Foundations of Natural Language Processing

Lecture 2: Ambiguity and Probabilistic Models

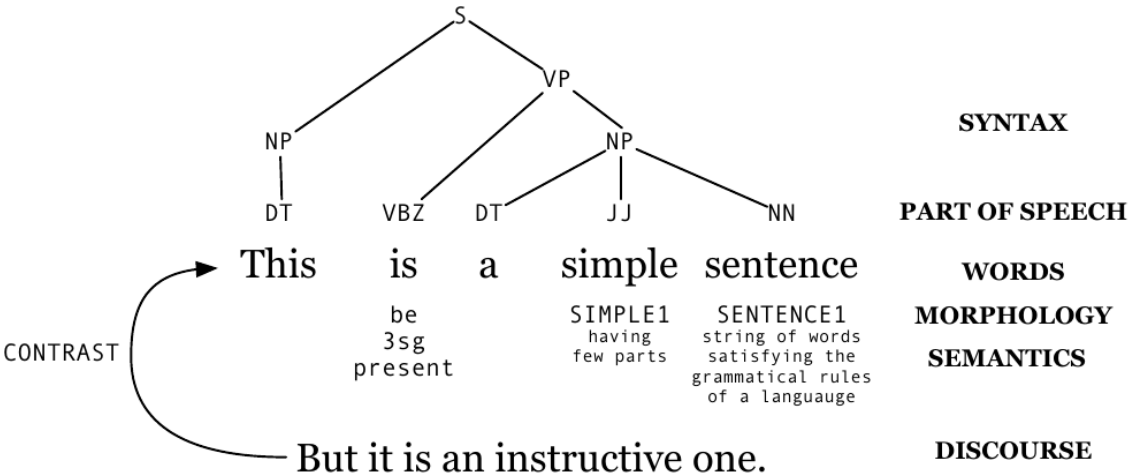
Mirella Lapata
School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

Slides based on content from: Philipp Koehn, Alex Lascarides, Sharon Goldwater, Shay Cohen, Khalil Sima'an, Ivan Titov

Plan for today

- Recap, this is FNLP!
- Continue discussing why NLP is hard
- Why should we use probabilistic models (and machine learning) for NLP?

Levels of Structure

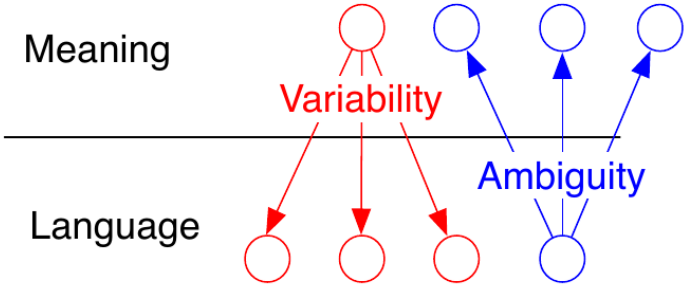


Manifest themselves in applications

London Bridge really is falling down. The bridge is being taken apart and moved. Its new home will be a small town in Arizona. This bridge is hundreds of years old. It stretches across the Thames River. Robert McCulloch saw this bridge and decided to bring it to the USA. He paid more than two million dollars. It will cost him more than three million dollars to move it. Each stone will be marked. The pieces must fit when they reach their new home. All that work will not take place overnight. The job will take six years. The bridge is not small. It is longer than three football fields. It is almost as wide as one football field. In time, the London Bridge will stand high above a new river. Flags will be placed at both ends. Cars will cross it. A small town will be built next to the bridge. Most people in Arizona will never see London. But they will see a part of it in their own state.

1. Who bought a bridge?
2. Where will the bridge be re-built?
3. How long will it take?

Why is NLP hard? Ambiguity at many levels



Variability:

He drew the house
He made a sketch of the house
He showed me his drawing of the house
He portrayed the house in his paintings
He drafted the house in his sketchbook

Ambiguity:

She drew a picture of herself
cart drawn by two horses...
He drew crowds wherever he went ...
The driver slowed as he drew even with me
The officer drew a gun and pointed it at ...

