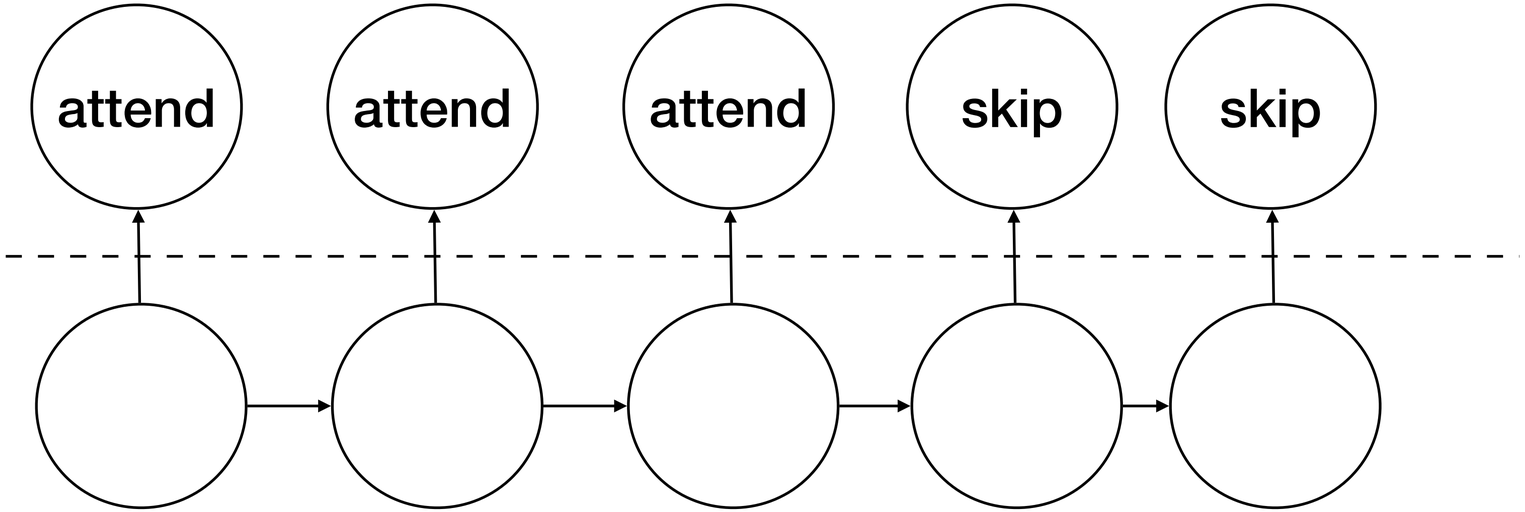
Hidden Markov Models



Hidden Markov Models

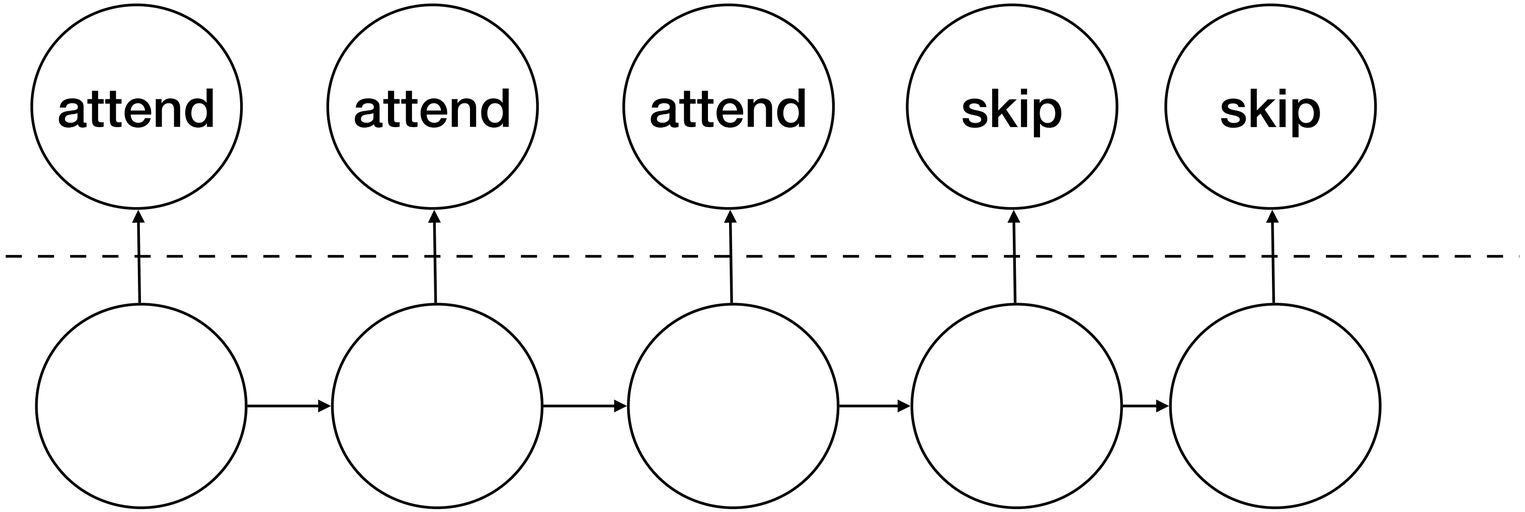


Hidden Markov Models



Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

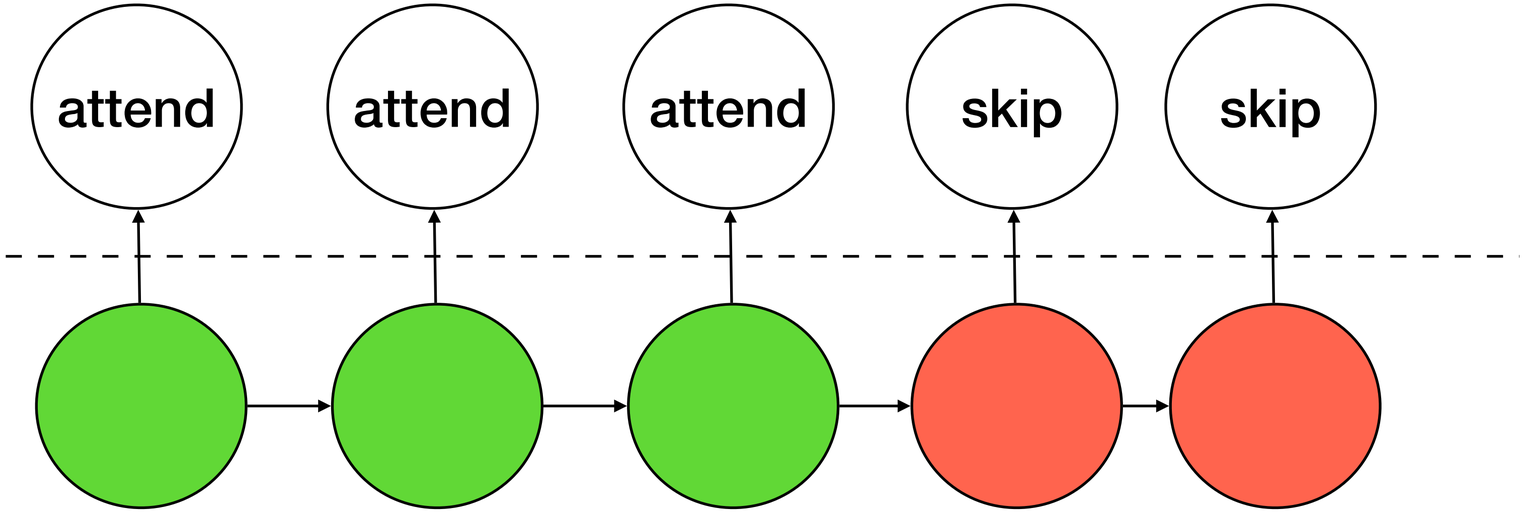
Hidden Markov Models



Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

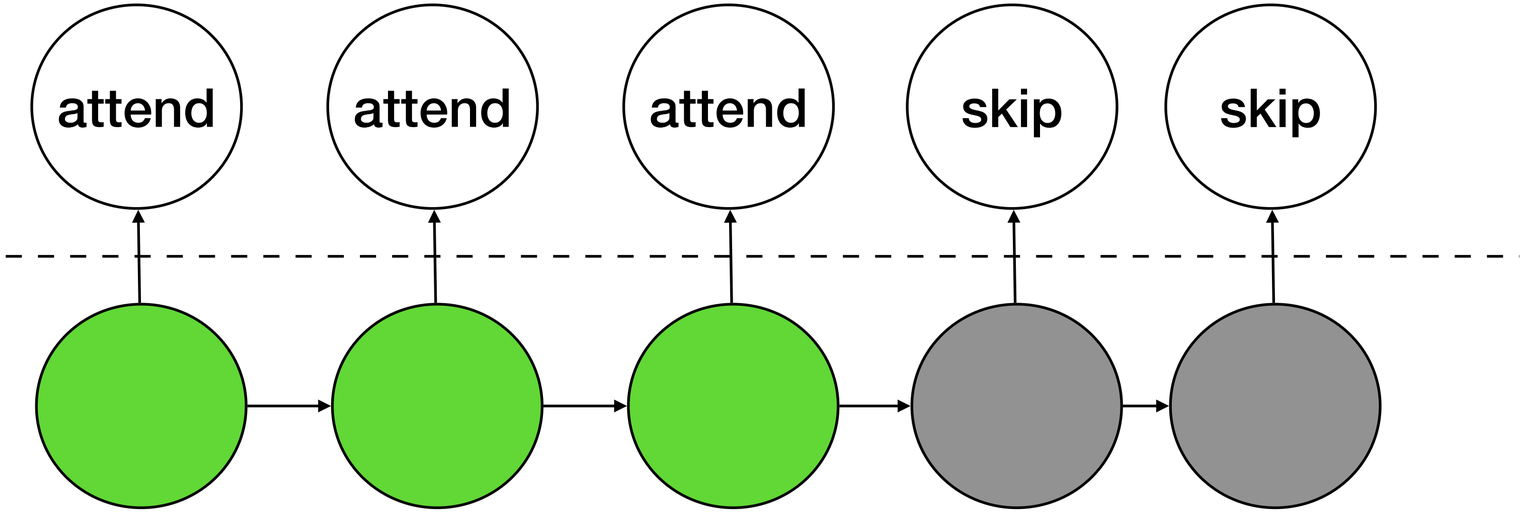


Is it more likely that they got sick?

