Natural Language Processing

Text classification

Yulia Tsvetkov

yuliats@cs.washington.edu

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Announcements

- HW1 overview
  - https://gitlab.cs.washington.edu/cse447-au22/internal/assignment-1-public-ready/-/blob/main/pset1.ipynb
- Quiz 1 is on Wednesday

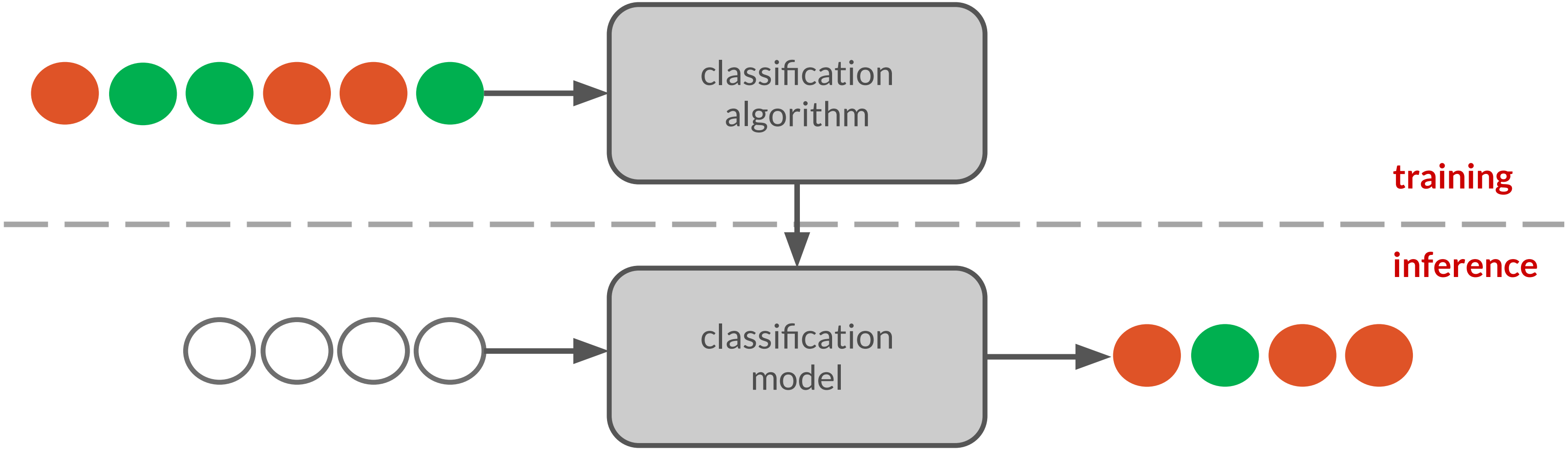
Compute resources for completing HW

- Access to non-CS students
  - Response from CSE support: Based on day 5 enrollment, I have finished catching up on this, this morning. Students should check their email, maybe spam folder for the CSE account email. There was a user or 2 that previously had a non-major account that I reactivated and they may need to reset their password at password.cs.washington.edu. If anyone has not had an account created (very late add, etc.) please let me know and I'll make the accounts for them.
- Google cloud credits
  - Email TAs

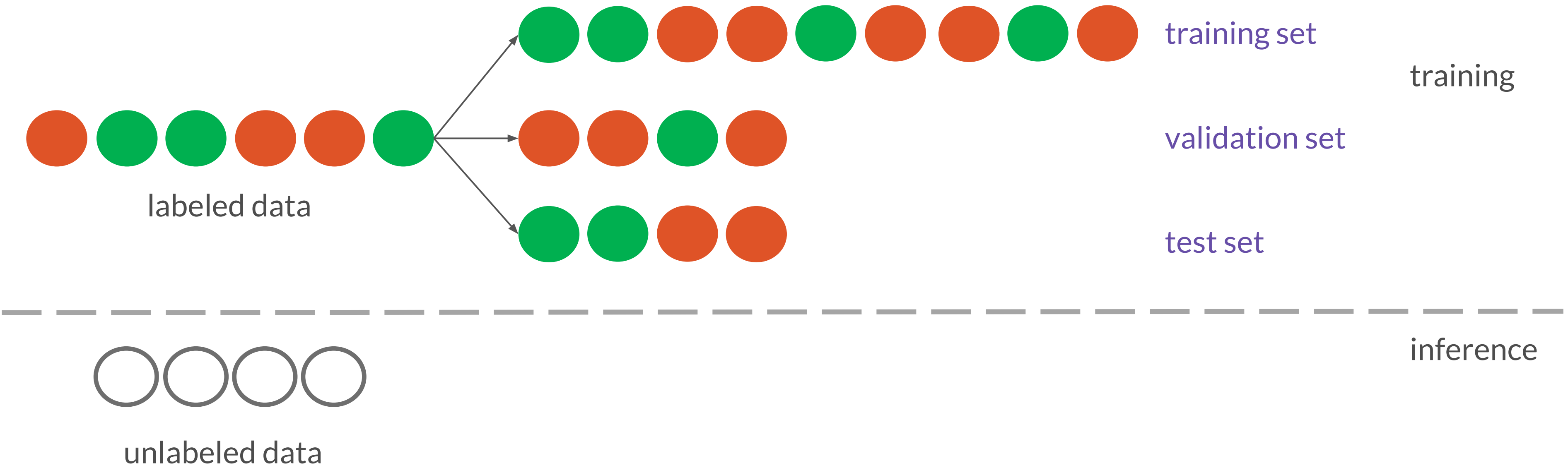
Readings

- Eis 2 https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
- J&M III 4 https://web.stanford.edu/~jurafsky/slp3/4.pdf
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of EMNLP, 2002
- Andrew Y. Ng and Michael I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, In Proceedings of NeurIPS, 2001.

