INTERACT

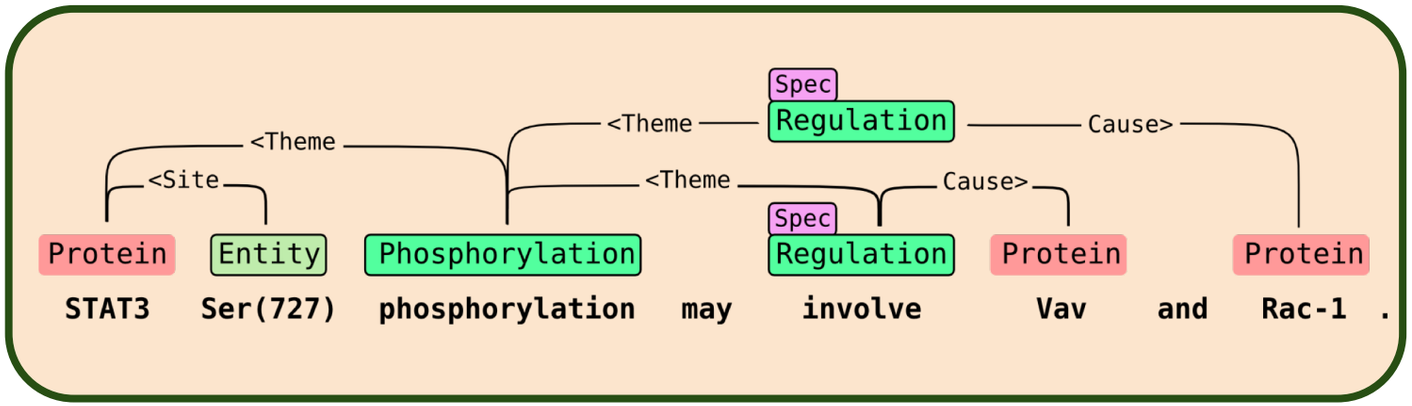
Interactive Machine Learning for Compositional Models of Natural Language

Ariadna Quattoni

Natural Language Understanding (NLU)

plain text → **NLU** → linguistic structure

Both input and outputs are structured
NLU tasks are **structured prediction** tasks



entity and relation extraction

- State-of-the-art supervised learning in NLU:
  - Supervised machine learning for structured prediction, great progress in three decades
  - **Good performance**
  - But it requires **unrealistic amounts of annotated data.**
  - Good advances on **active learning** approaches
  - But main focus on classical text classification and **not specialized for structured prediction**

NLU for Specific Tasks

The 7+ and 7 are water resistant. The 6S is generally not considered much improved from the 6. None of these phones have wireless charging. Conclusion: IMO **7+** is worth it if the wireless headphones aren't too much of an inconvience ….

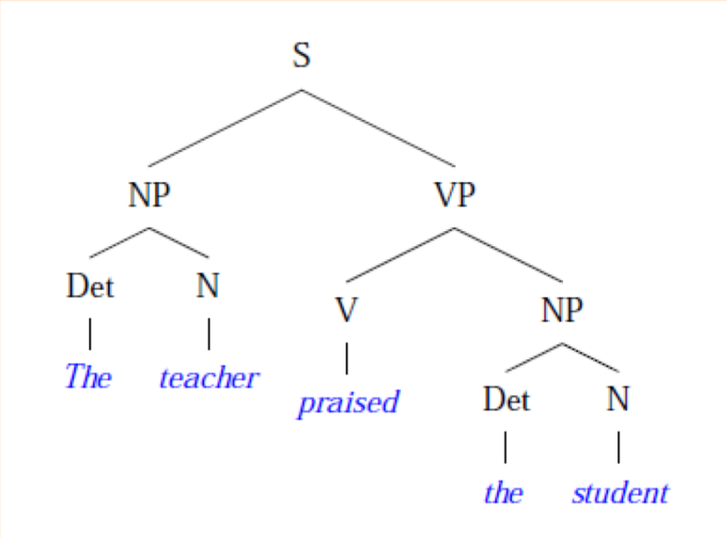
- Domain: web forum on electronic devices
- We want to collect:
  - What devices are mentioned
  - Features of these devices
  - Opinions, in what aspects?
  - Problems, in what parts?
  - Troubleshooting, fix procedures, ...