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

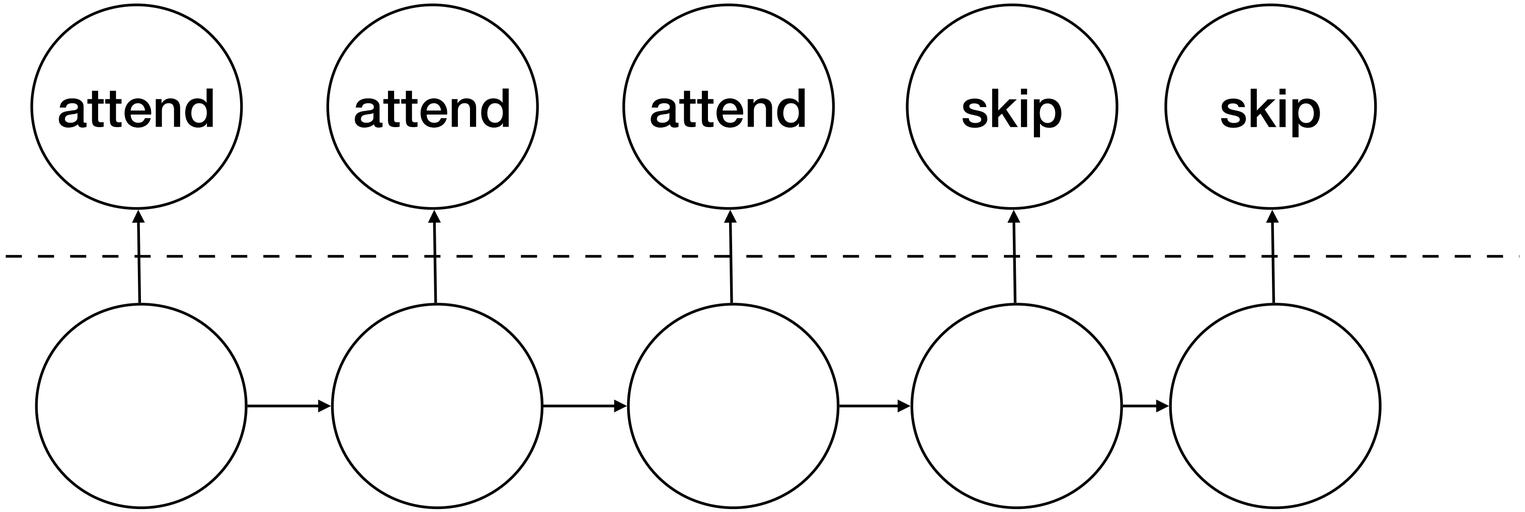


Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Is it more likely that they got sick?
Or that they lost interest?

Hidden Markov Models



What is the most likely chian of hidden states, given my observations?

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$\hat{q} = argmax_{q_1 \ldots q_n} P(q_1 \ldots q_n \,|\, o_1 \ldots o_n)$$

Hidden Markov Models

$$\hat{q} = argmax_{q_1 \ldots q_n} P(q_1 \ldots q_n \,|\, o_1 \ldots o_n)$$

$$\hat{q} = \frac{argmax_{q_1 \ldots q_n} P(o_1 \ldots o_n \,|\, q_1 \ldots q_n) P(q_1 \ldots q_n)}{P(o_1 \ldots o_n)}$$

$$\hat{q} = argmax_{q_1 \ldots q_n} P(o_1 \ldots o_n \,|\, q_1 \ldots q_n) P(q_1 \ldots q_n)$$

Hidden Markov Models

$$\hat{q} = argmax_{q_1 \ldots q_n} P(o_1 \ldots o_n \mid q_1 \ldots q_n) P(q_1 \ldots q_n)$$

Hidden Markov Models

$$\hat{q} = argmax_{q_1 \ldots q_n} P(o_1 \ldots o_n \mid q_1 \ldots q_n) P(q_1 \ldots q_n)$$

output independence

$$P(o_1 \ldots o_n \mid q_1 \ldots q_n) \approx \prod_{i=1}^{n} P(o_i \mid q_i)$$

Hidden Markov Models

$$\hat{q} = argmax_{q_1 \ldots q_n} P(o_1 \ldots o_n \mid q_1 \ldots q_n) P(q_1 \ldots q_n)$$

output independence

$$P(o_1 \ldots o_n \mid q_1 \ldots q_n) \approx \prod_{i=1}^{n} P(o_i \mid q_i)$$

Markov

$$P(t_1 \ldots t_n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1})$$

Hidden Markov Models

$$\hat{q} = argmax_{q_1 \ldots q_n} P(o_1 \ldots o_n \mid q_1 \ldots q_n) P(q_1 \ldots q_n)$$

output independence

$$P(o_1 \ldots o_n \mid q_1 \ldots q_n) \approx \prod_{i=1}^{n} P(o_i \mid q_i)$$

Markov

$$P(t_1 \ldots t_n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1})$$

$$\hat{q} \approx \text{argmax}_{q_1 \ldots q_n} \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$

Hidden Markov Models

$$\hat{q} = argmax_{q_1 \ldots q_n} P(o_1 \ldots o_n \mid q_1 \ldots q_n) P(q_1 \ldots q_n)$$

output independence

$$P(o_1 \ldots o_n \mid q_1 \ldots q_n) \approx \prod_{i=1}^{n} P(o_i \mid q_i)$$

Markov

$$P(t_1 \ldots t_n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1})$$

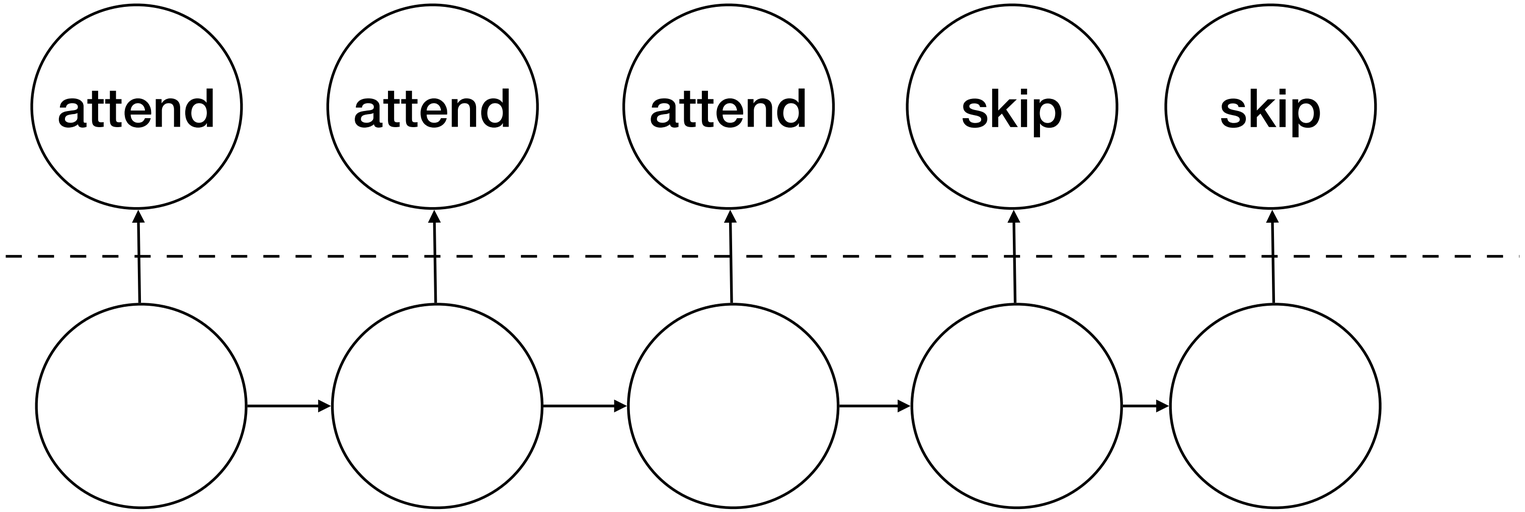
Emission

Transition

$$\hat{q} \approx \text{argmax}_{q_1 \ldots q_n} \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