Supervised classification



Training, validation, and test sets



We consider alternative models for classification

- Supervised text classification
- Rule-based
- Probabilistic
- Generative models
- Discriminative models
- Naïve Bayes
- Linear models
  - Multinomial logistic regression (aka MaxEnt)
- Non-linear models
  - Multilayer perceptron

Generative and discriminative models

imagenet

imagenet

Generative model

- Build a model of what's in a cat image
  - Knows about whiskers, ears, eyes
  - Assigns a probability to any image:
  - how cat-y is this image?
- Also build a model for dog images

imagenet

imagenet

Now given a new image:

Run both models and see which one fits better

Discriminative model

Just try to distinguish dogs from cats

Oh look, dogs have collars! Let's ignore everything else

Generative and discriminative models

- Generative model: a model that calculates the probability of the input data itself

$$P(X, Y)$$

joint

- Discriminative model: a model that calculates the probability of a latent trait given the data

$$P(Y \mid X)$$

conditional

Generative and discriminative models

- Generative text classification: Learn a model of the joint P(X, y), and find

$$\hat{y} = \underset{\tilde{y}}{\operatorname{argmax}}\; P(X, \tilde{y})$$

- Discriminative text classification: Learn a model of the conditional P(y | X), and find

$$\hat{y} = \underset{\tilde{y}}{\operatorname{argmax}}\; P(\tilde{y}|X)$$

Andrew Y. Ng and Michael I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, Advances in Neural Information Processing Systems 14 (NIPS), 2001.

Finding the correct class c from a document d in Generative vs Discriminative Classifiers

- Naive Bayes

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

- Logistic Regression

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overset{\text{posterior}}{P(c|d)}$$

We’ll consider alternative models for classification

- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
      - Naïve Bayes
    - Discriminative models
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron

Generative text classification: Naïve Bayes

$$C_{NB} = \underset{c}{\text{argmax}} P(c|d) = \underset{c}{\text{argmax}} \frac{P(d|c)P(c)}{P(d)} \propto$$

Bayes rule

$$\underset{c}{\text{argmax}} P(d|c)P(c) =$$

same denominator

$$\underset{c}{\text{argmax}} P(w_1, w_2, \ldots, w_n|c)P(c) =$$

representation

$$\underset{c_j}{\text{argmax}}\ P(c_j) \prod_i P(w_i|c)$$

conditional independence

Underflow prevention: log space

- Multiplying lots of probabilities can result in floating-point underflow
- Since log(xy) = log(x) + log(y)
  - better to sum logs of probabilities instead of multiplying probabilities
- Class with highest un-normalized log probability score is still most probable

$$C_{NB} = \underset{c_j}{\operatorname{argmax}}\ P(c_j) \prod_i P(w_i|c)$$

$$C_{NB} = \underset{c_j}{\operatorname{argmax}}\ log(P(c_j)) + \sum_i log(P(w_i|c))$$

- Model is now just max of sum of weights

Learning the multinomial naïve Bayes

- How do we learn (train) the NB model?

Learning the multinomial naïve Bayes

- How do we learn (train) the NB model?
- We learn $P(c)$ and $P(w_i|c)$ from training (labeled) data

$$C_{NB} = \underset{c_j}{\operatorname{argmax}}\ log(\underline{P(c_j)}) + \sum_i log(\underline{P(w_i|c)})$$

Parameter estimation for NB

- Parameter estimation during training
- Concatenate all documents with category c into one mega-document
- Use the frequency of w_i in the mega-document to estimate the word probability

$$\mathrm{C}_{NB} = \underset{c_j}{\operatorname{argmax}}\, log(P(c_j)) + \textstyle\sum_i log(P(w_i|c))$$

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}} \qquad \hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Parameter estimation for NB

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

- fraction of times word w_i appears among all words in documents of topic c_j

- Create mega-document for topic j by concatenating all docs in this topic
  - Use frequency of w in mega-document

Problem with Maximum Likelihood

- What if we have seen no training documents with the word “fantastic” and classified in the topic positive?

