Learning with Partially Ordered Representations

Jonathan Rawski
Department of Linguistics

July 24, 2019

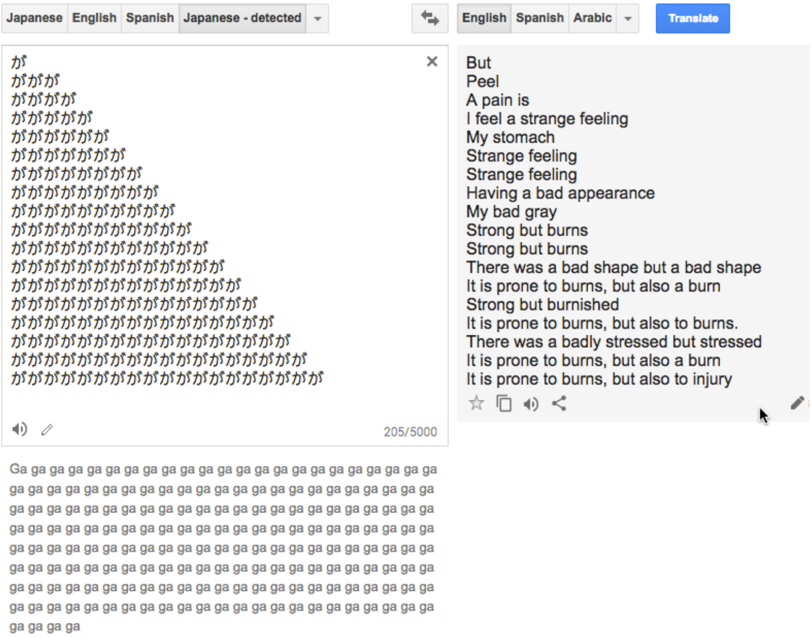
Thank you for the support!

Output in 2018-2019

- Journal Articles: 1 published, 1 in review
- Paper-Reviewed proceedings: 2
- Abstract-reviewed proceedings: 2
- Invited Talks: 6
- Conference Talks: 11
- Conference Posters: 1

The Main Idea

Learning is eased when shared properties of the domain structure the space of hypotheses


Japanese English Spanish Japanese - detected
English Spanish Arabic
Translate
が
ががが
がががが
ががががが
がががががが
ががががががが
がががががががが
ががががががががが
がががががががががが
ががががががががががが
がががががががががががが
ががががががががががががが
がががががががががががががが
ががががががががががががががが
がががががががががががががががが
ががががががががががががががががが
がががががががががががががががががが
ががががががががががががががががががが
205/5000
Ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga
ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga
ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga
ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga
ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga
ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga
ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga
ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga ga
ga ga ga ga
But
Peel
A pain is
I feel a strange feeling
My stomach
Strange feeling
Strange feeling
Having a bad appearance
My bad gray
Strong but burns
Strong but burns
There was a bad shape but a bad shape
It is prone to burns, but also a burn
Strong but burnished
It is prone to burns, but also to burns.
There was a badly stressed but stressed
It is prone to burns, but also a burn
It is prone to burns, but also to injury

Poverty of the Stimulus and Data Sparsity

Number of English words: $N \sim 10000$
Possible English 2-grams: $N^2 = 100000000$
Possible English 3-grams: $N^3 = 1000000000000$
Possible English 4-grams: $N^4 = 10000000000000000$
$\ldots$
easy learning if normal distribution

Poverty of the Stimulus and Data Sparsity

BUT:

In the million-word Brown corpus of English:
45% of words,
80% of 2-grams
95% of 3-grams
appear EXACTLY ONCE
Bad for learning: Huge long-tailed distribution

How can a machine know that new sentences like
"nine and a half turtles yodeled" is good?
"turtles half nine a the yodeled" is bad?

