Phrase Clustering with Posterior Regularization

CLSP Summer Workshop 2010
SMT Team
Desai Chen
joint work with Trevor Cohn

Outline

- clustering problem
- EM with posterior regularization
- results and future experiments

Phrase clustering

Phrases are defined as contiguous spans aligned with each other

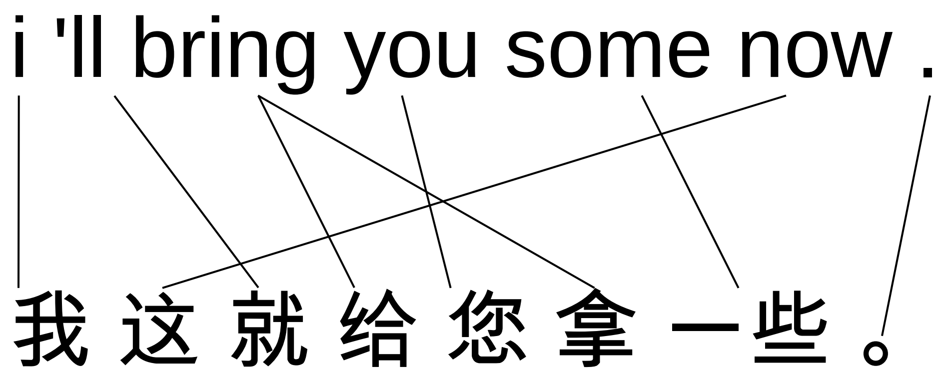
i 'll bring you some now .

我 这 就 给 您 拿 一些 。

Example from btec

Phrase clustering

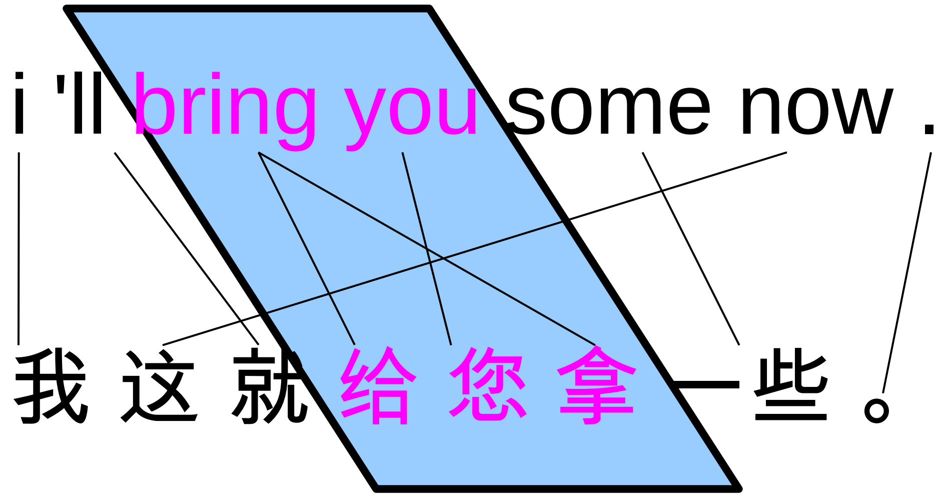
Phrases are defined as contiguous spans aligned with each other



Example from btec

Phrase clustering

Phrases are defined as contiguous spans aligned with each other



Example from btec

Phrase clustering

Phrases are defined as contiguous spans aligned with each other

i 'll bring you some now .

我 这 就 给 您 拿 一些 。

Phrase clustering

Contexts are words before or after the phrase:

target side context

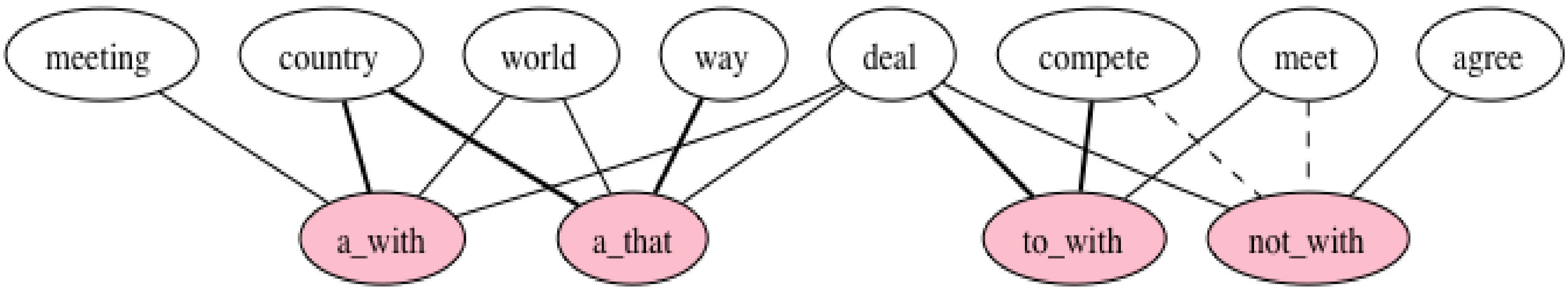
i 'll bring you some now .

我 这 就 给 您 拿 一些 。

source side context

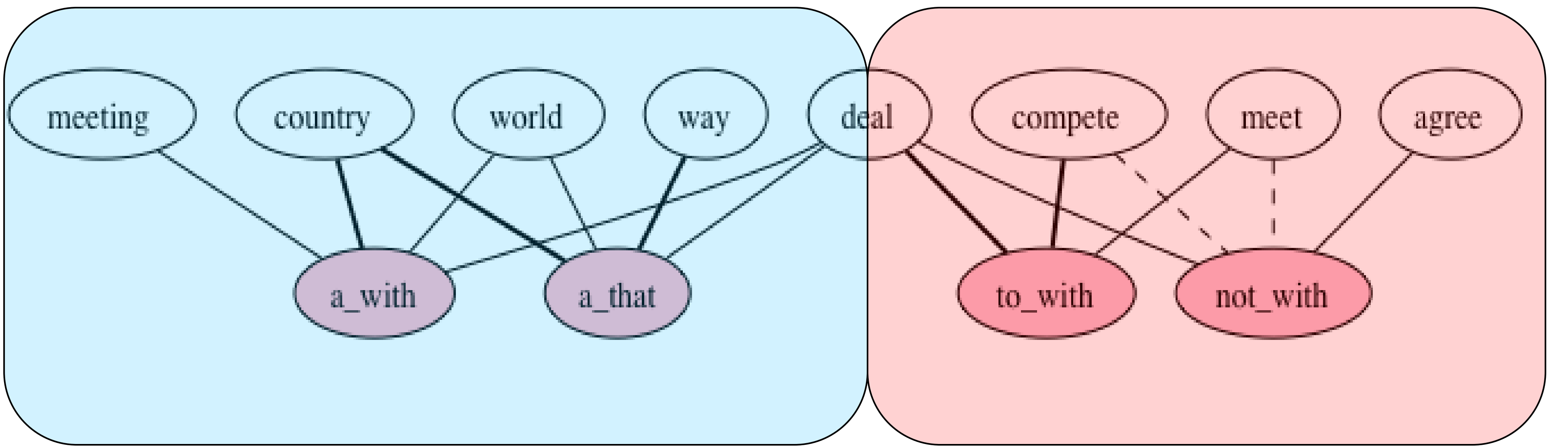
Objective

Put all phrase-context pairs into categories



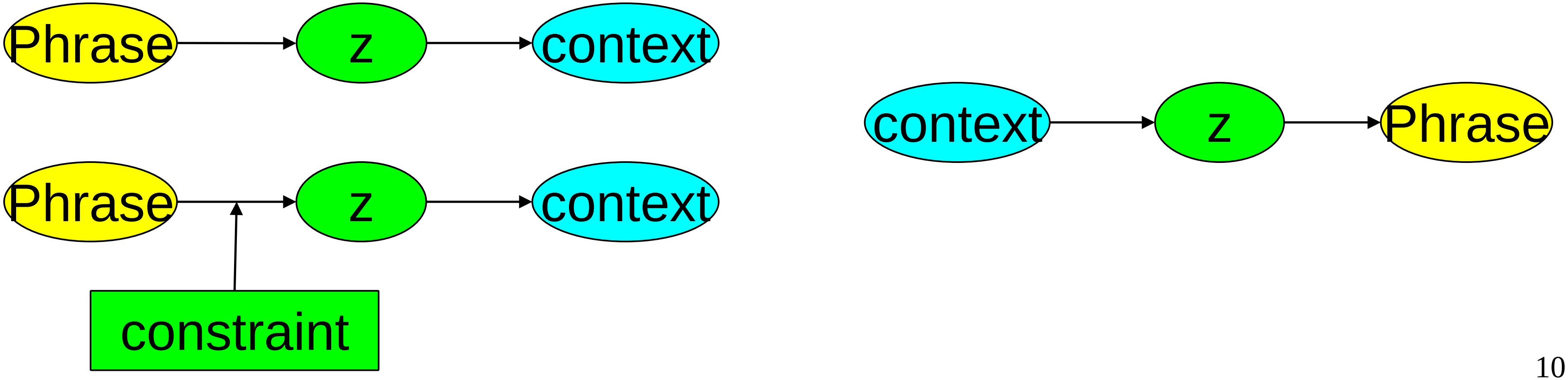
Objective

Put all phrase-context pairs into categories



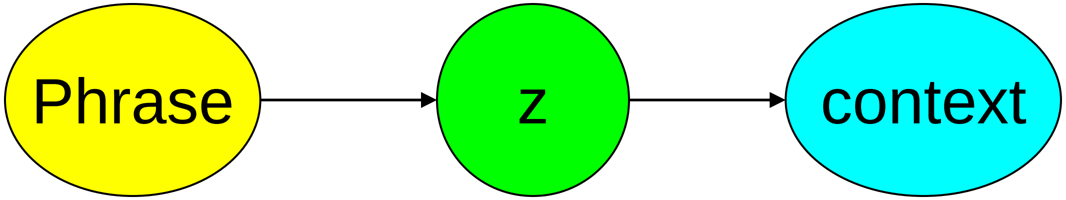
Outline

- Where do phrases come from?
- EM with posterior regularization
- results and future experiment