Problem with Maximum Likelihood

- What if we have seen no training documents with the word “fantastic” and classified in the topic positive?

$$\hat{P}(\text{“}fantastic\text{”}|c = \text{positive}) = \frac{count(\text{“}fantastic\text{”}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$\underset{c_j}{\text{argmax}}\ P(c_j) \prod_i P(w_i|c)$$

Laplace (add-1) smoothing for naïve Bayes

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)}$$

Laplace (add-1) smoothing for naïve Bayes

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)}$$

$$= \frac{count(w_i, c_j) + 1}{(\sum_{w \in V}(count(w, c_j)) + |V|}$$

- Note about log space

Example

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Example

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

Priors:

$P(c)= \frac{3}{4}$

$P(j)= \frac{1}{4}$

Example

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|c) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|c) = (0+1) / (8+6) = 1/14

P(Japan|c) = (0+1) / (8+6) = 1/14

P(Chinese|j) = (1+1) / (3+6) = 2/9

P(Tokyo|j) = (1+1) / (3+6) = 2/9

P(Japan|j) = (1+1) / (3+6) = 2/9

Example

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Priors:

$P(c)= \frac{3}{4}$

$P(j)= \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

P(c|d5) $\propto$ 3/4 * $(3/7)^3$ * 1/14 * 1/14
≈ 0.0003

P(j|d5) $\propto$ 1/4 * $(2/9)^3$ * 2/9 * 2/9
≈ 0.0001

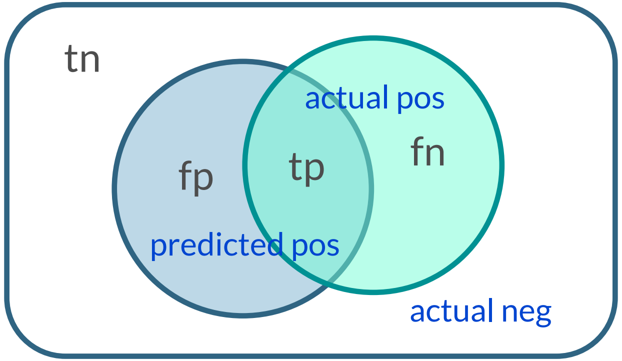
Summary: naïve Bayes is not so naïve

- Naïve Bayes is a probabilistic model
- Naïve because is assumes features are independent of each other for a class
- Very fast, low storage requirements
- Robust to Irrelevant Features
- Very good in domains with many equally important features
- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification
  - But we will see other classifiers that give better accuracy

Classification evaluation

- Contingency table: model's predictions are compared to the correct results
  - a.k.a. confusion matrix

	actual pos	actual neg
predicted pos	true positive (tp)	false positive (fp)
predicted neg	false negative (fn)	true negative (tn)



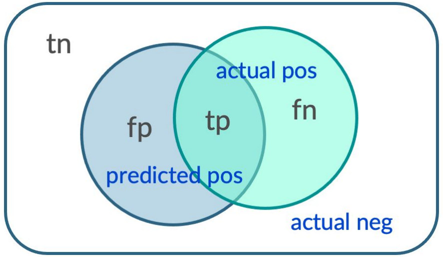
Classification evaluation

- Borrowing from Information Retrieval, empirical NLP systems are usually evaluated using the notions of precision and recall

Classification evaluation

- Precision (P) is the proportion of the selected items that the system got right in the case of text categorization
  - it is the % of documents classified as “positive” by the system which are indeed “positive” documents
- Reported per class or average

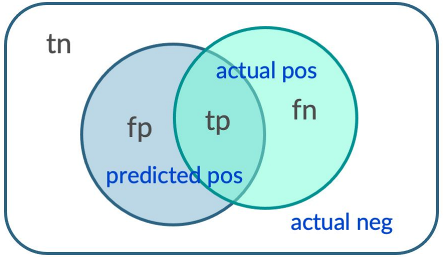
$$\text{precision} = \frac{\textit{true positives}}{\textit{true positives} + \textit{false positives}} = \frac{tp}{tp + fp}$$



Classification evaluation

- Recall (R) is the proportion of actual items that the system selected in the case of text categorization
  - it is the % of the “positive” documents which were actually classified as “positive” by the system
- Reported per class or average

$$\text{recall} = \frac{\textit{true positives}}{\textit{true positives} + \textit{false negatives}} = \frac{tp}{tp + fn}$$



Classification evaluation

- We often want to trade-off precision and recall
  - typically: the higher the precision the lower the recall
  - can be plotted in a precision-recall curve
- It is convenient to combine P and R into a single measure
  - one possible way to do that is F measure

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R} \quad \text{for } \beta=1,\ F_1 = \frac{2PR}{P+R}$$

Classification evaluation

- Additional measures of performance: accuracy and error
  - accuracy is the proportion of items the system got right
  - error is its complement

	actual pos	actual neg
predicted pos	true positive (tp)	false positive (fp)
predicted neg	false negative (fn)	true negative (tn)

