ECE594: Mathematical Models of Language

Spring 2022

Lecture 6: Sequence-level Models and Low-Resource NLP

Slide credits to Graham Neubig

Logistics

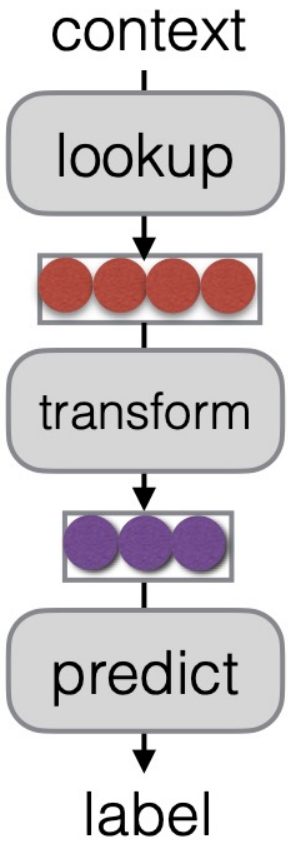
- Presentation slots created
- Assignment 2 out
  - due 2/25
  - post issues on Piazza
  - Start early!

From Words to Word Sequences

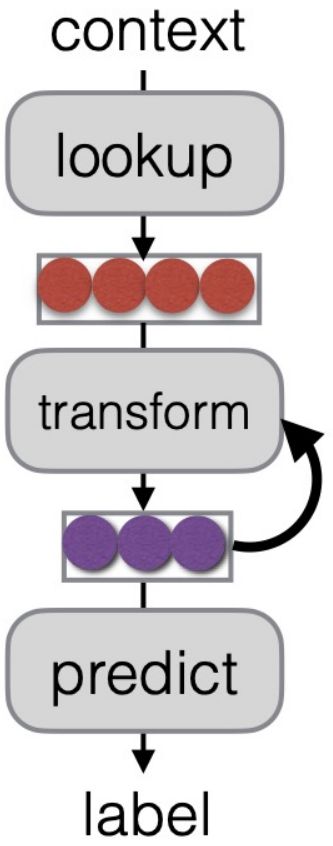
- Words as units of text
  - Word level models for text classification
- Relations between words
  - Word meaning and similarity
- Words as sequences
  - Language modeling

Recurrent Neural Networks (Elman 1990)

Feed-forward NN

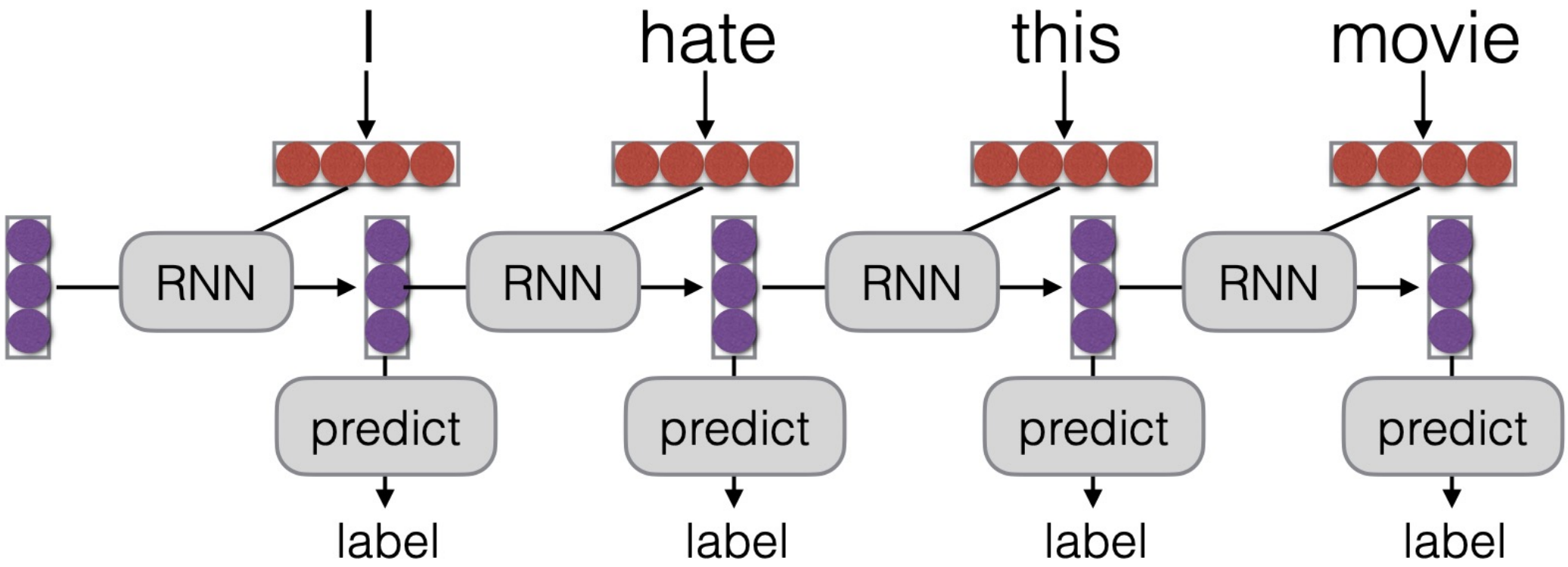


Recurrent NN

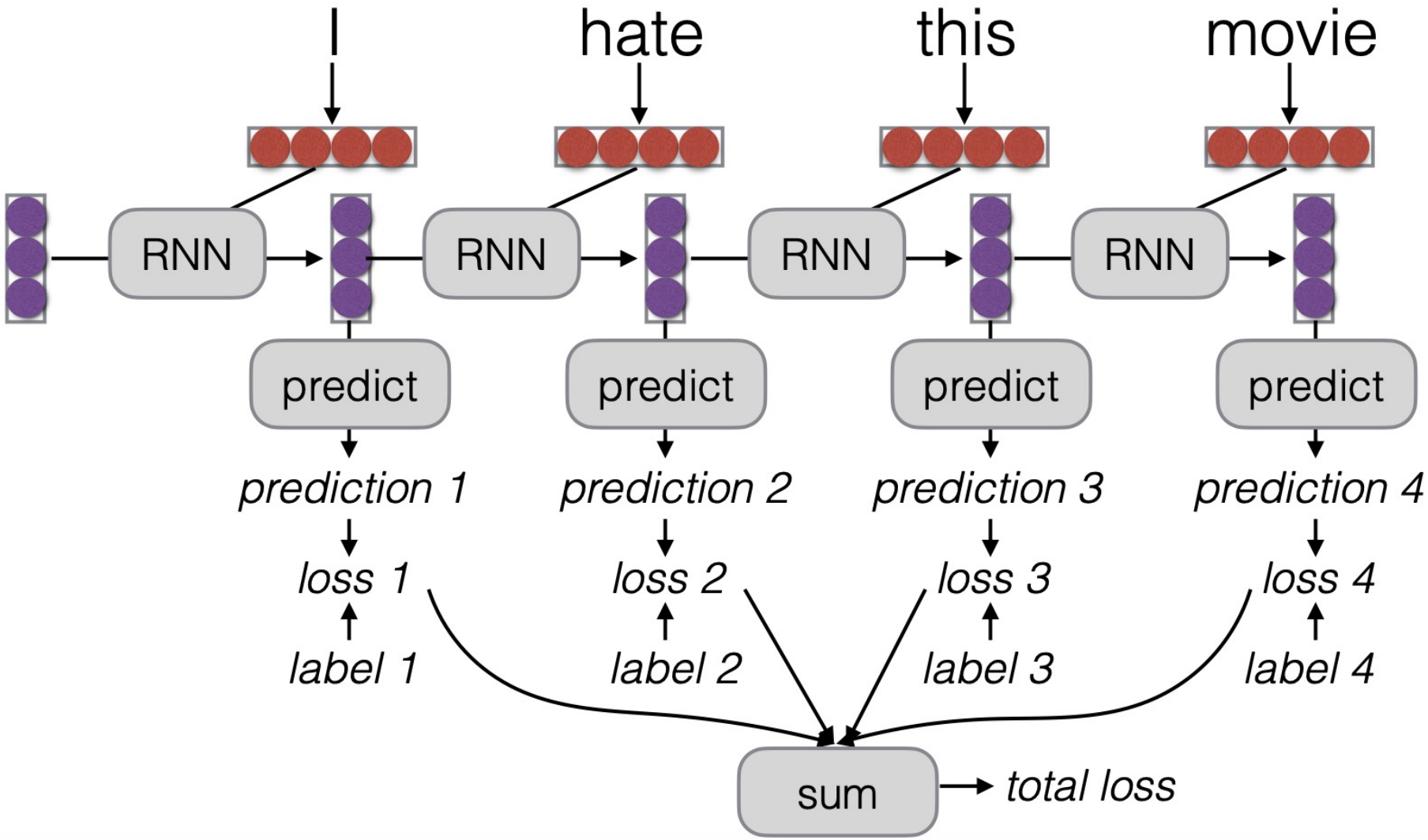


Recurrent Neural Networks (Elman 1990)

- What does processing a sequence look like?



RNN Training



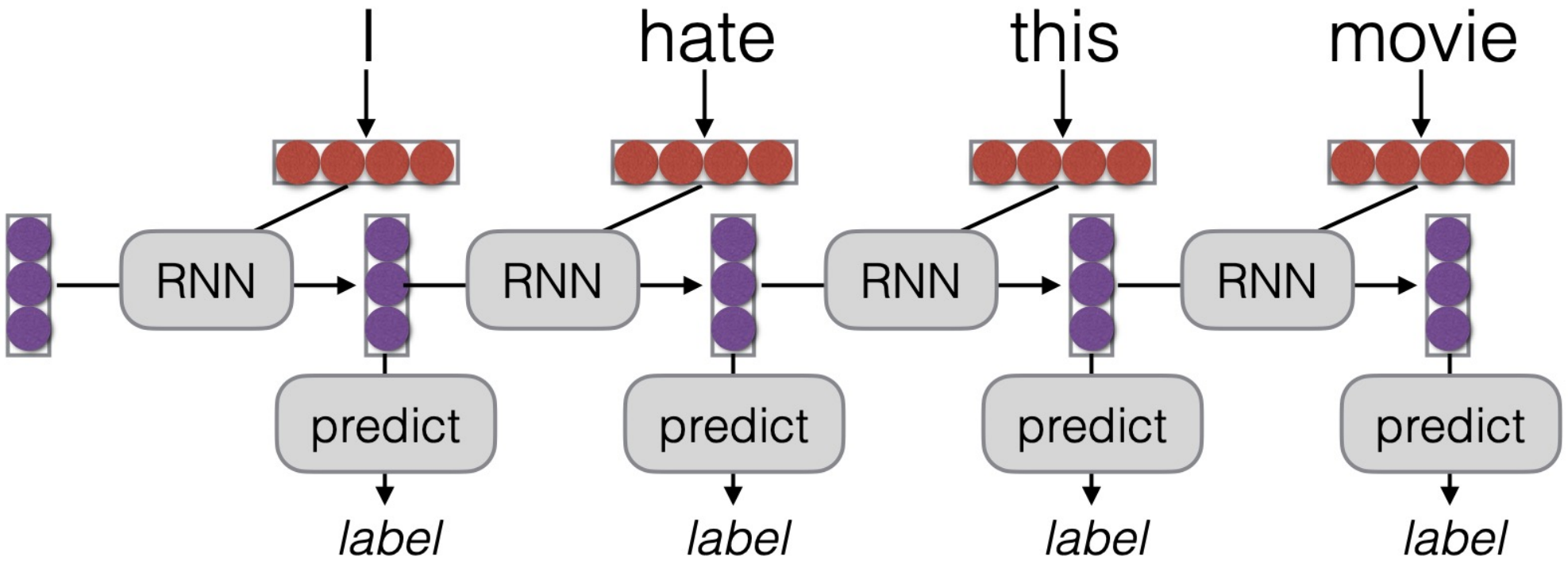
RNN Advantage

- Represent a sentence
  - Read whole sentence, make a prediction
- Represent a context within a sentence
  - Read context up until that point

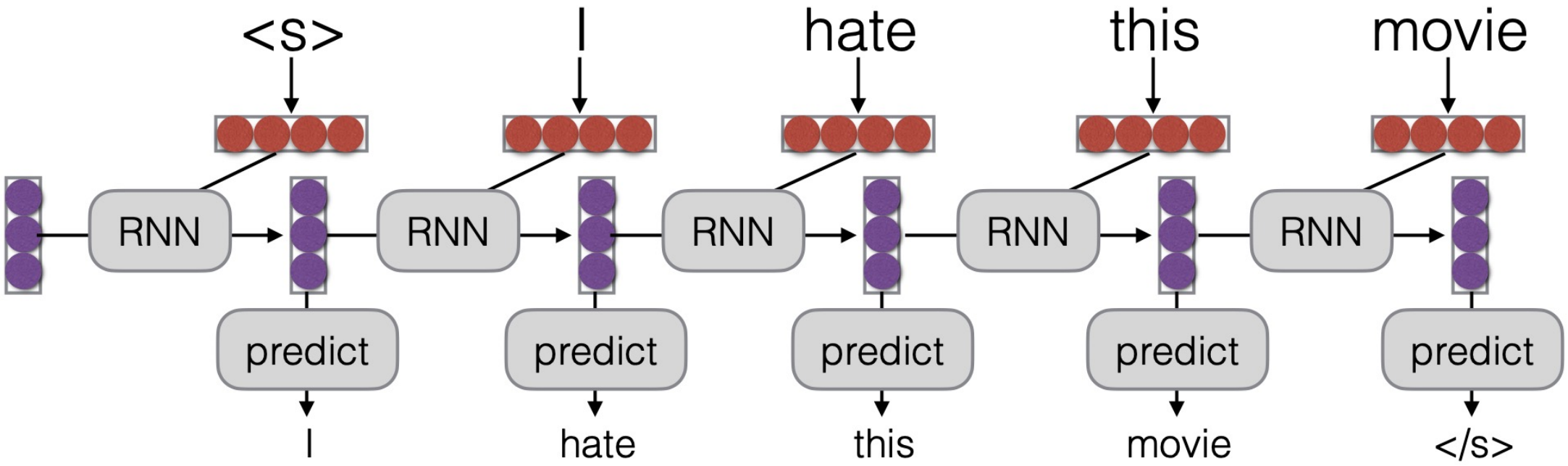
RNN Advantage

- Represent a sentence
  - Read whole sentence, make a prediction
- Represent a context within a sentence
  - Read context up until that point