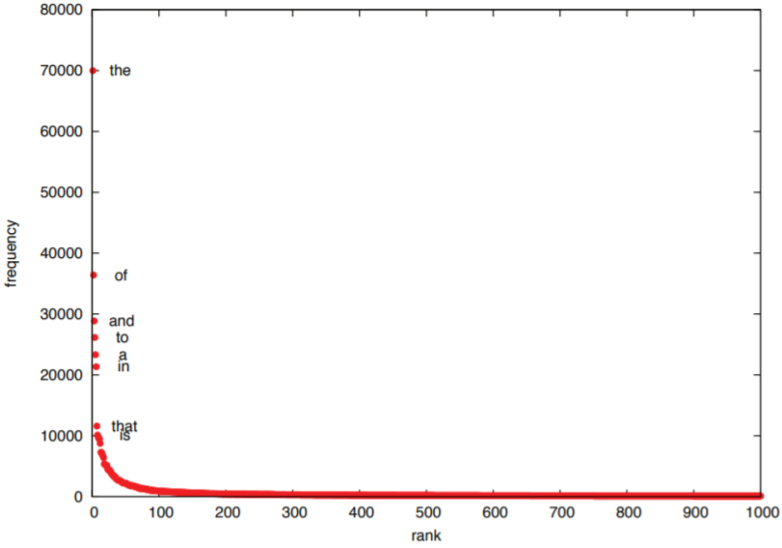
Poverty of the Stimulus and Data Sparsity

BUT:

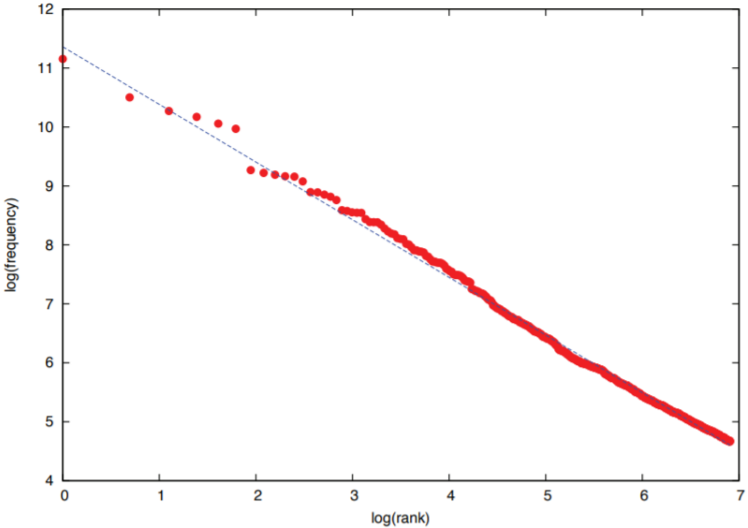
In the million-word Brown corpus of English:
45% of words,
80% of 2-grams
95% of 3-grams
appear EXACTLY ONCE
Bad for learning: Huge long-tailed distribution

How can a machine know that new sentences like
"nine and a half turtles yodeled" is good?
"turtles half nine a the yodeled" is bad?

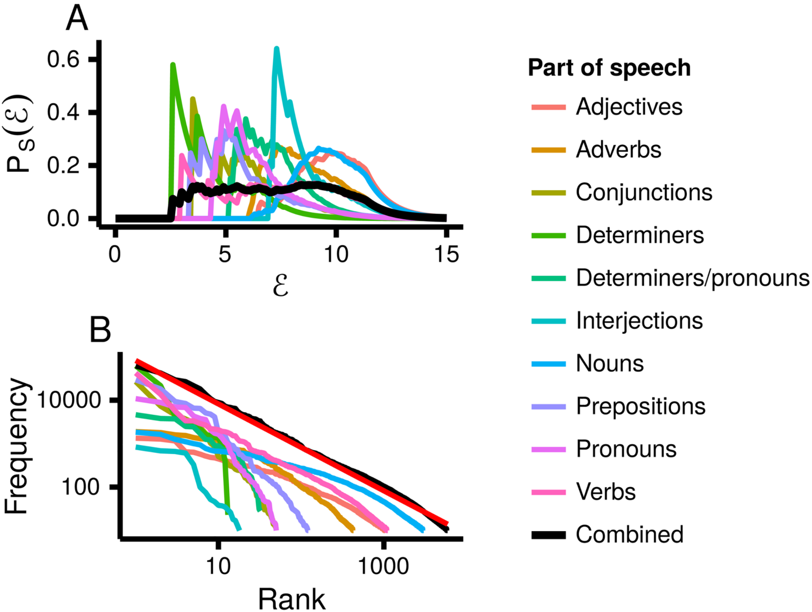
The Zipf Problem



The Zipf Problem



Zipf Emerges from Latent Features



NLP Example

In many NLP applications, text symbols are treated independently

Alphabet $= \{\mathtt{a}, \ldots, \mathtt{z}, \mathtt{A}, \ldots, \mathtt{Z}\} = 52$ symbols

Forbidding maybe all capitals $\rightarrow$ Explosion!