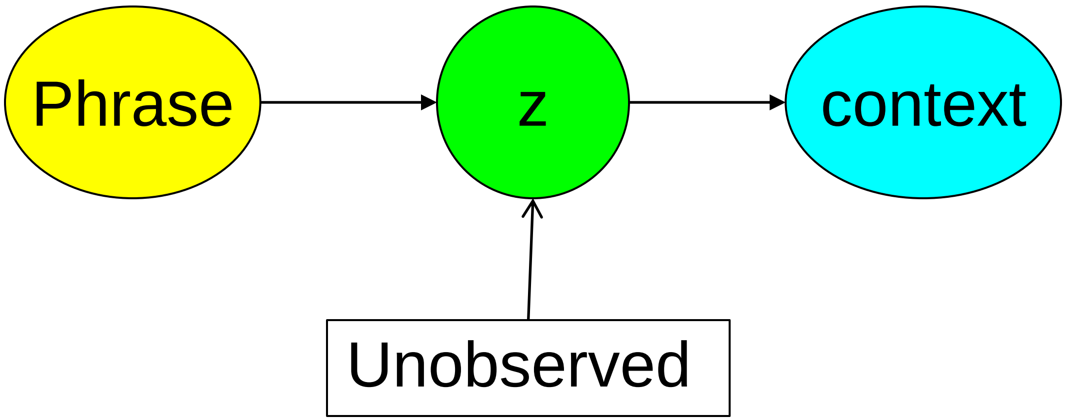
Expectation-Maximization

- naïve Bayes model for phrase labeling



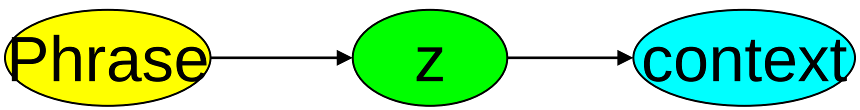
EM clustering

- naïve Bayes model for phrase labeling



EM clustering

- naïve Bayes model for phrase labeling

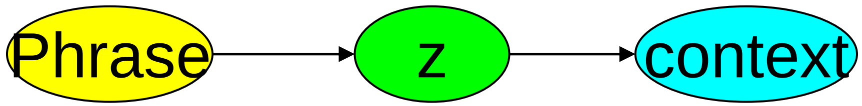


E-step

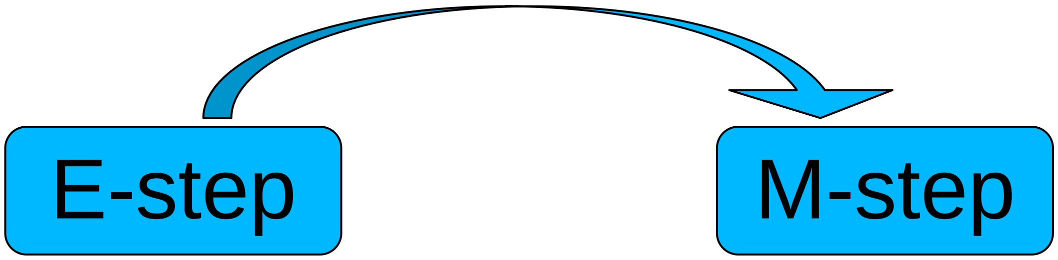
M-step

EM clustering

- naïve Bayes model for phrase labeling



$$q(z|p,c) = P_\theta(z|p,c)$$

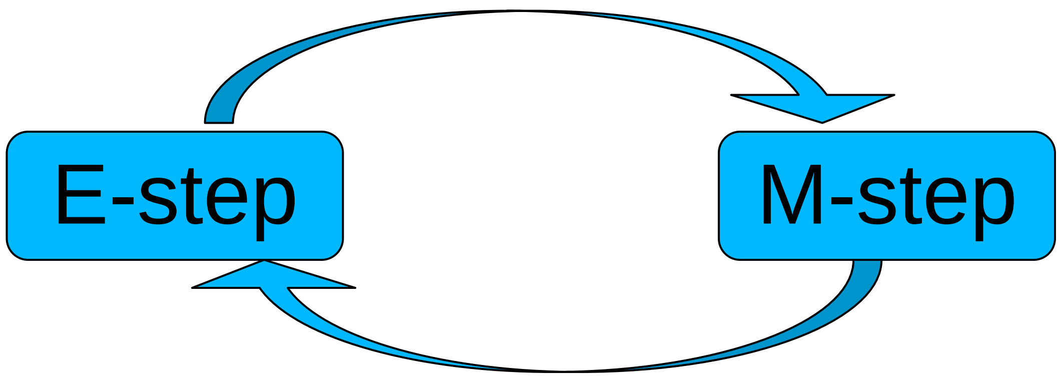


EM clustering

- naïve Bayes model for phrase labeling

Phrase → z → context

$$q(z|p,c) = P_\theta(z|p,c)$$



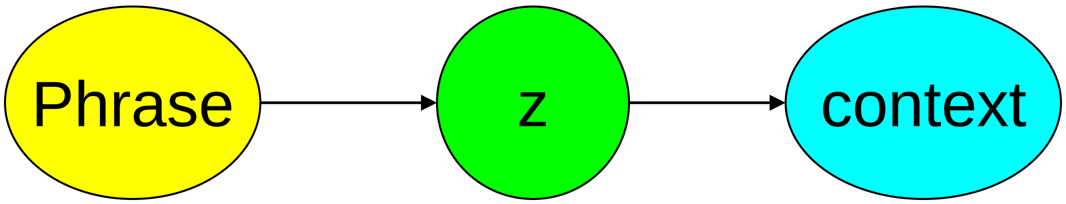
$$\theta = MLE\ q(z|p,c)$$

Problem with EM

- Problem: EM uses as many categories as it wants for each phrase.
- We want to limit the number of categories associated with each phrase.

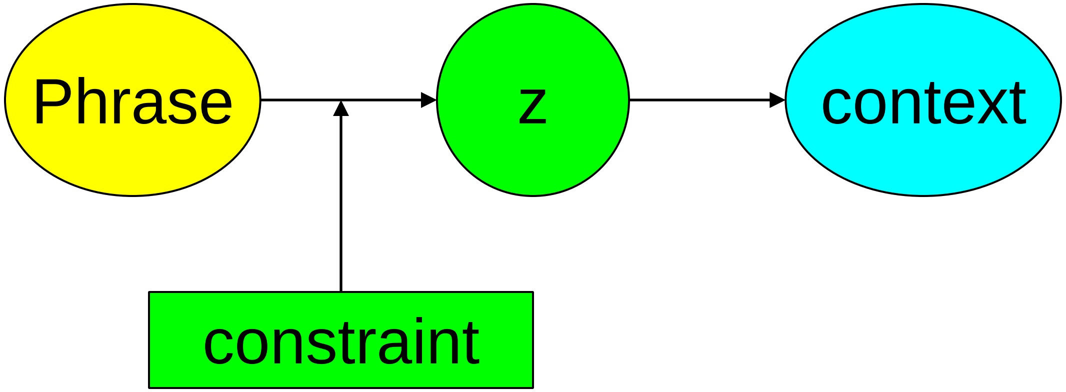
Sparsity constraints

- Sparsity:Each phrase/context should be labeled with fewer kinds of labels.



Sparsity constraints

- Sparsity:Each phrase/context should be labeled with fewer kinds of labels.



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

Contexts:

i understand there are some sightseeing bus tours here , is that right ?

there are only a few seats left in the dress circle .

well , of course there are fine restaurants .

your hotel brochure shows there are some tennis counts at your hotel .

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

Contexts:

i understand there are some sightseeing bus tours here , is that right ?

there are only a few seats left in the dress circle .

well , of course there are fine restaurants .

your hotel brochure shows there are some tennis counts at your hotel .

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants
brochure shows _ some tennis

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

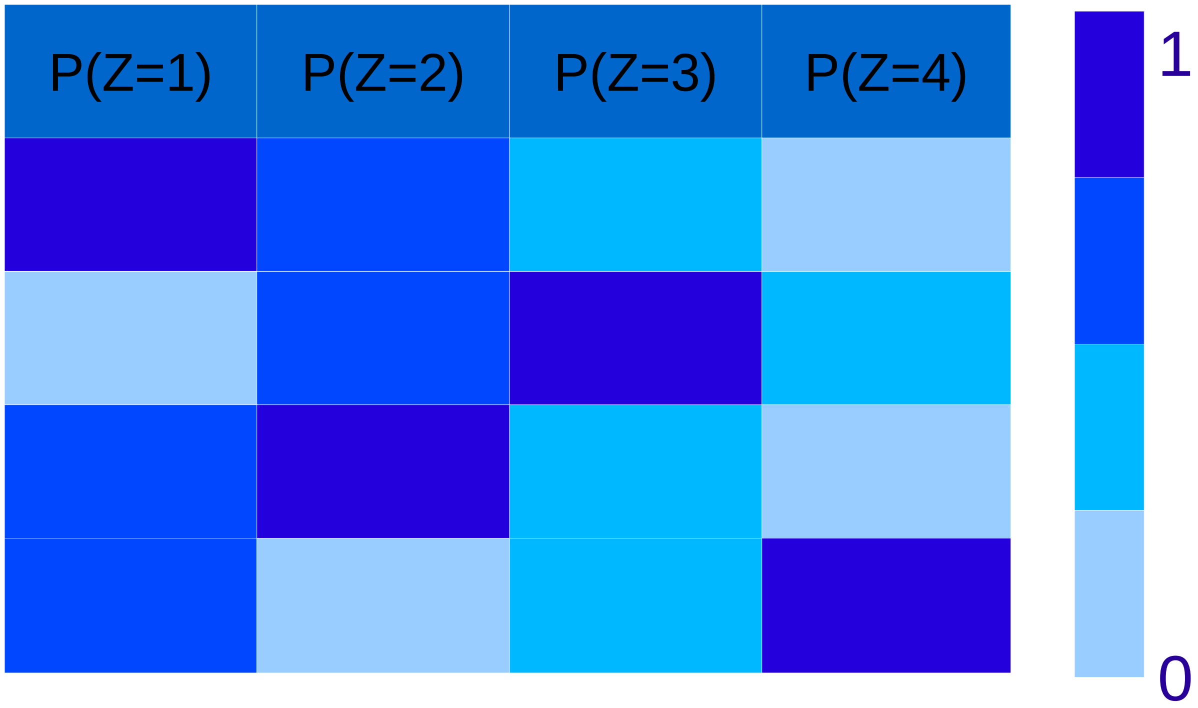
Phrase: there are

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

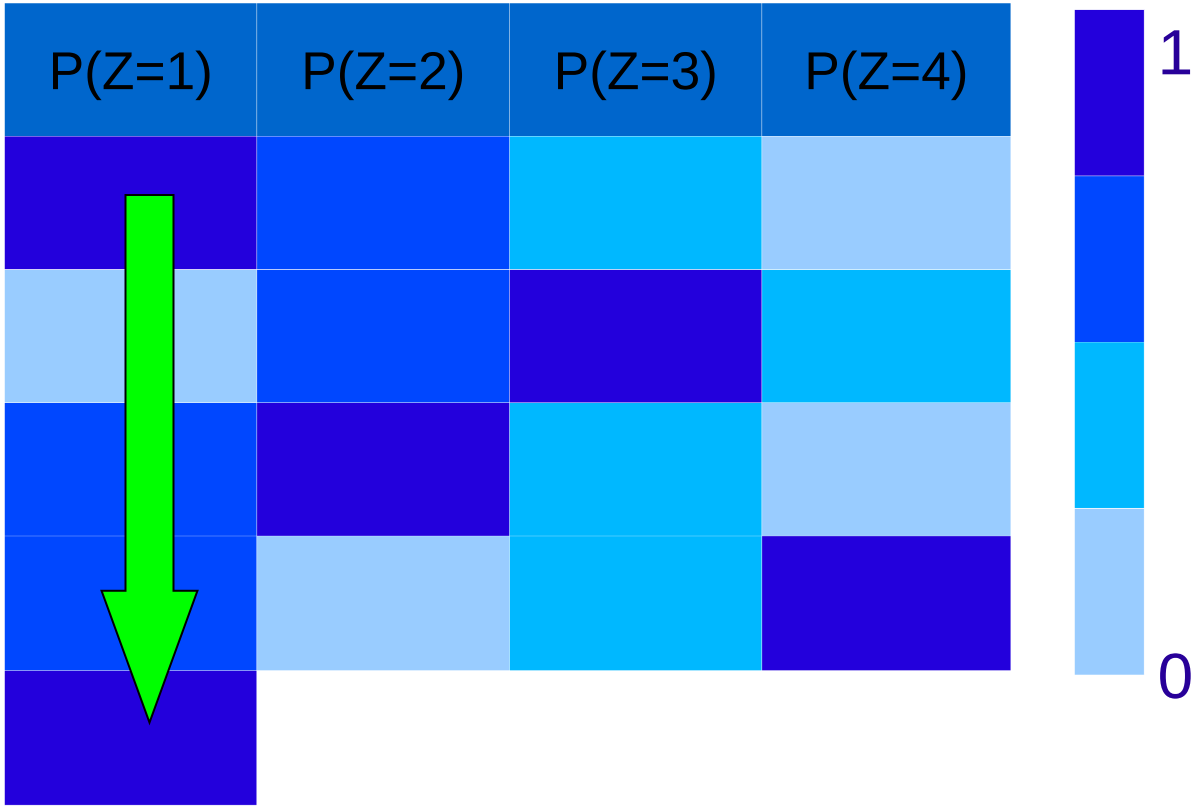
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