Represent Contexts



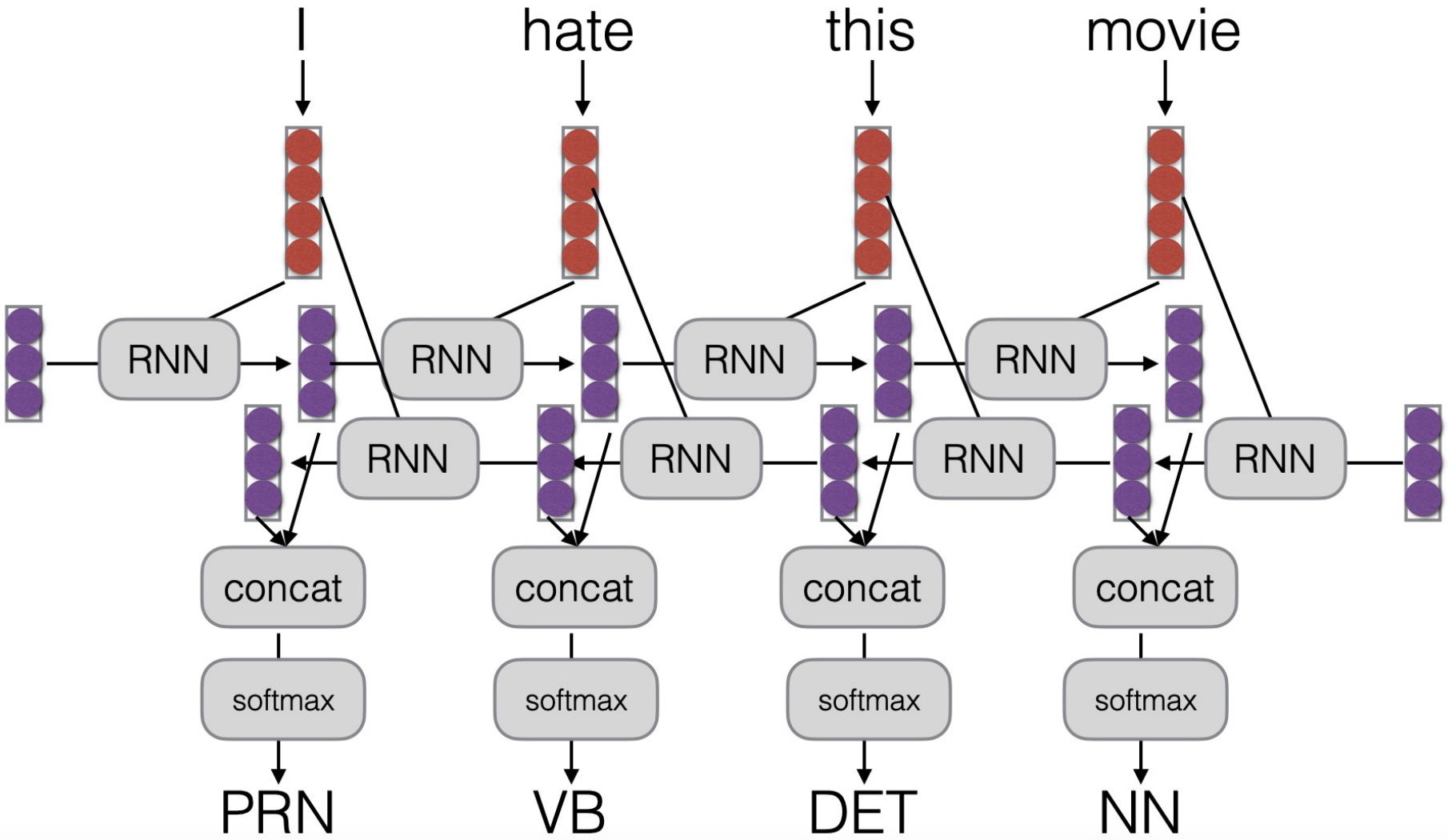
Represent Contexts: Language Modeling



- Language modeling is like a tagging task, where each tag is the next word!

Bidirectional RNNs

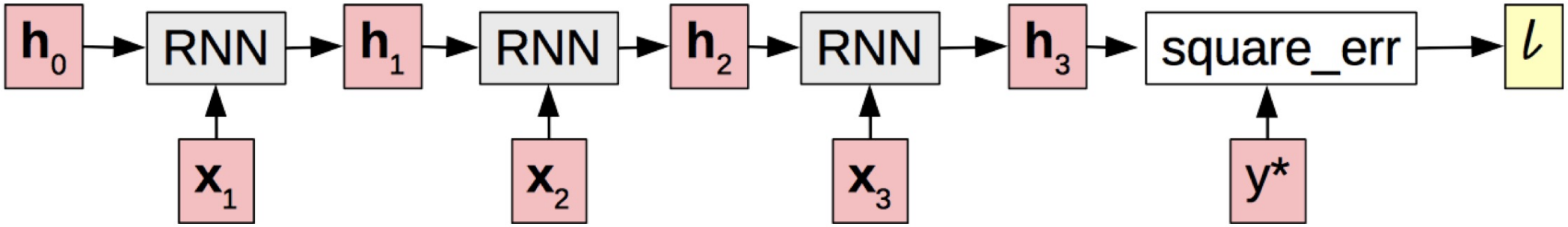
- A simple extension, run the RNN in both directions



Vanishing Gradients

- Gradients decrease as they get pushed back

$$\frac{dl}{d_{h_0}} = \text{tiny} \qquad \frac{dl}{d_{h_1}} = \text{small} \qquad \frac{dl}{d_{h_2}} = \text{med.} \qquad \frac{dl}{d_{h_3}} = \text{large}$$



- Why? “Squashed” by non-linearities or small weights in matrices.

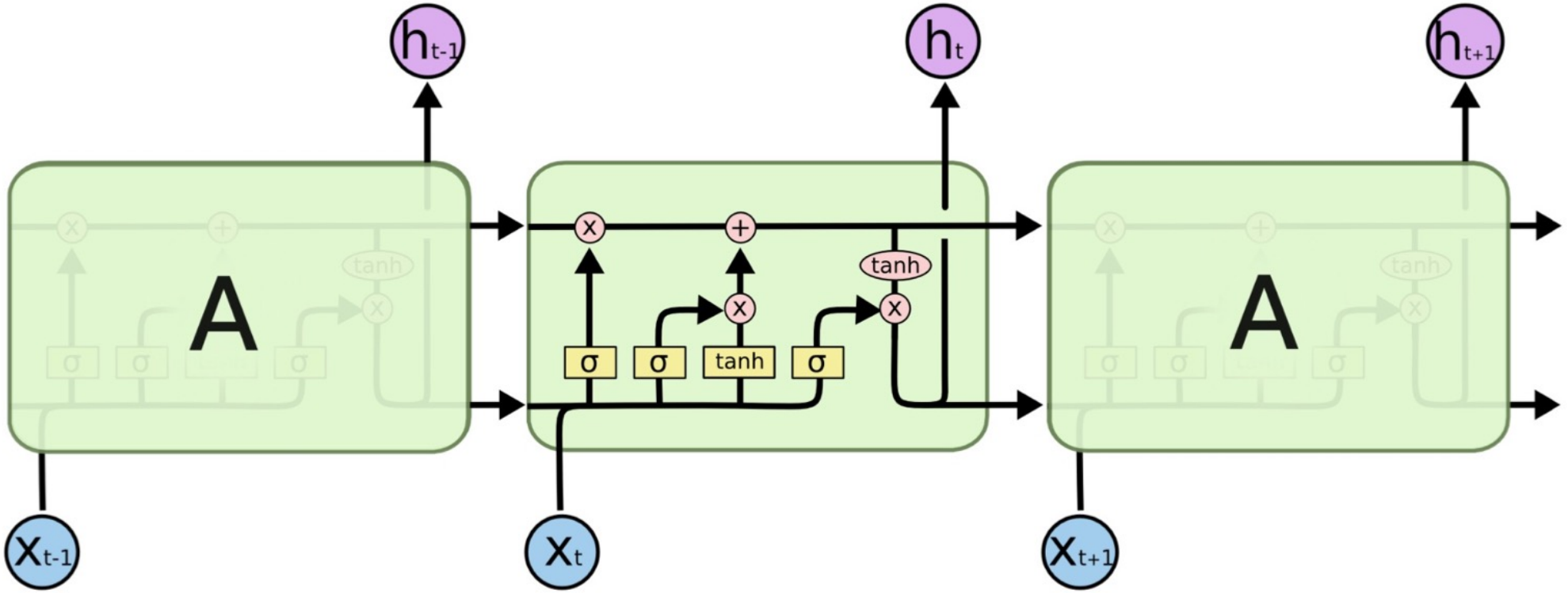
Long Short-Term Memory

(Hochreiter and Schmidhuber 1997)

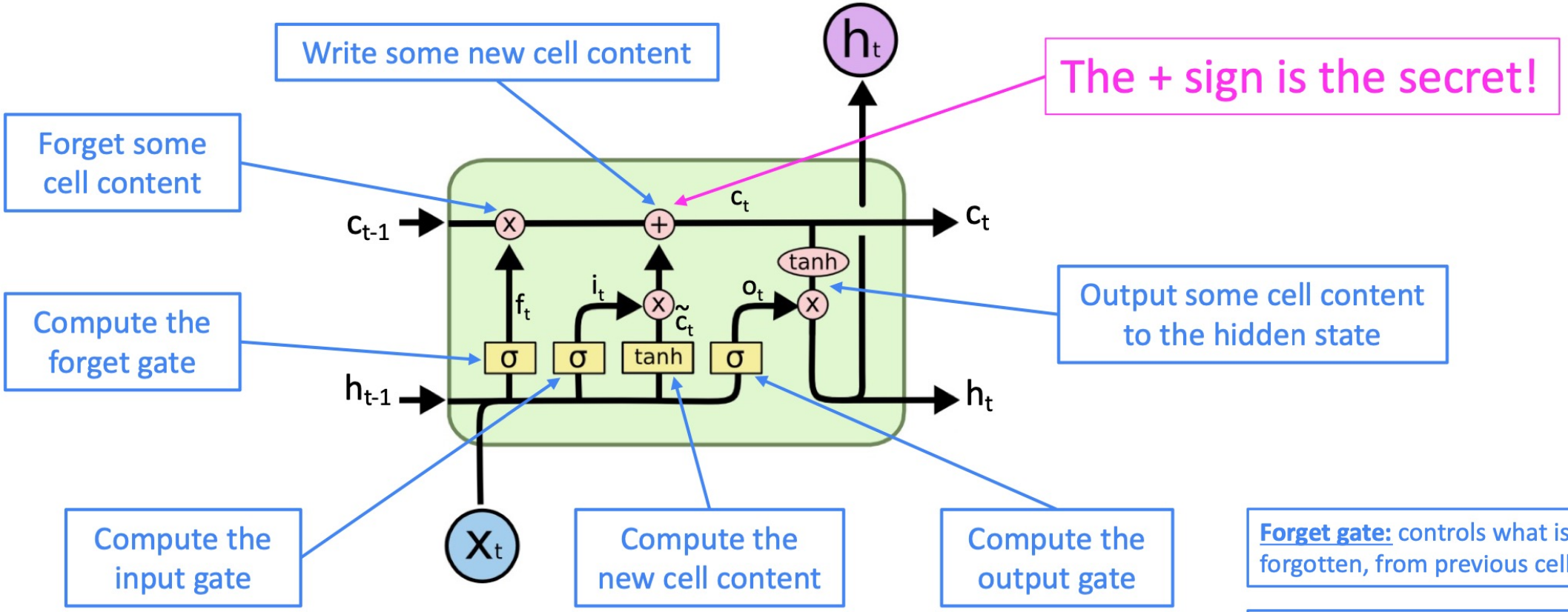
Key idea:

- Additive connections between time steps
- Addition of gradients solves vanishing gradient
- Control information flow using gates

LSTM Structure



LSTM Structure: An RNN



Forget gate: controls what is kept vs forgotten, from previous cell state

Input gate: controls what parts of the new cell content are written to cell

Output gate: controls what parts of cell are output to hidden state

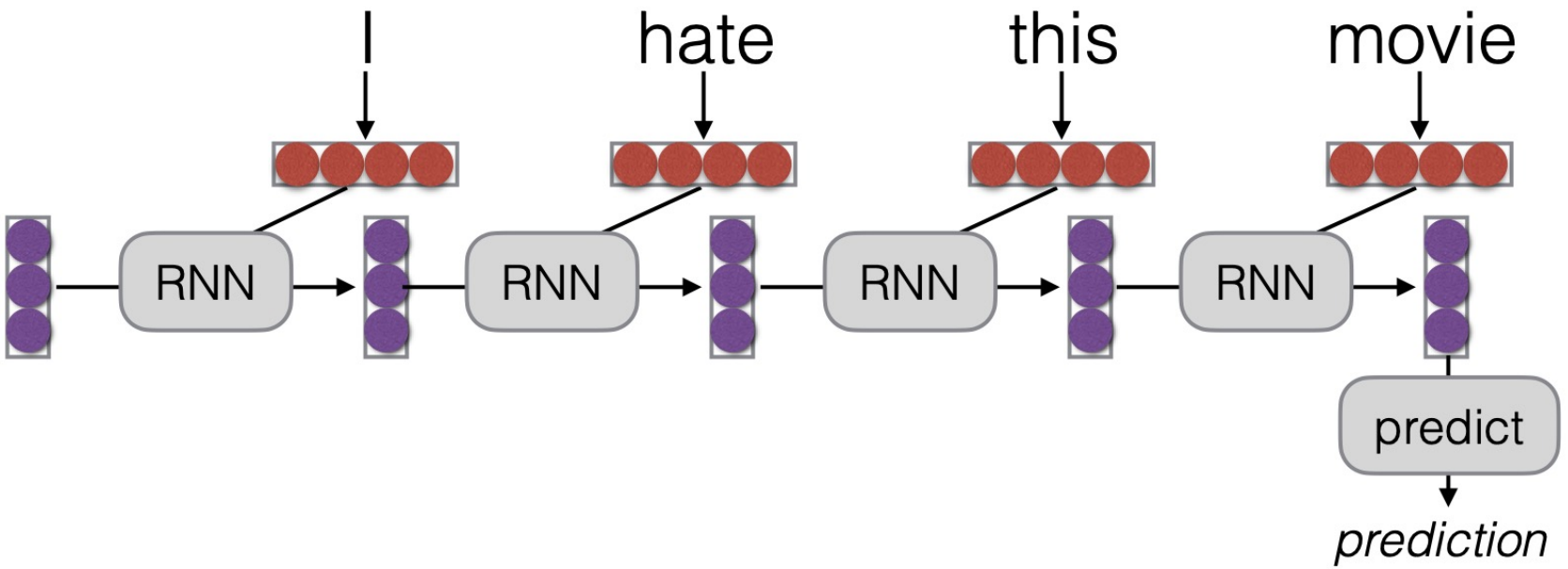
Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM and GRUs