If we use feature [`capital`], only 27! 26 letters + [`capital`]

Learning Algorithm (Chandlee et al 2018)

What have we done so far?

- Provably correct relational learning algorithm
- Prunes Hypothesis space according to ordering relation
- Provably identifies correct constraints for sequential data
- Uses data sparsity to its advantage!

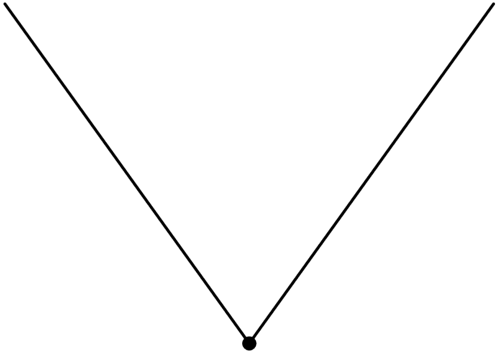
Collaborative work with:

Jane Chandlee (Haverford)

Jeff Heinz (SBU)

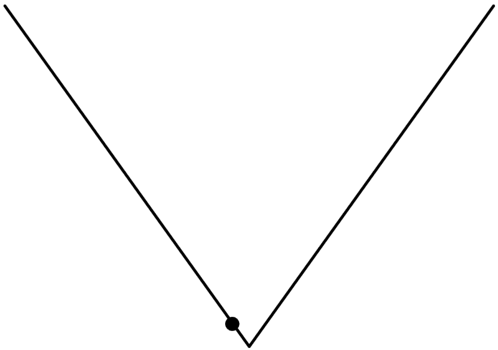
Adam Jardine (Rutgers)

2D Schema of a Semilattice of Constraints



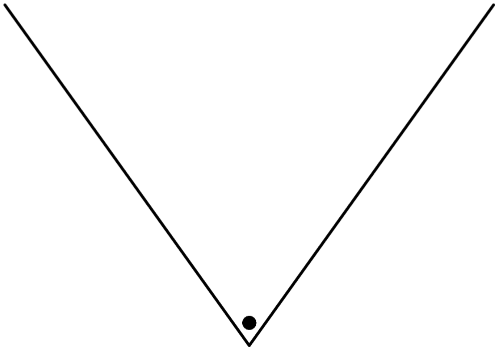
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



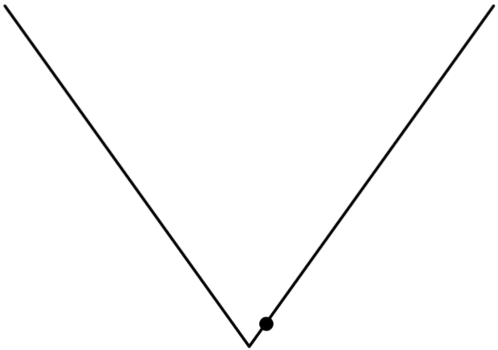
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



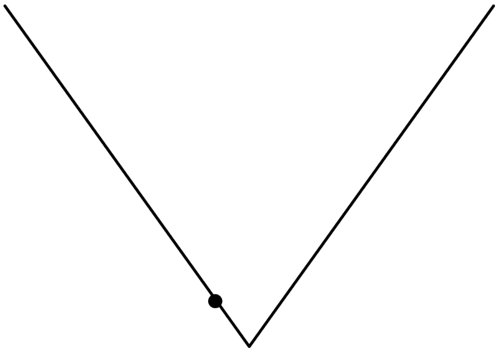
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



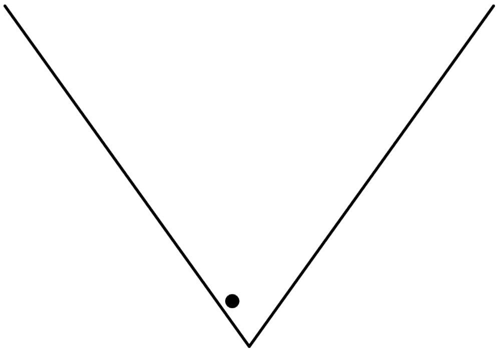
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