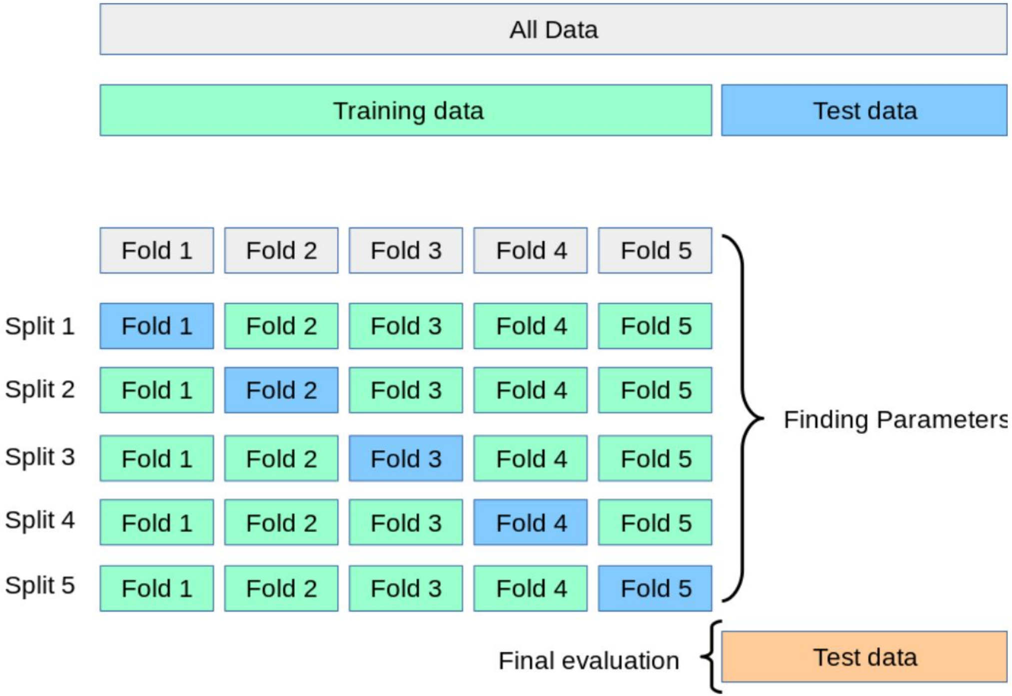
$$\text{accuracy} = \frac{tp + tn}{tp + fp + tn + fn}$$

Micro- vs. macro-averaging

If we have more than one class, how do we combine multiple performance measures into one quantity?

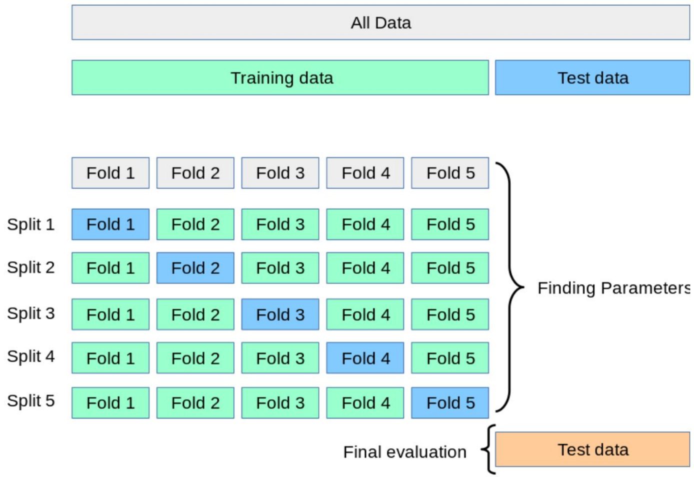
- Macroaveraging
  - Compute performance for each class, then average.
- Microaveraging
  - Collect decisions for all classes, compute contingency table, evaluate.

K-fold cross-validation



K-fold cross-validation

- Metric: P/R/F1 or Accuracy
- Unseen test set
  - avoid overfitting ('tuning to the test set')
  - more conservative estimate of performance
- Cross-validation over multiple splits
      - Handles sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance

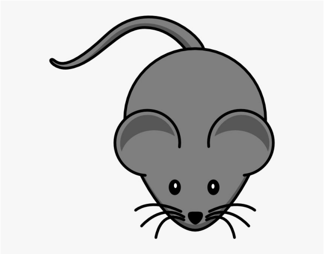


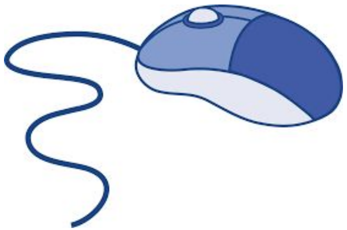
Back to the introduction topics...

Why is language interpretation hard?

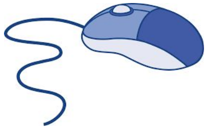
1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Ambiguity: word sense disambiguation





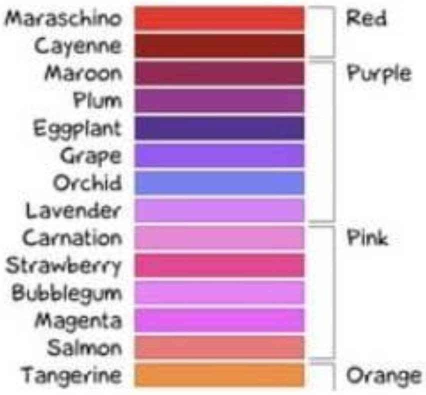
Ambiguity



- Ambiguity at multiple levels:
  - Word senses: **bank** (finance or river?)
  - Part of speech: **chair** (noun or verb?)
  - Syntactic structure: **I can see a man with a telescope**
  - Multiple: **I saw her duck**

Semantic analysis

- Every language sees the world in a different way
  - For example, it could depend on cultural or historical conditions