NLU on Target Textual Domains

- Many different textual domains and unique information needs
- Need tools for task-specific semantics
  - Need to annotate from scratch?
- Can we minimize human annotation effort?
  - Minimize annotation redundancy
- Can we leverage human expertise?
  - More powerful feedback

<u>Challenge 1:</u> Unique information needs of each NLU application

Natural Language is Structured and Complex



Complexity of Natural Language

- Meaning in Natural Language is Compositional
  - Word meaning: determined by morphology units
  - Sentence meaning: determined by composing the meaning of its parts.
- Multiple wordings for the same underlying meaning.

<u>Challenge 2:</u> Compositionality and Expressivity of Natural Language

INTERACT Grand Goal

New interactive machine learning methods for natural language understanding that:

- Enable efficient training of NLU models by optimizing the **collaboration** between the **learning algorithm** and the **human teacher**
- Active feedback strategy specifically designed to handle the **compositionality** and **expressivity** of natural language

Dimensions of the problem:

- **Machine learning aspects**, minimize annotation redundancy
- **Human aspects**, what type of feedback should humans provide?
- **Application aspects**, drastically reduce supervision in NLU tasks

Novelty and impact:

- New methods for **interactive learning for compositional models** that go beyond current active learning approaches for structured prediction
- **Increase the reach of NLU technologies**
- Bring the annotator to the center, leverage the knowledge of domain experts

Objective 1: Minimize Annotation Redundancy

Goal: Minimize Annotation Redundancy

Approach:
Empower the learning algorithm to better leverage the human teacher:

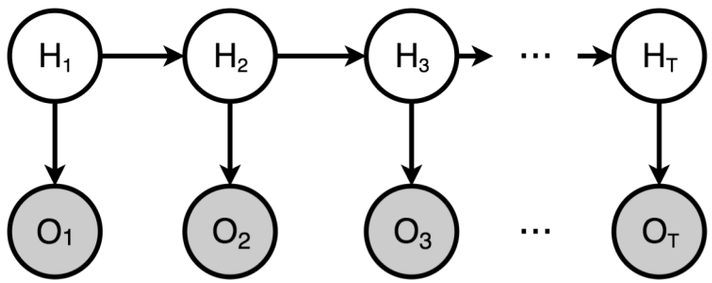
- Request a variety of feedback
- Efficient exploration of compositional domains

INTERACT solution:

- Active feedback + representation learning:
  - Analogous to inducing clusters for active learning of classifiers
  - Induce latent structure of the underlying parts of a composite structure