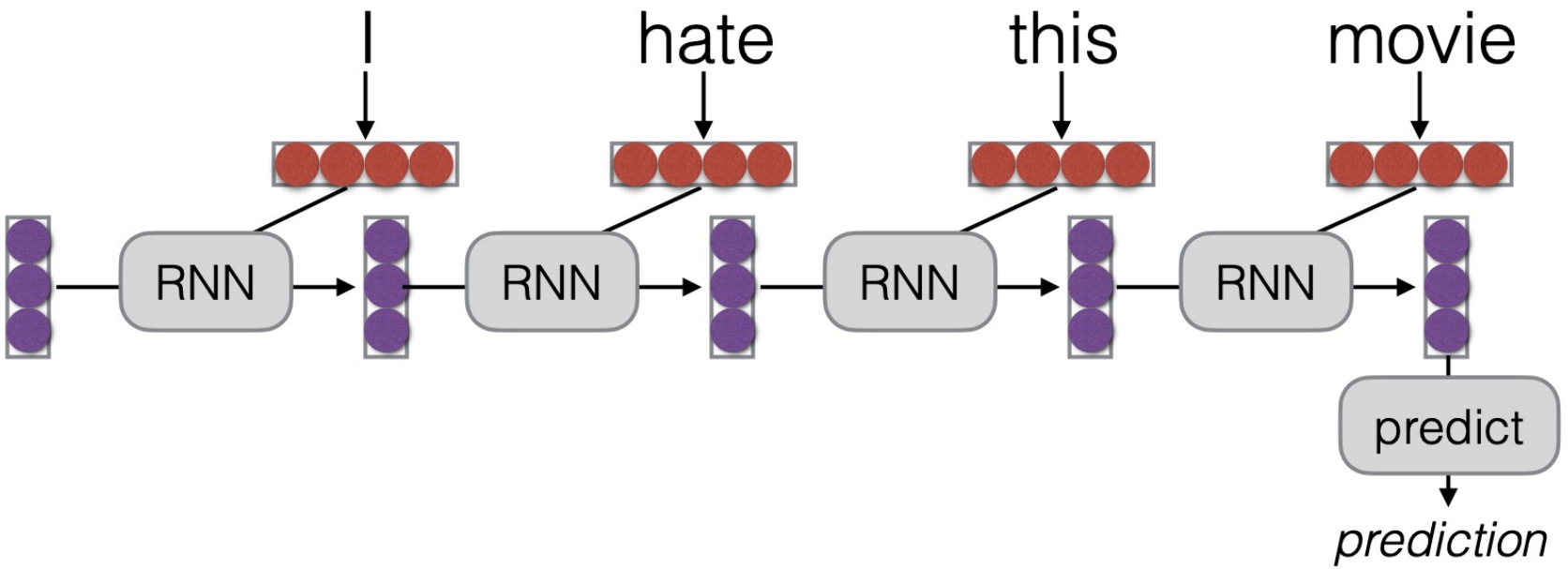
- Many gated RNN variants
  - LSTM and GRU are most widely-used

  - Rule of thumb:
    - LSTM is a good default choice
    - Lots of data, especially long dependencies
    - Switch to GRUs for speed and fewer parameters

Represent Sentences



Represent Sentences



- Sentence classification
- Conditioned generation
- Retrieval

Conditioned Generation

- Generate text according to some specification

Input *X*	Output *Y* (**Text**)	Task
Structured Data	NL Description	NL Generation
English	Japanese	Translation
Document	Short Description	Summarization
Utterance	Response	Response Generation
Image	Text	Image Captioning
Speech	Transcript	Speech Recognition

Conditioned Generation

- Generate text according to some specification; conditional language model

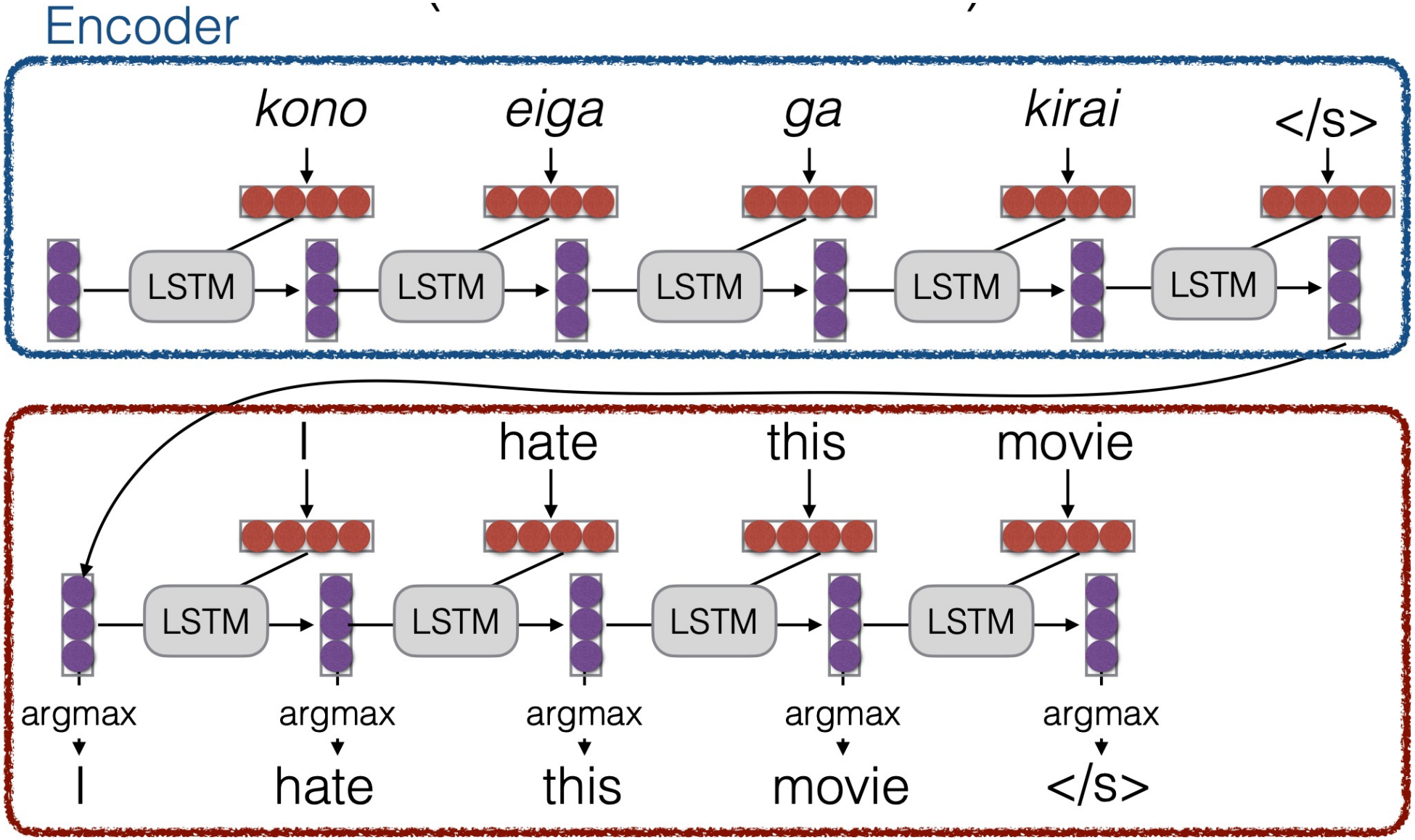
$$P(X) = \prod_{i=1}^{I} P(x_i \mid x_1, \ldots, x_{i-1})$$

Next Word Context

$$P(Y|X) = \prod_{j=1}^{J} P(y_j \mid X, y_1, \ldots, y_{j-1})$$

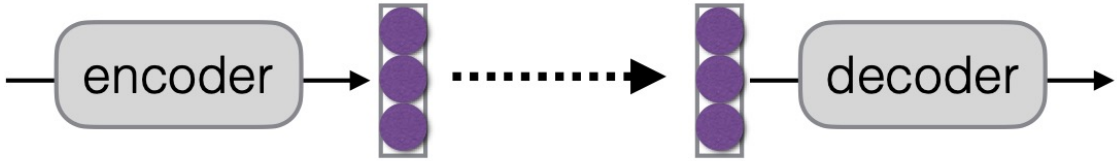
Added Context!

Conditioned Generation



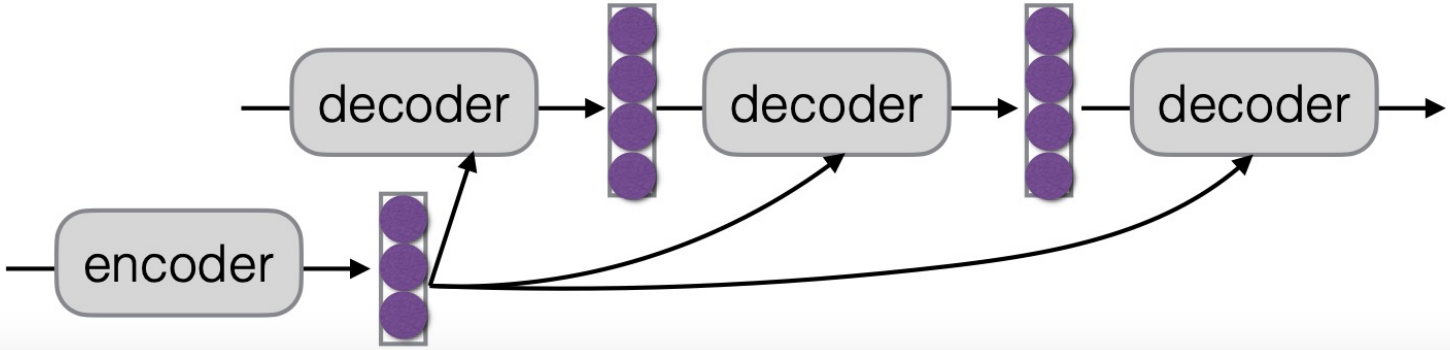
Passing Hidden State

- Initialize decoder w/ encoder (Sutskever et al. 2014)



- Transform (can be different dimensions)

- Input at every time step (Kalchbrenner & Blunsom 2013)



Beyond RNNs

- In 2013–2015, LSTMs started achieving state-of-the-art results
  - Successful tasks include handwriting recognition, speech recognition, machine translation, parsing, and image captioning, as well as language models
  - LSTMs became the dominant approach for most NLP tasks

- Now (2019–2022), other approaches (e.g., Transformers) have become dominant for many tasks
  - For example, in **WMT** (a Machine Translation conference + competition):
    - In WMT 2014, there were 0 neural machine translation systems (!)
    - In WMT 2016, the summary report contains "RNN" 44 times (and these systems won)
    - In WMT 2019: "RNN" 7 times, "Transformer" 105 times

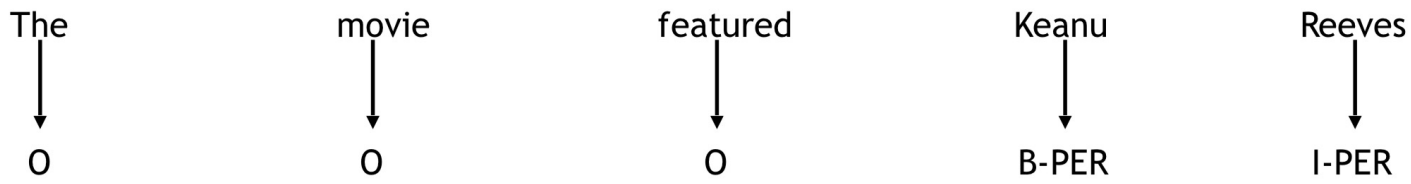
Other Sequence Models

- Label depends on context, so classification of words using context

Sequence Labeling

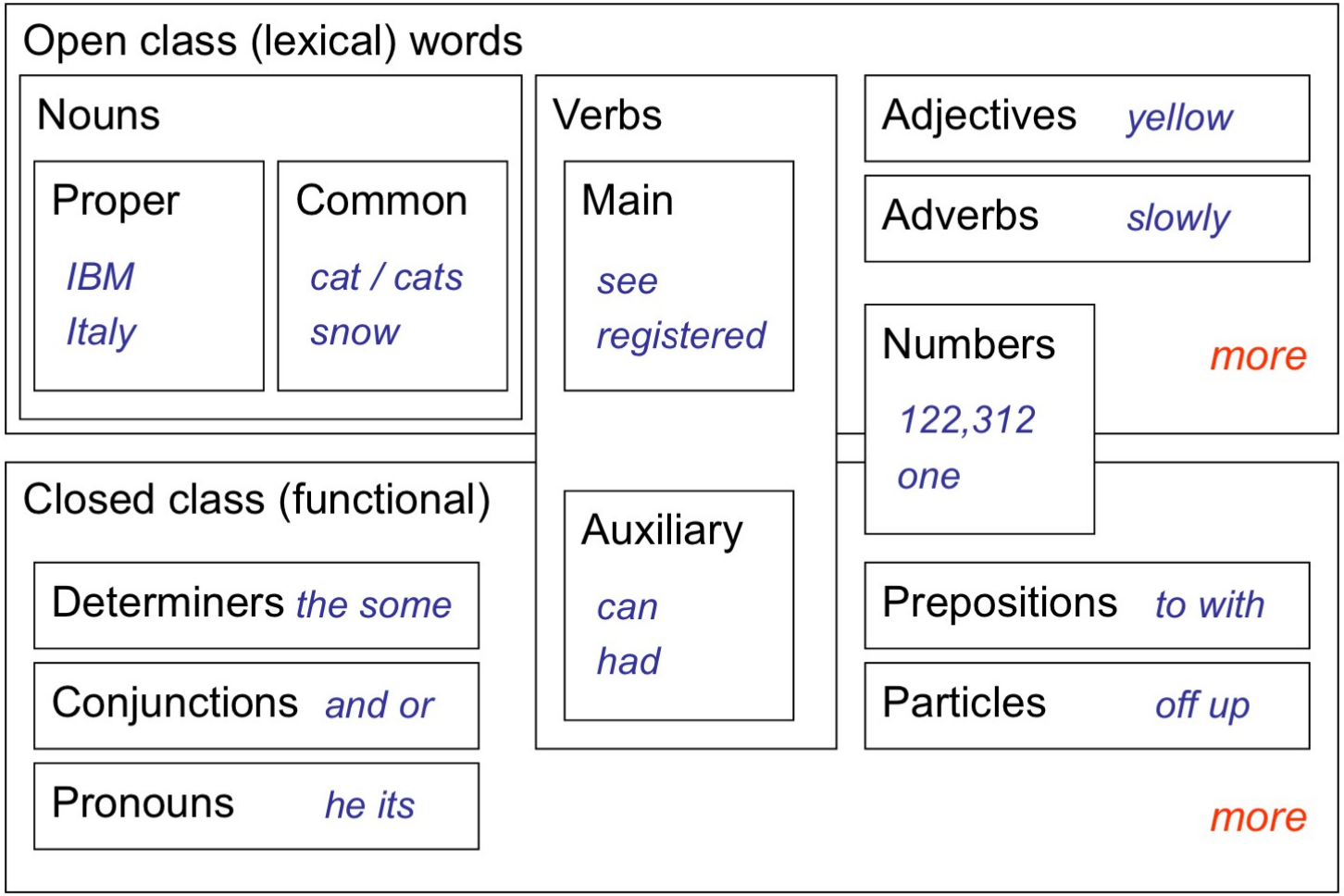
- One tag for one word
- e.g. Part of speech tagging

- e.g. Named entity recognition



Part of Speech Tagging

- Tag words in a sequence with syntactic categories



Part of Speech Ambiguities

A word can have multiple parts of speech

VBD		VB			
VBN	VBZ	VBP	VBZ		
NNP	NNS	NN	NNS	CD	NN

Fed raises interest rates 0.5 percent

Mrs./NNP Shaefer/NNP never/RB got/VBD **around/RP** to/TO joining/VBG

All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB **around/IN** the/DT corner/NN

Chateau/NNP Petrus/NNP costs/VBZ **around/RB** 250/CD

Disambiguating features: lexical identity (word), context, morphology (suffixes, prefixes), capitalization, gazetteers (dictionaries), …

Why Part of Speech Tagging?