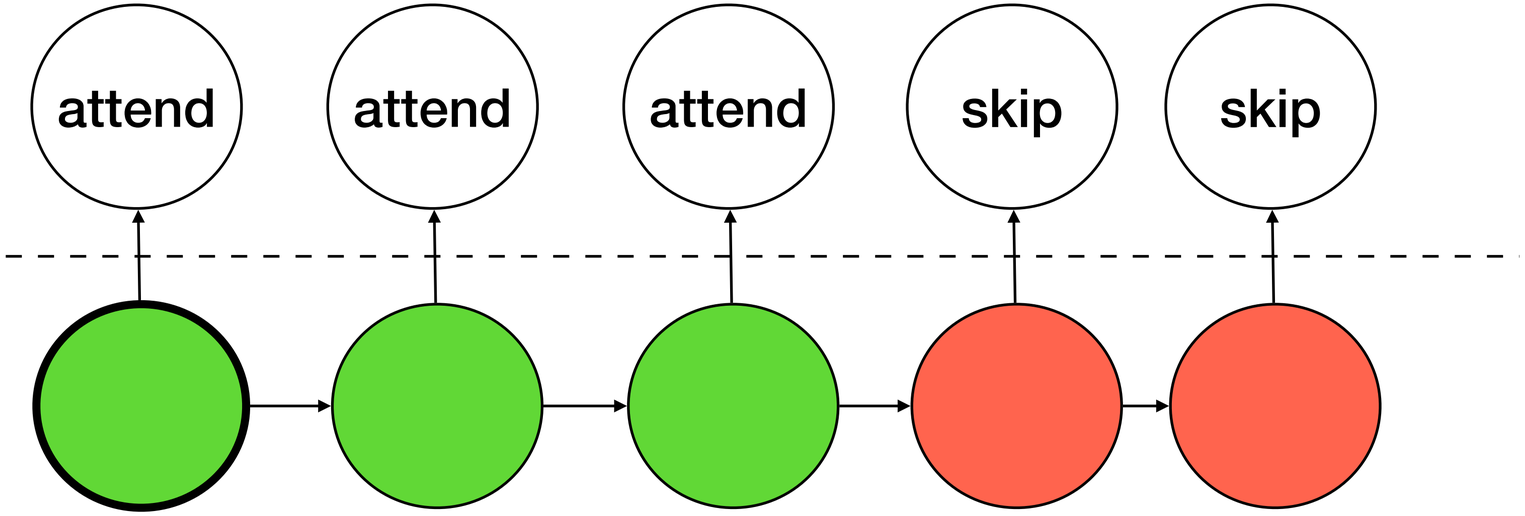
Is it more likely that they got sick?
Or that they lost interest?

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



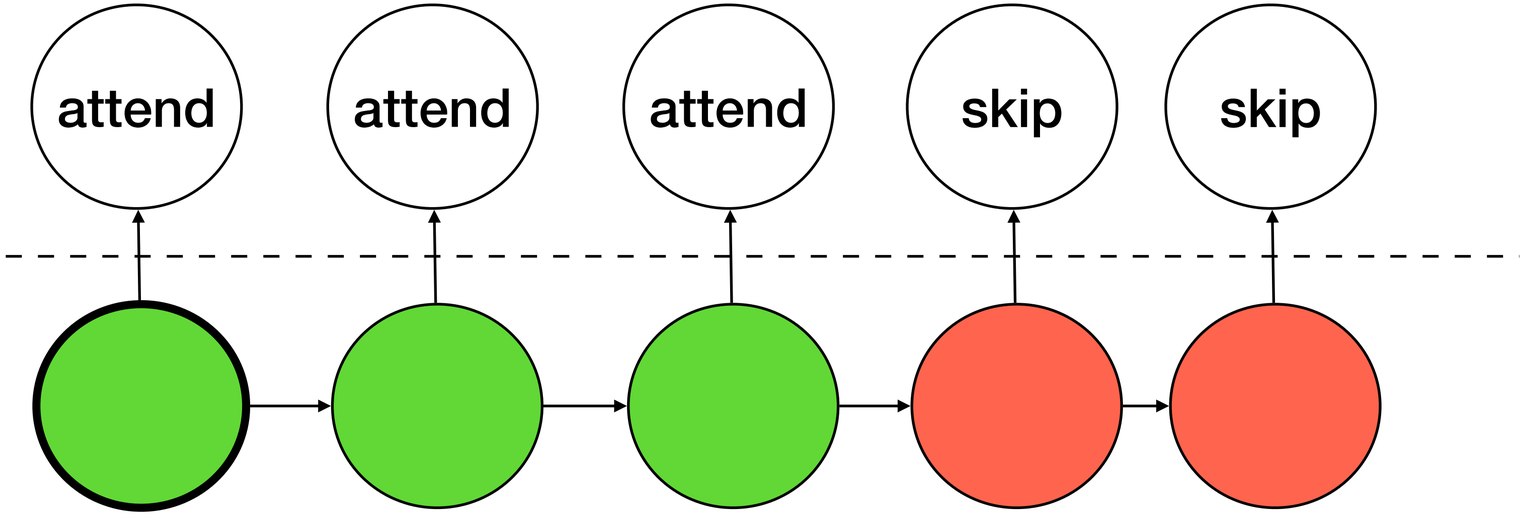
P(IH IH IH IU IU) = π(IH)P(attend | IH) x

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



P(IH IH IH IU IU) = (0.25 x 1.0) x

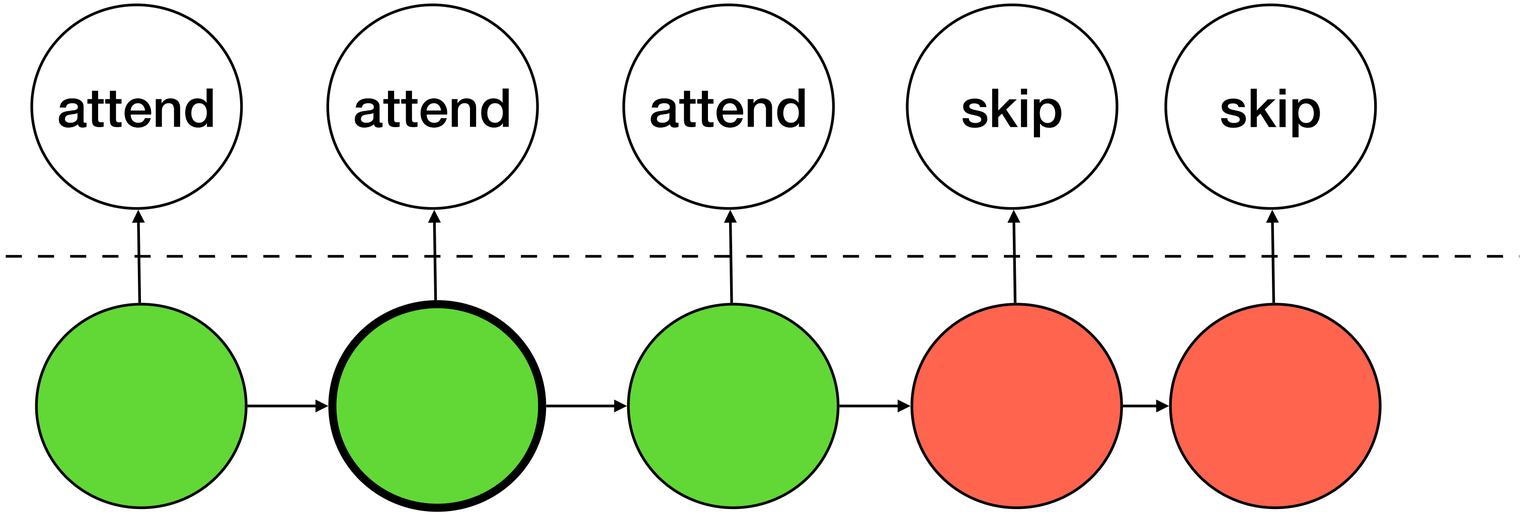
(assume uniform start probabilities)