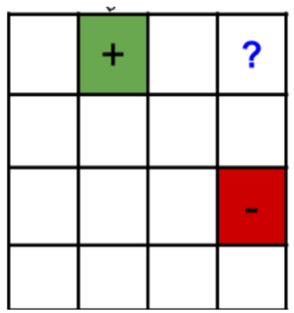
Technical Novelty

Interactive learning of compositional latent state models



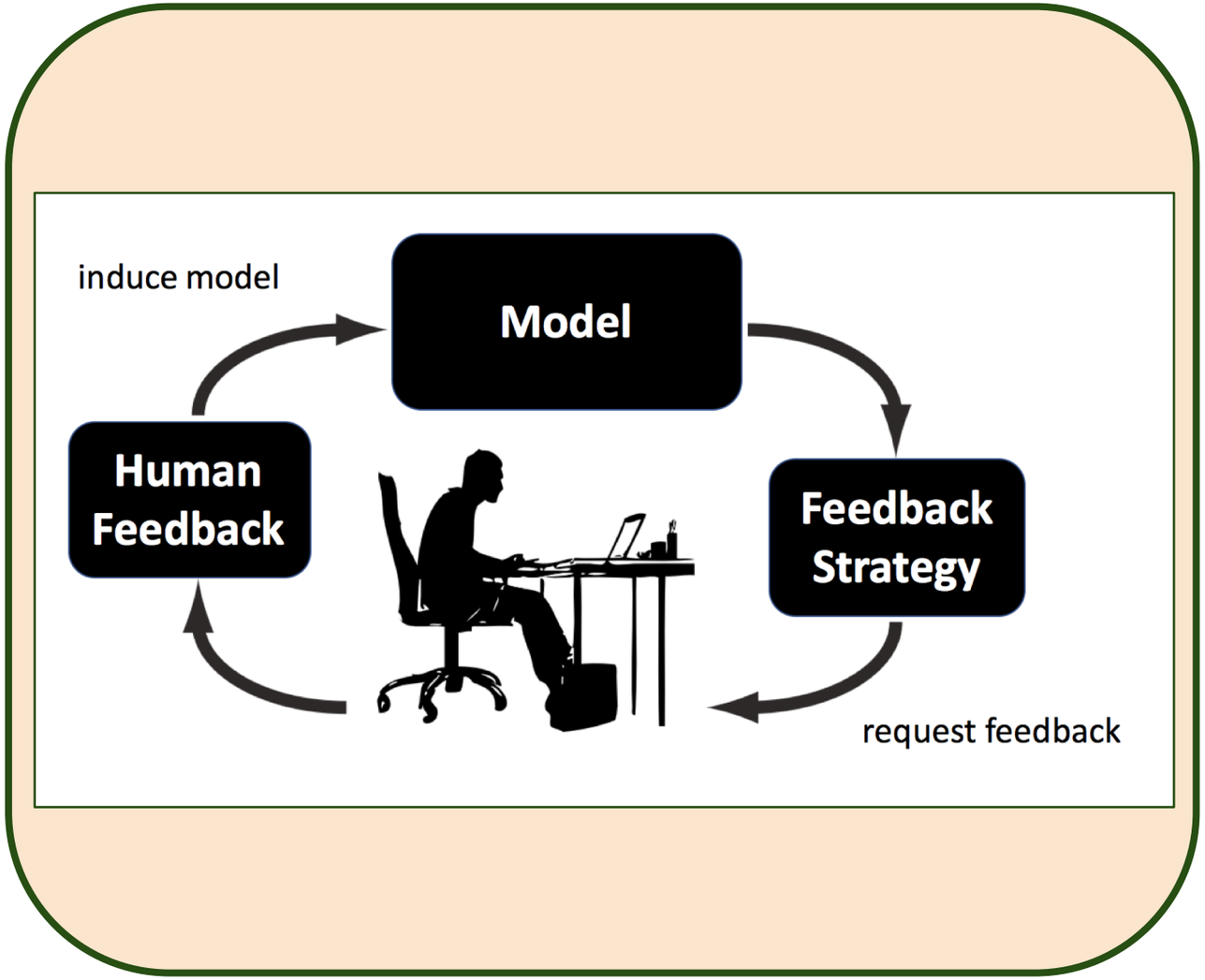
reduction

Interactive low rank matrix completion



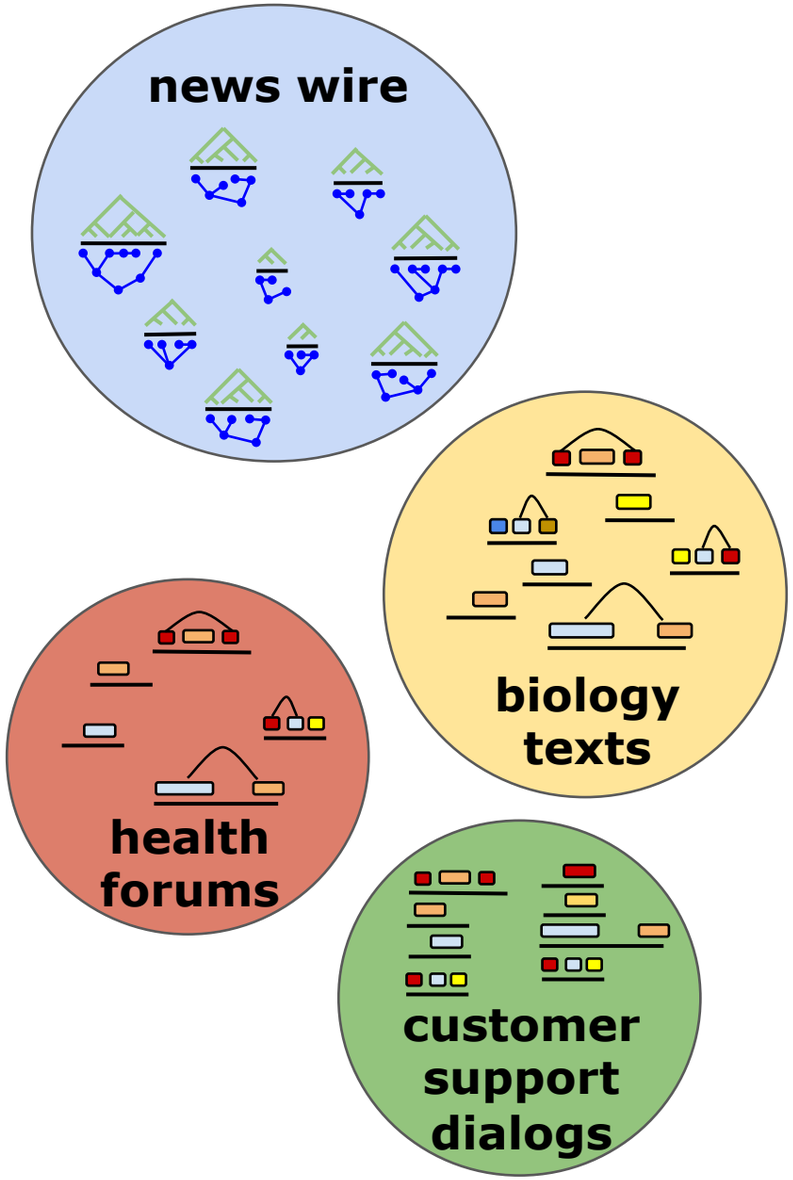
Exploiting ideas from spectral learning of non-deterministic weighted automata