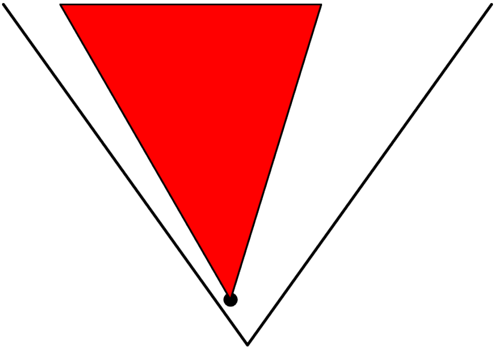
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



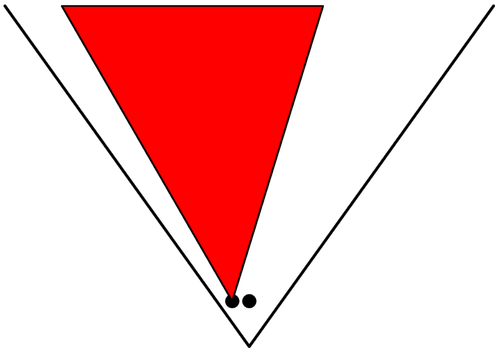
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



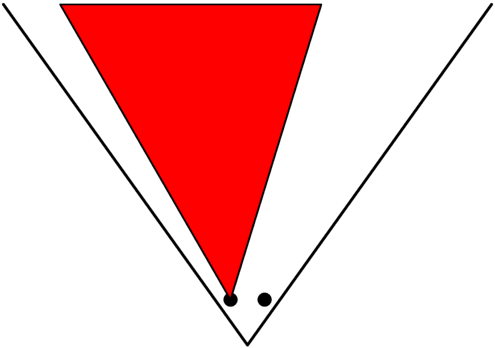
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



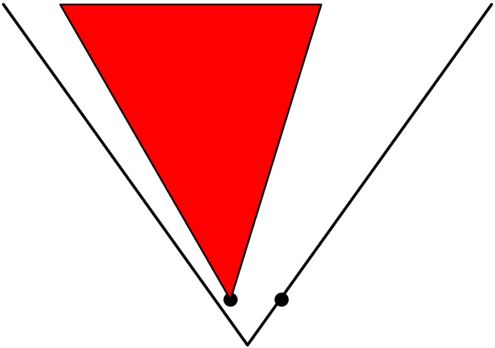
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



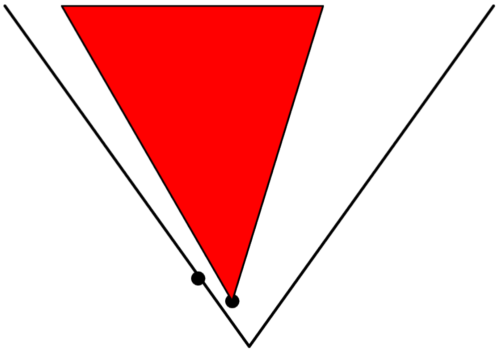
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



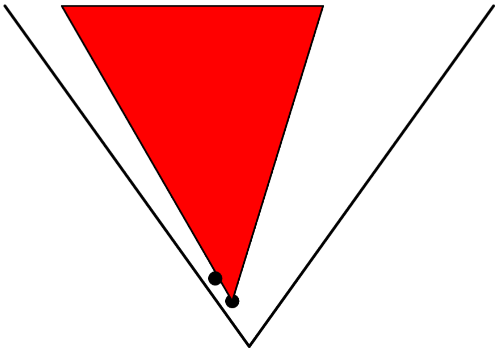
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



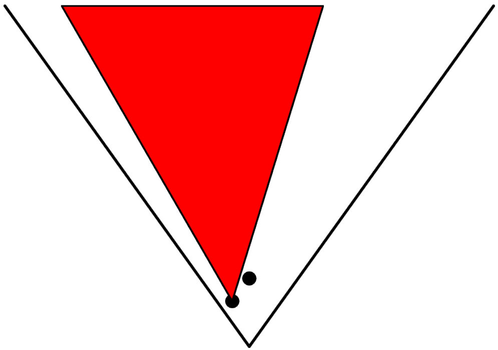
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