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



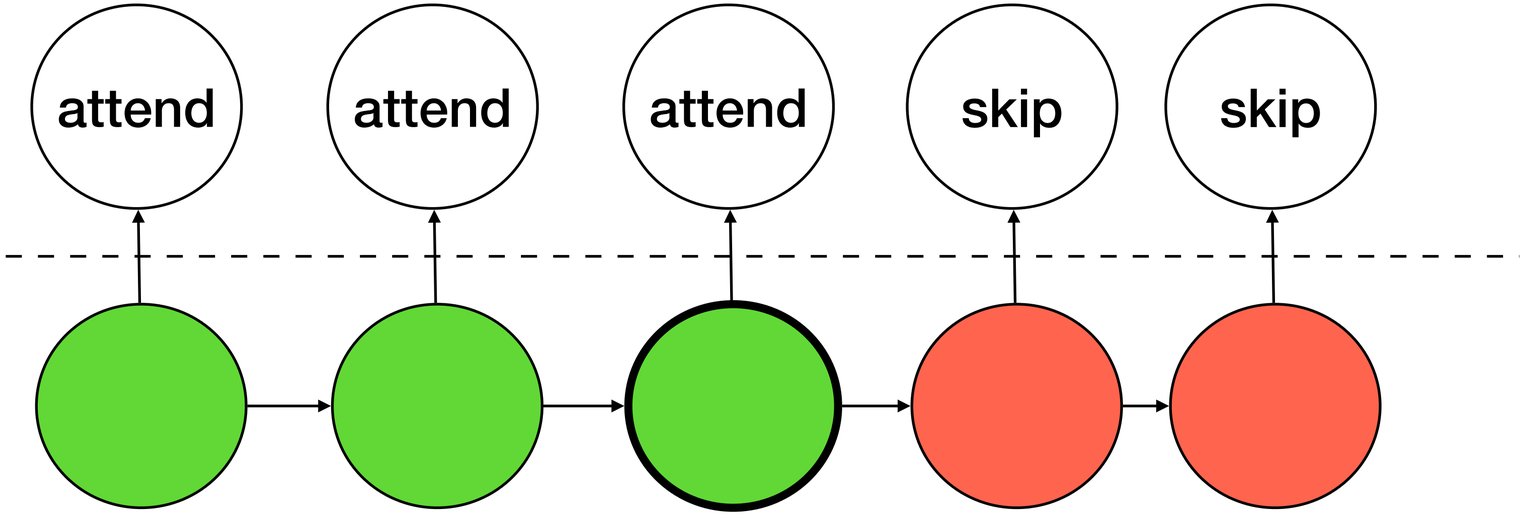
P(IH IH IH IU IU) = (0.25 x 1.0) x
(0.7 x 1.0) x

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



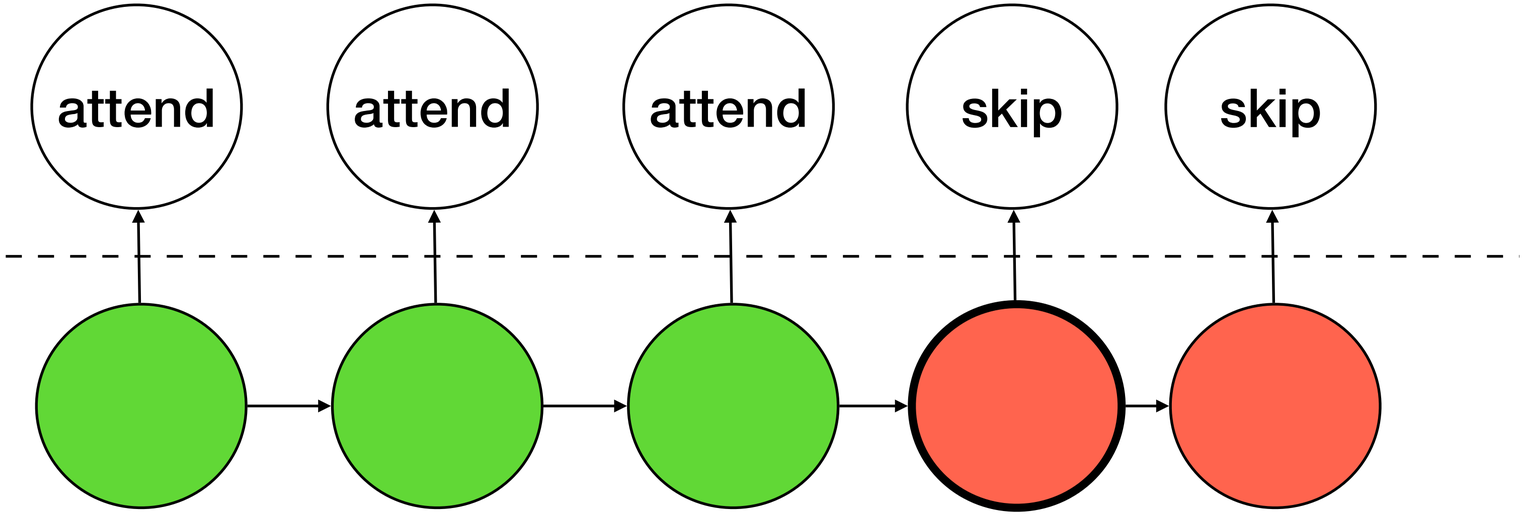
P(IH IH IH IU IU) = (0.25 x 1.0) x
(0.7 x 1.0) x
(0.7 x 1.0) x

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



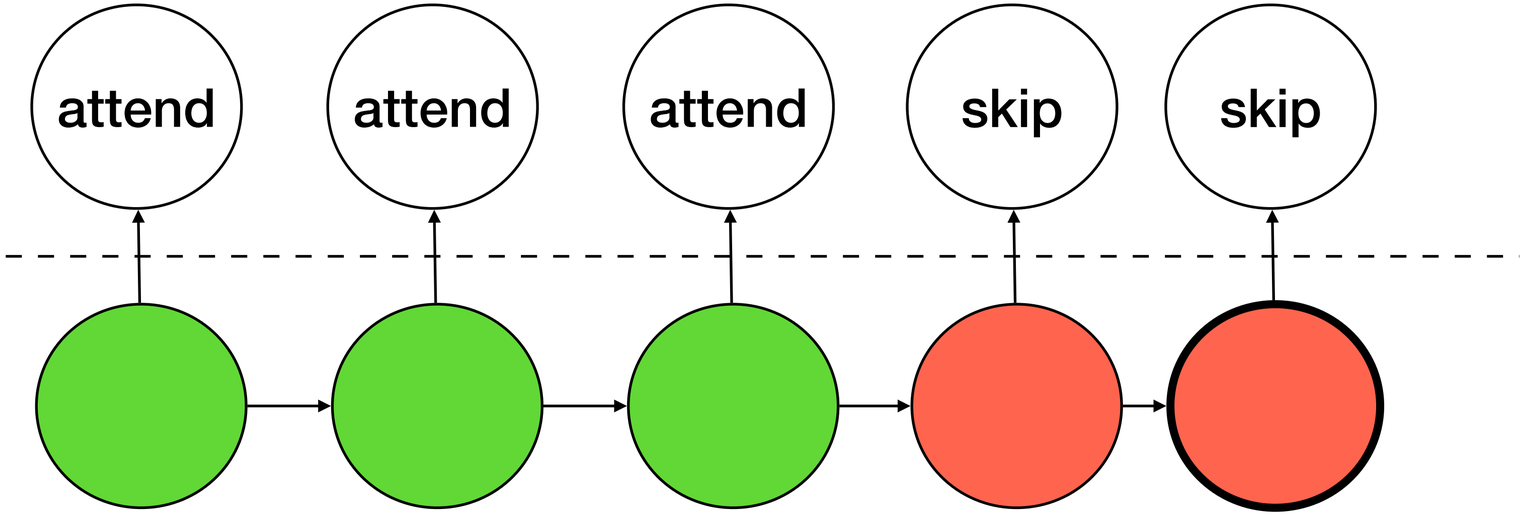
P(IH IH IH IU IU) = (0.25 x 1.0) x
(0.7 x 1.0) x
(0.7 x 1.0) x
(0.2 x 0.9) x

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



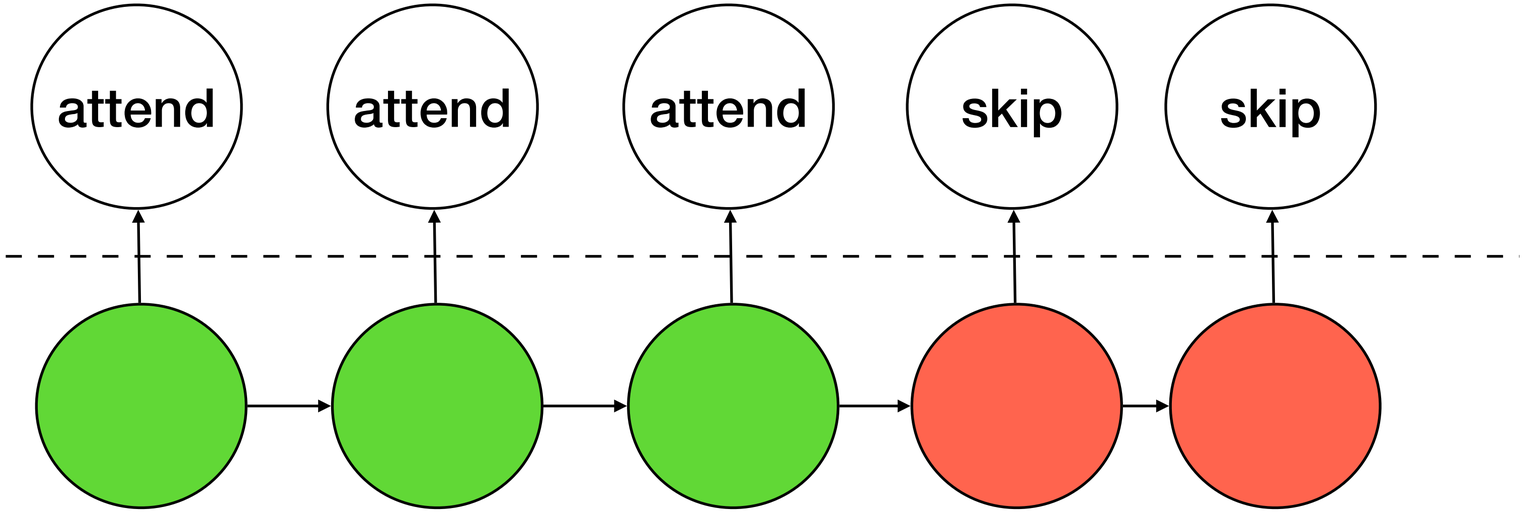
P(IH IH IH IU IU) = (0.25 x 1.0) x
(0.7 x 1.0) x
(0.7 x 1.0) x
(0.2 x 0.9) x
(0.5 x 0.9) x