Why is NLP hard? Ambiguity at many levels

- Ambiguity takes place at different levels (syntax, semantics, co-reference, discourse).
- Leads to combinatorial explosion of possible analyses!

Example with 3 preposition phrases, 5 interpretations:

- Put the block ((in the box on the table) in the kitchen)
- Put the block (in the box (on the table in the kitchen))
- Put ((the block in the box) on the table) in the kitchen
- Put (the block (in the box on the table)) in the kitchen
- Put (the block in the box) (on the table in the kitchen)

Why is NLP hard? Sparsity

any word	
Frequency	Type
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	Type
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

- Let's look at at the frequencies of different words in a large text corpus.
- Assume a "word" is a string of letters separated by spaces
- Most frequent word types in the English Europarl corpus (out of 24M words).

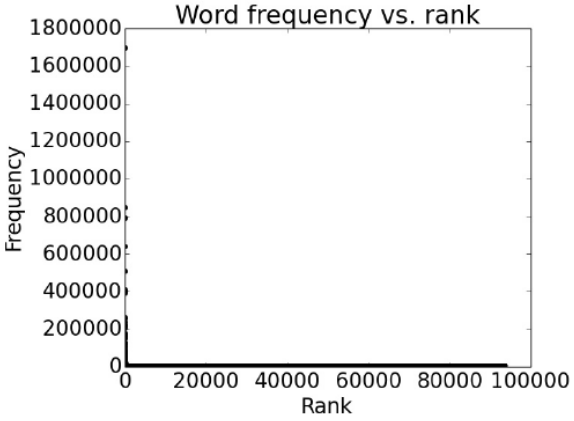
Word Counts

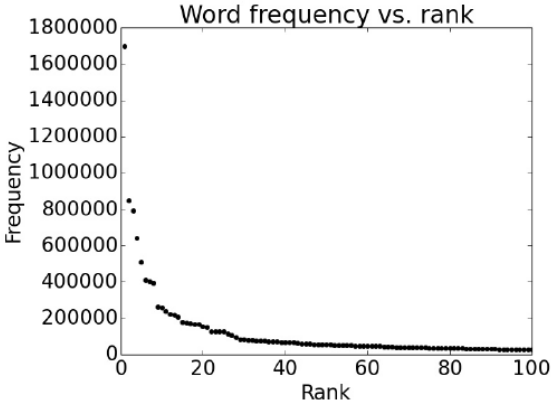
But also, out of 93,638 distinct word types 36,231 (almost 40%) occur only once.

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

Plotting Word Frequencies

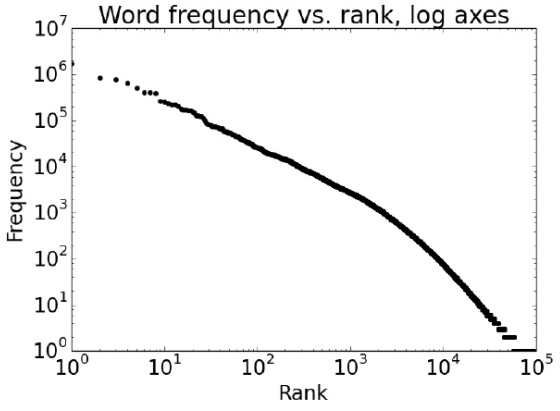
Order words by frequency. What is the frequency of nth ranked word?



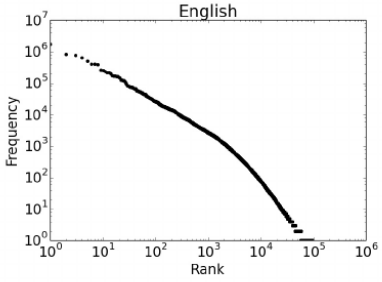


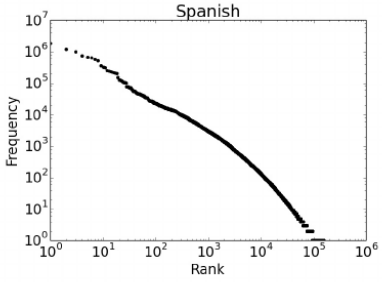
Rescaling the Axes

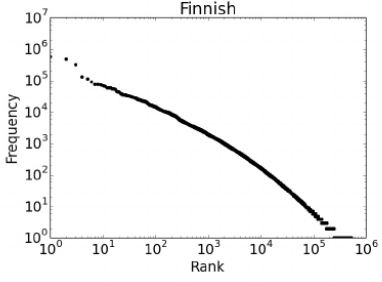
To really see what's going on, use logarithmic axes:

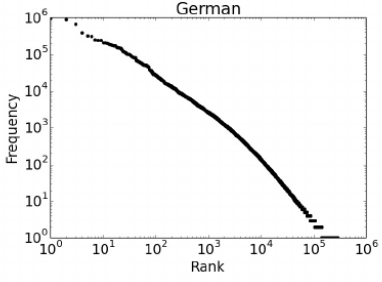


Rescaling the Axes









Zipf's Law

Summarizes the behavior we just saw:

$$f \times r \approx k \quad \text{or equivalently} \quad f(r) = \frac{k}{r}$$

- f = frequency of a word
- r = rank of a word (if sorted by frequency)
- k = a constant

George Kingsley Zipf
(1902–1950)

Why a line in log-scales?

$$f \times r = k \;\Rightarrow\; f = \frac{k}{r} \;\Rightarrow\; \log f = \log\left(\frac{k}{r}\right) \;\Rightarrow\; \log f = \log k - \log r$$

Implications of Zipf's Law

- Any guesses?

Implications of Zipf's Law

- Any guesses?
- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words.

Implications of Zipf's Law

- Any guesses?
- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words.
- In fact, the same holds for many other levels of linguistic structure (e.g., syntactic rules in a CFG).

Implications of Zipf's Law

- Any guesses?
- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words.
- In fact, the same holds for many other levels of linguistic structure (e.g., syntactic rules in a CFG).
- Why is this a problem?

Implications of Zipf's Law

- Any guesses?
- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words.
- In fact, the same holds for many other levels of linguistic structure (e.g., syntactic rules in a CFG).
- Why is this a problem?
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen during training.

Why is NLP hard? Robustness

Will our NLP systems work across styles, domains, registers, genres?

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

> Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsevier/NNP N.V./NNP ,/, the/DT Dutch/NNP publishing/VBG group/NN ./.

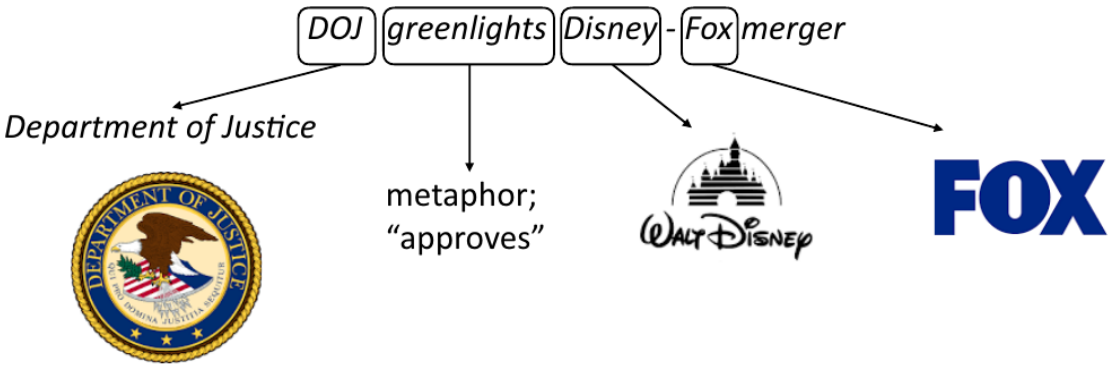
- What will happen if we try to use this tagger for social media??

> Pikr smh he asked fir yo last name

- New words will always be created; NLP algorithms must be robust to new observations.