Useful in itself:

- Text-to-speech: *read, lead, record*
- Lemmatization: *saw*[v] → *see*, *saw*[n] → *saw*
- Shallow Chunking: grep {JJ | NN}* {NN | NNS} Shallow information extraction

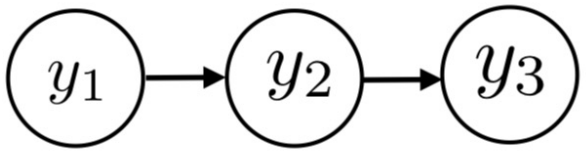
Useful for downstream tasks (e.g., in parsing, and as features in various word/text classification tasks)

Preprocessing step in parsing: allows fewer parse options if less tag ambiguity (but some cases still decided by parser)

Demos: http://nlp.stanford.edu:8080/corenlp/

Classic Solution: Hidden Markov Model

- Input $\mathbf{x} = (x_1, ..., x_n)$ Output $\mathbf{y} = (y_1, ..., y_n)$
- Model the sequence of tags ***y*** over words ***x*** as a Markov process
- Markov property: future is conditionally independent of the past given the present



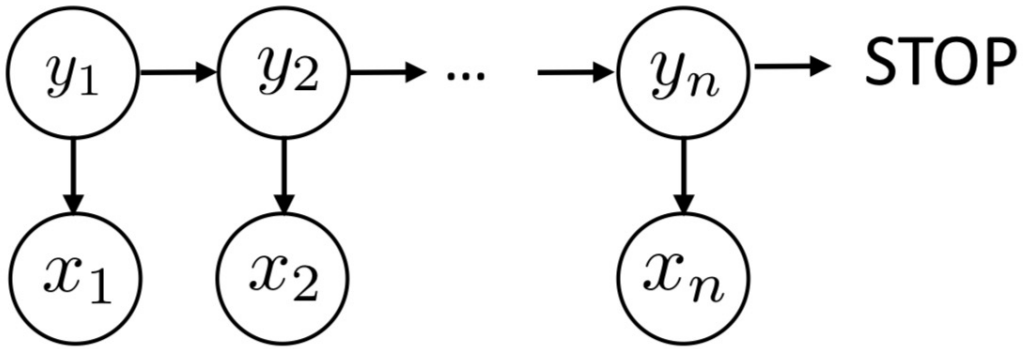
$$P(y_3|y_1, y_2) = P(y_3|y_2)$$

- If ***y*** are tags, this roughly corresponds to assuming that the next tag only depends on the current tag, not anything before

Classic Solution: Hidden Markov Model

- Input $\mathbf{x} = (x_1, ..., x_n)$ Output $\mathbf{y} = (y_1, ..., y_n)$

y ∈ T = set of possible tags (including STOP);
x ∈ V = vocab of words



$$P(\mathbf{y}, \mathbf{x}) = \underbrace{P(y_1)}_{\text{Initial distribution}} \underbrace{\prod_{i=2}^{n} P(y_i|y_{i-1})}_{\text{Transition probabilities}} \underbrace{\prod_{i=1}^{n} P(x_i|y_i)}_{\text{Emission probabilities}}$$

- Observation (*x*) depends only on current state (*y*)

Transitions in POS Tagging

VBD		VB				
VBN	VBZ	VBP	VBZ			
NNP	NNS	NN	NNS	CD	NN	

Fed raises interest rates 0.5 percent

- $P(y_1 = \text{NNP})$ likely because start of sentence
- $P(y_2 = \text{VBZ} | y_1 = \text{NNP})$ likely because verb often follows noun
- $P(y_3 = \text{NN} | y_2 = \text{VBZ})$: direct object can follow verb

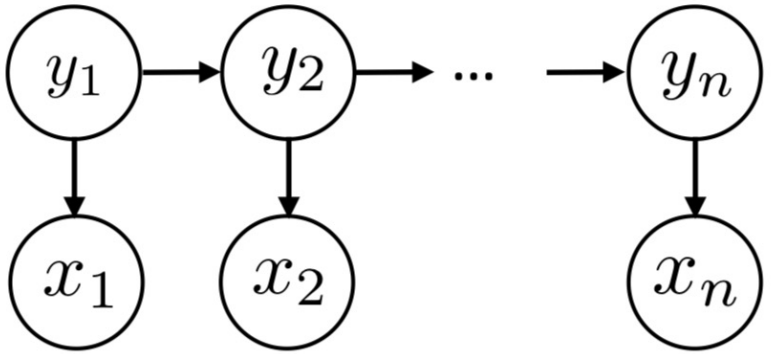
- How are these probabilities learned?

Learning

- Transitions
  - Count up all pairs (y_i, y_{i+1}) in the training data
  - Count up occurrences of what tag *T* can transition to
  - Normalize to get a distribution for P(next tag|*T*)
  - Need to *smooth* this distribution, won't discuss here
- Emissions: similar scheme, but trickier smoothing!

Inference

- Input $\mathbf{x} = (x_1, ..., x_n)$ Output $\mathbf{y} = (y_1, ..., y_n)$



$$P(\mathbf{y}, \mathbf{x}) = P(y_1) \prod_{i=2}^{n} P(y_i|y_{i-1}) \prod_{i=1}^{n} P(x_i|y_i)$$

- Inference problem: $\text{argmax}_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) = \text{argmax}_{\mathbf{y}} \frac{P(\mathbf{y}, \mathbf{x})}{\cancel{P(\mathbf{x})}}$

  Given observation sequence what is the most likely state sequence?

- Exponentially many possible ***y*** here!

- Solution: dynamic programming (possible because of Markov structure!)

HMM Tagging Performance

- Baseline of most frequent tag to each word: 90% accuracy
- Trigram HMM tagging: ~95% accuracy/ ~55% on unknown words
- State-of-the-art: BiLSTM-CRF: 97.5 accuracy/ 89% on unknown words

Other Languages

Language	CRF+	CRF
Bulgarian	97.97	97.00
Czech	98.38	98.00
Danish	95.93	95.06
German	93.08	91.99
Greek	97.72	97.21
English	95.11	94.51
Spanish	96.08	95.03
Farsi	96.59	96.25
Finnish	94.34	92.82
French	96.00	95.93
Indonesian	92.84	92.71
Italian	97.70	97.61
Swedish	96.81	96.15
AVERAGE	96.04	95.41

Other Sequence Labeling Tasks

- Named Entity Recognition
- Spelling Correction
- Word Alignment
- Noun Phrase Chunking
- Supersense Tagging
- Multiword Expressions

ECE594: Mathematical Models of Language

Spring 2022

Lecture 6: Low-Resource NLP

UNIT 1

- Modeling language at different levels
  - Words
  - Word sequences + Sequence labeling
  - Meaning

UNIT 2

- Low-Resource NLP
- NLP Applications

Natural Language Processing

Core technologies

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic role labelling
- ....

Applications

- Machine Translation
- Information Extraction
- Question Answering
- Dialogue Systems
- Summarization
- Sentiment Analysis
- ...

The multilingual world

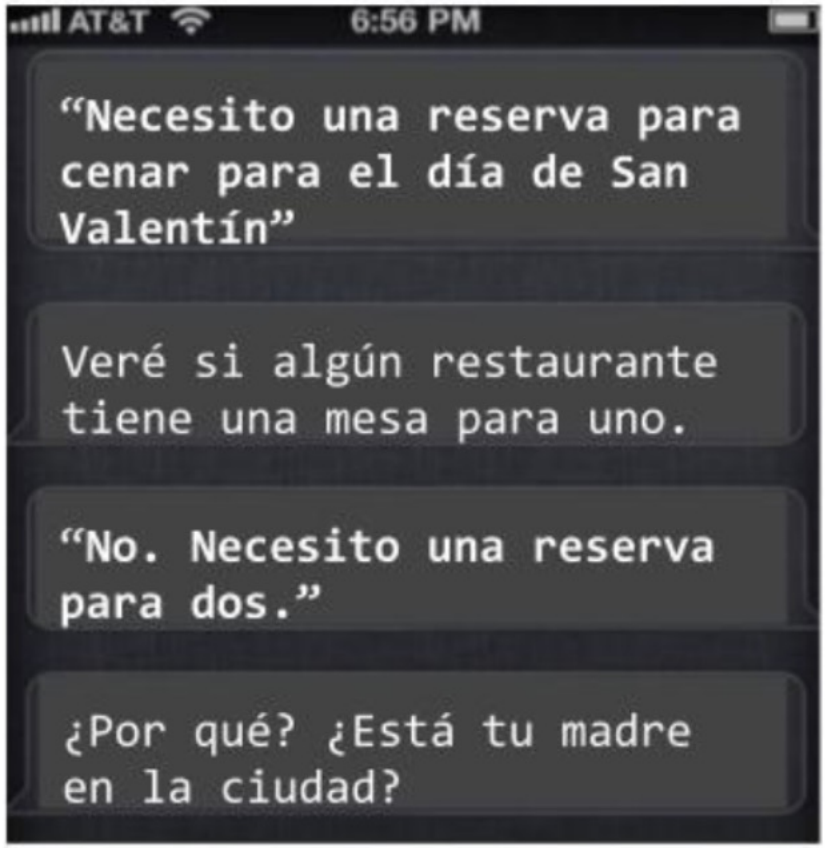
Language diversity: language families

www. enthologue.com

- Niger-Congo (1538 languages) (20.6%)
- Austronesian (1257 languages) (16.8%)
- Trains-New Guinea (480 languages) (6.4%)
- Sino-Tibetian (457 languages) (6.1%)
- Indo-European (444 languages) (5.9%)
- Australian (378 languages) (5.1%)
- Afro-Asiatic (375 languages) (5.0%)
- Nilo-Saharan (205 languauges) (2.7%)
- Oto-Manguean (177 languages) (2.4%)
- Austroasiatic (169 languages) (2.3%)
- Volta Congo (108 languages) (1.5%)
- Tai-Kadai (95 languages) (1.3%)
- Dravidian (85 languages) (1.1%)
- Tupian (76 languages (1.0%)

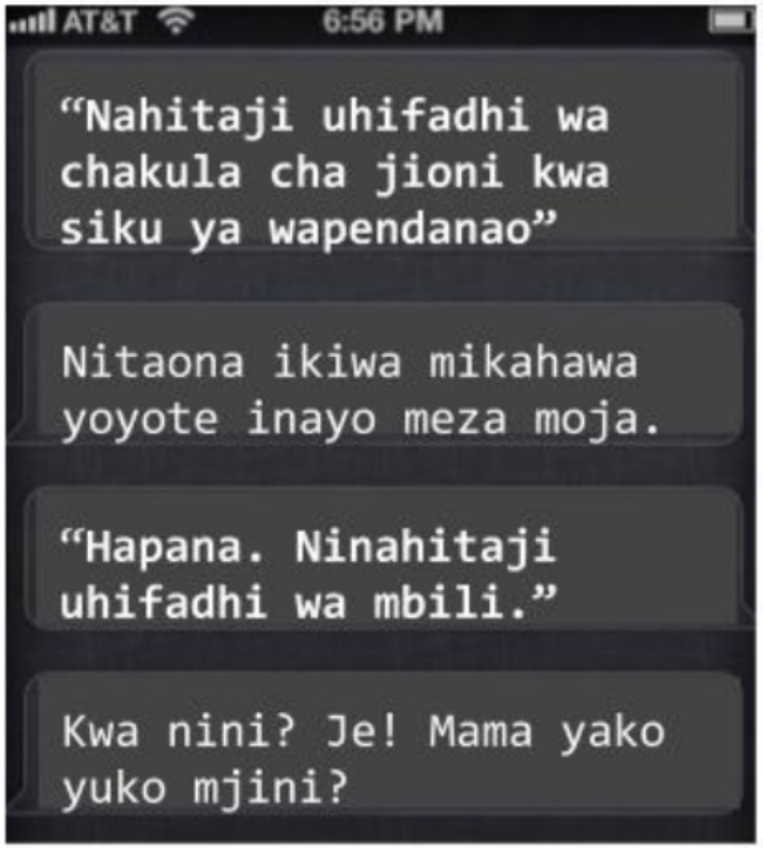
Ideal situation

Spanish
534 million speakers



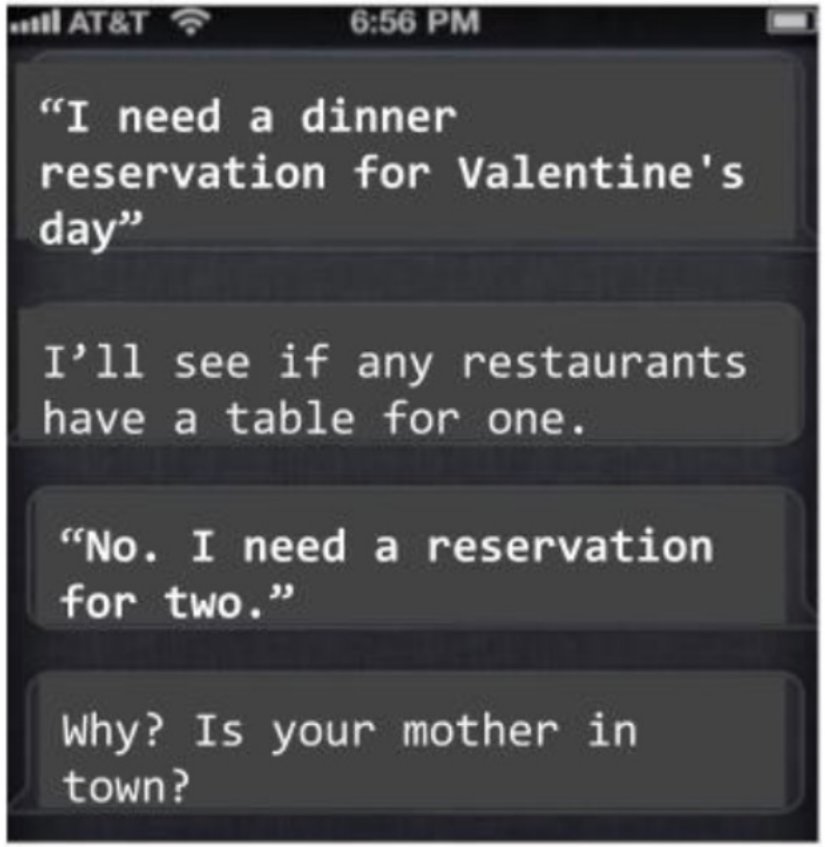
Hindi
615 million speakers

Swahili
100 million speakers



World Englishes

Your English, My English…



American English

Scottish English

Our English

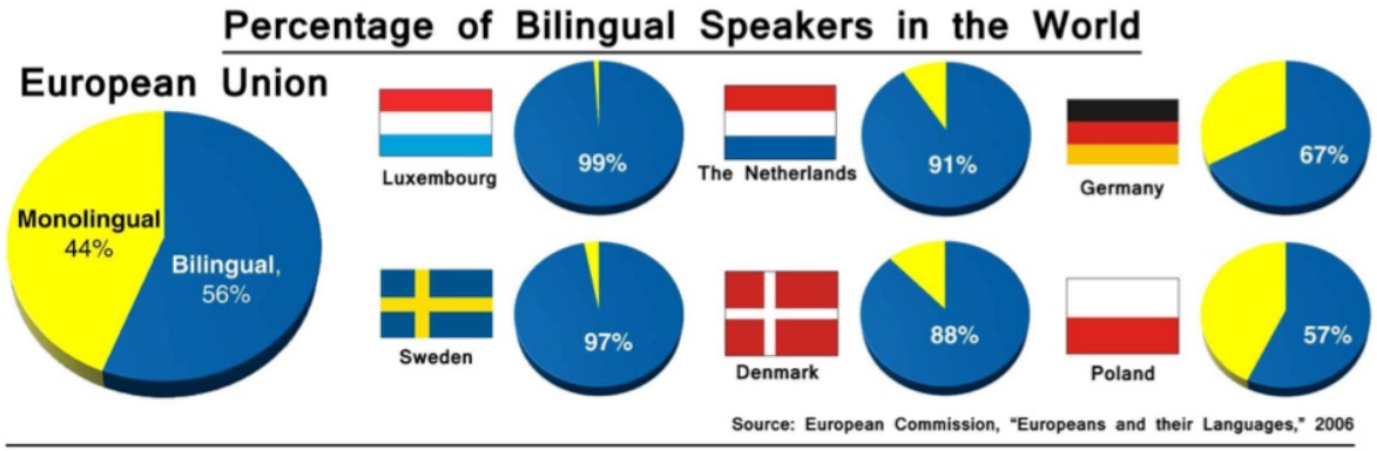

The Royal Family
@RoyalFamily
Follow
Taking place this week on the river Thames is 'Swan Upping' – the annual census of the swan population on the Thames.