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



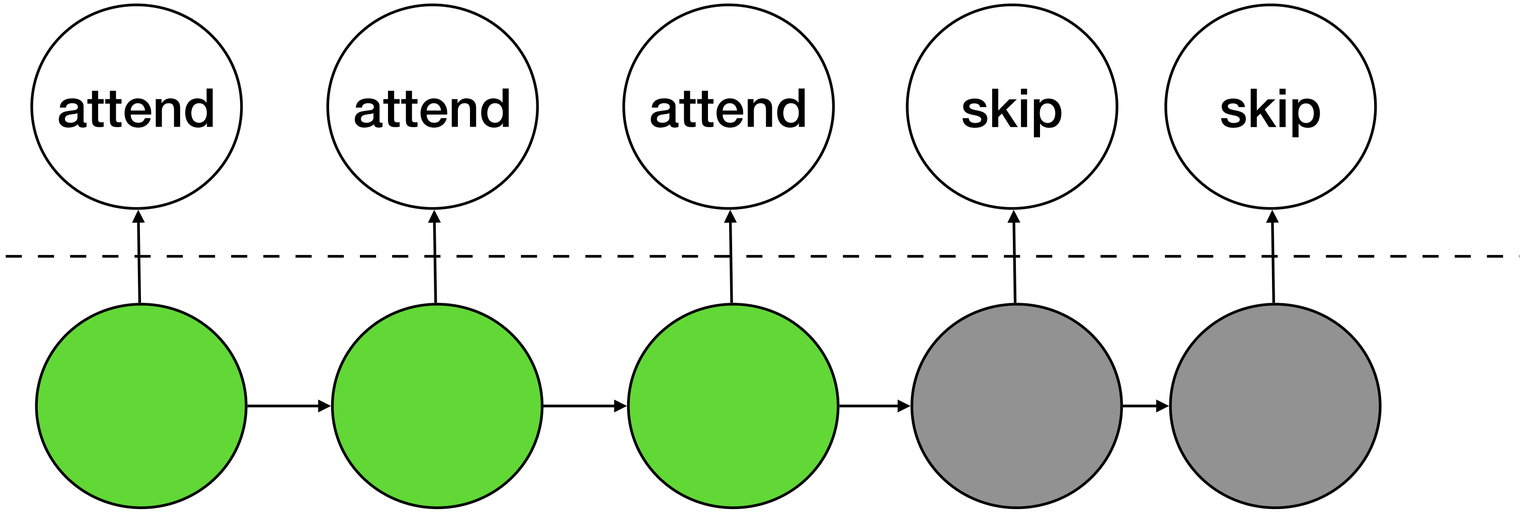
P(IH IH IH IU IU) = 0.0099

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



P(IH IH IH IU IU) = 0.0099

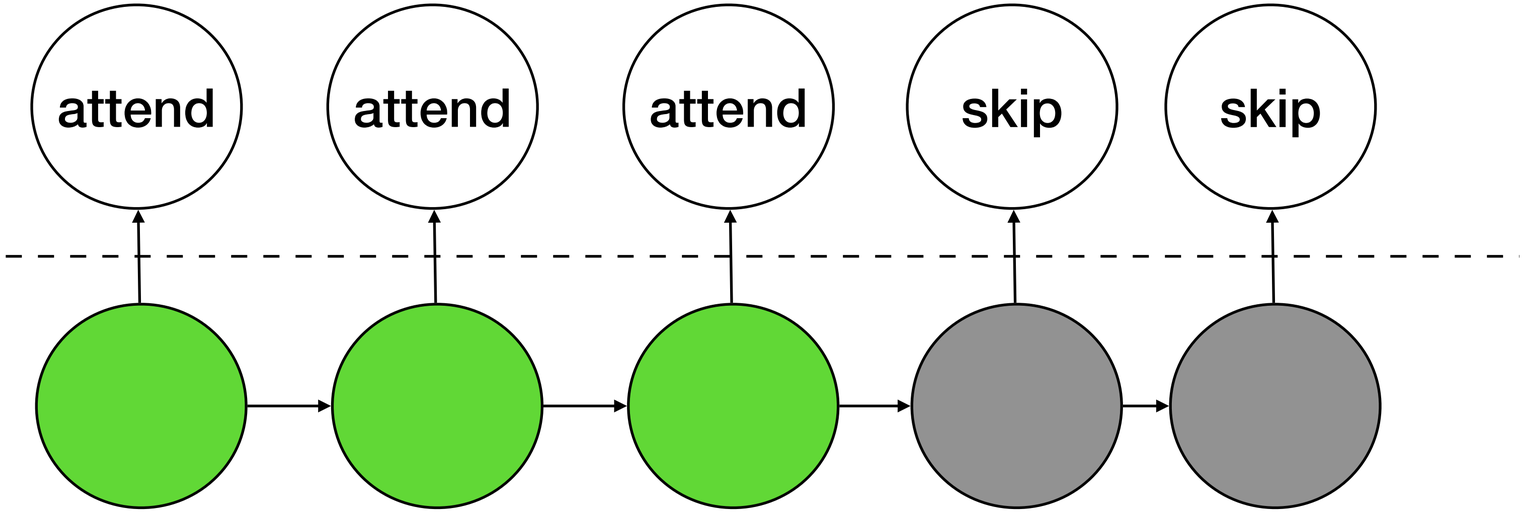
P(IH IH IH UH UH) = (0.25 x 1.0) x (0.7 x 1.0) x (0.7 x 1.0) x (0.1 x 0.8) x (0.7 x 0.8)

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

$$P(q_1 \ldots q_n) = \prod_{i}^{n} P(o_i \mid q_i) P(q_i \mid q_{i-1})$$



P(IH IH IH IU IU) = 0.0099

P(IH IH IH UH UH) = 0.0055

Emissions	attend	skip
interested, healthy	1.0	0.0
interested, unwell	0.1	0.9
uninterested, healthy	0.2	0.8
uninterested, unwell	0.0	1.0

Trans.	IH	IU	UH	UU
IH	0.7	0.2	0.1	0.0
IU	0.3	0.5	0.1	0.1
UH	0.1	0.0	0.7	0.2
UU	0.1	0.1	0.3	0.5

Hidden Markov Models

Part of Speech Tagging

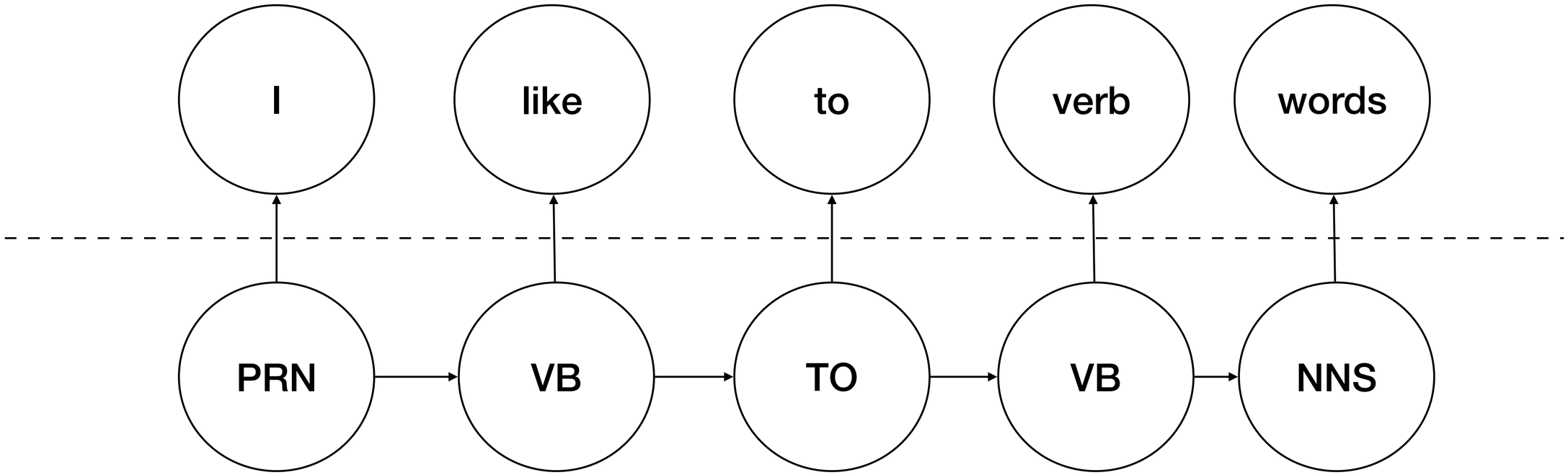
- In language, POS is *latent*, it is not observed
- So, we can use an HMM to find the most likely POS sequence for a sentence