Twitter example from Noah Smith

Why is NLP hard? World Knowledge and Context



Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)

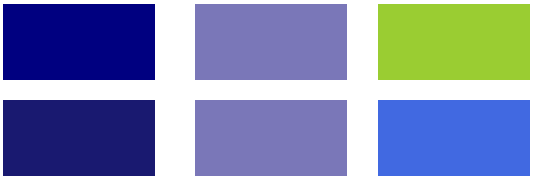
Context	Utterance
	blue

Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)

Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)

Context	Utterance
	blue
	The darker blue one

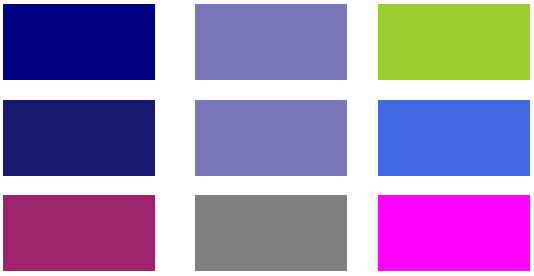
Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)

Context	Utterance
	blue
	The darker blue one

Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)

Context	Utterance
	blue
	The darker blue one
	dull pink not the super bright one

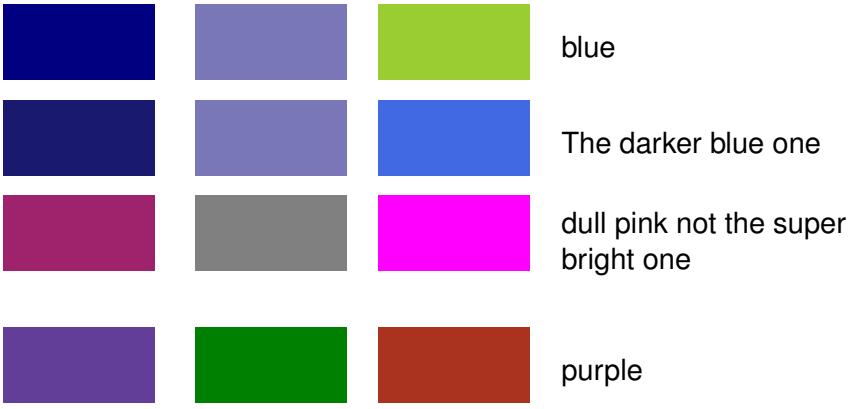
Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)

Context	Utterance
	blue
	The darker blue one
	dull pink not the super bright one

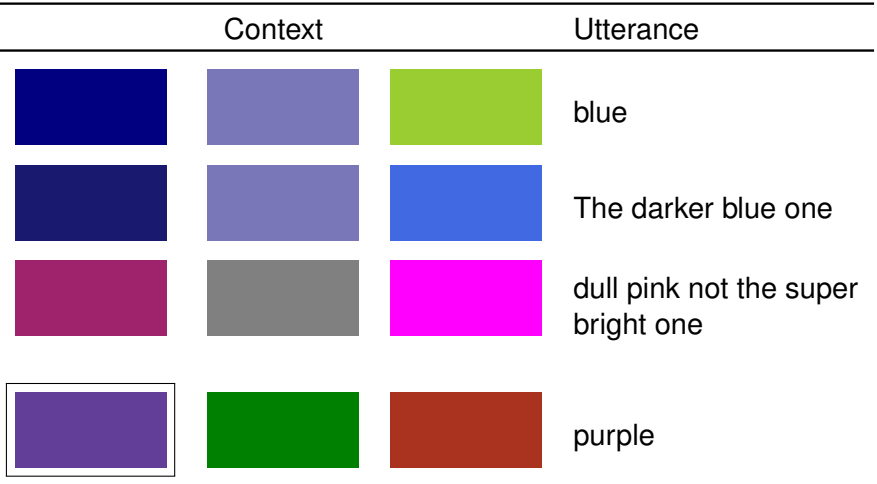
Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)