Objective 2: Oracles for Human Teachers



- Inspired by query learning algorithms for NWA
- Goal: Develop strategies to **approximate a function oracle** with feedback provided by a human teacher.
- Feedback strategy should take into account what is optimal for the algorithm and what is **optimal for the human teacher**
- Human-Computer Interaction Challenges

Objective 3: Natural Language Understanding with Interactive Learning



The main objective is to develop tools for fast training and deployment of NLU models for any textual domain and information need.

Four specific NLU tasks:

- Entity extraction
- Language modeling
- Syntactic parsing
- Task-specific semantic parsing

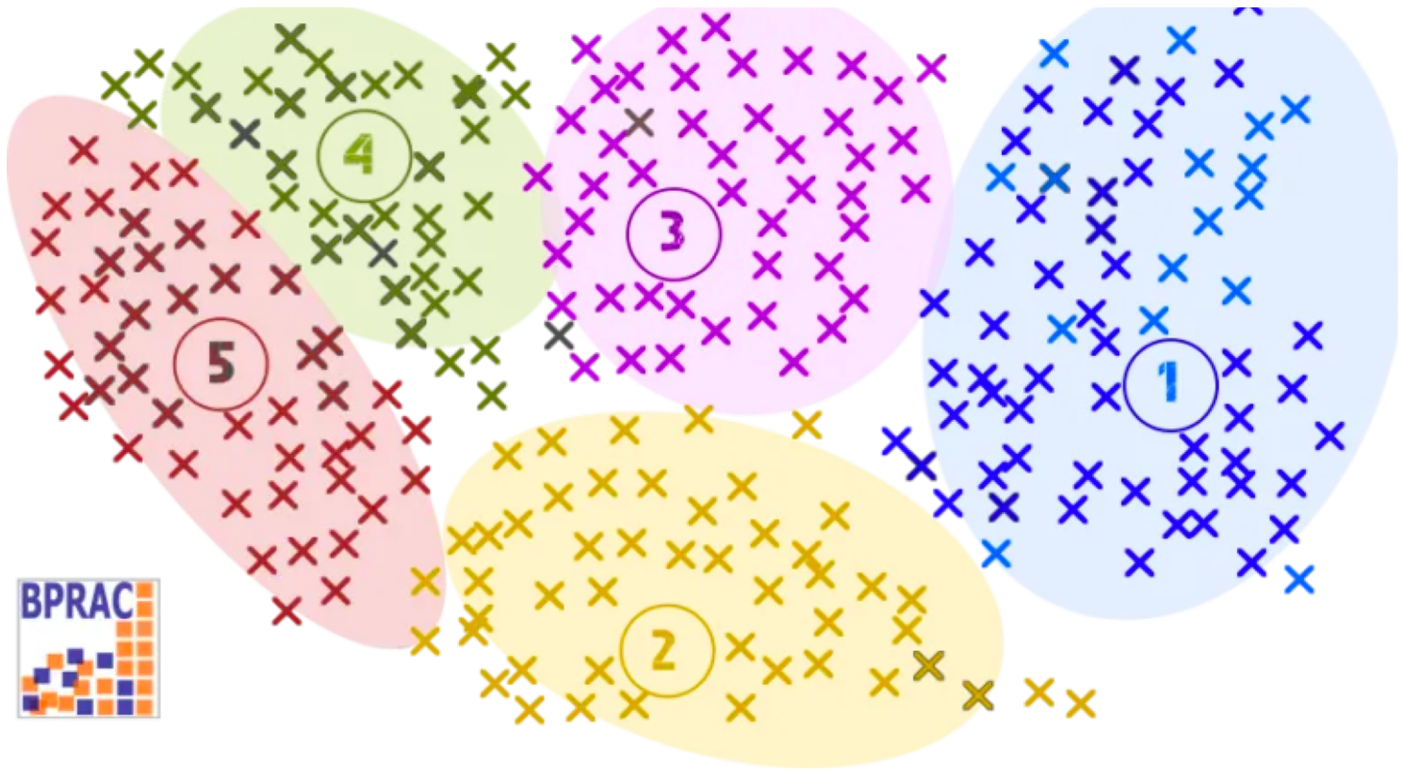
We want to:

- **Vastly reduce the amount of supervision** required to train viable systems.

Learning sequence models with an annotation budget constraint

- If I can only annotate n tokens which ones would be optimal?
- Different from active learning → interested in the performance of the initial model (cold-start problem)
- Approach: Minimize annotation redundancy.
- Annotate partial samples (i.e. phrases)

Diversity Sampling



Hankel Matrix

Two equivalent representations of a WA

- Functional: $f : \Sigma^\star \to \mathbb{R}$
- Matricial: $\mathbf{H}_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$, the *Hankel matrix* of f

Definition: p prefix, s suffix $\quad \Rightarrow \quad \mathbf{H}_f(p, s) = f(p \cdot s)$

$$\mathbf{H}_f = \begin{array}{c} \\ \lambda \\ a \\ b \\ c \\ \vdots \\ ca \\ cat \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} \lambda & a & \cdots & t & at & cat & \cdots \end{array} \\ \begin{bmatrix} \ddots & \cdots & & & & \frac{3}{100} & \cdots \\ \vdots & & & & & & \\ & & & & \frac{3}{100} & & \\ & & & & & & \\ & & & \frac{3}{100} & & & \\ \frac{3}{100} & & & & & & \\ \vdots & & & & & & \ddots \end{bmatrix} \end{array}$$

$f(cat) = \frac{3}{100}$

$\mathbf{H}_f(\lambda, cat) = \mathbf{H}_f(c, at) = \mathbf{H}_f(ca, t) = \mathbf{H}_f(cat, \lambda) = \frac{3}{100}$

- In practice for language modeling:
  - Use ngrams in data as prefixes and suffixes
  - set $f(x)$ to be the expected frequency of an ngram x

Hankel Matrix

$$
\begin{array}{c}
\\ \lambda \\ a \\ b \\ aa \\ \vdots
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \lambda & a & b & aa & \cdots \end{array} \\
\begin{bmatrix}
0 & 1 & 0 & 2 & \cdots \\
1 & 2 & 1 & 3 & \\
0 & 1 & 0 & 2 & \\
2 & 3 & 2 & 4 & \\
\vdots & & & & \ddots
\end{bmatrix}
\end{array}
$$