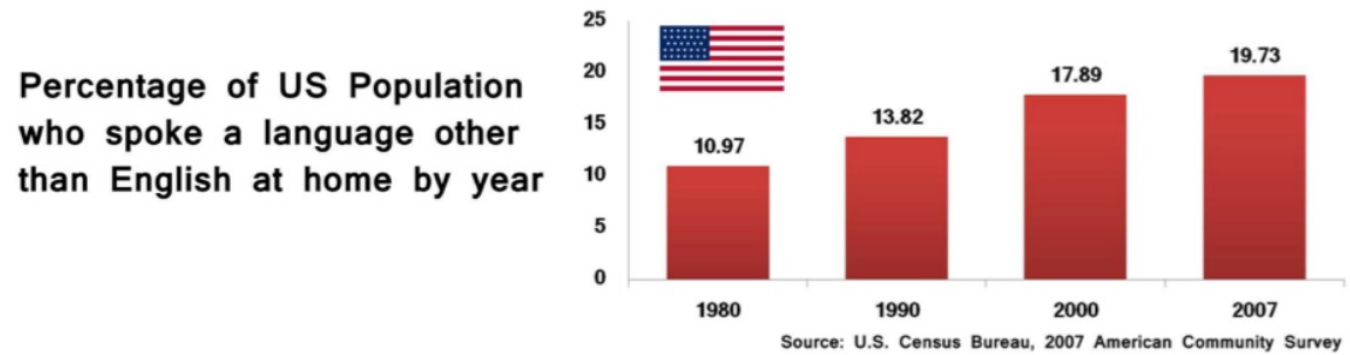

da'Rah-zingSun
@TIME7SS
Follow
@kimguilfoyle prblm I hve wit ur reportng is its 2 literal, evry1 knos pple tlk diffrnt evrywhere, u kno wut she means jus like we do!

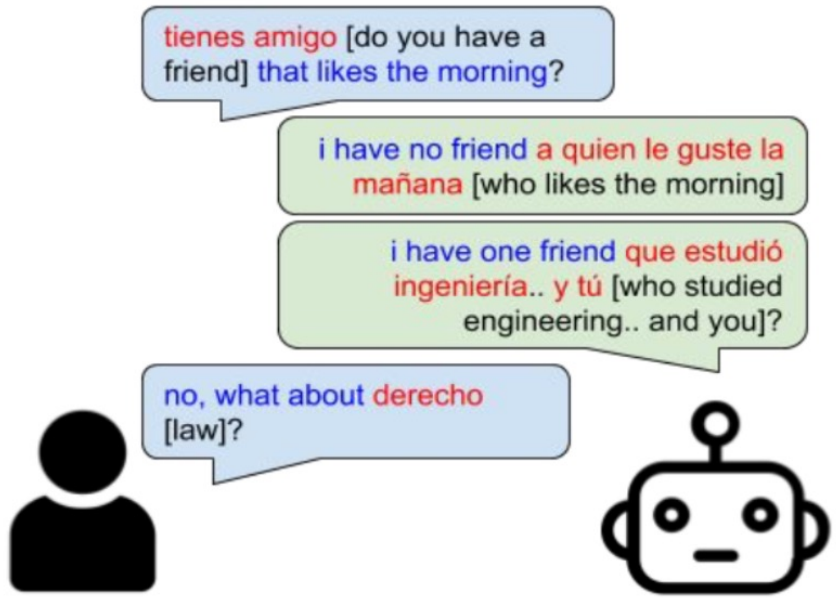

Mooktar
@bossmukky
Follow
"@Ecstatic_Mi: @bossmukky Ebi like say I wan dey sick sef wlh 'Flu' my whole body dey weak"uw gee...


Ebenezer·
@Physique_cian
Follow
@Tblazeen R u a wizard or wat gan sef : in d mornin- u tweet, afternoon - u tweet, nyt gan u dey tweet.beta get ur IT placement wiv twitter

Code-switching







Source: US Census Bureau

Low-Resource NLP

- Bringing language technology to languages lacking large monolingual or parallel corpora
- Help preserve languages
- Increase participation of speakers in digital world

Low-Resource NLP

- Bringing language technology to domains/tasks lacking large corpora
- Biomedical domain
- Scientific domain
- Natural language inference

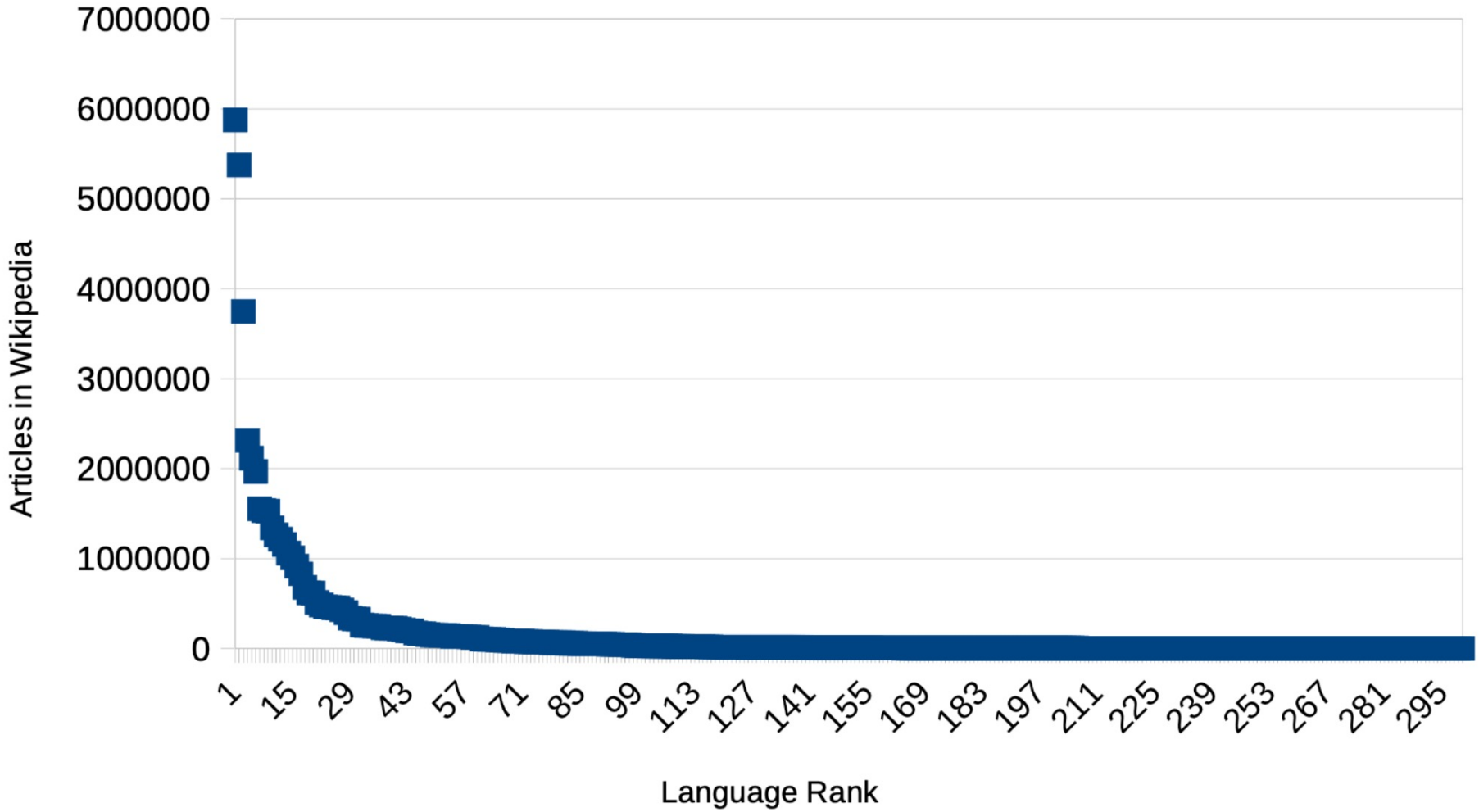
Why is low-resource NLP hard?

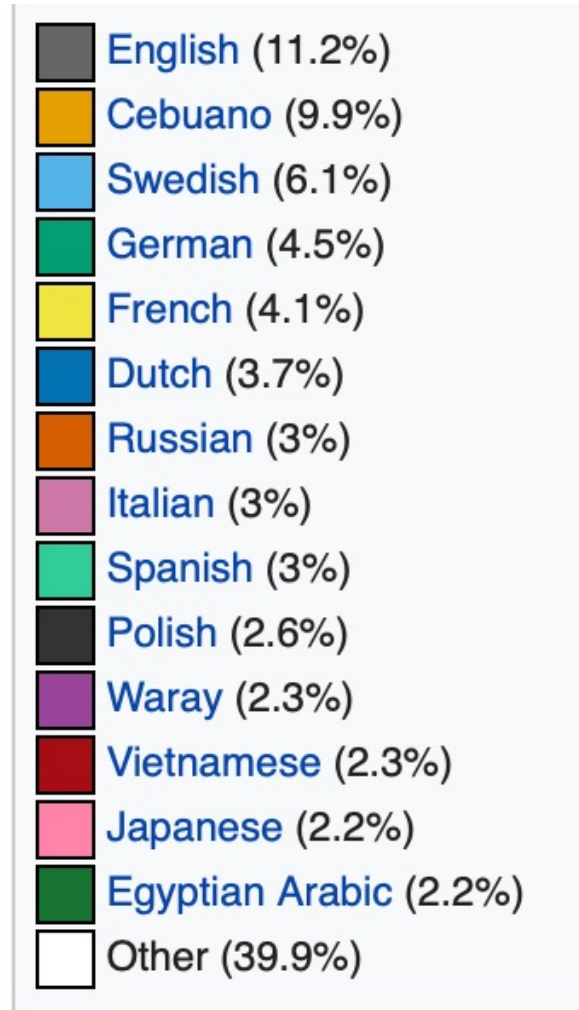
- Low-resource languages
  - Insufficient data (raw + labeled)
  - Languages may not have a written form
  - Need trained linguists for language-specific engineering

Why is low-resource NLP hard?

- Low-resource domains
  - Insufficient data
  - Need experts for domain-specific engineering

The Long Tail of Data





NLP System Needs

- Ability to process language at different levels
  - Sounds
  - Words
  - Structure
  - Meaning

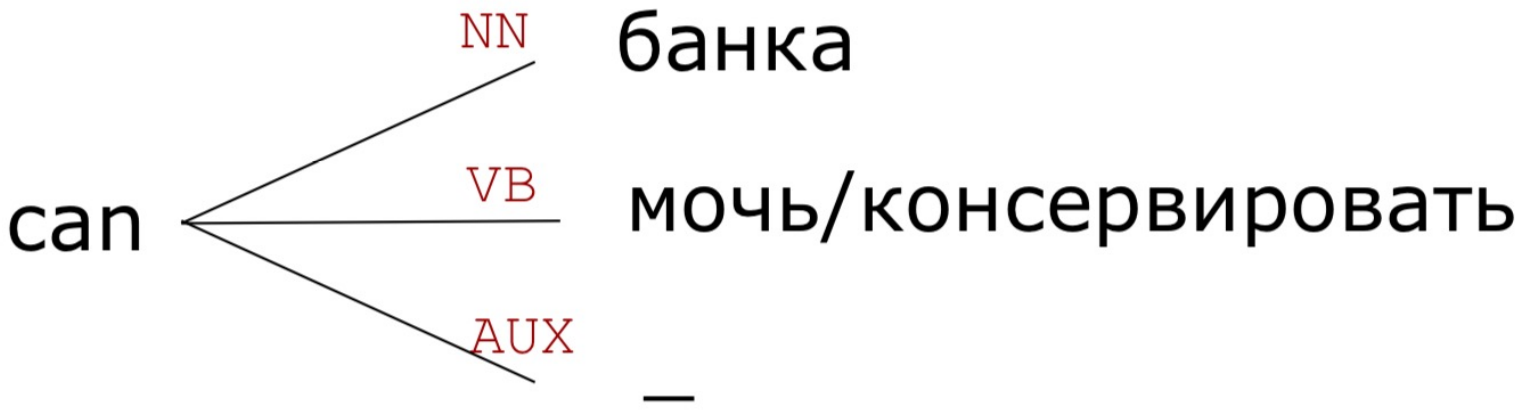
Key Questions

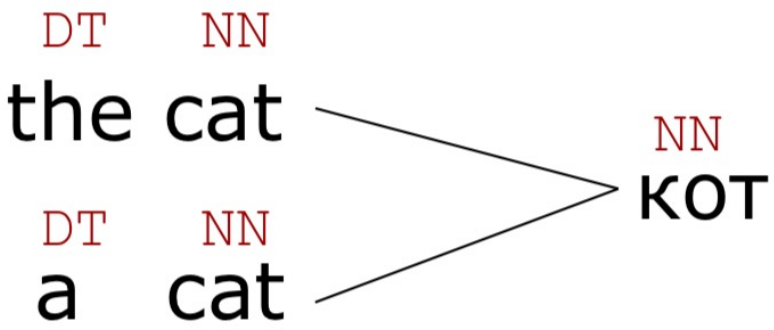
- Tokenization --- what are units?
- Parts of speech do not map easily from English
- Morphology can be very different
- Syntax varied

Kannada Morphology

	Neuter Singular
Nominative	ಮನೆ
Accusative	ಮನೆಯನ್ನು
Instrumental	ಮನೆಯಿಂದ
Dative	ಮನೆಗೆ
Genitive	ಮನೆಯ
Locative	ಮನೆಯಲ್ಲಿ