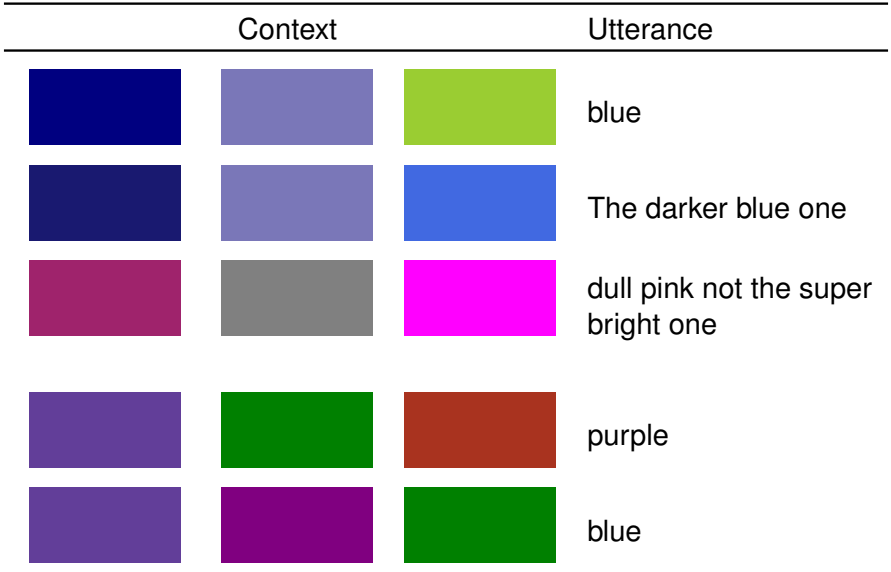
Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)



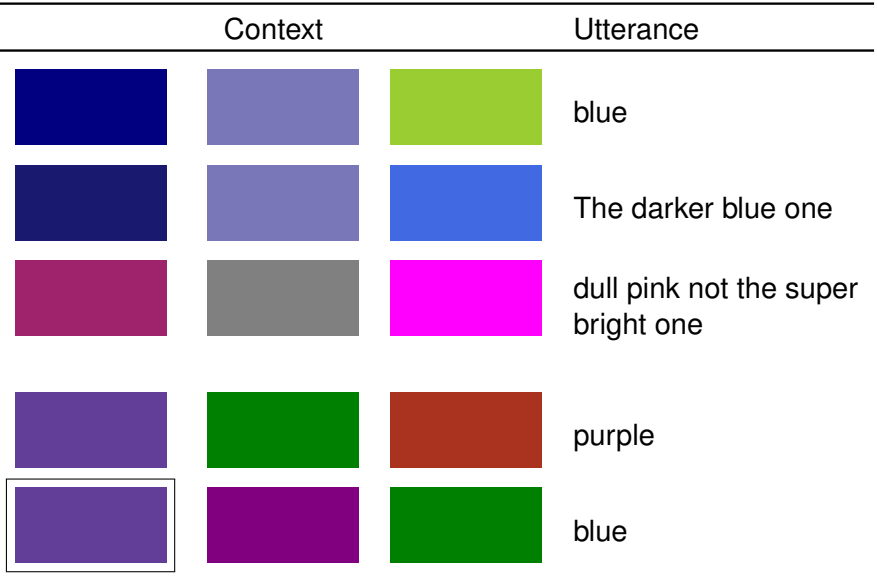
Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)



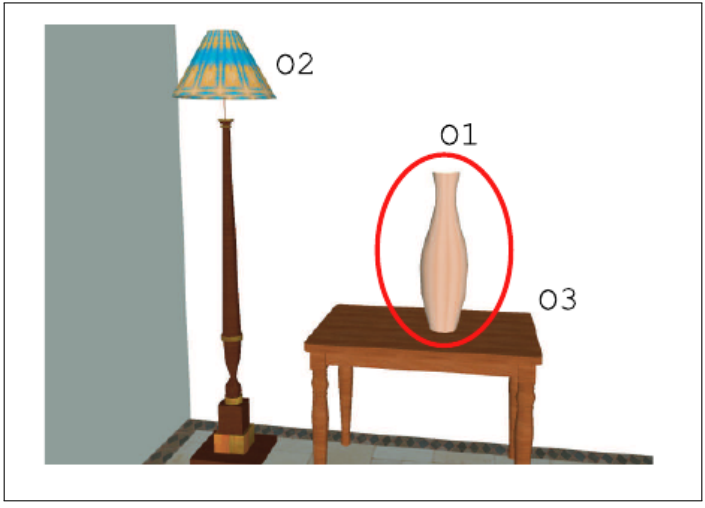
Colors in Context

Stanford Colors in Context corpus (Monroe et al. 2017)