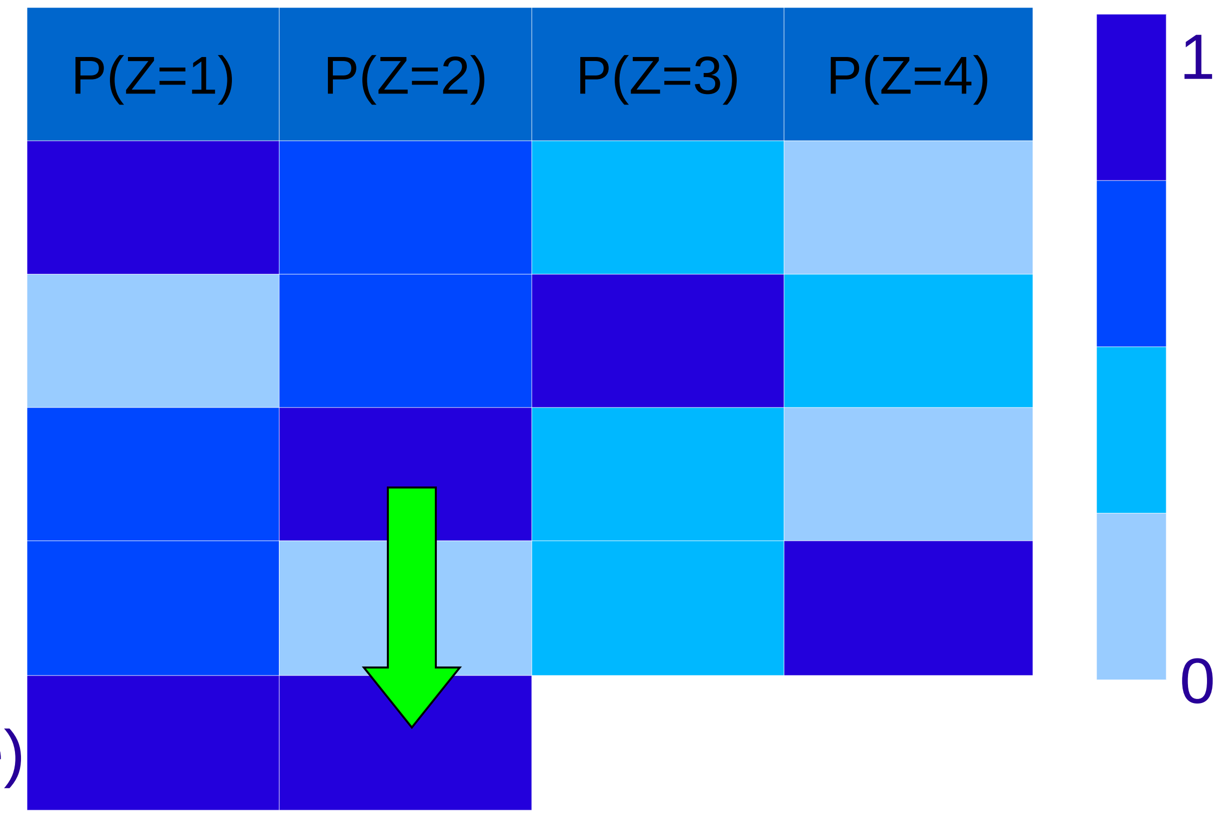
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

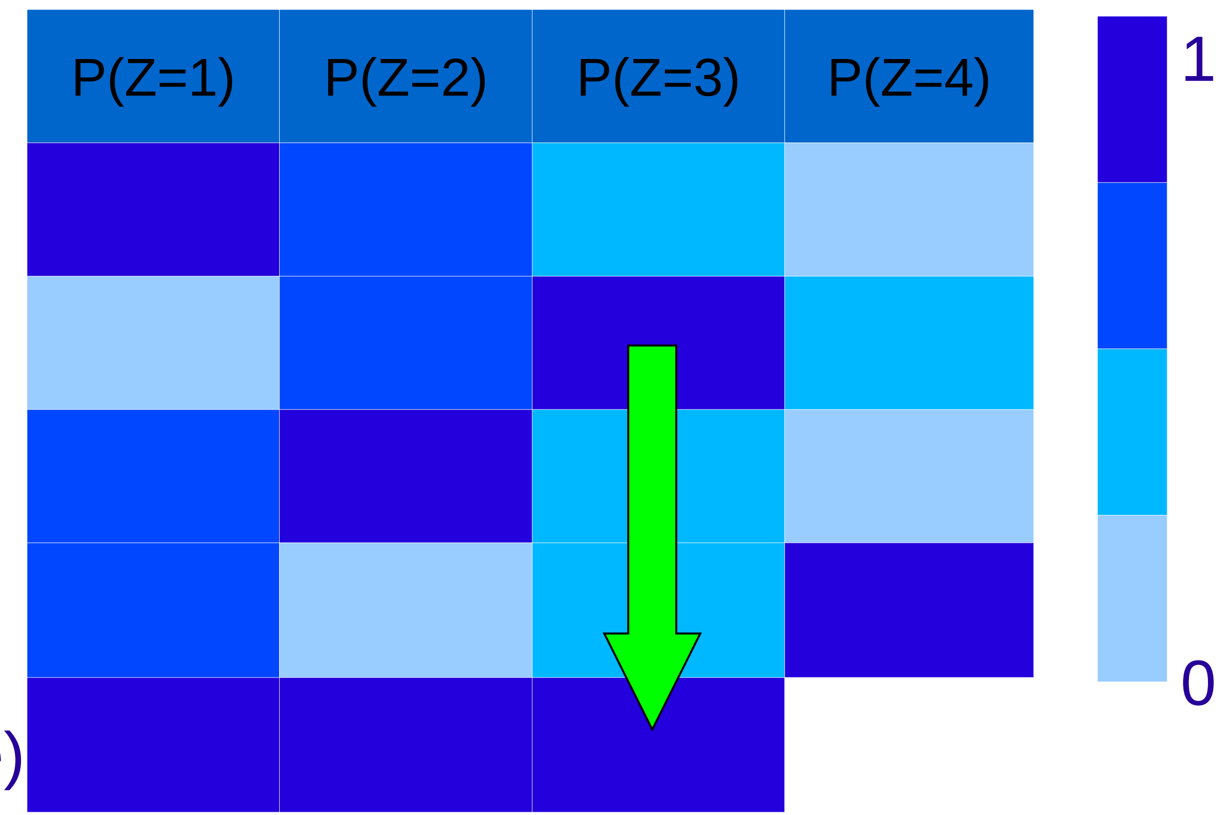
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

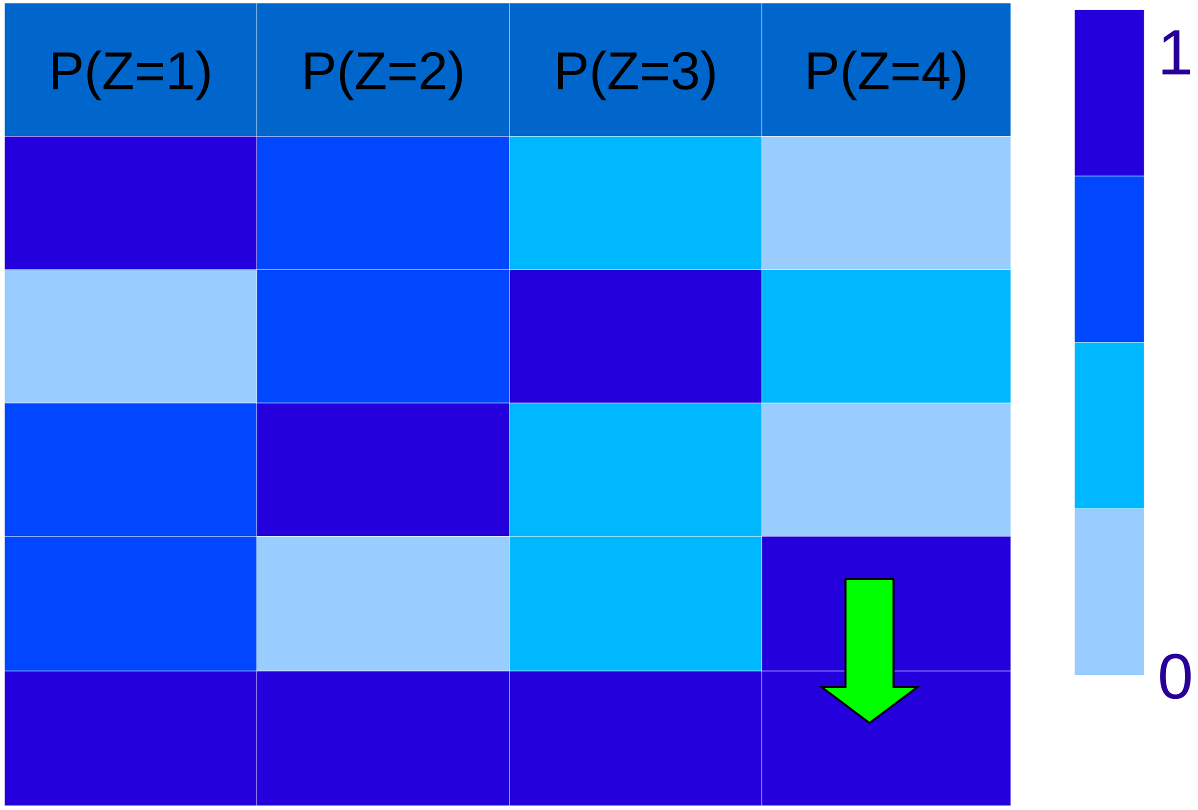
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