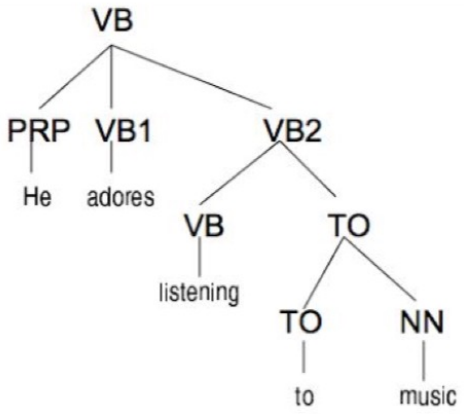
Part of Speech Tags Map Differently



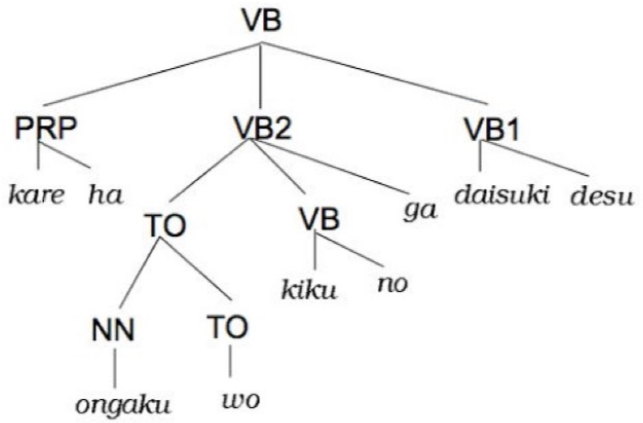


Different Syntax

SVO

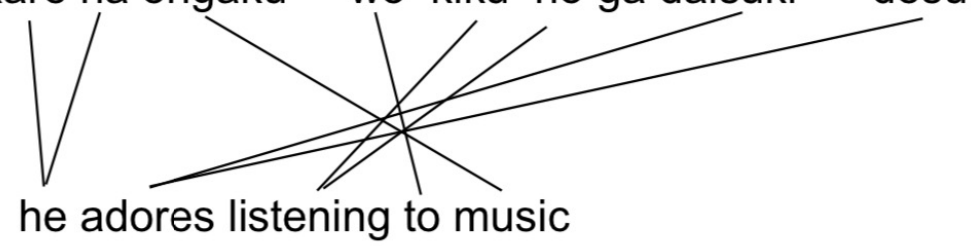


SOV



he adores listening to music

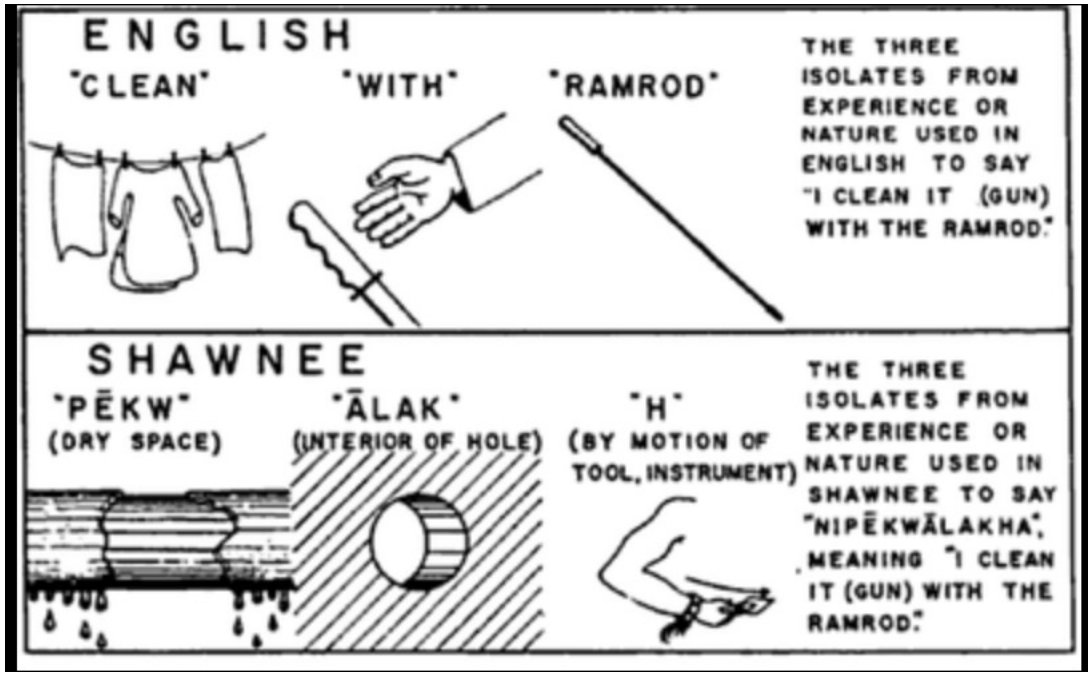
かれ は おんがく を きく の が だいすき です
kare ha ongaku wo kiku no ga daisuki desu



(Yamada & Knight '02)

Diversity in Semantics

- Meaning is about associating language with world
  - Every language describes world in unique way
  - Sapir-Whorf Hypothesis
    - Structure of language shapes our thinking and behavior
  - Culture and History



Diversity in Semantics

- Meaning is about associating language with world

Yellow series [edit]

Name	Romanized	English	RGB	Hex triplet	Name	Romanized	English	RGB	Hex triplet
浅黄	Asagi	Light yellow	247,187,125	#F7BB7D	山吹色	Yamabuki-iro	Golden yellow (and a particular rose varietal)	255,164,0	#FFA400
玉子色	Tamago-iro	Egg-colored	255,166,49	#FFA631	櫨染	Hajizome	Sumac-dyed	224,138,30	#E08A1E
山吹茶	Yamabukicha	Gold-brown	203,126,31	#CB7E1F	桑染	Kuwazome	Mulberry-dyed	197,127,46	#C57F2E
生壁色	Namakabe-iro	The color of an undried wall	120,94,73	#785E49	梔子	Kuchinashi	Cape jasmine or gardenia	255,185,90	#FFB95A
玉蜀黍色	Tōmorokoshi-iro	Corn-colored	250,169,69	#FAA945	白橡	Shirotsurubami	White oak	206,159,111	#CE9F6F
黄橡	Kitsurubami	Golden oak	187,129,65	#BB8141	藤黄	Tō'ō	Gamboge	255,182,30	#FFB61E
花葉色	Hanaba-iro or kayou-iro	Floral leaf-colored	255,185,78	#FFB94E	鳥の子色	Torinoko-iro	Eggshell paper-colored	226,190,159	#E2BE9F
鬱金色	Ukon-iro	Turmeric-colored	230,155,58	#E69B3A	黄朽葉	Kikuchiba	Golden fallen leaves	226,156,69	#E29C45

Diversity in Semantics

- Meaning is about associating language with world

- Guugu Yimithirr spoken in North Queensland, Australia
  - No use of “left” or “right”
  - Describe locations and directions using the cardinal directions, which do not rotate with them as they turn.

For “you’re standing in front of the best ice cream shop in town,”

they’d say, “you’re standing north of it.”

Diversity in Semantics

- Meaning is about associating language with world
- Multiword expressions
  - It’s raining cats and dogs
  - Life is a journey

Linguistic Diversity

1. Niger–Congo (1,538 languages) (20.6%)
2. Austronesian (1,257 languages) (16.8%)
3. Trans–New Guinea (480 languages) (6.4%)
4. Sino-Tibetan (457 languages) (6.1%)
5. Indo-European (444 languages) (5.9%)
6. Australian (378 languages) (5.1%)
7. Afro-Asiatic (375 languages) (5.0%)
8. Nilo-Saharan (205 languages) (2.7%)
9. Oto-Manguean (177 languages) (2.4%)
10. Austroasiatic (169 languages) (2.3%)
11. Volta Congo (108 languages) (1.5%)
12. Tai–Kadai (95 languages) (1.3%)
13. Dravidian (85 languages) (1.1%)
14. Tupian (76 languages) (1.0%)

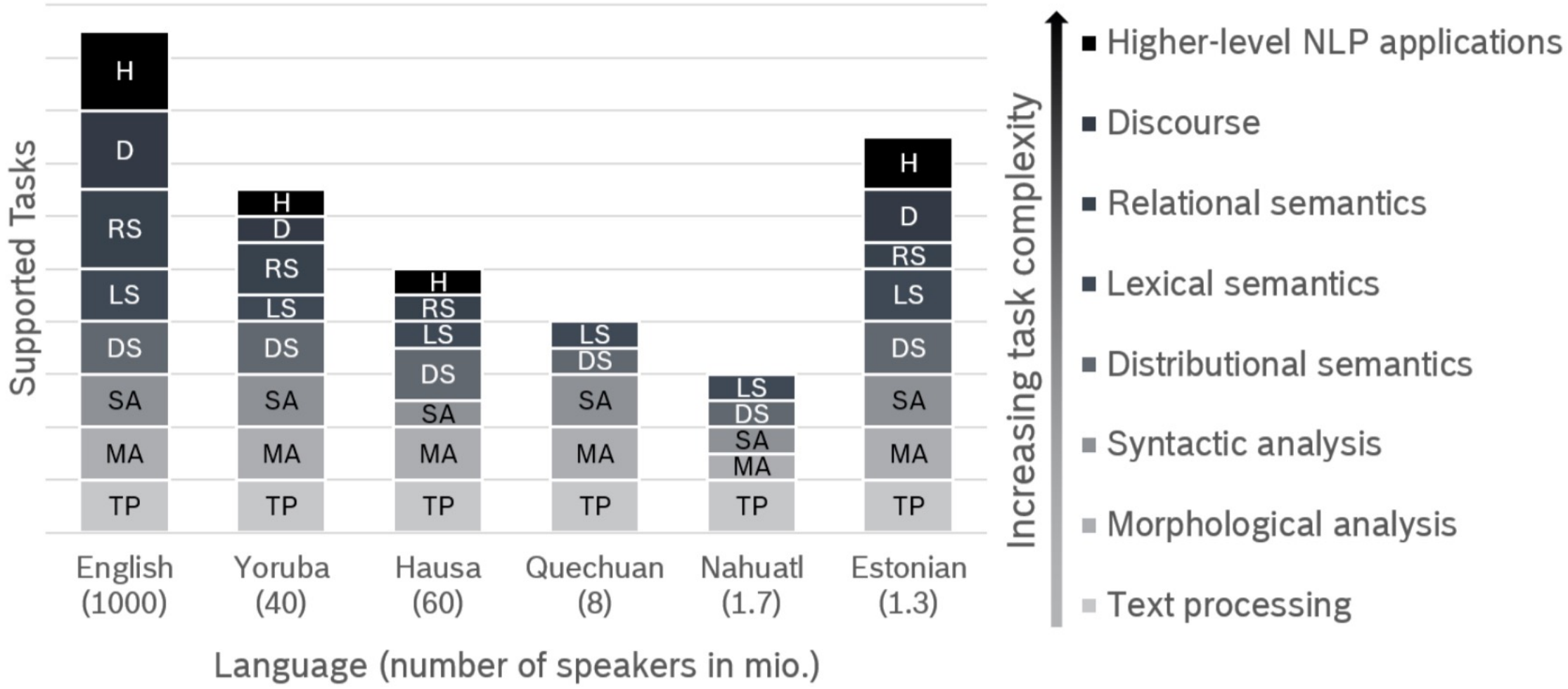
No Generic NLP Model

- Languages diverge across all levels
- Need data/resources for statistical models

Key Questions

- What language is it?
  - www.ethnologue.com, wals.info
  - Words: Tokenization --- what are units?
  - Morphology: How do you lemmatize it?
  - Syntax: How are sentences structured?
  - Typology: Who is a rich cousin?

Supported NLP tasks and Languages



Google Translate

DETECT LANGUAGE | ENGLISH | KOREAN | SPANISH | KOREAN | SPANISH | ENGLISH

Search languages

Detect language	Danish	Hmong	Lithuanian	Romanian	Telugu
Afrikaans	Dutch	Hungarian	Luxembourgish	Russian	Thai
Albanian	✓ English	Icelandic	Macedonian	Samoan	Turkish
Amharic	Esperanto	Igbo	Malagasy	Scots Gaelic	Turkmen
Arabic	Estonian	Indonesian	Malay	Serbian	Ukrainian
Armenian	Filipino	Irish	Malayalam	Sesotho	Urdu
Azerbaijani	Finnish	Italian	Maltese	Shona	Uyghur
Basque	French	Japanese	Maori	Sindhi	Uzbek
Belarusian	Frisian	Javanese	Marathi	Sinhala	Vietnamese
Bengali	Galician	Kannada	Mongolian	Slovak	Welsh
Bosnian	Georgian	Kazakh	Myanmar (Burmese)	Slovenian	Xhosa
Bulgarian	German	Khmer	Nepali	Somali	Yiddish
Catalan	Greek	Kinyarwanda	Norwegian	Spanish	Yoruba
Cebuano	Gujarati	Korean	Odia (Oriya)	Sundanese	Zulu
Chichewa	Haitian Creole	Kurdish (Kurmanji)	Pashto	Swahili	
Chinese	Hausa	Kyrgyz	Persian	Swedish	
Corsican	Hawaiian	Lao	Polish	Tajik	
Croatian	Hebrew	Latin	Portuguese	Tamil	
Czech	Hindi	Latvian	Punjabi	Tatar	

Key Questions

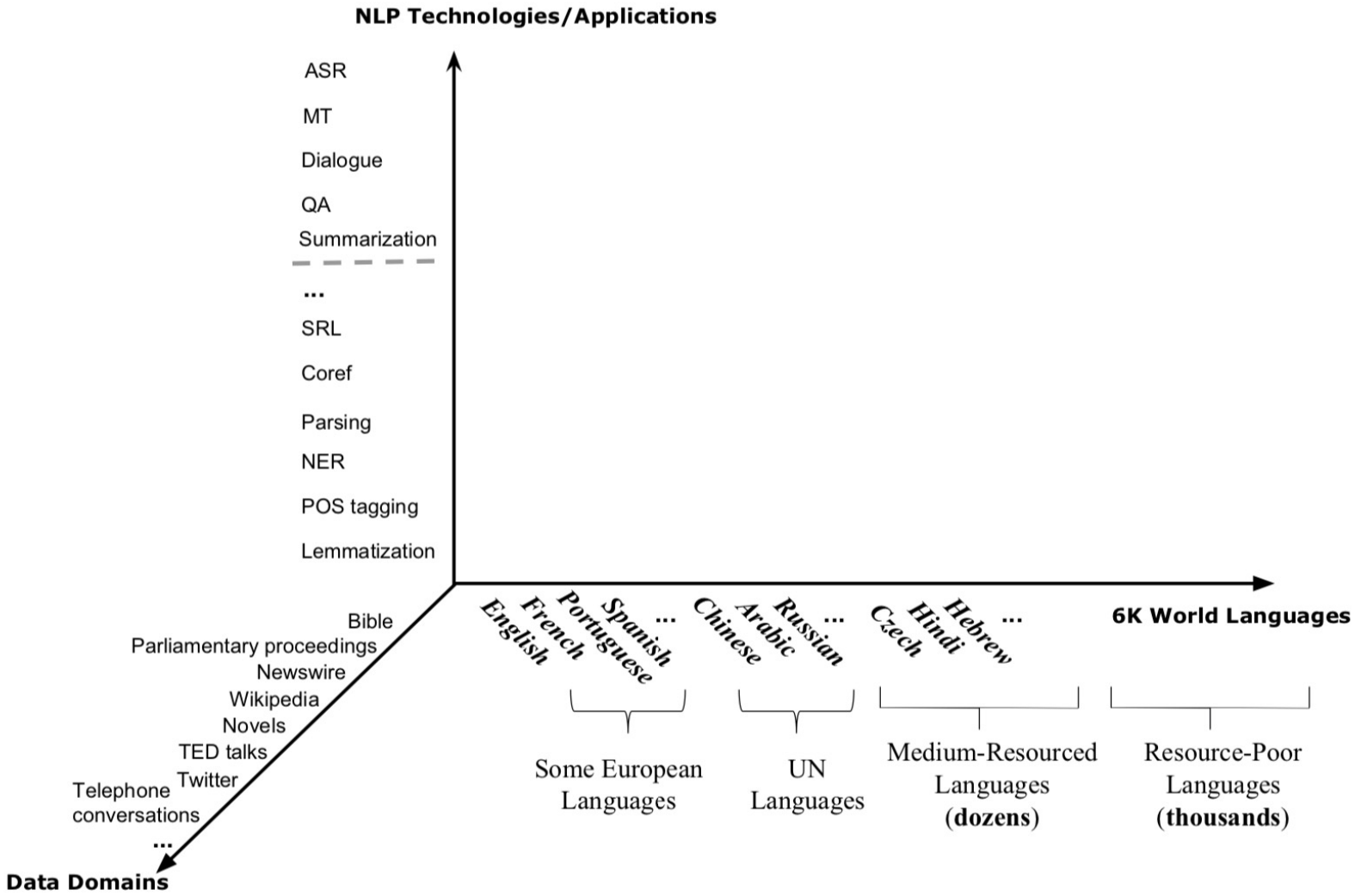
You will just have to find a way of getting over it.