Natural Language is Grounded

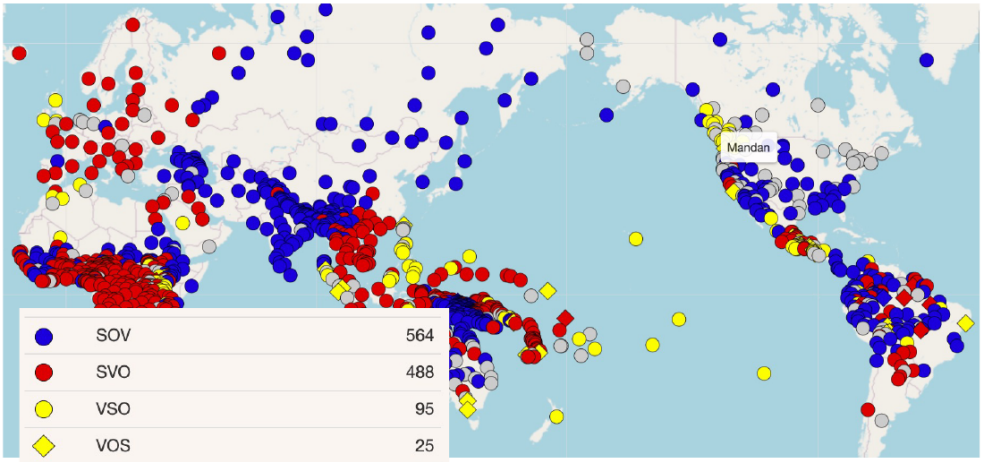
How do you refer to object 01? (a) right of 02 (b) on top of 03



Why is NLP hard? Language Diversity

- In this class, we will focus on English
- ... as most of the NLP community does
- However, we would like to develop technologies which work for other languages
- One example: morphology and syntax

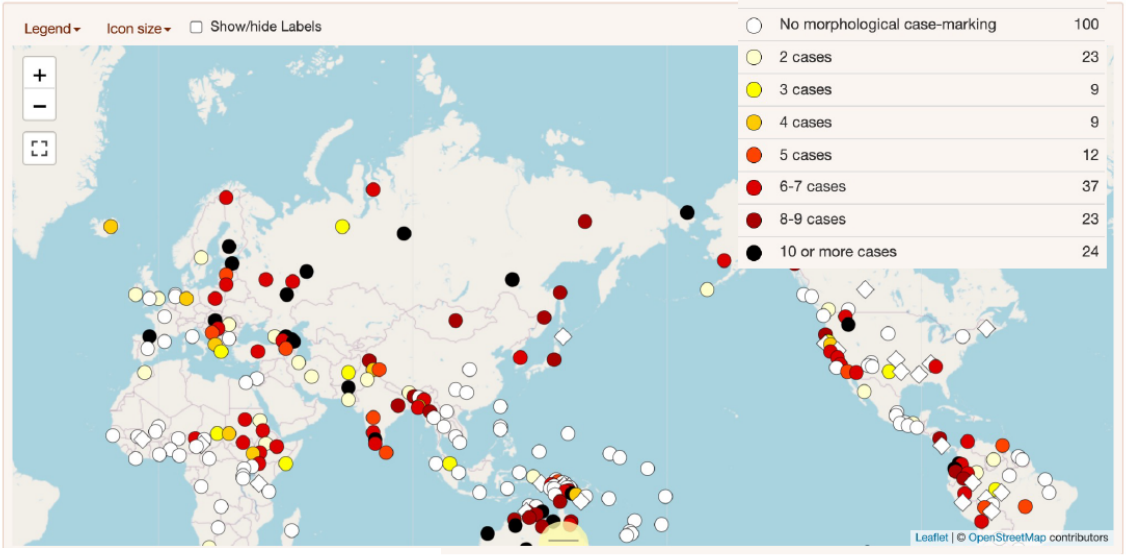
Syntactic Diversity



(S)ubject, (V)erb, (O)bject order

World Atlas of Language Structures (wals.info)

Number of Cases



World Atlas of Language Structures (wals.info)

Number of Cases

Word order freedom and morphology are inter-related. The more freedom in word order:

- The less information is conveyed by word positions
- The more information should be included in the “tokens”
- The richer morphology

English vs Russian (object marked with ей)

- Cats eat mice
- Кошки едят мышей
- Мышей едят кошки
- Едят кошки мышей
- Едят мышей кошки.