N

N

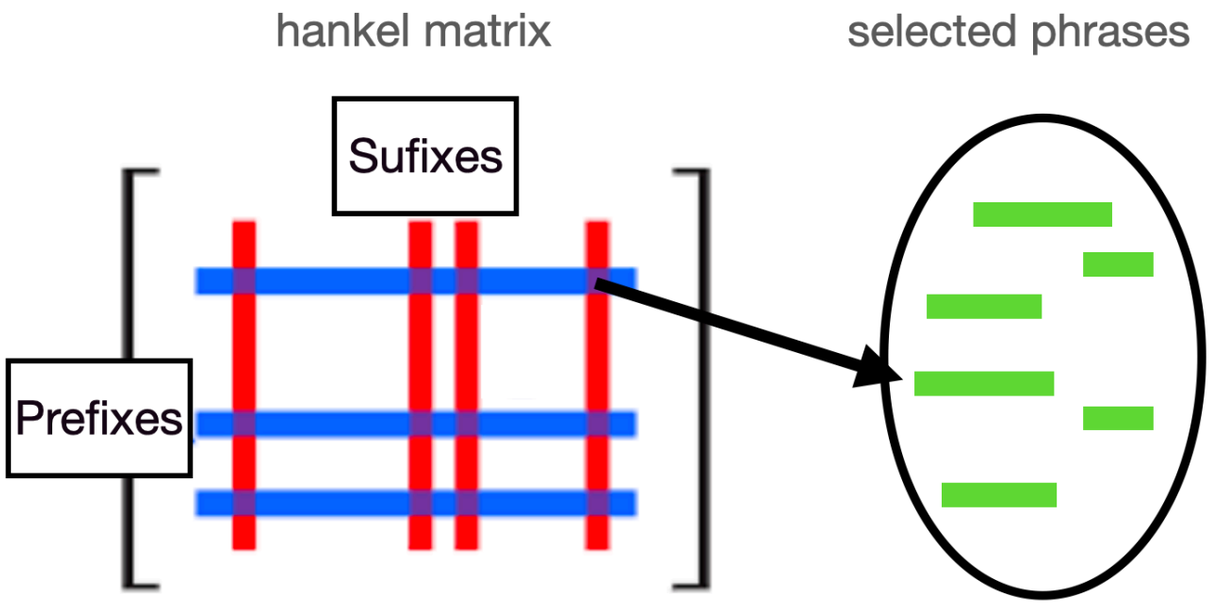
There exists an n by n sub-block that contains all the necessary information.

The longest string in the sunblock can be though as is an upper-bound on the length of the longest dependency.

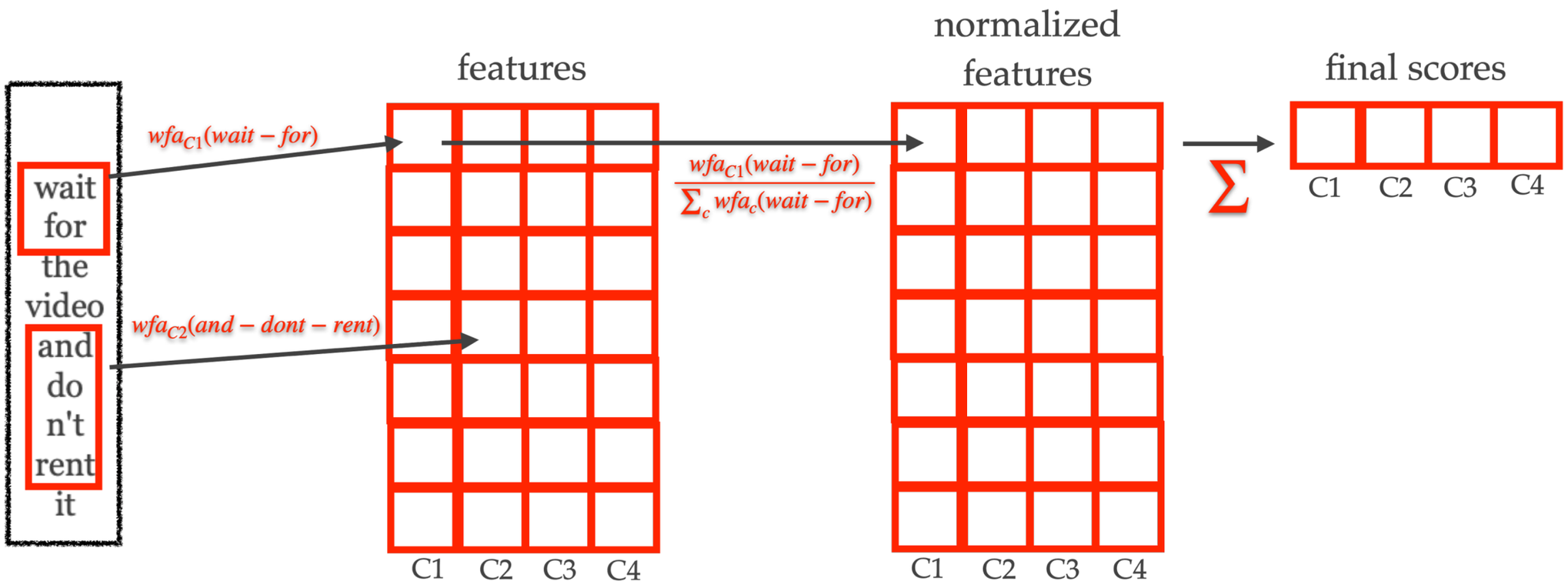
Hankel Matrix Language Representation

- Distributional Hypothesis: phrases of similar meaning appear in similar contexts