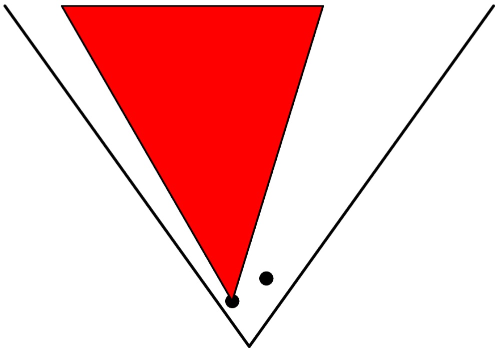
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



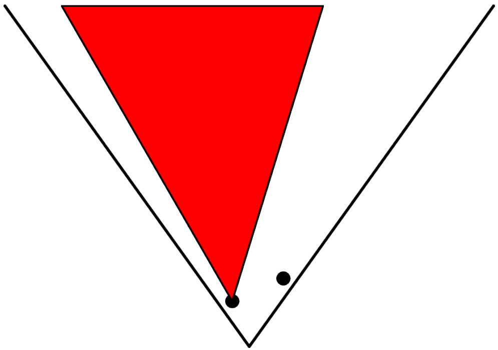
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



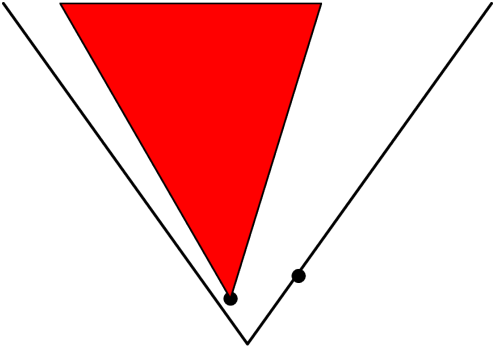
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



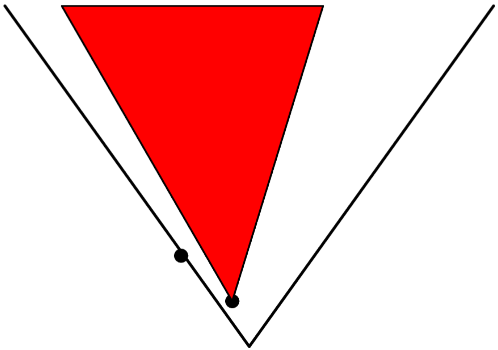
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



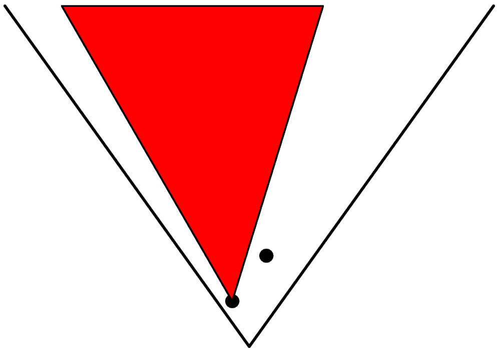
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



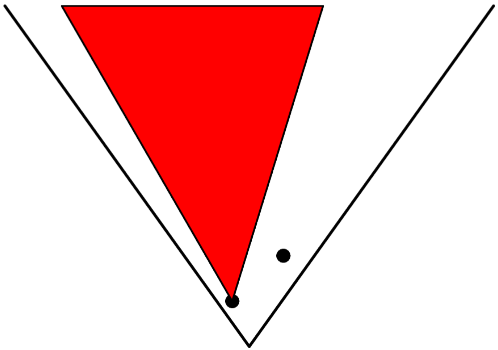
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



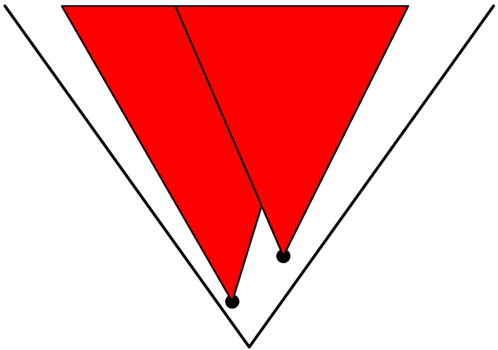
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



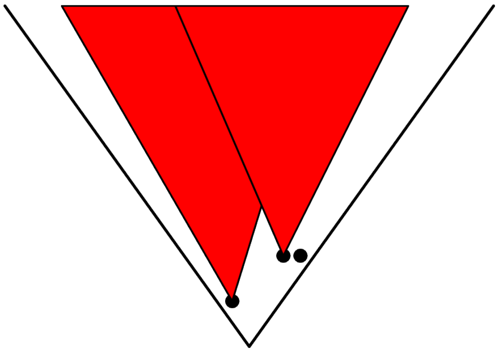
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