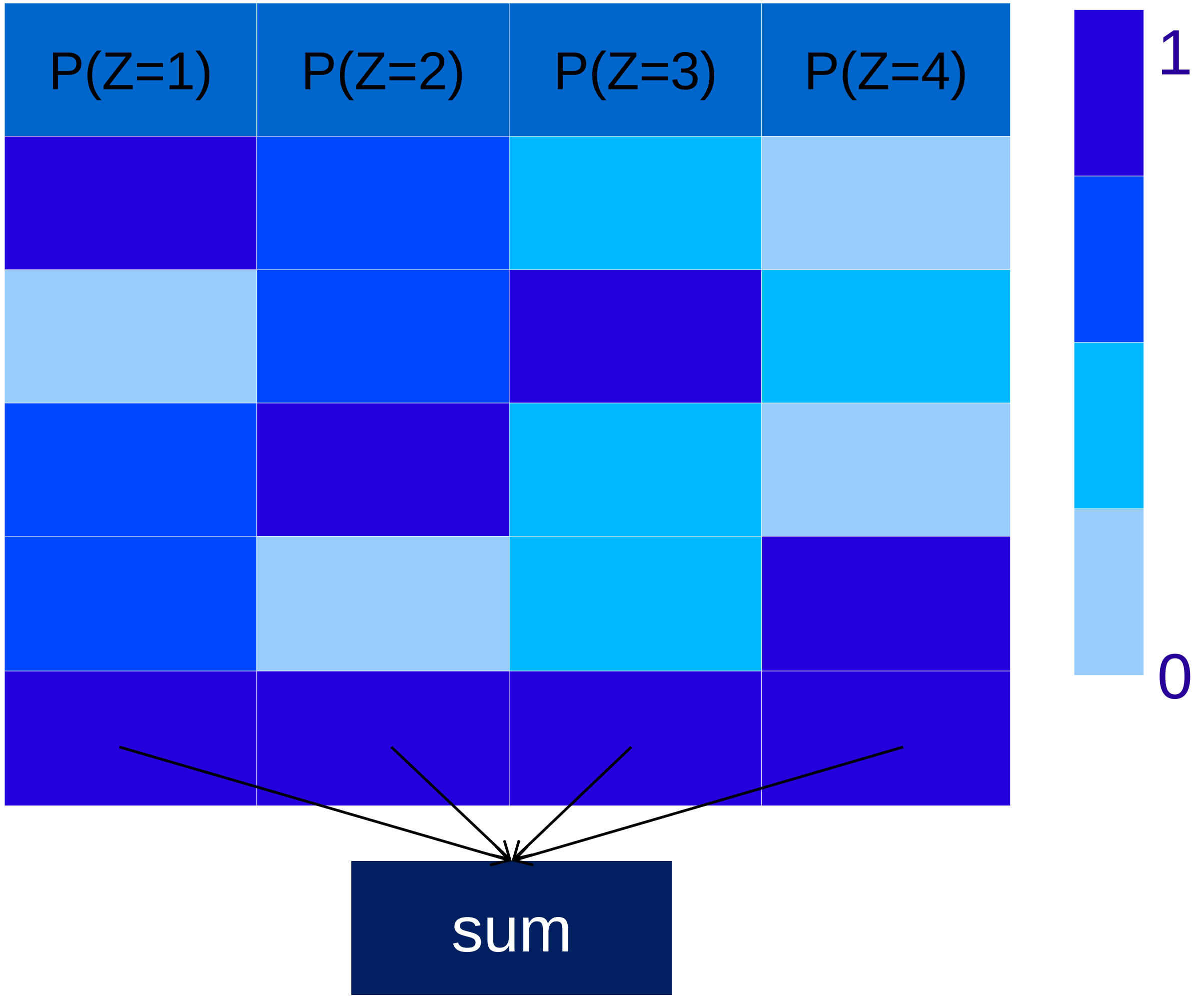
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: there are

Contexts:

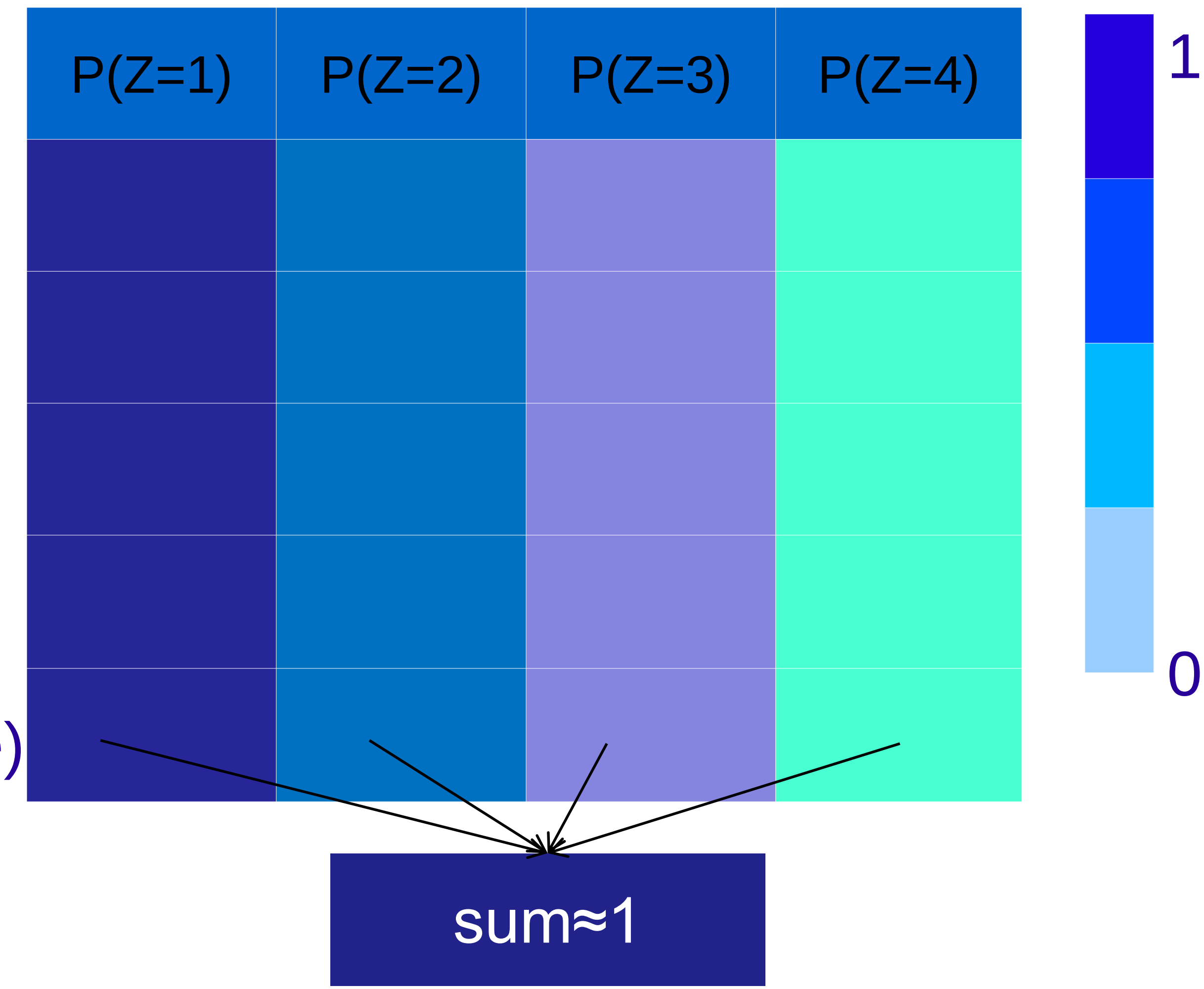
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)

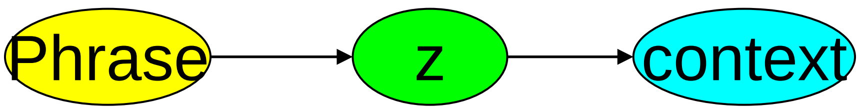


Posterior Regularization

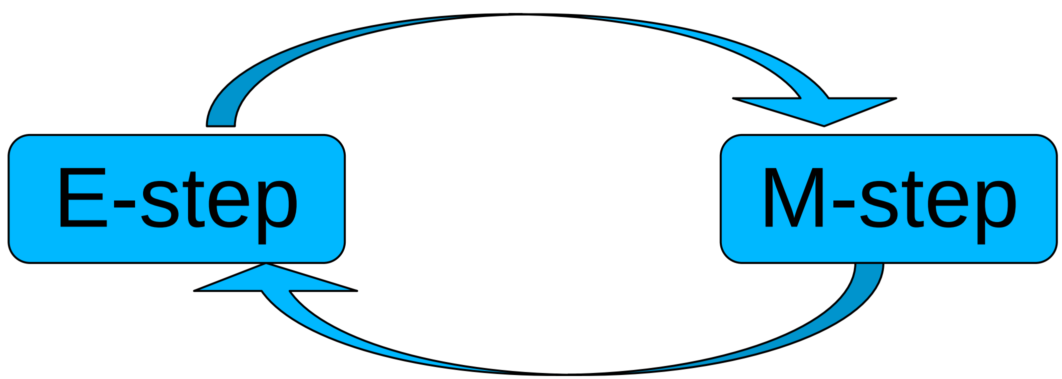
- Follows *Posterior Regularization for Structured Latent Variable Models,* Ganchev et al., 2009
- During E-step, impose constraints on the posterior q to guide the search

Posterior Regularization

- impose constraints on the posterior q



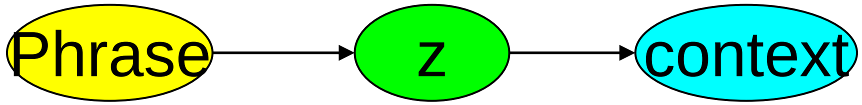
$$q(z|p,c) = \underset{q \in Q}{\arg\min}\, KL(q||P_\theta)$$