- Implicitly Cluster Phrases



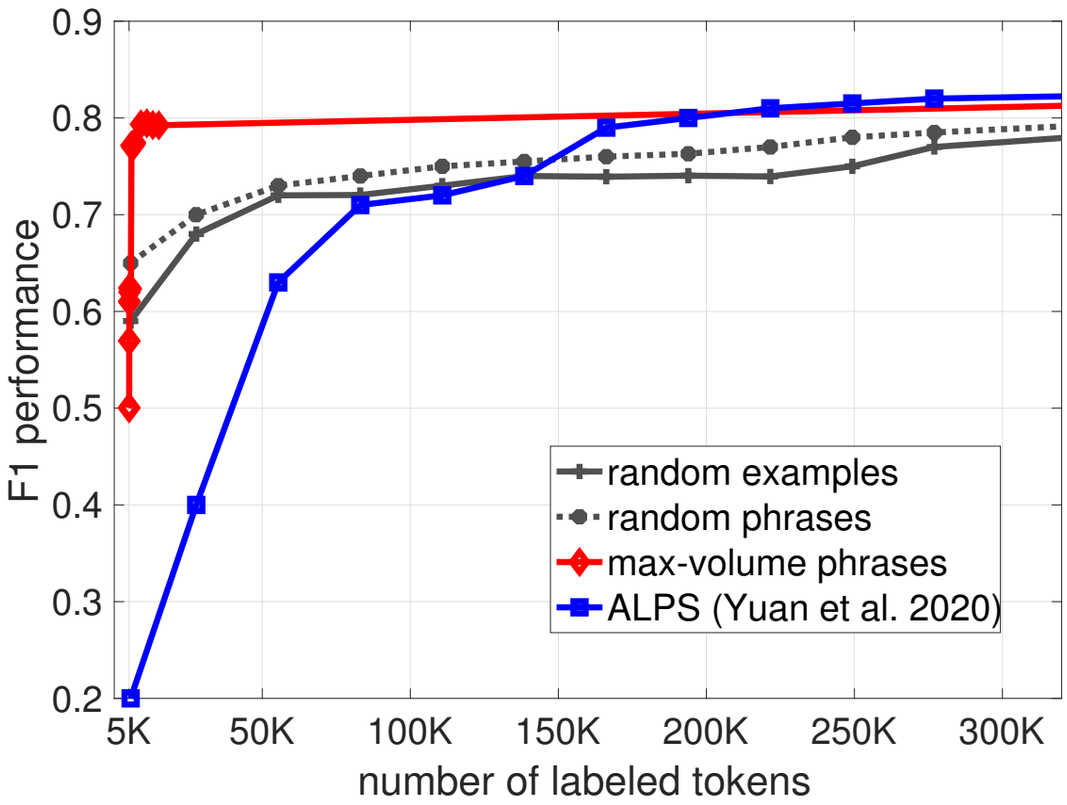
WFA Classifier Ensembles



Max-volumen phrase sampling for training sequence classifiers under budget constraints

- 1- Compute the domain hankel matrix
- 2- Compute a max-volume subblock
- 3- Ask feedback for all non-zero phrases that appear in the subblock
  - ‘Multilabel feedback’ Can this phrase appear in an example of class c?

Some results: Sentiment Classification (IMDB)



basis size	#tokens	F1
11	5,056	50%
30	5,151	57%
50	5,251	61%
70	5,351	62%
200	6,001	62%
400	7,001	77%
1,000	10,001	77%
2,000	15,001	79%
5,000	30,121	79%

Take home:

- Latent state sequence models have low-rank signatures → only a few key statistics matter
- Annotation strategies should exploit this fact → cleverly select the smalest set that contains the key stats.
- If you have an annotation budget pick a simple model (i.e WFA with few states) and select your samples wisely.
- Questions:
  - Could the strategy work for other latent state models?
  - Can we use it as initial model for active learning?

Thanks!!