  - Russian has very few words for colors, Japanese has hundreds
  - Multiword expressions, e.g. happy as a clam, it's raining cats and dogs or wake up and metaphors, e.g. love is a journey are very different across languages

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Scale

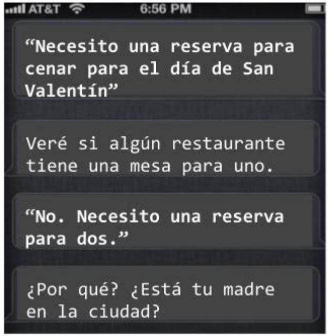
- ~7K languages
- Thousands of language varieties

Englishes

Africa is a continent with a very high linguistic diversity: there are an estimated 1.5-2K African languages from 6 language families. 1.33 billion people

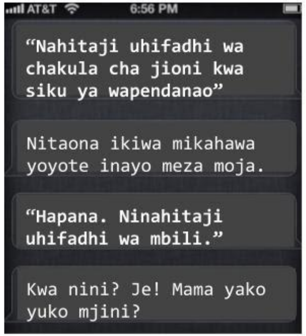
NLP beyond English

- ~7,000 languages
- thousands of language varieties



Spanish
534 million speakers

Hindi
615 million speakers



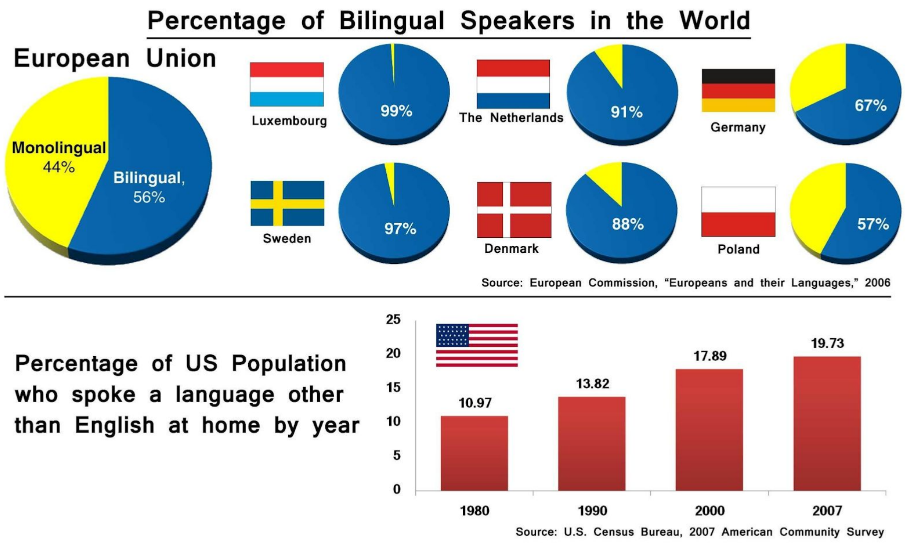
Swahili
100 million speakers

American English

Scottish English

Hinglish

Most of the world today is multilingual



Source: US Census Bureau

Source: Ethnologue

Tokenization

这是一个简单的句子

WORDS — This is a simple sentence

זה משפט פשוט

Tokenization + disambiguation

in tea her daughter	בתה

- most of the vowels unspecified

in tea	בתה
in the tea	בהתה
that in tea	שבתה
that in the tea	שבהתה
and that in the tea	ושבהתה

ושבתה

and her saturday	ו+שבת+ה
and that in tea	ו+ש+ב+תה
and that her daughter	ו+ש+בת+ה

- most of the vowels unspecified
- particles, prepositions, the definite article, conjunctions attach to the words which follow them
- tokenization is highly ambiguous

Tokenization + morphological analysis

- Quechua

Much’ananayakapushasqakupuniñataqsunamá

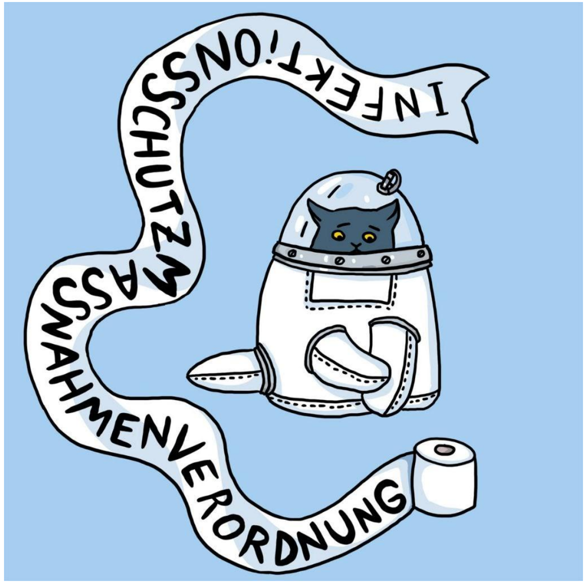
Much’a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

“So they really always have been kissing each other then”

Much’a	to kiss
-na	expresses obligation, lost in translation
-naya	expresses desire
-ka	diminutive
-pu	reflexive (kiss *eachother*)
-sha	progressive (kiss*ing*)
-sqa	declaring something the speaker has not personally witnessed
-ku	3rd person plural (they kiss)
-puni	definitive (really*)
-ña	always
-taq	statement of contrast (...then)
-suna	expressing uncertainty (So...)
-má	expressing that the speaker is surprised