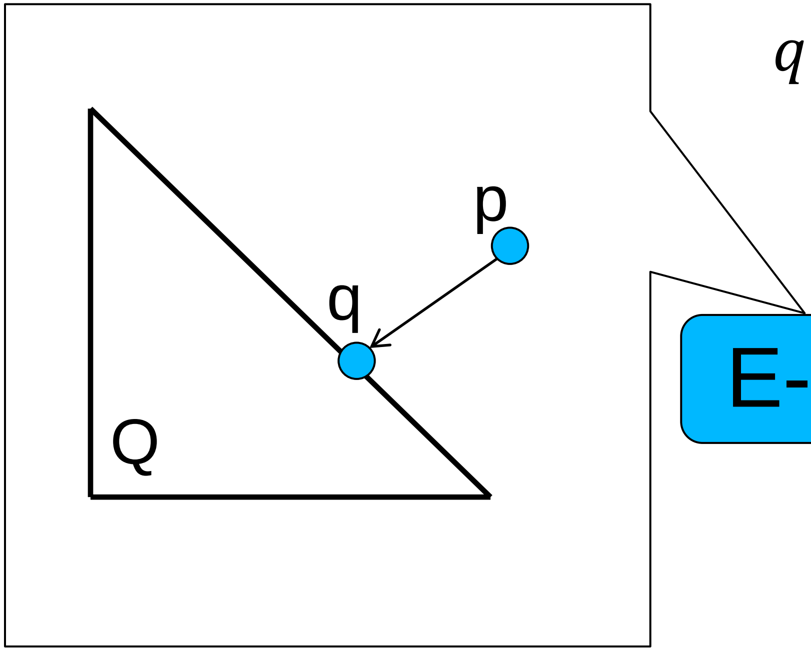


$$\theta = MLE\ q(z|p,c)$$

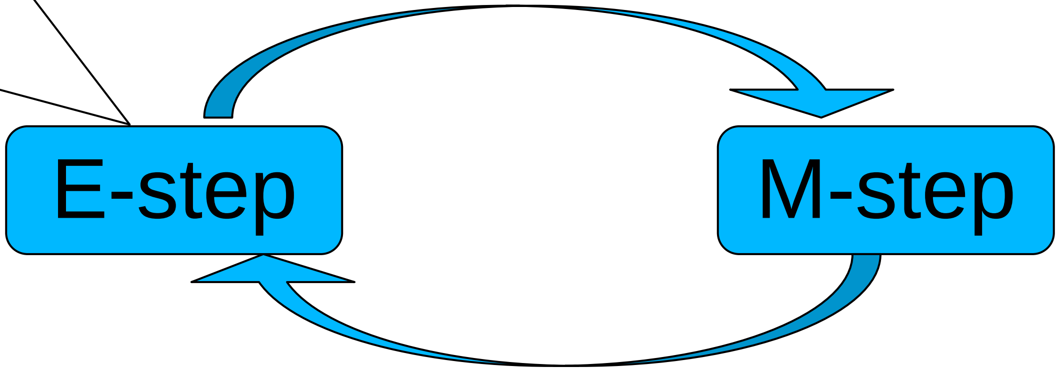
Posterior Regularization

- impose constraints on the posterior q





$$q(z|p,c) = \arg\min_{q \in Q} KL(q||P_\theta)$$



$$\theta = MLE\ q(z|p,c)$$

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

Phrase: like this

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants
brochure shows _ some tennis

Define feature functions:

$$\phi_{i,j}(p,z) = \begin{cases} 1 \; if \; p = i \; and \; z = j \\ 0 \quad\quad\quad\quad otherwise \end{cases}$$

Sparsity constraints

Minimize $\sum_{p,z} max_i P(z|p_i)$

- Soft constraint. Softness controlled by σ.
- During E-step, find q distribution:

$$\min_{q,c_{p,z}} KL(q||P_\theta) + \sigma \sum_{p,z} c_{p,z}$$

$$\text{s.t. } E_q\left[\phi_{p,z}\right] \leq c_{p,z}$$

where “c”s are maximums of expectation for each word tag pair by definition.

Primitive results

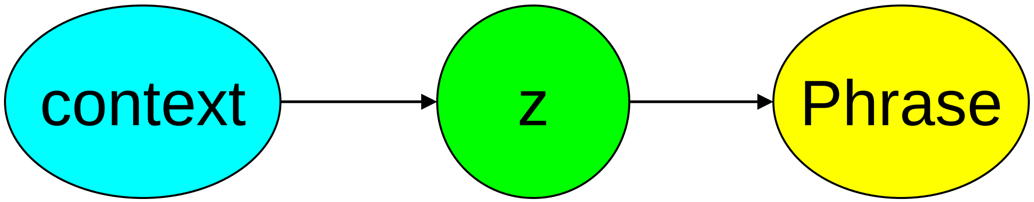
- Constrained model gives clustering that's more sparse
- Clustering for a few phrases with 25 tags on BTEC ZH-EN

Phrase/Word	**Count of the most used tag**		**Number of tags used**	
the	1194	**1571**	11	**4**
there is	**53**	50	5	**4**
'd like	723	**873**	5	**2**

More experiments

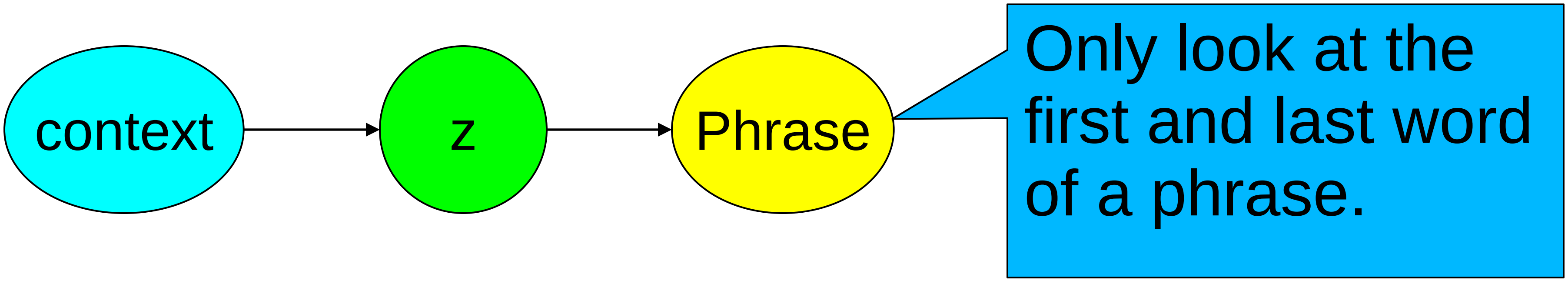
- agreement constraint: different “good” models should agree on posterior distribution
- what model to agree with: another naïve Bayes model in the reverse direction or in the other language.

Agreement model



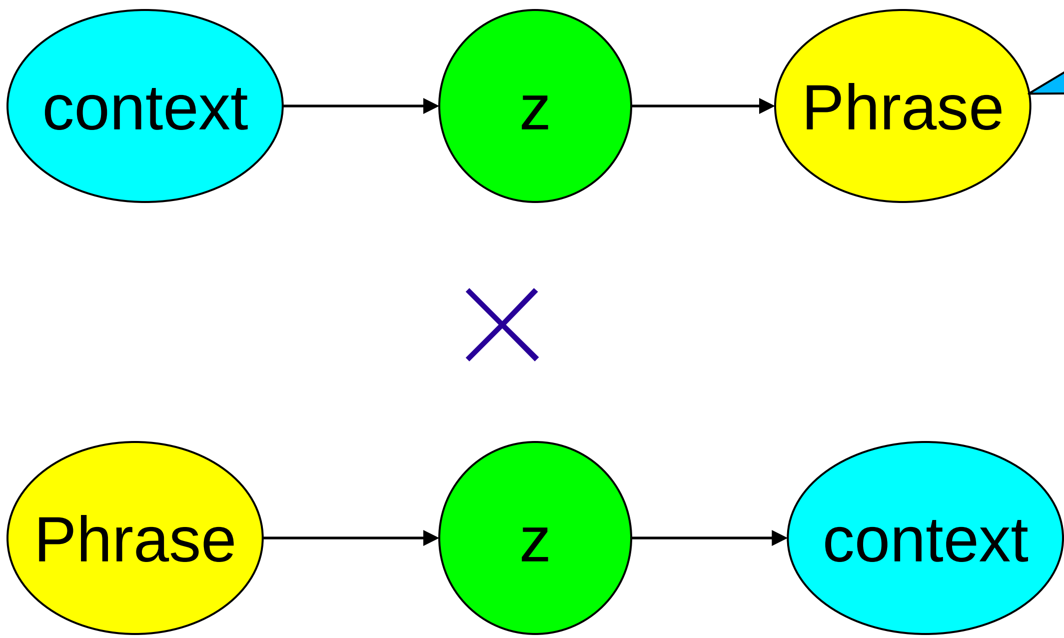
- implementation: multiply posteriors of two models together.

Agreement model



- implementation: multiply posteriors of two models together.

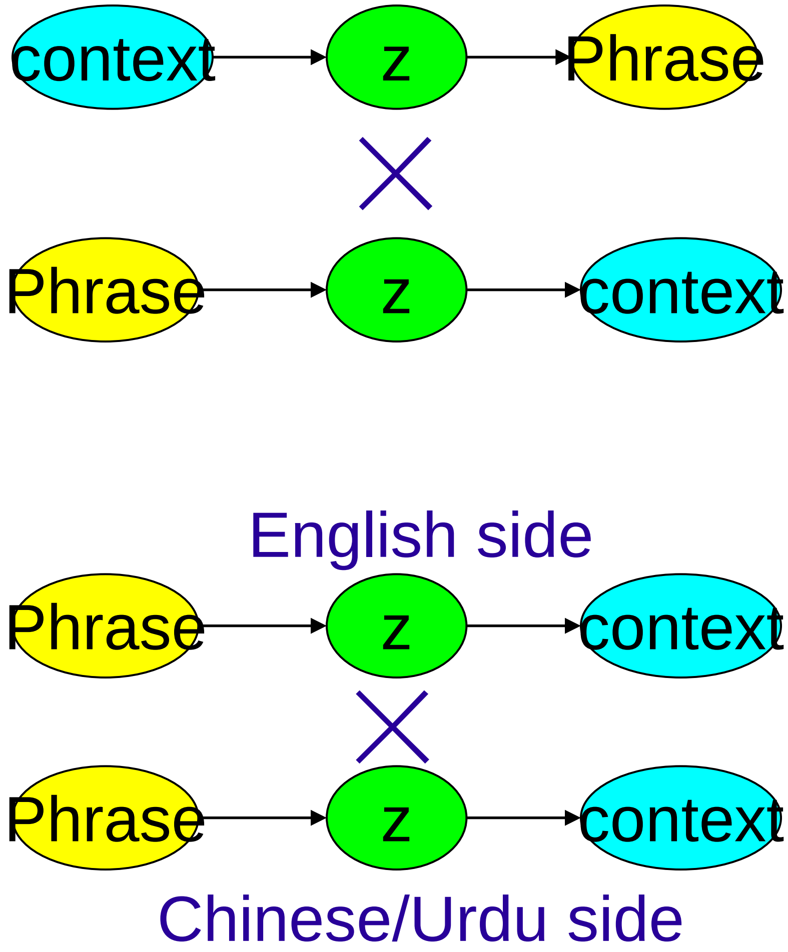
Agreement model



Only look at the first and last word of a phrase.

- implementation: multiply posteriors of two models together.