Hidden Markov Models

Part of Speech Tagging

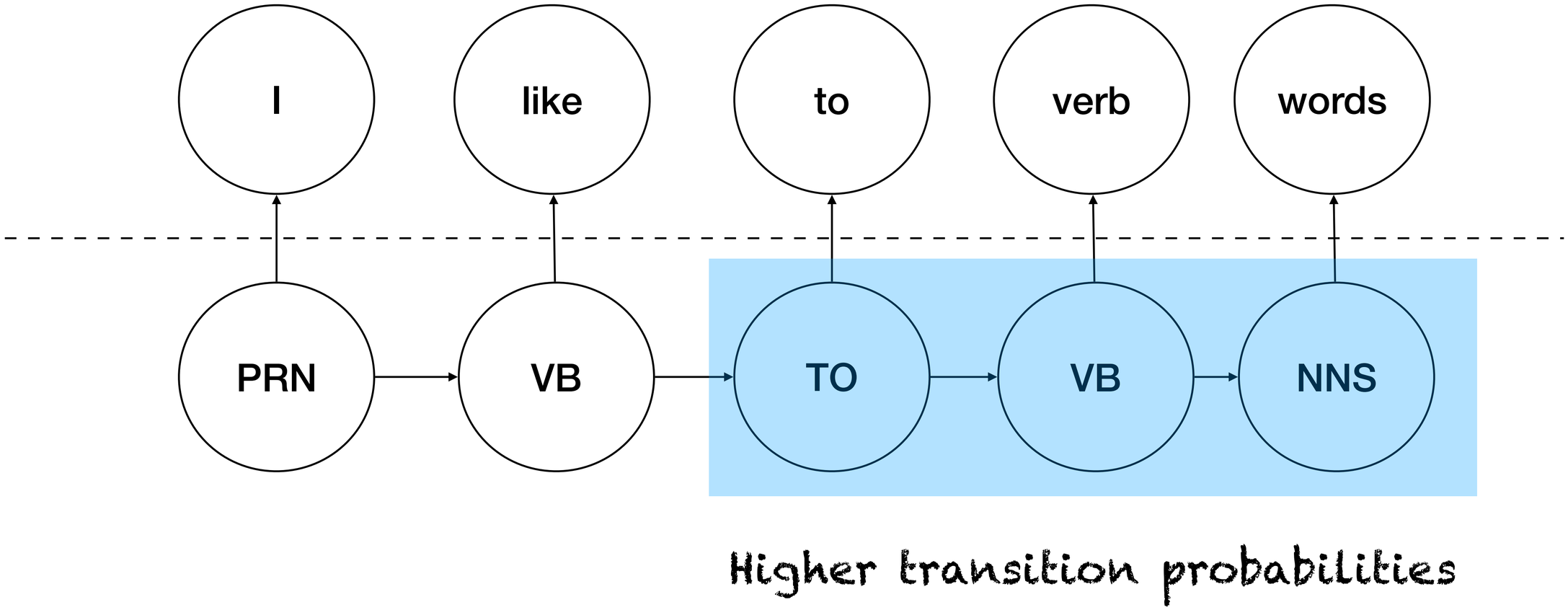
Hidden Markov Models

Part of Speech Tagging



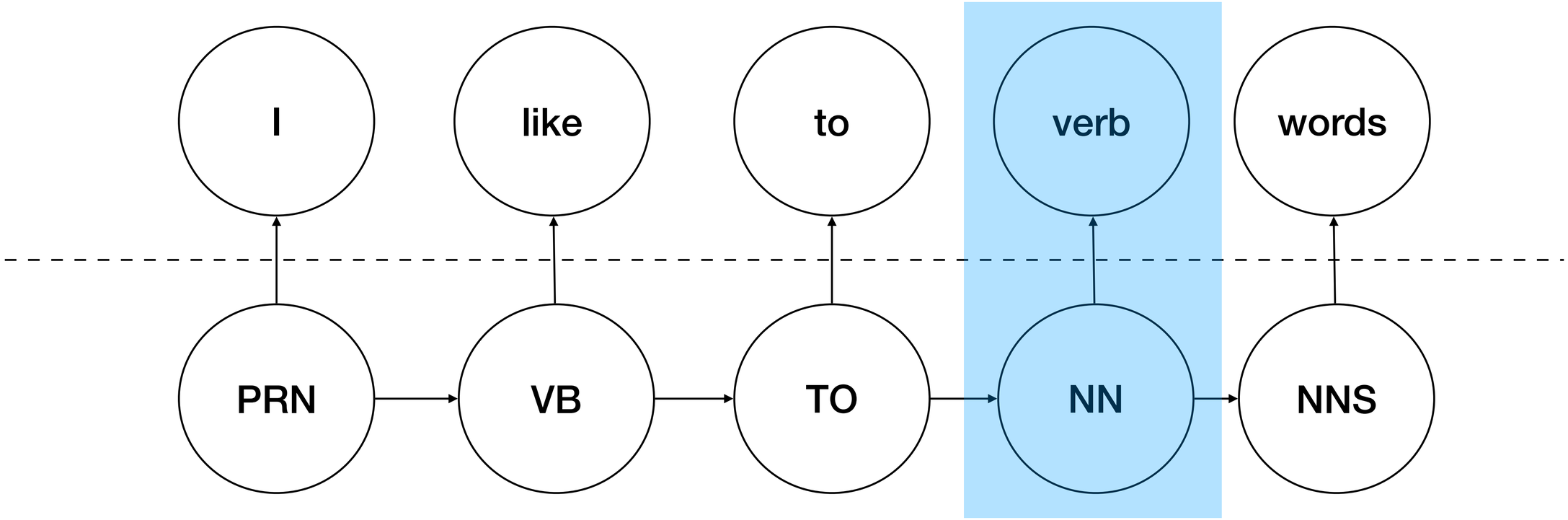
Hidden Markov Models

Part of Speech Tagging



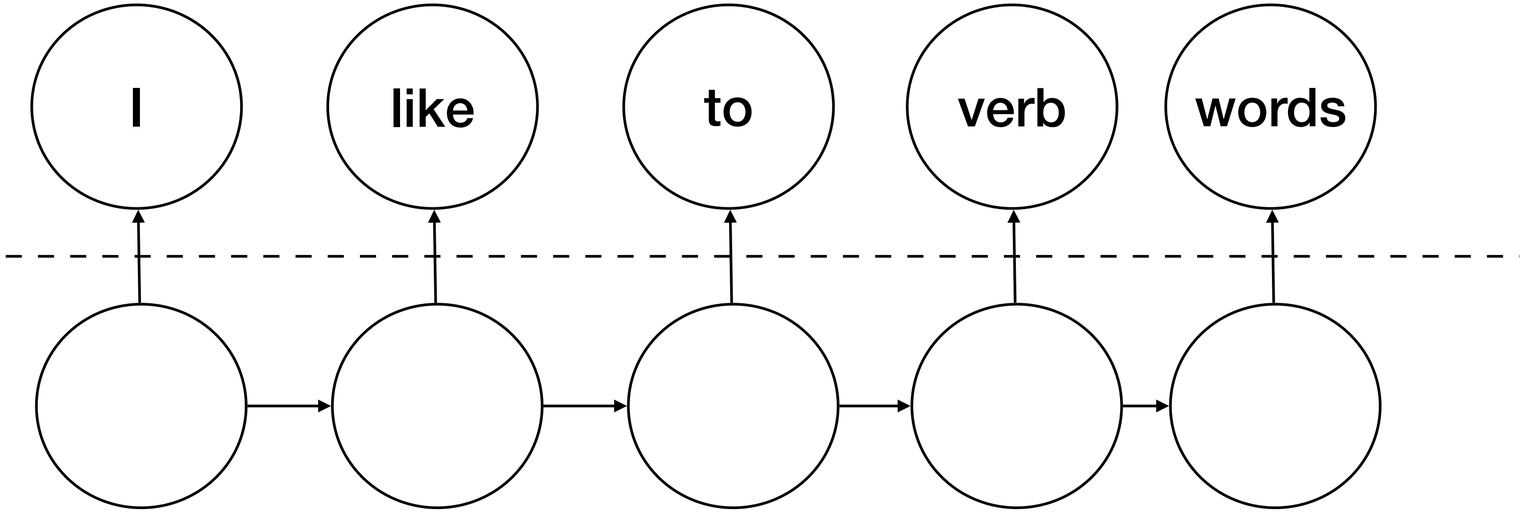
Hidden Markov Models

Part of Speech Tagging



Higher emission probabilities

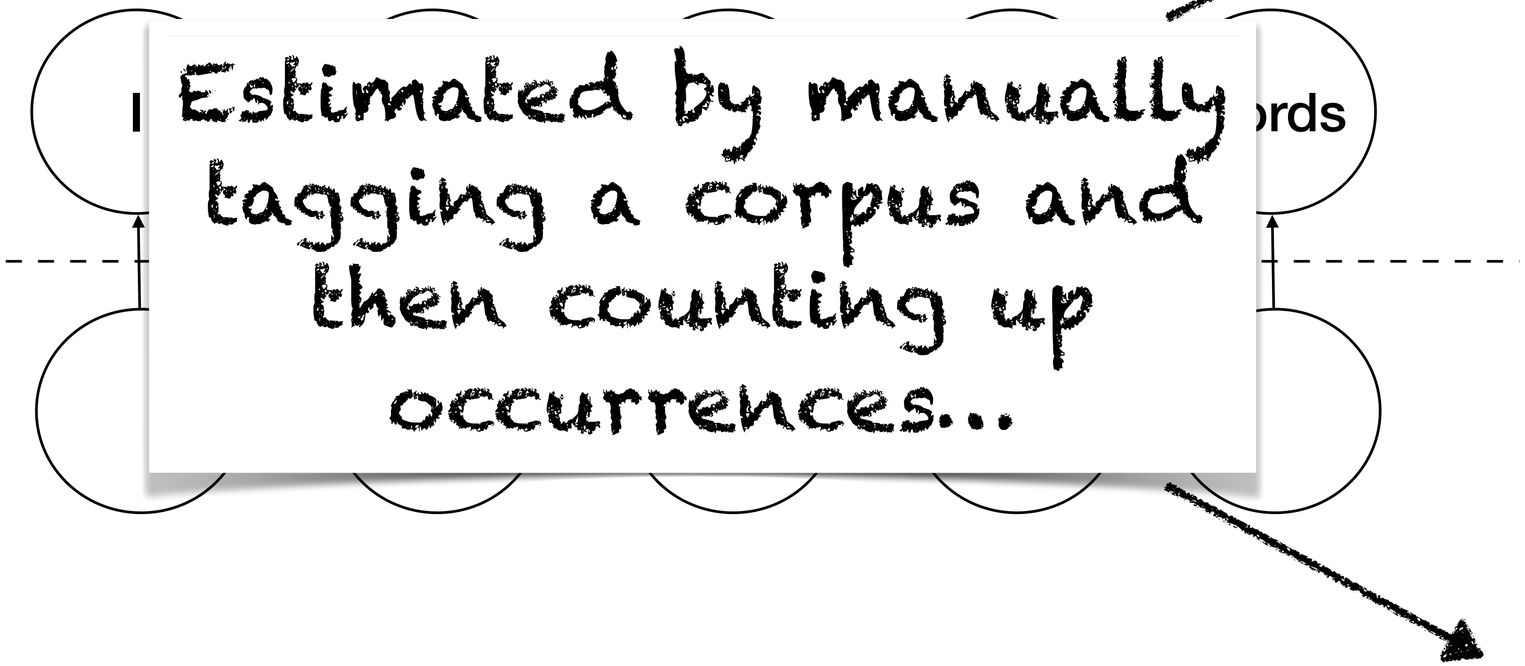
Hidden Markov Models



Emissions	aardvark	…	Zyzzyva
CC			
CD			
…			
WRB			

Trans.	CC	CD	…	WRB
CC				
CD				
…				
WRB				

Hidden Markov Models



Emissions	aardvark	…	Zyzzyva
CC			
CD			
…			
WRB			

Trans.	CC	CD	…	WRB
CC				
CD				
…				
WRB				

Hidden Markov Models

Viterbi Decoding

- We want to maximize P(oi|qi)P(qi|qi-1), but how to we actually do that?
- Like in MT—we call this ***decoding***: finding the most likely sequence obeying according to probability model
- ***Viterbi Algorithm*** is a dynamic programming algorithm for finding the most likely output sequece

Hidden Markov Models

Viterbi Decoding

- Maintain M: a matrix for #states x #outputs
- M[i][j] represents the probability of being in state i after seeing outputs 0…j (and having passed through the most probable sequence in order to get there)
- j = current state (tag)
- i = previous state (tag)
- t = current observation (word)
- A = transitions; B = emissions

$$M[j][t] = max_i^n M[i][t-1]\ A[i][j]\ B[j][t]$$

Transitions

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00	0.01	0.12	0.02	0.01
RB	0.01	0.01	0.10	0.10	0.01	0.07	0.05
DT	0.11	0.00	0.00	0.22	0.47	0.01	0.00

Emissions

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
RB			0.01		
DT				0.51	

Transitions

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00	0.01	0.12	0.02	0.01
RB	0.01	0.01	0.10	0.10	0.01	0.07	0.05
DT	0.11	0.00	0.00	0.22	0.47	0.01	0.00

Emissions

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
RB			0.01		
DT				0.51	

	Janet	will	back	the	bill
q7: DT					
q6: RB					
q5: NN					
q4: JJ					
q3: VB					
q2: MD					
q1: NNP					

Transitions

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00	0.01	0.12	0.02	0.01
RB	0.01	0.01	0.10	0.10	0.01	0.07	0.05
DT	0.11	0.00	0.00	0.22	0.47	0.01	0.00

Emissions

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
RB			0.01		
DT				0.51	

	Janet	will	back	the	bill
q7: DT	0.20x0				
q6: RB					
q5: NN					
q4: JJ					
q3: VB					
q2: MD					
q1: NNP					

P(DT|<s>)*P(Janet|DT)

Transitions

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00	0.01	0.12	0.02	0.01
RB	0.01	0.01	0.10	0.10	0.01	0.07	0.05
DT	0.11	0.00	0.00	0.22	0.47	0.01	0.00

Emissions

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
RB			0.01		
DT				0.51	

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN					
q4: JJ					
q3: VB					
q2: MD					
q1: NNP					

P(RB|<s>)*P(Janet|RB)

Transitions

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00	0.01	0.12	0.02	0.01
RB	0.01	0.01	0.10	0.10	0.01	0.07	0.05