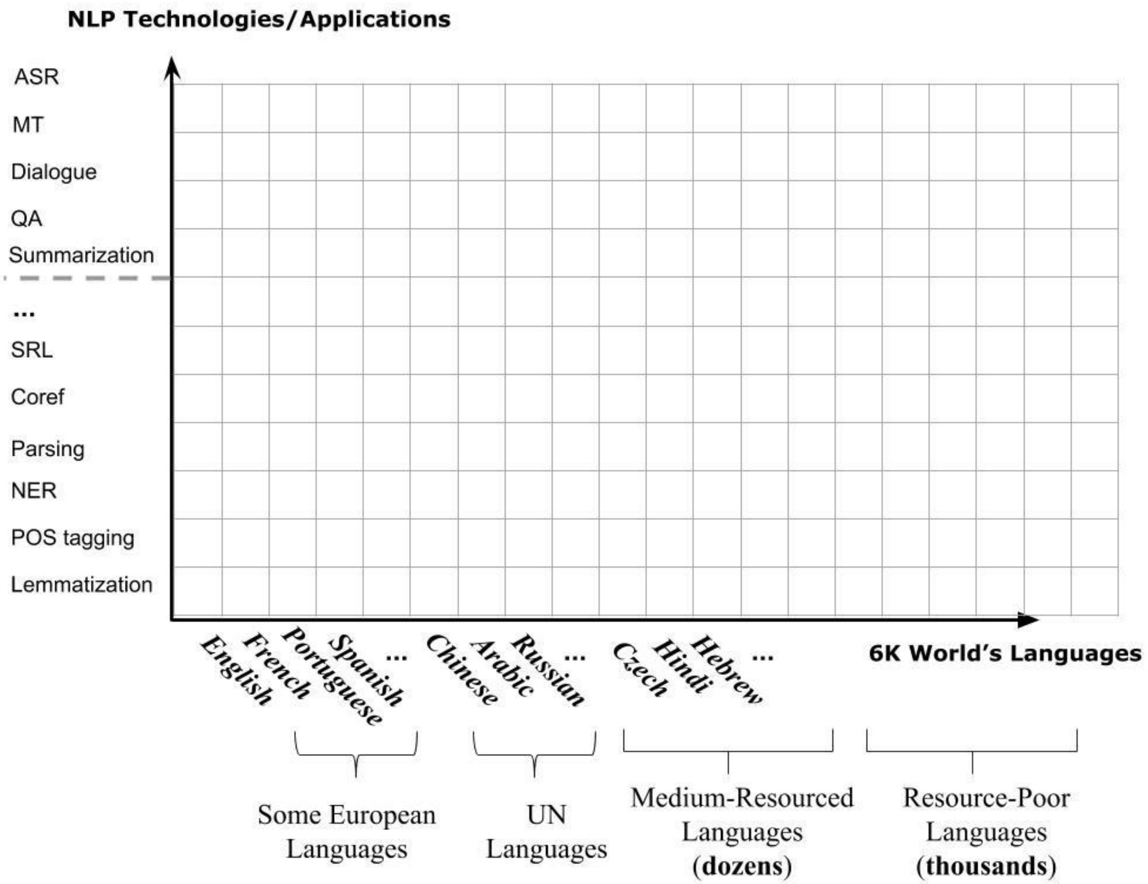
Tokenization + morphological analysis

- German



Infektionsschutzmaßnahmenverordnung

“Infection Protection Measures Ordinance”


NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
6K World's Languages
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

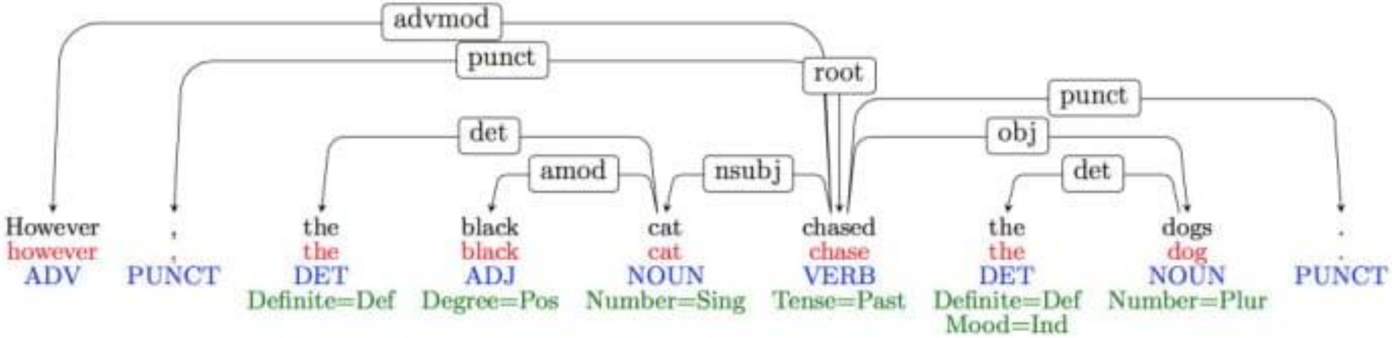
Linguistic variation

- Non-standard language, emojis, hashtags, names

chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂 🤷‍♀️ Note the cute little crab indicator in the 2nd pic 🦀 💕