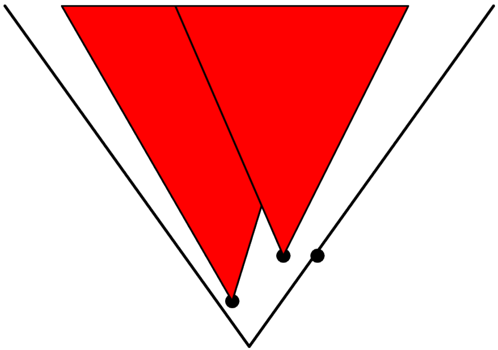
Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



Is this sub-structure in the data sample?

2D Schema of a Semilattice of Constraints



Is this sub-structure in the data sample?

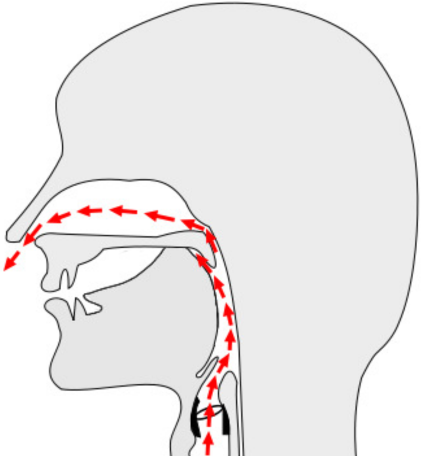
Example: Features in Linguistics

sing

ring

bling

ng = [+Nasal,+Voice,+Velar]



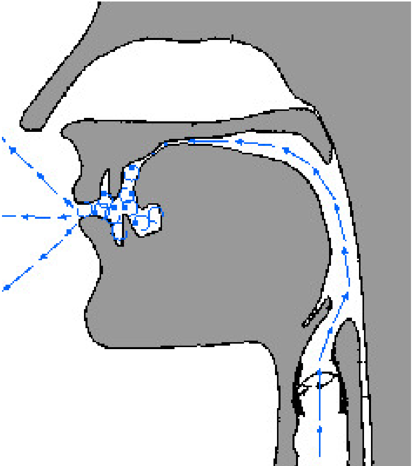
Example: Features in Linguistics

sand

sit

cats

s= [-Nasal,-**Voice**,- **Velar**]