Agreement model

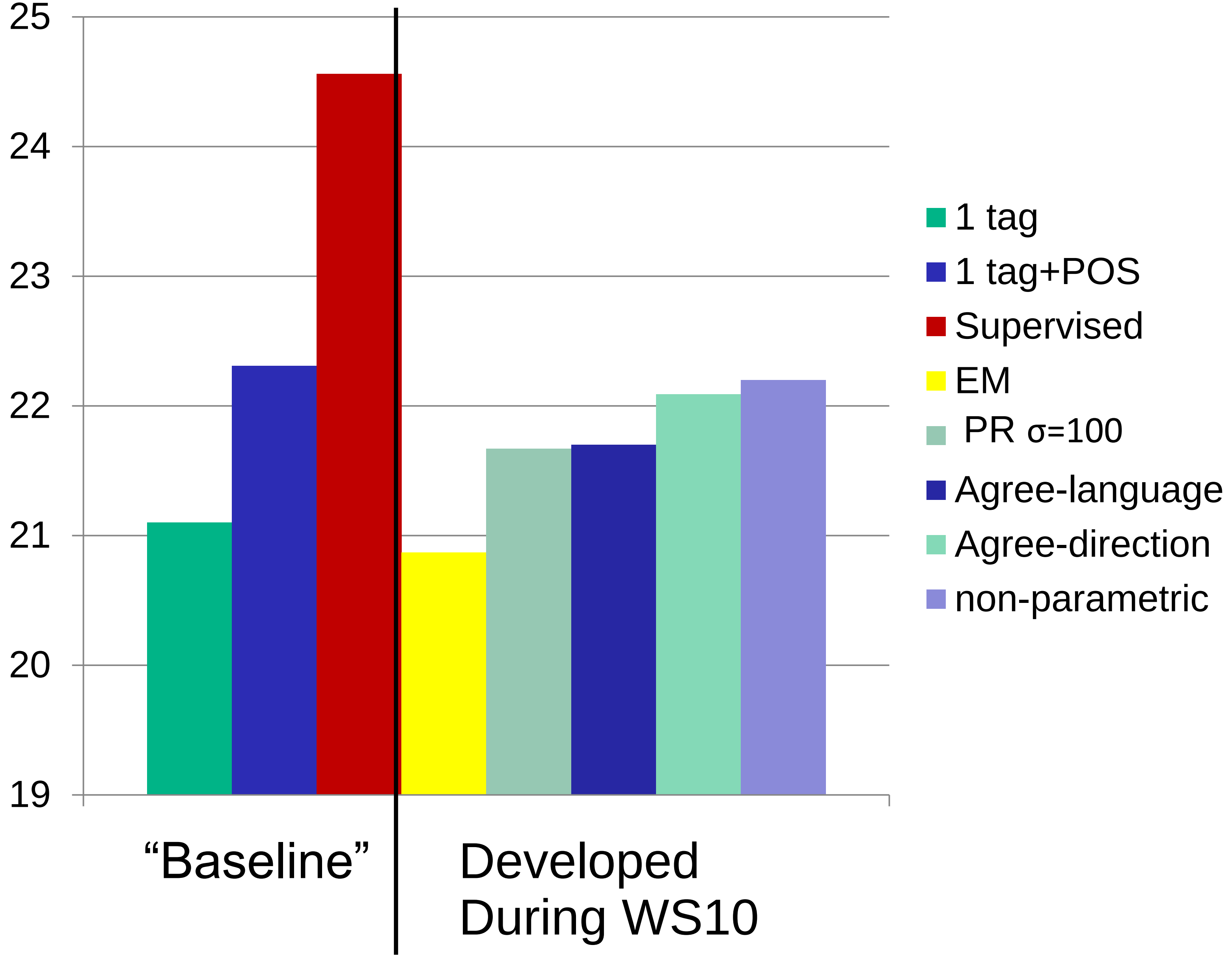


- implementation: multiply posteriors of two models together.

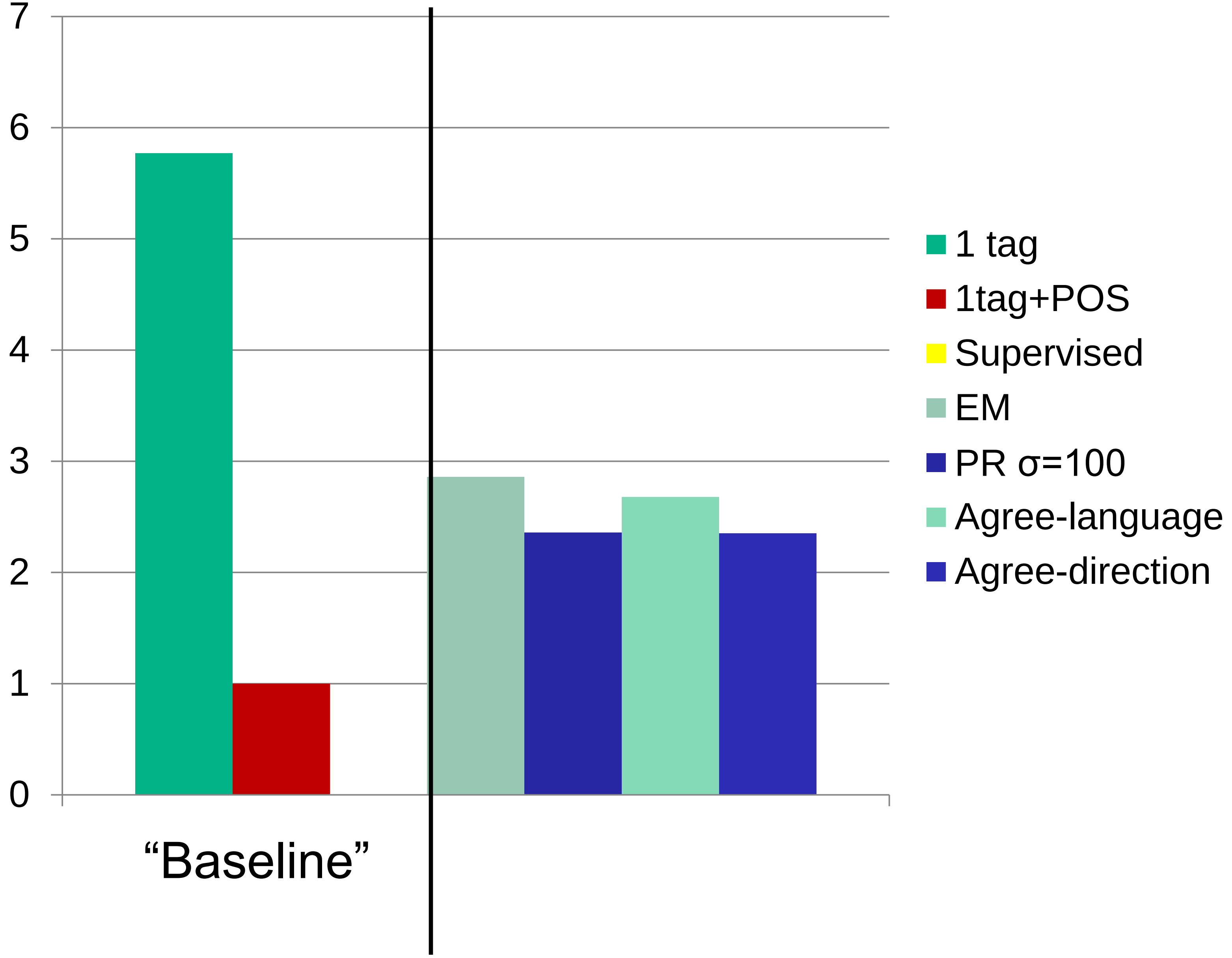
Outline

- Where do phrases come from?
- EM with posterior regularization
- results and future experiments

Evaluation through the translation pipeline on Urdu-English data
BLEU score, higher is better

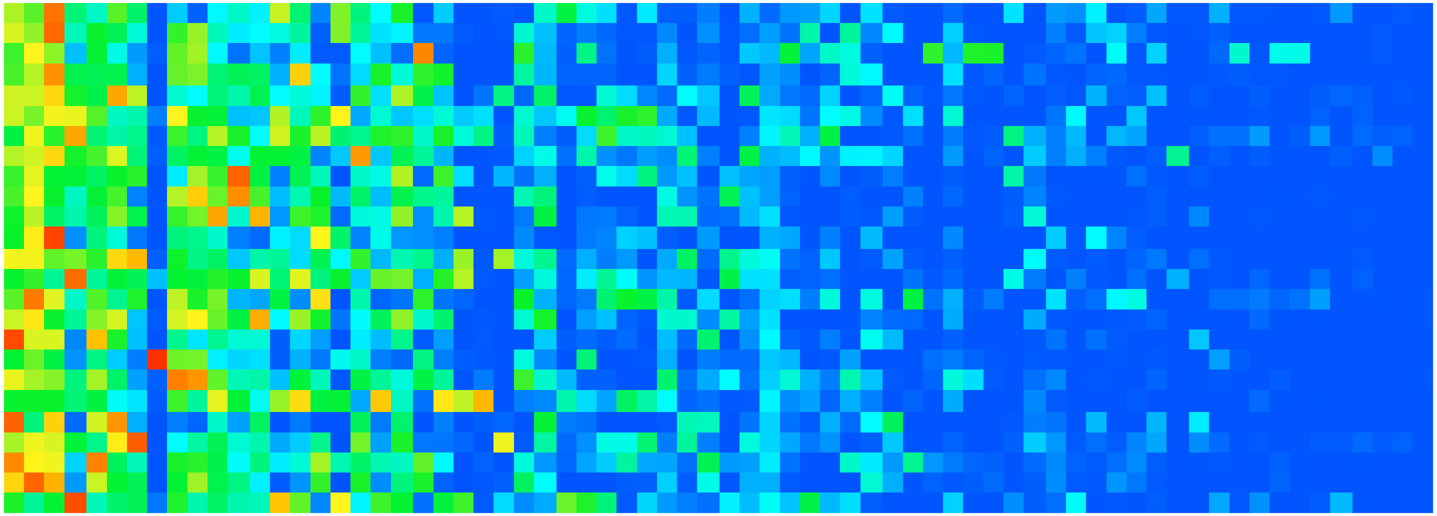


Evaluation against supervised grammar (Conditional Entropy, lower is better)

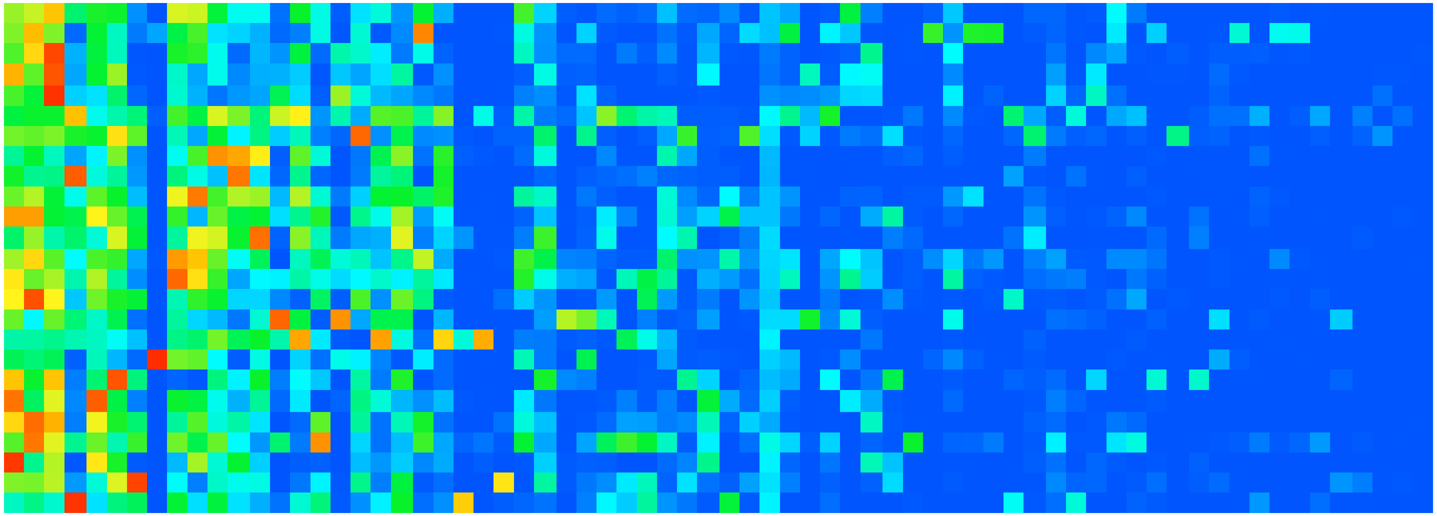


Confusion matrix against supervised labeling

EM



Agreement model between languages



Things we didn't have time to get working

- Semi-supervised training with POS tags.
- Label single-word phrases with their POS tags.