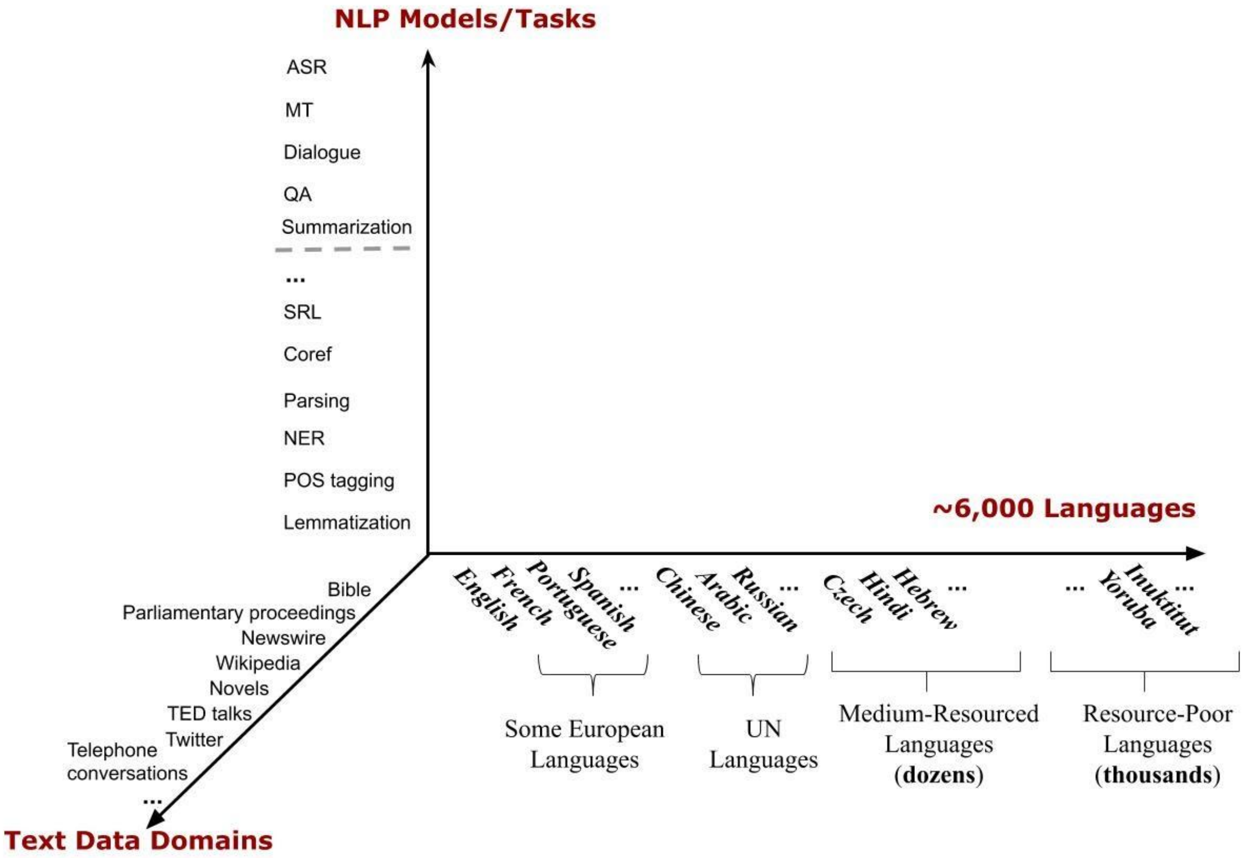
Variation

- Suppose we train a part of speech tagger or a parser on the Wall Street Journal



- What will happen if we try to use this tagger/parser for social media??

@_rkpntrnte hindi ko alam babe eh, absent ako
kanina I'm sick rn hahaha 😌🙌


NLP Models/Tasks
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
~6,000 Languages
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
...
Yoruba
Inuktitut
...
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)
Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations
...
Text Data Domains

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Sparsity

Sparse data due to Zipf's Law

- To illustrate, let's look at the frequencies of different words in a large text corpus
- Assume "word" is a string of letters separated by spaces

Word Counts

Most frequent words in the English Europarl corpus (out of 24m word tokens)

any word

Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

nouns

Frequency	Token
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

Word Counts

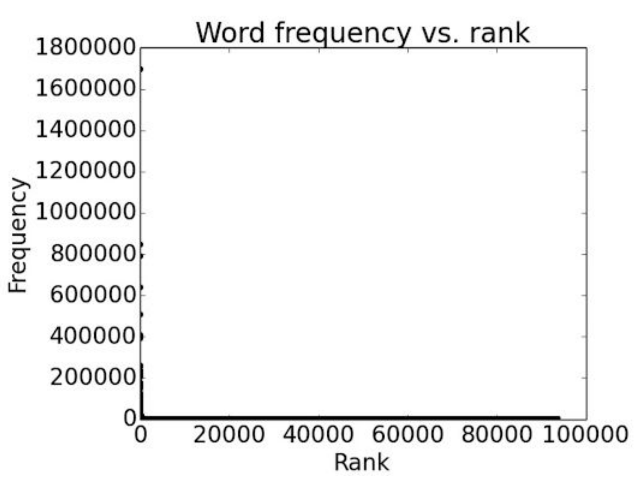
But also, out of 93,638 distinct words (word types), 36,231 occur only once.