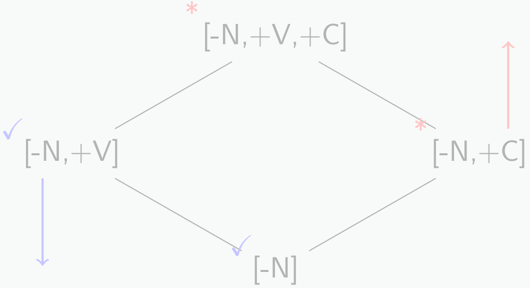


Structuring the Hypothesis Space: Feature Matrix Ideals

Feature Inventory

- $\pm$N = Nasal
- $\pm$V = Voiced
- $\pm$C = Consonant

Example

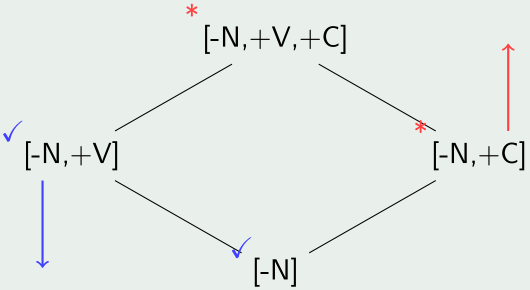


Structuring the Hypothesis Space: Feature Matrix Ideals

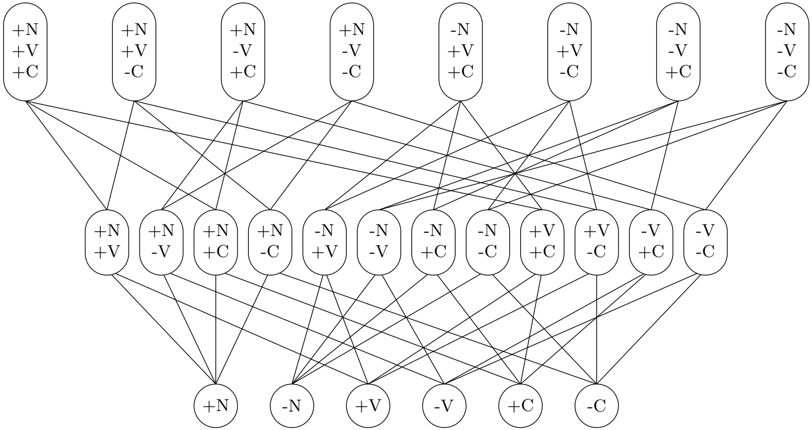
Feature Inventory

- $\pm$N = Nasal
- $\pm$V = Voiced
- $\pm$C = Consonant

Example

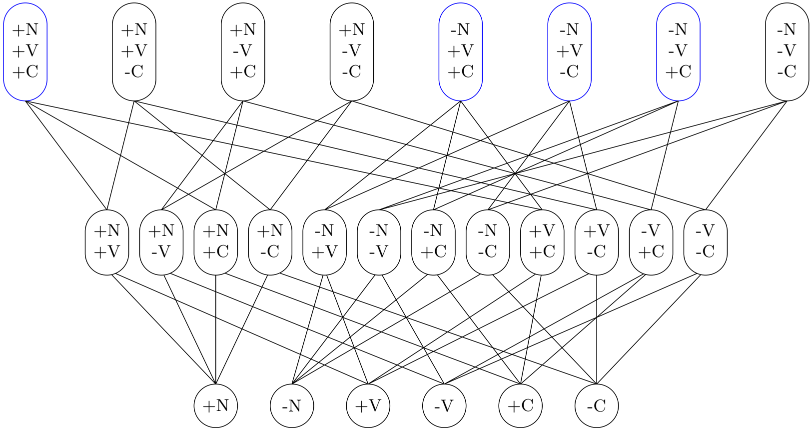


Example



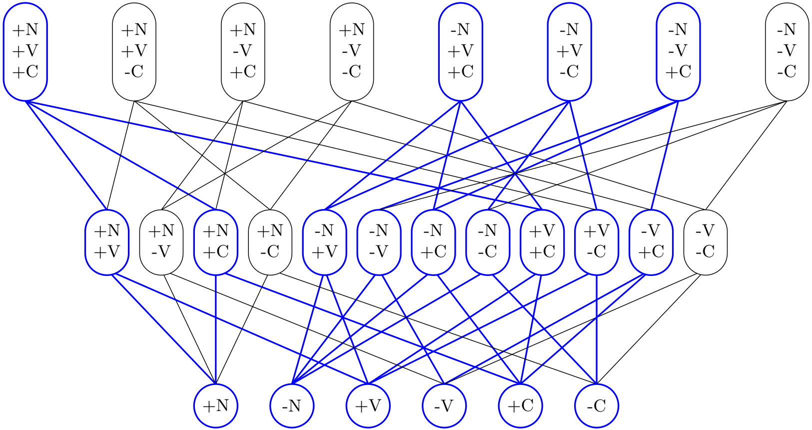
+N +V +C
+N +V -C
+N -V +C
+N -V -C
-N +V +C
-N +V -C
-N -V +C
-N -V -C
+N +V
+N -V
+N +C
+N -C
-N +V
-N -V
-N +C
-N -C
+V +C
+V -C
-V +C
-V -C
+N
-N
+V
-V
+C
-C

Example



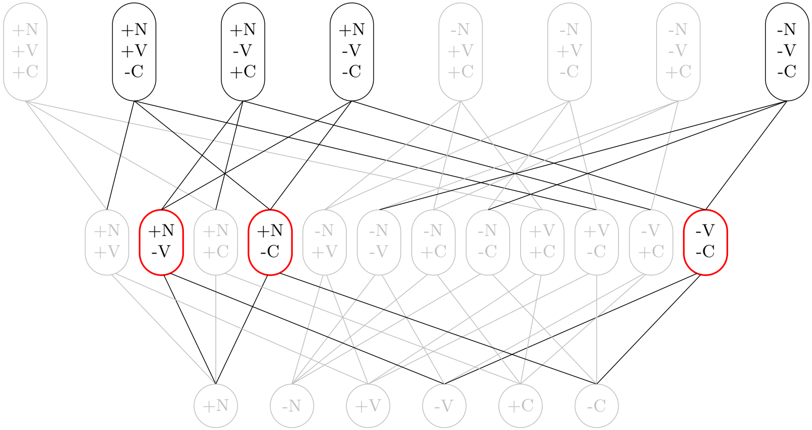
+N +V +C
+N +V -C
+N -V +C
+N -V -C
-N +V +C
-N +V -C
-N -V +C
-N -V -C
+N +V
+N -V
+N +C
+N -C
-N +V
-N -V
-N +C
-N -C
+V +C
+V -C
-V +C
-V -C
+N
-N
+V
-V
+C
-C