Constrained word order
Limited or no morphological marking

(Relatively) free word order
Rich morphology

Recap: Summary of Challenges

1. variability
2. ambiguity
3. sparsity
4. robustness
5. context dependence
6. language diversity

Uncertainty

Most challenges we discussed can be regarded as manifestations of uncertainty!

Ambiguity: uncertainty with respect to interpretation

Variability: uncertainty in a specific realization for a semantic concept

Robustness: uncertainty with respect to potential inputs

Lack of knowledge (cf sparsity / context dependence): uncertainty!

We need: probabilistic models / machine learning

Sketch of a probabilistic model

Many natural language processing problems can be written mathematically as:

$$\hat{\mathbf{y}} = \underset{\mathbf{y} \in \mathcal{Y}(\mathbf{x})}{\operatorname{argmax}} P(\mathbf{x}, \mathbf{y})$$

- $\mathbf{x}$ is the input, which is an element of a set $\mathcal{X}$
- $\mathbf{y}$ is the output, which is an element of a set $\mathcal{Y}$
- $P(\mathbf{x}, \mathbf{y})$ is the **model**, which maps from the set $\mathcal{X} \times \mathcal{Y} \rightarrow [1, 0]$;
- $\hat{\mathbf{y}}$ is the predicted output, which is chosen to maximize $P(\mathbf{x}, \mathbf{y})$.
- finding $\hat{\mathbf{y}}$ is often a **search** problem
- estimating $P(\mathbf{x}, \mathbf{y})$ is a **learning** problem.

Sketch of a probabilistic model: Spam Classification

- Input $\mathcal{X}$ is a set of emails.
- Output $\mathcal{Y}$ is 1 (spam) or 0 (not spam).
- Model sees lots of $(\mathbf{x}, \mathbf{y})$ pairs (emails labeled with 0/1)
- Goal is to learn to predict labels of new, future emails.
- How do we obtain probabilities $P(\mathbf{x}, \mathbf{y})$?
- How do we measure success?

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

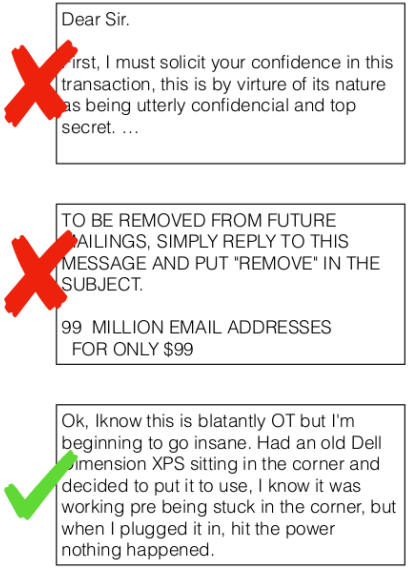
TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Sketch of a probabilistic model: Spam Classification

- Input $\mathcal{X}$ is a set of emails.
- Output $\mathcal{Y}$ is 1 (spam) or 0 (not spam).
- Model sees lots of $(\mathbf{x}, \mathbf{y})$ pairs (emails labeled with 0/1)
- Goal is to learn to predict labels of new, future emails.
- How do we obtain probabilities $P(\mathbf{x}, \mathbf{y})$?
- How do we measure success?



Take-home message

- There are many challenges which make NLP hard
- Ambiguity is (?) the most fundamental one
  - happens at many levels
  - can lead to a combinatorial explosion in a number ‘interpretations’
- Probabilistic modeling is a way to deal with many of these challenges
- Requires decisions at different levels (probability models, algorithms, ...)

Next lecture: we discuss corpora and evaluation methods.