Things we didn't have time to get working

Bayesian Bayesian Bayesian

- variational Bayes inference

Bayesian Bayesian Bayesian

Bayesian Bayesian Bayesian

Bayesian Bayesian Bayesian

Things we didn't have time to get working

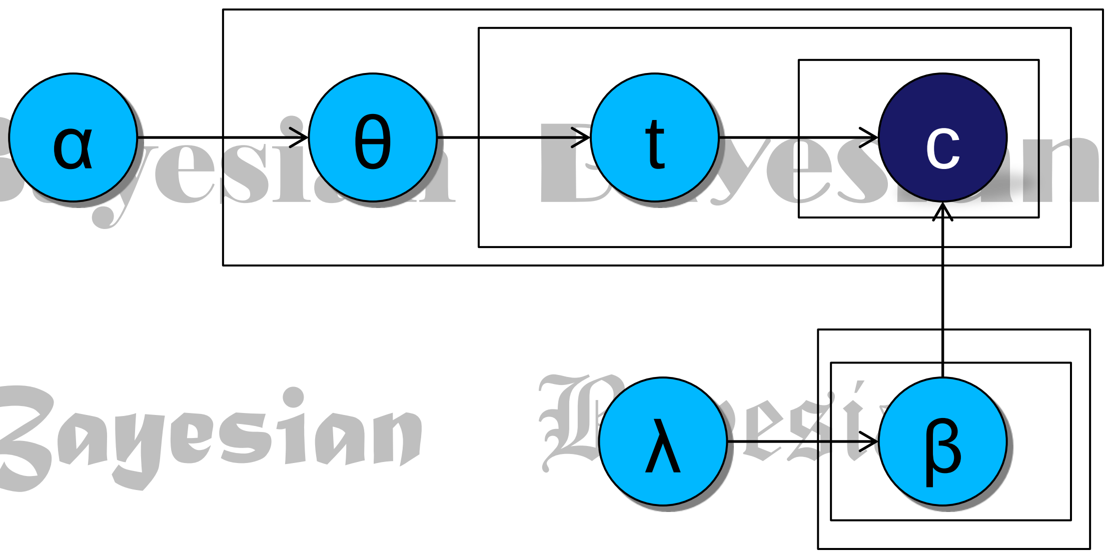
Bayesian Bayesian Bayesian

- variational Bayes inference

Bayesian Bayesian Bayesian

Bayesian Bayesian Bayesian

Bayesian Bayesian Bayesian

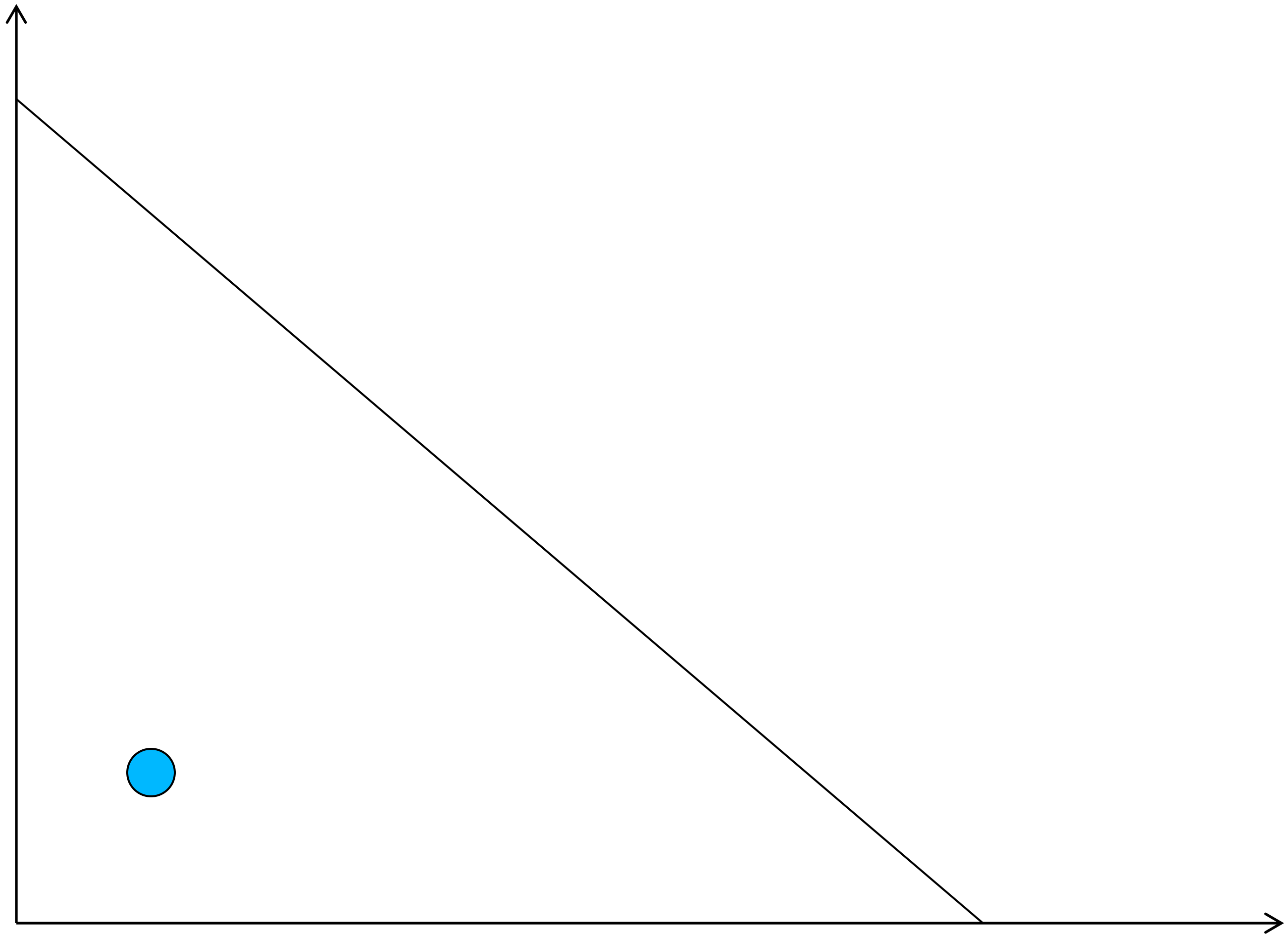


Outline

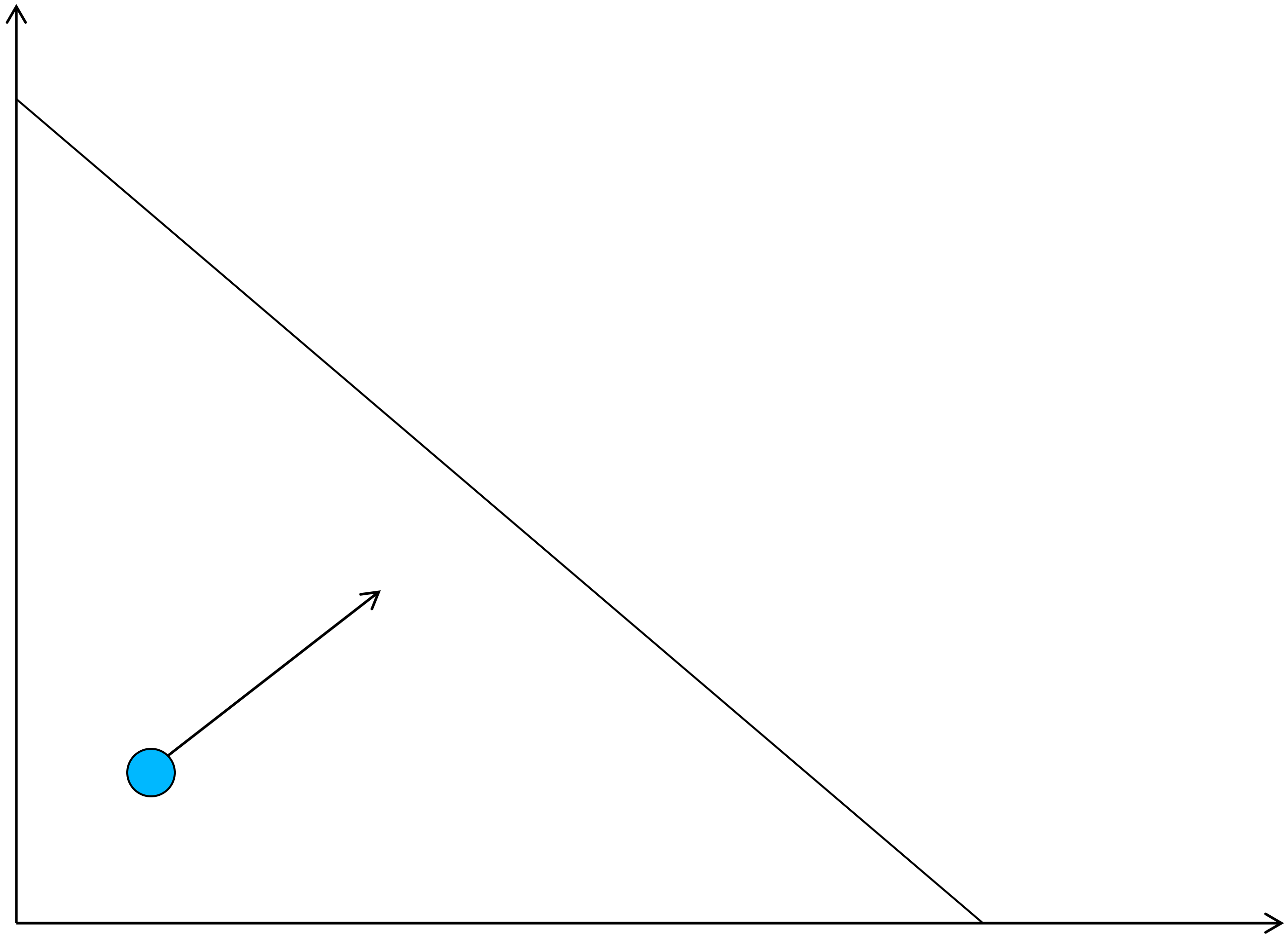
- Where do phrases come from?
- EM with posterior regularization
- results and future experiments

Thanks!