Example



+N +V +C
+N +V -C
+N -V +C
+N -V -C
-N +V +C
-N +V -C
-N -V +C
-N -V -C
+N +V
+N -V
+N +C
+N -C
-N +V
-N -V
-N +C
-N -C
+V +C
+V -C
-V +C
-V -C
+N
-N
+V
-V
+C
-C

Example



+N +V +C
+N +V -C
+N -V +C
+N -V -C
-N +V +C
-N +V -C
-N -V +C
-N -V -C
+N +V
+N -V
+N +C
+N -C
-N +V
-N -V
-N +C
-N -C
+V +C
+V -C
-V +C
-V -C
+N
-N
+V
-V
+C
-C

Two Ways to Explore the Space

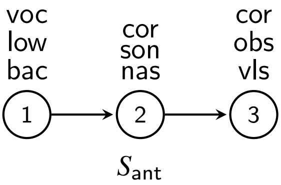
Top-Down Induction

- Start at the most specific points (highest) in the space
- Remove all the substructures that are present in the data.
- Collect the most general substructures remaining.

Bottom-Up Induction

- Beginning at the lowest element in the spave,
- Check whether this structure is present in the input data.
- If so, move up the space, either to a point with an adjacent underspecified segment, or a feature extension of a current segment, and repeat.

Semilattice Explosion



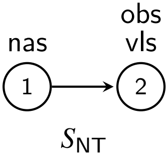
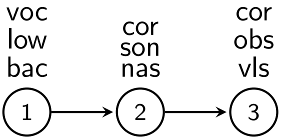


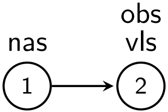
Table 2
Number of possible constraints for various values of $|C|$ and n

		$\lvert C\rvert$			
		30	**100**	**200**	**400**
	1	30	100	200	400
	2	900	10,000	40,000	160,000
n	**3**	27,000	1,000,000	8 million	64 million
	4	810,000	100 million	1.6 billion	26 billion
	5	24 million	10 billion	320 billion	10 trillion

Semilattice Explosion



S_{ant}



S_{NT}

Table 2
Number of possible constraints for various values of $|C|$ and n

		$\vert C\vert$			
		30	**100**	**200**	**400**
	1	30	100	200	400
	2	900	10,000	40,000	160,000
n	**3**	27,000	1,000,000	8 million	64 million
	4	810,000	100 million	1.6 billion	26 billion
	5	24 million	10 billion	320 billion	10 trillion

Plan of the project

What has been done

Provably correct bottom-up learning algorithm

Goals of the Project

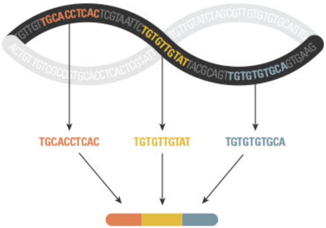
- Model Efficiency
- Model Implementation
- Model Testing - large linguistic datasets
- Model Comparison: UCLA Maximum Entropy Learner

Broader Impacts

- Learner that takes advantage of data sparsity
- applicable on any sequential data (language, genetics, robotic planning, etc.)
- implemented, open-source code

INSTRUCTIONS IN THE CODE

Healthy



DNA
Along with genes (shown here in **orange**, **yellow**, and **blue**), which produce the components for proteins, the genome contains non-coding instructions (**gray**) that direct how these components are assembled.

ASSEMBLY
The cell transcribes specific parts of the code according to the instructions.

PROTEIN
The parts are then assembled into a healthy protein.

Diseased

DNA
A mutation (**red**) in the non-coding instructions causes one gene segment to be ignored.

ASSEMBLY
This variation makes the cell skip over a protein-coding segment of the genome.

PROTEIN
The error in the instruction set leads to an altered protein, which may raise the risk for disease.