DT	0.11	0.00	0.00	0.22	0.47	0.01	0.00

Emissions

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
RB			0.01		
DT				0.51	

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0				
q1: NNP	0.28 x 0.00003				

P(NNP|<s>)*P(Janet|NNP)

Transitions

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00	0.01	0.12	0.02	0.01
RB	0.01	0.01	0.10	0.10	0.01	0.07	0.05
DT	0.11	0.00	0.00	0.22	0.47	0.01	0.00

Emissions

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
RB			0.01		
DT				0.51	

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0				
q1: NNP	0.00001				

P(NNP|<s>)*P(Janet|NNP)

Transitions (A)

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00	0.01	0.12	0.02	0.01
RB	0.01	0.01	0.10	0.10	0.01	0.07	0.05
DT	0.11	0.00	0.00	0.22	0.47	0.01	0.00

Emissions (B)

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
RB			0.01		
DT				0.51	

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0				
q1: NNP	0.00001				

Transitions (A)

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00				
RB	0.01	0.01	0.10				
DT	0.11	0.00	0.00				

Emissions (B)

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
				0.51	

$$M[j][t] = max_i^n M[i][t-1]\ A[i][j]\ B[j][t]$$

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0				
q1: NNP	0.00001				

Transitions (A)

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00				
RB	0.01	0.01	0.10				
DT	0.11	0.00	0.00				

Emissions (B)

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
				0.51	

$$M[j][t] = max_i^n M[i][t-1]\ A[i][j]\ B[j][t]$$

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0				
q1: NNP	0.00001				

cell[DT][t-1] x P(MD|DT) x P(will|MD)

Transitions (A)

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00				
RB	0.01	0.01	0.10				
DT	0.11	0.00	0.00				

Emissions (B)

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
RB					
DT				0.51	

$$M[j][t] = max_i^n M[i][t-1]\ A[i][j]\ B[j][t]$$

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0				
q1: NNP	0.00001				

cell[DT][t-1] x P(MD|DT) x P(will|MD)

cell[RB][t-1] x P(MD|RB) x P(will|RB)

.

.

.

cell[NNP][t-1] x P(MD|NNP) x P(will|MD)

Transitions (A)

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00				
RB	0.01	0.01	0.10				
DT	0.11	0.00	0.00				

Emissions (B)

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
				0.51	

$$M[j][t] = max_i^n M[i][t-1]\ A[i][j]\ B[j][t]$$

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0				
q1: NNP	0.00001				

0 x 0.002 x 0.31

0 x 0.007 x 0.31

.

.

.

0.00001 x 0.01 x 0.31

Transitions (A)

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00				
RB	0.01	0.01	0.10				
DT	0.11	0.00	0.00				

Emissions (B)

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
				0.51	

$$M[j][t] = max_i^n M[i][t-1]\ A[i][j]\ B[j][t]$$

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0	2.7×10^{-8}			
q1: NNP	0.00001				

~~0 x 0.002 x 0.31~~

~~0 x 0.007 x 0.31~~

.

.

.