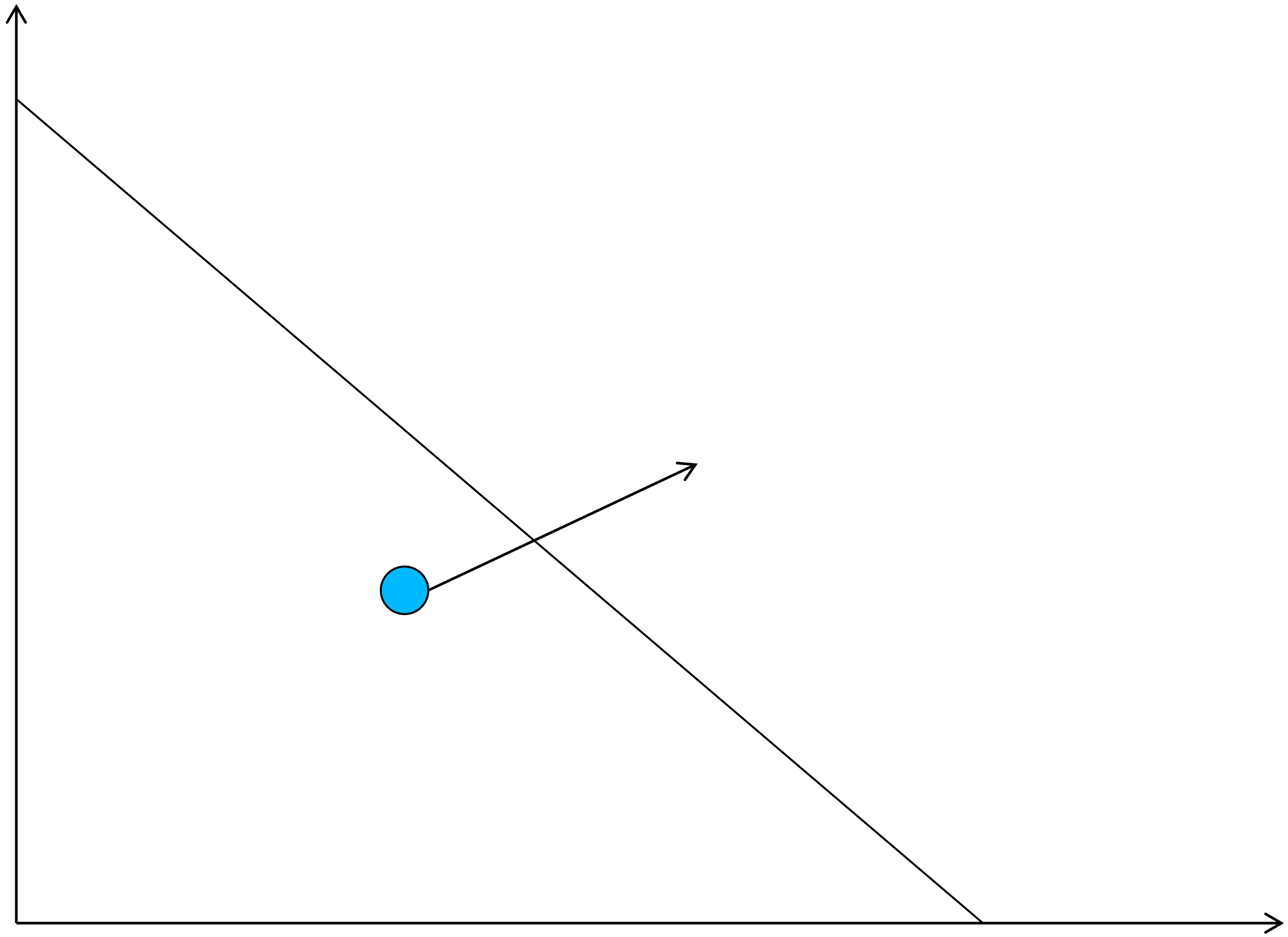
Projected Gradient Descent



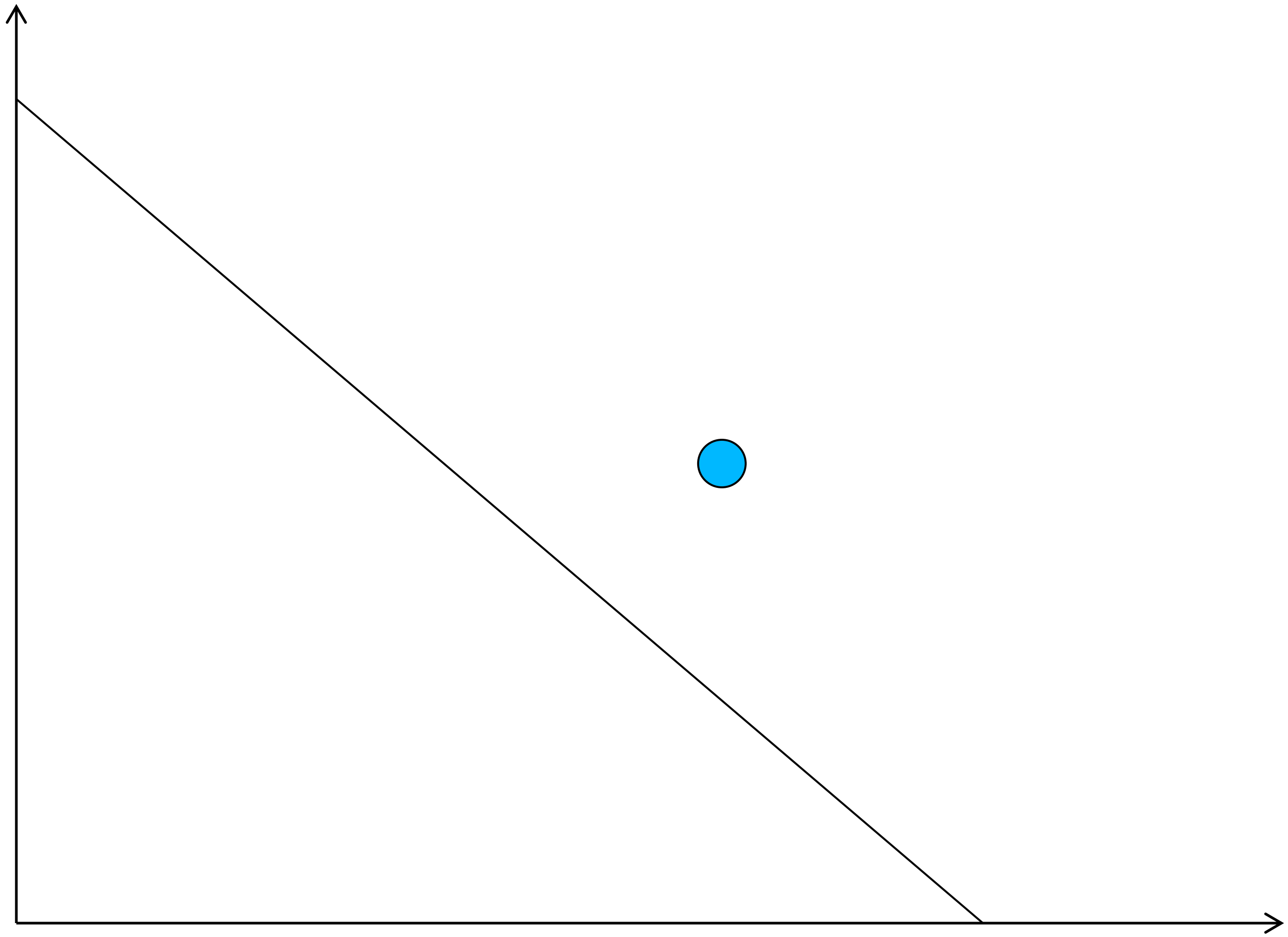
Projected Gradient Descent



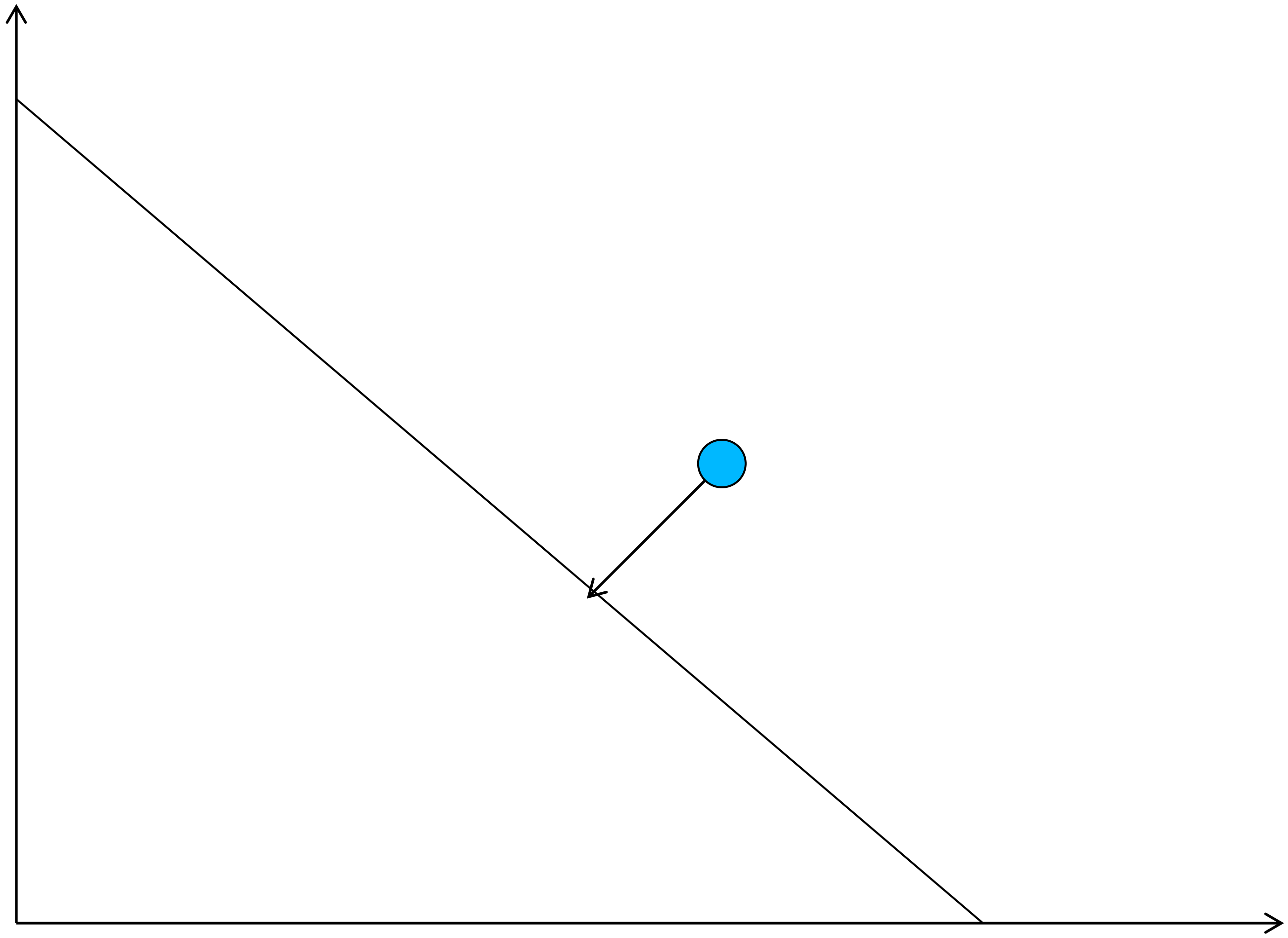
Projected Gradient Descent



Projected Gradient Descent



Projected Gradient Descent



Projected Gradient Descent