あなたはそれを乗り越える方法を見つける必要があるでしょう。

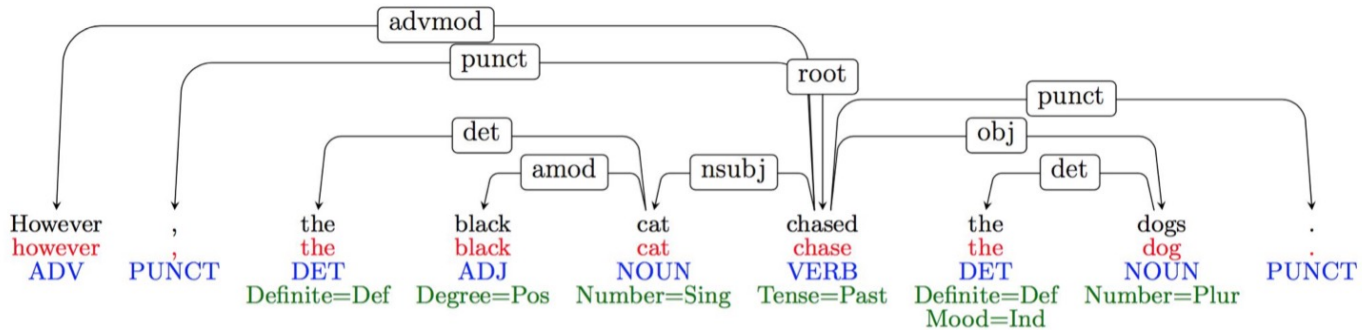
ನೀವು ಅದನ್ನು ಮೀರುವ ಮಾರ್ಗವನ್ನು ಕಂಡುಹಿಡಿಯಬೇಕಾಗಿದೆ.

سيكون عليك فقط إيجاد طريقة للتغلب عليها.

Low-Res NLP Not Just Multi-linguality



- Parsing models trained using WSJ treebanks do not work for spoken language domain



- Spoken language is riddled with verbal disfluencies that interrupt the flow of speech, including long pauses, repeated words or phrases, restarts, and revisions of content:

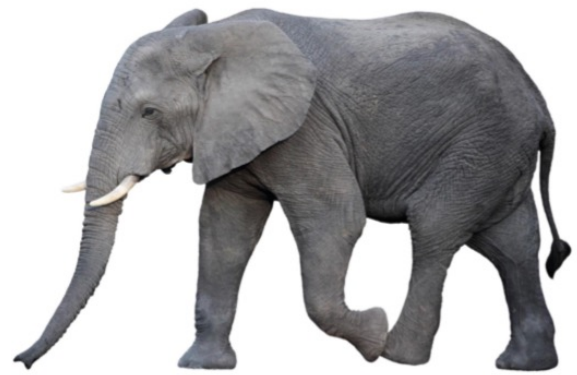
Um, the black the black cat ch- chased the dogs.

- Much of the world knowledge is not in text, corpora contain what people said, but not what they meant, or how they understood things, or what they did in response to the language

This is milk

?

<



Resource Characterization

- Task-specific labeled data
- Unlabeled text
  - For word embeddings
- Auxiliary data
  - Related resource-rich language
  - Knowledge bases

What do Resources Look Like?

- Task-specific labeled data
  - POS tagging
  - Non-neural methods > Neural methods

Language		Treebank Data (test)	
code	family	sentences	tokens
am	AA	1,095	10k
be	IE	68	1.3k
br	IE	888	10.3k
fo	IE	1,208	10.0k
hsb	IE	623	10.7k
hy	IE	514	11.4k
kmr	IE	734	10.1k
lt	IE	55	1.0k
mr	IE	47	0.4k
mt	AA	100	2.3k
bxr	Mo	908	10.0k
kk	Tu	1,047	10.1k
ta	Dr	120	2.2k
te	Dr	146	0.7k
tl	Au	55	0.2k
de	IE	1,000	21.3k
es	IE	1,000	23.3k
it	IE	1,000	23.7k
pt	IE	1,000	23.4k
sv	IE	1,000	19.1k

Generating More Data

- Manual
  - Expert linguists or human-in-the-loop
    - Quality labels

- Automatic
  - Noisy labels

Generating More Data Automatically

- Data augmentation (task-specific)
  - Transforming instances without changing label
    - Replacing words with equivalents
      - synonyms (Wei and Zou, 2019), entities of the same type (Raiman and Miller, 2017; Dai and Adel, 2020) or words that share the same morphology (Gulordava et al., 2018; Vania et al., 2019)

Generating More Data Automatically

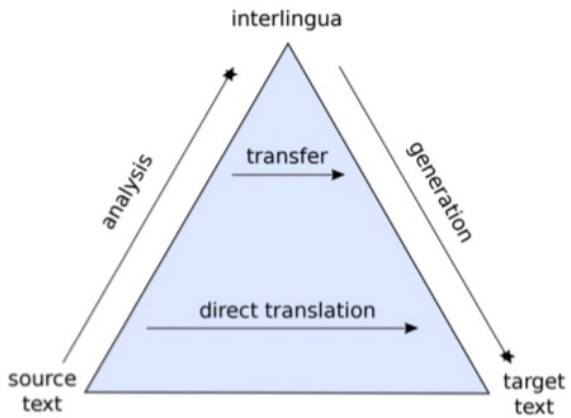
- Distant supervision (NER, relation extraction)
  - Unlabeled data labels obtained from an external source (knowledge bases, gazetteers, dictionaries)
  - Non-native speakers or non-expert labels
  - Helpful for resource-rich languages

Cross-lingual Projections

- Task-specific classifier trained for high-resource language (POS-tagging, MT)
  - Using parallel corpora unlabeled low-resource data (corresponding labels) aligned
  - Limitations on availability of ways of aligning (dictionaries)

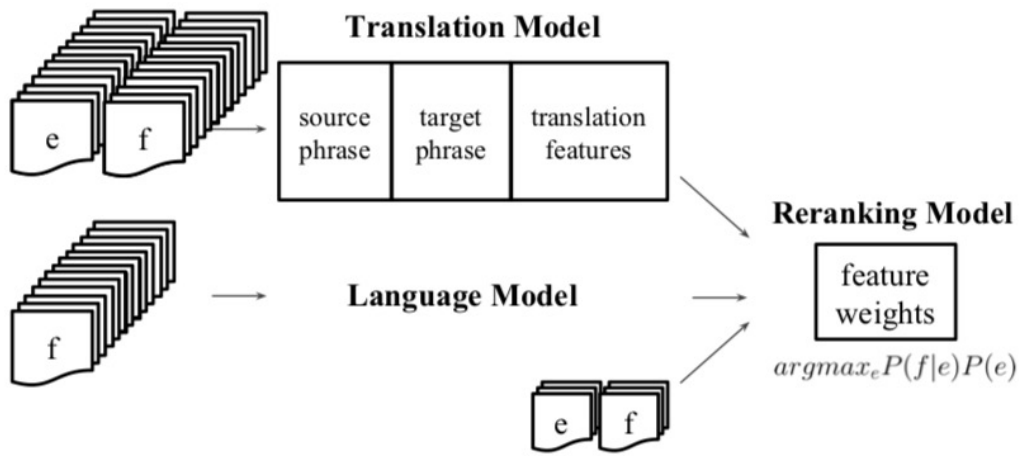
Paradigm Shifts in NLP

Logic-based/Rule-based NLP



~90s

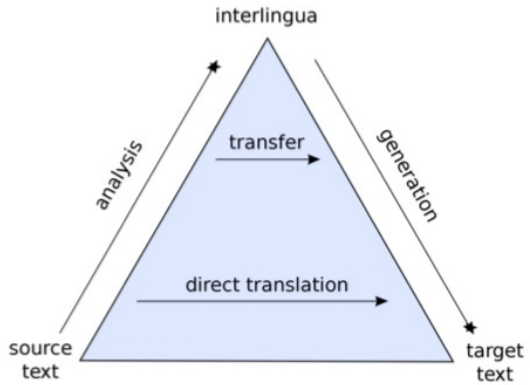
Statistical NLP



- Rule-based models: high precision but very low recall

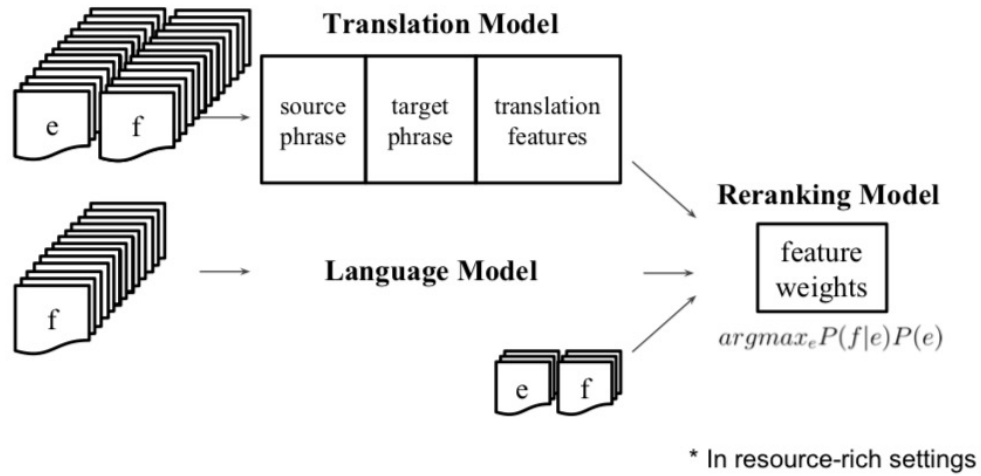
Paradigm Shifts in NLP

Logic-based/Rule-based NLP



<

Statistical NLP

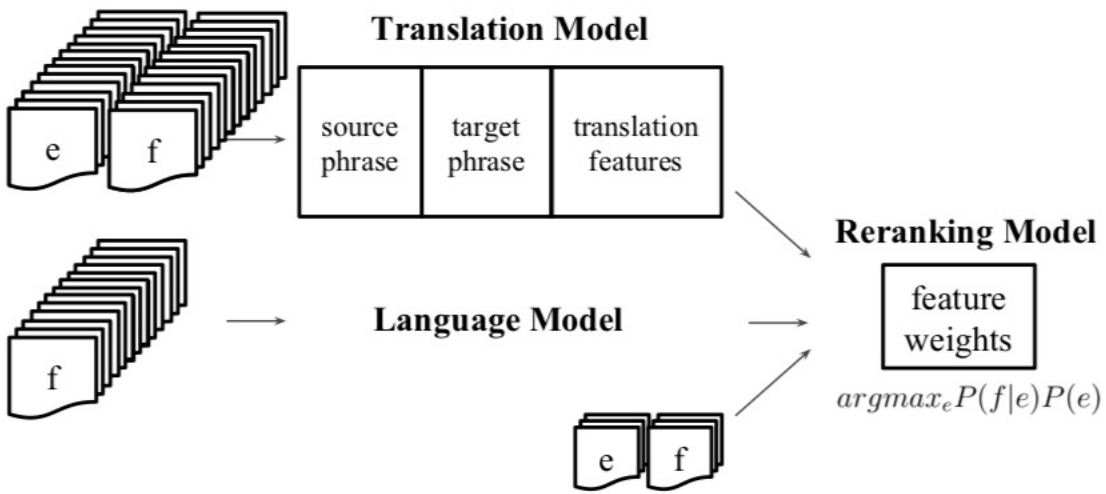


* In resource-rich settings

- Statistical models: robust in the face of real-world data
- Better performance
- Less engineering of hand-crafted rules/knowledge