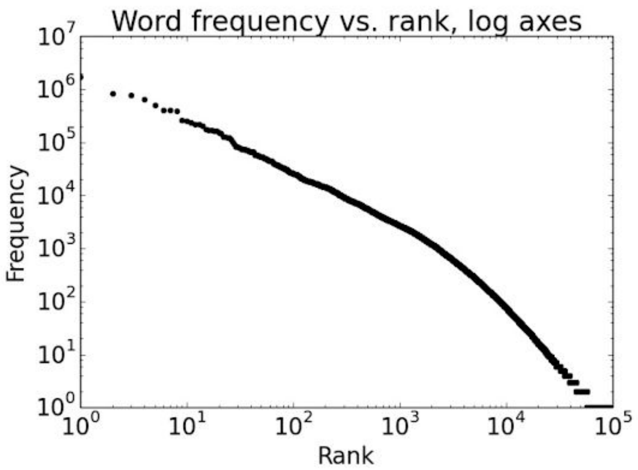
Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a

Plotting word frequencies

Order words by frequency. What is the frequency of nth ranked word?

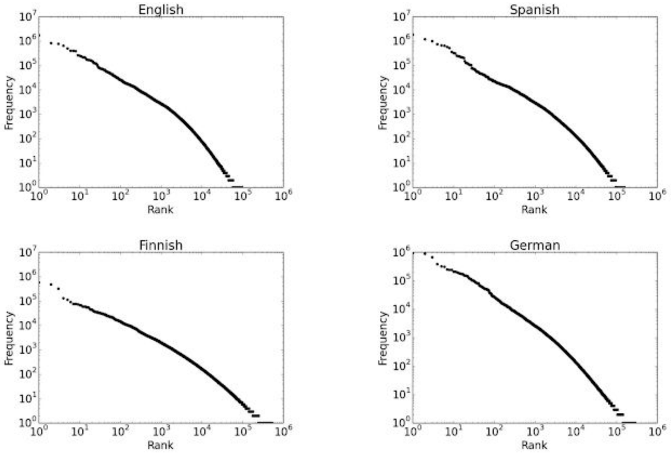




Zipf's Law

Implications

- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen



Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Expressivity

Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:

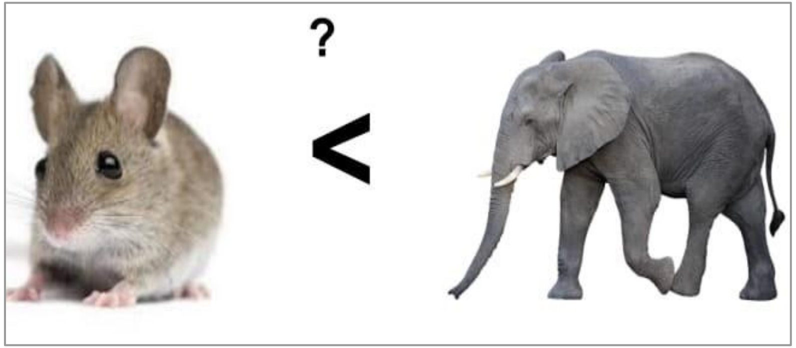
She gave the book to Tom	vs.	She gave Tom the book
Some kids popped by	vs.	A few children visited
Is that window still open?	vs.	Please close the window

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Unmodeled variables

"Drink this milk"



World knowledge

- I dropped the glass on the floor and it broke
- I dropped the hammer on the glass and it broke

Why is language interpretation hard?

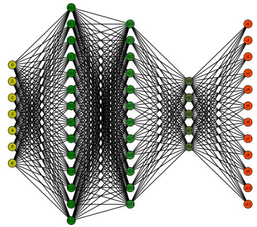
1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation $\mathcal{R}$

Unknown representation

- Very difficult to capture what is $\mathcal{R}$, since we don't even know how to represent the knowledge a human has/needs:
  - What is the "meaning" of a word or sentence?
  - How to model context?
  - Other general knowledge?

Dealing with ambiguity

- How can we model ambiguity and choose the correct analysis in context?
  - non-probabilistic methods (FSMs for morphology, CKY parsers for syntax) return *all possible analyses*.
  - probabilistic models (HMMs for part-of-speech tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return *the best possible analysis,* i.e., the most probable one according to the model
  - Neural networks, pretrained language models now provide end-to-end solutions



- But the “best” analysis is only good if our probabilities are accurate. Where do they come from?

Next class: Logistic regression

- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
      - Naïve Bayes
    - Discriminative models
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron

Logistic regression classifier

- Important analytic tool in natural and social sciences
- Baseline supervised machine learning tool for classification
- Is also the foundation of neural networks