0.00001 x 0.01 x 0.31

Transitions (A)

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00				
RB	0.01	0.01	0.10				
DT	0.11	0.00	0.00				

Emissions (B)

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
				0.51	

$$M[j][t] = max_i^n M[i][t-1]\ A[i][j]\ B[j][t]$$

	Janet	will	back	the	bill
q7: DT	0				~~0 x 0.002 x 0.31~~
q6: RB	0				~~0 x 0.007 x 0.31~~
q5: NN	0				.
q4: JJ	0				.
q3: VB	0				. 0.00001 x 0.01 x 0.31
q2: MD	0	2.7×10^{-8}			
q1: NNP	0.00001				

Transitions (A)

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.28	0.00	0.00	0.04	0.04	0.05	0.20
NNP	0.38	0.01	0.00	0.01	0.06	0.00	0.00
MD	0.00	0.00	0.80	0.00	0.00	0.17	0.00
VB	0.03	0.00	0.00	0.08	0.06	0.05	0.22
JJ	0.04	0.00	0.00	0.07	0.45	0.00	0.00
NN	0.00	0.02	0.00				
RB	0.01	0.01	0.10				
DT	0.11	0.00	0.00				

Emissions (B)

	Janet	will	back	the	bill
NNP	3.2×10^{-5}			4.8×10^{-5}	
MD		0.31			
VB		2.8×10^{-5}	6.7×10^{-4}		2.8×10^{-5}
JJ			3.4×10^{-4}		
NN		2.0×10^{-4}	2.0×10^{-4}		2.0×10^{-3}
				0.51	

$$M[j][t] = max_i^n M[i][t-1]\ A[i][j]\ B[j][t]$$

	Janet	will	back	the	bill
q7: DT	0				
q6: RB	0				
q5: NN	0				
q4: JJ	0				
q3: VB	0				
q2: MD	0	2.7×10^{-8}			
q1: NNP	0.00001				

Probability of "best case scenario" that would result in word j having state i

Topics

- Follow up: Multlingual LMs and Zero-Shot
- Part of Speech Tags
- Hidden Markov Models
- **Conditional Random Fields**

Sequence Labeling Tasks

- cah
- You’re being a cah!
- She’s cahing at the camera
- You *would* have a cah daughter

Sequence Labeling Tasks

- cah
- You're being a cah!
- She's cahing at the camera
- You *would* have a cah daughter

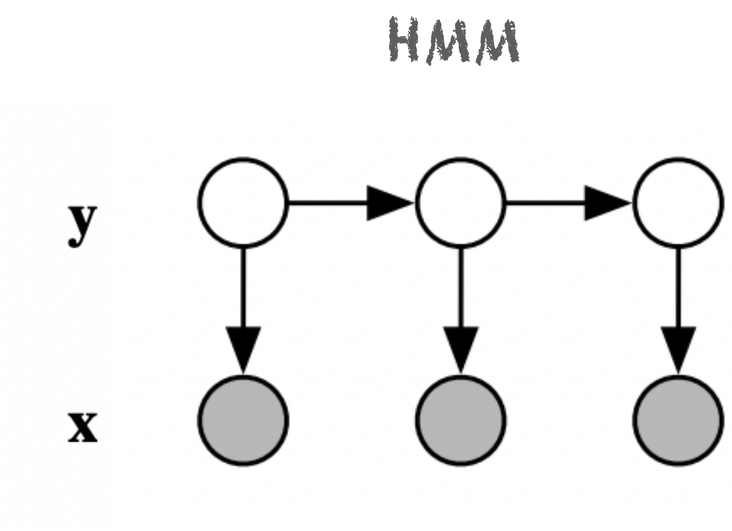
Conditional Random Fields (CRFs)

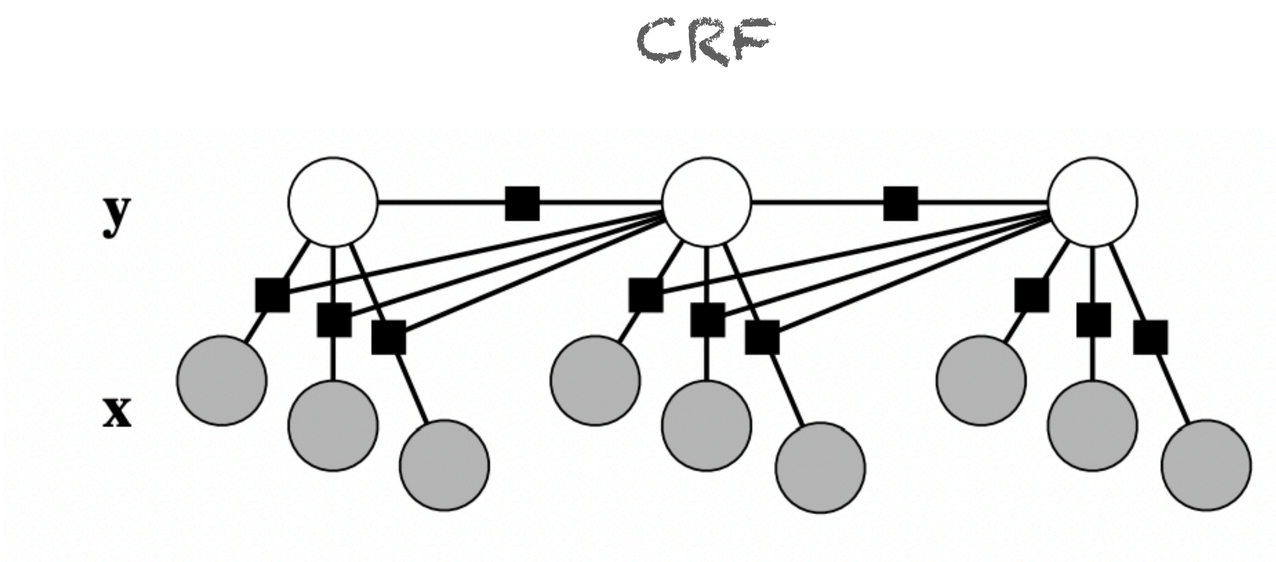
- HMMs struggle with unknown words
  - She's cahing at the camera
    - {ends-with-ing, occurs-before-"at", occurs-after-"'s", …}
- Hard to incorporate all these features into an HMM
  - (Need to either model dependencies between all Xs or else make extreme independence assumptions—like in NB—which will lead to bad performance)

Conditional Random Fields (CRFs)

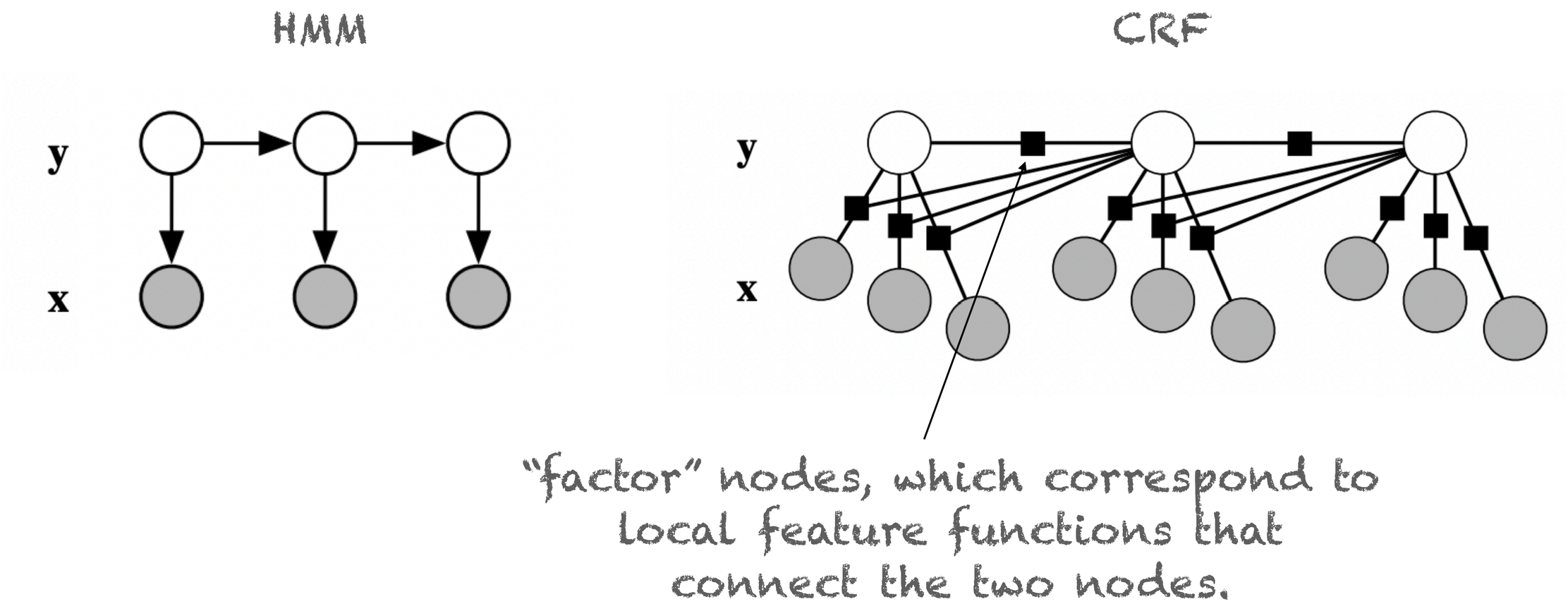
- Instead: Use discriminative (log-linear) models
  - Recap:
    - Generative models joint distribution $P(X,Y)$
    - Discriminative models the conditional distribution $P(Y|X)$ directly
  - So: Naive Bayes : Logistic Regression :: HMM : CRF

Conditional Random Fields (CRFs)





Conditional Random Fields (CRFs)



All done!
Questions?