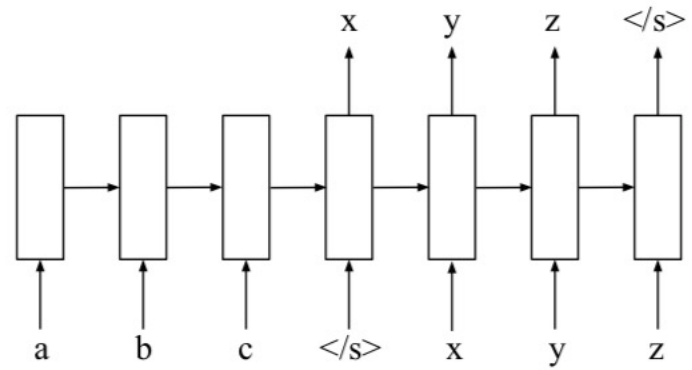
Paradigm Shifts in NLP

Statistical NLP



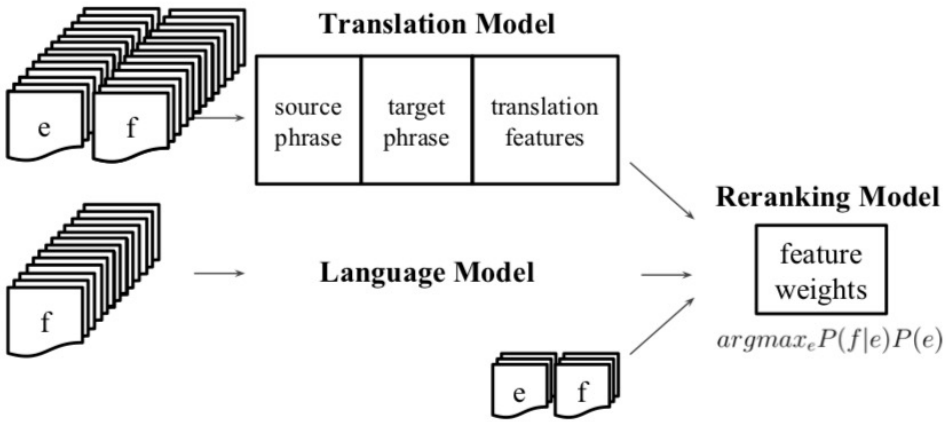
~mid 2010s

Statistical Neural NLP



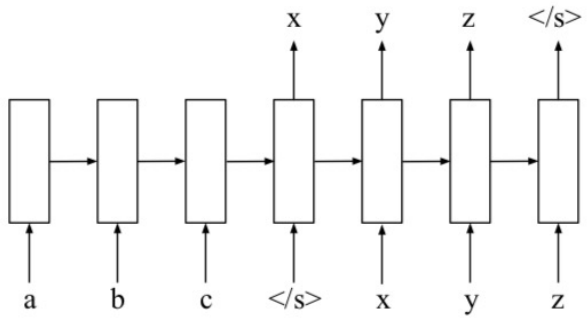
Paradigm Shifts in NLP

Statistical NLP



<

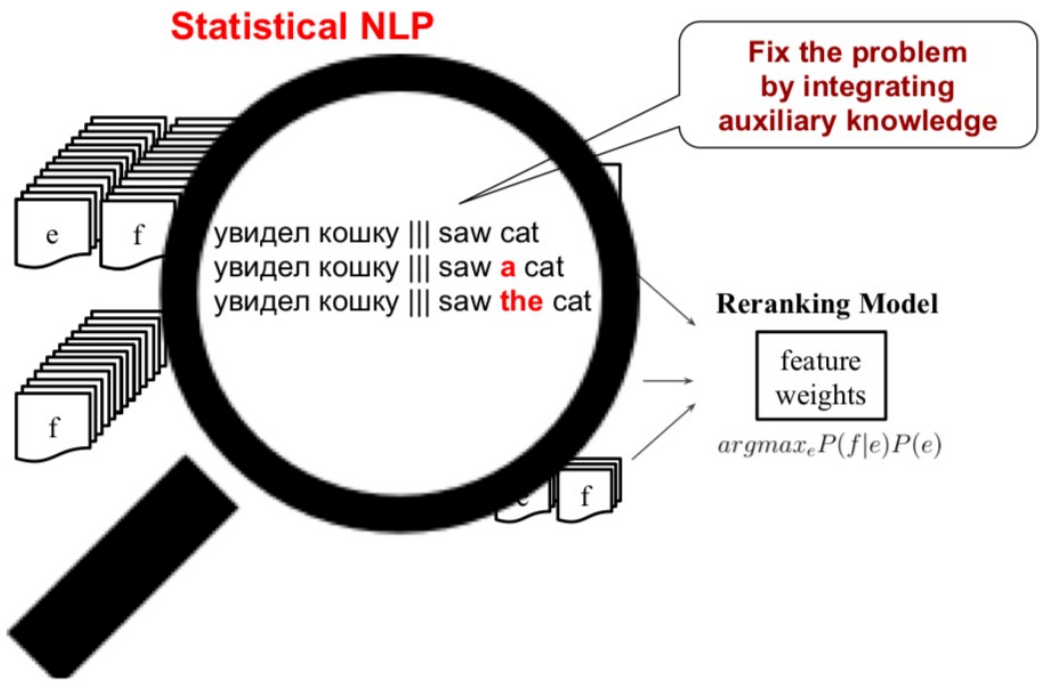
Statistical Neural NLP



* In resource-rich settings

- Robustness in the face of real-world data
- Better performance
- Less engineering of hand-crafted rules/knowledge

Hybrid Statistical NLP Models

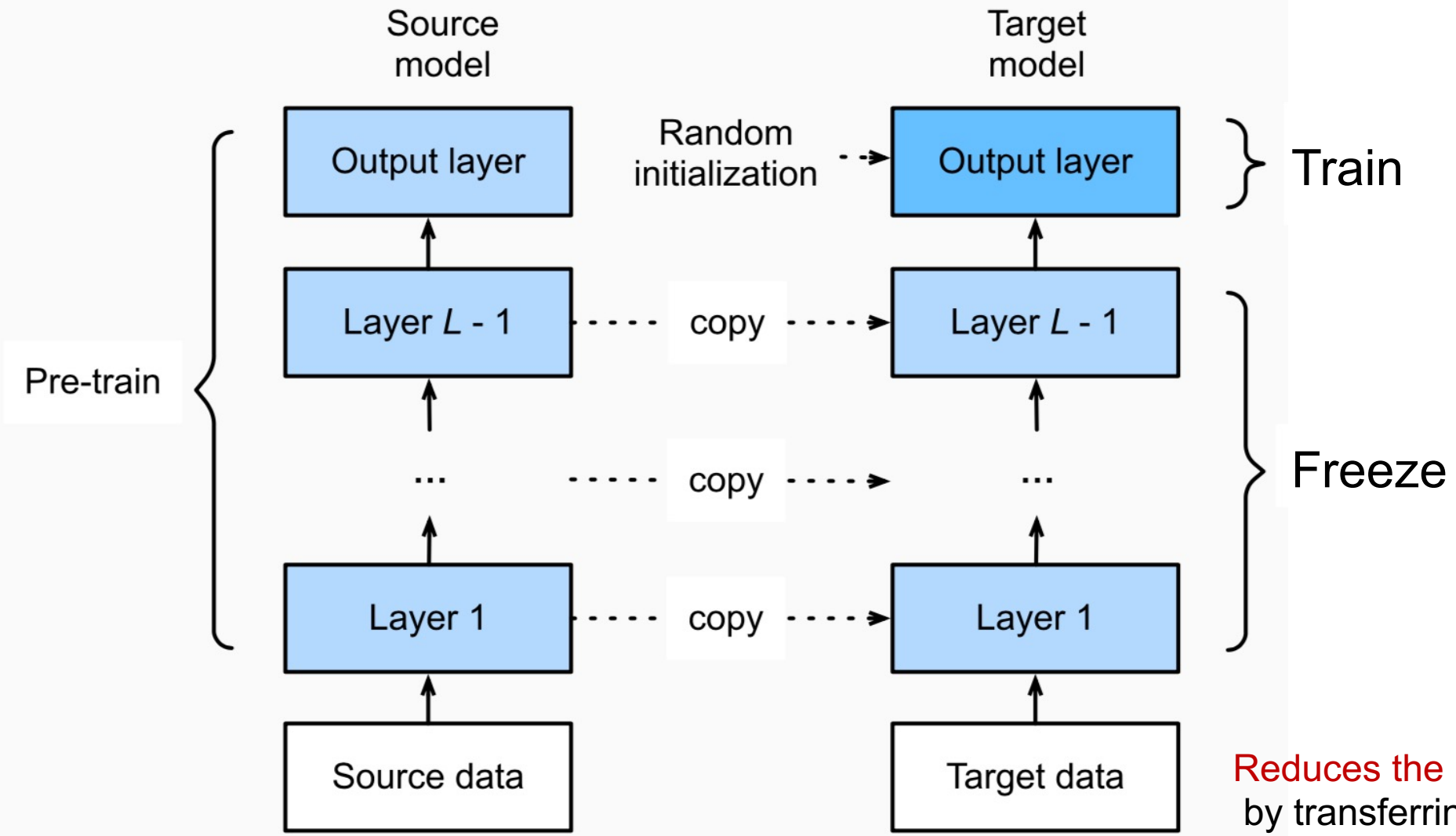


- To train high-quality models we need large amounts of training data
- We can partially compensate data scarcity with more sophisticated models that combine statistical learning with linguistic knowledge

Supervised, Unsupervised, Semi-supervised

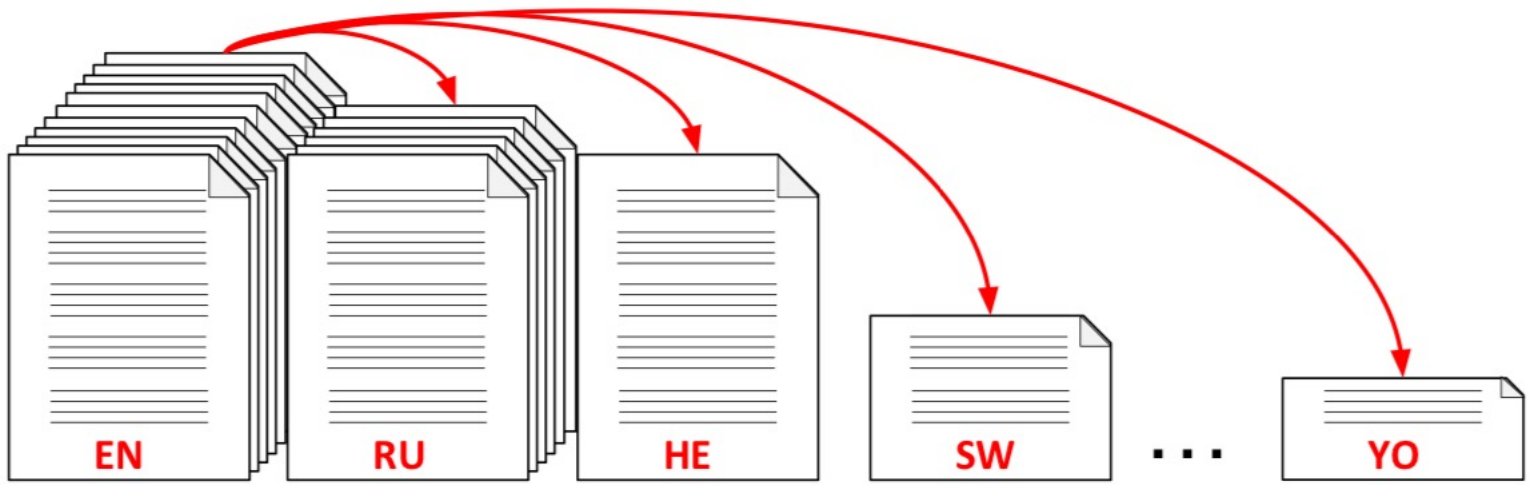
- Most models handled here are **supervised** learning
  - Model P(Y|X), at training time given both
- Sometimes we are interested in **unsupervised** learning
  - Model P(Y|X), at training time given only X
- Or **semi-supervised** learning
  - Model P(Y|X), at training time given both or only X

Transfer Learning



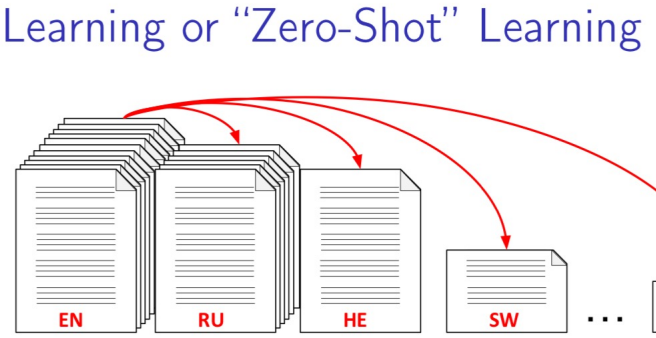
Reduces the need for labeled target data by transferring learned representations and models

Transfer Learning or "Zero-Shot" Learning



Transfer Learning

- Unlabeled data for pre-trained language representations
  - BERT
- Training of domain-specific or multilingual representations
- Auxiliary data
  - Train and transfer models from related tasks in the same language, or the same (or similar) task from other domains or languages.



Cross-Lingual Representations

- Cross-lingual representations of words
  - Reasoning about multi-lingual word meaning
  - Help cross-lingual transfer
  - Sample-efficient because require word translation pairs or only monolingual data
  - Aligning embedding spaces good for coarse-grained tasks like topic classification
  - Not sufficient for fine-grained tasks like MT

Cross-Lingual Zero-Shot Learning

- Learned parameters for seen classes along with their class representations
  - Rely on representational similarity among class labels so that, during inference, instances can be classified into new classes

- Useful for learning labels in low-resource languages
  - Leverage labeled data from a high-resource language when no task-specific labeled data is available in the low-resource target language.

Multilingual Language Models

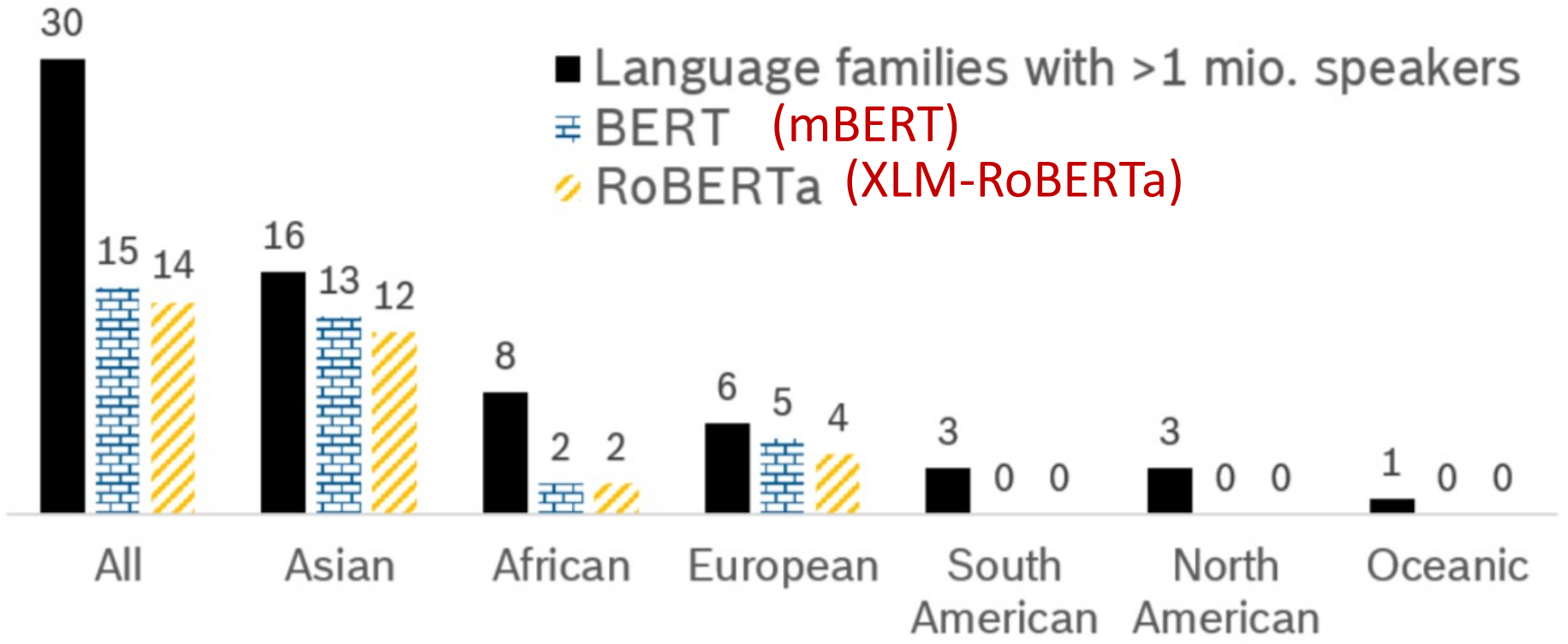


Figure 2: Language families with more than 1 million speakers covered by multilingual transformer models.

Domain-Specific Pretraining

- Models pre-trained on general-domain data, can be successfully transferred to texts from unseen domains, such as the clinical domain

- Continuing the training of an already pre-trained model with additional domain-adaptive and task-adaptive pre-training with unlabeled data leads to performance gains for both high- and low-resource task settings for English

Why standard techniques used in NLP cannot simply be applied to low-resource languages?

- State-of-the-art NLP models require large amounts of training data and/or sophisticated language-specific engineering
- Large amounts of training data are unavailable for most languages
  - an extreme case is languages that don't have a written form, e.g. Shanghainese spoken by 14 million people
  - or languages that just don't have online presence, e.g. Chichewa, a Bantu language spoken by 12 million people
- Language-specific engineering is expensive, requires linguistically trained speakers of the language

Solutions

- Transfer learning approaches
  - Use bridge language
- Unsupervised approaches
  - Use web as a corpus