Models of Synchronous Grammar Induction for SMT

Workshop 2010

The Center for Speech and Language Processing
Johns Hopkins University

June 28, 2010

Statistical machine translation

Urdu → English

اس حملہ کے بعد بڑی تعداد میں مقامی باشندوں نے علاقوں کو خالی کردیا ہے .

- Statistical machine translation: Learn how to translate from parallel corpora.

Statistical machine translation:

Urdu $\longrightarrow$ English

اس حملہ کے بعد بڑی تعداد میں مقامی باشندوں نے علاقوں کو خالی کردیا ہے .

$\Downarrow$

After this incident, a large number of local residents fled from these areas.

- Statistical machine translation: Learn how to translate from parallel corpora

Statistical machine translation: state-of-the-art

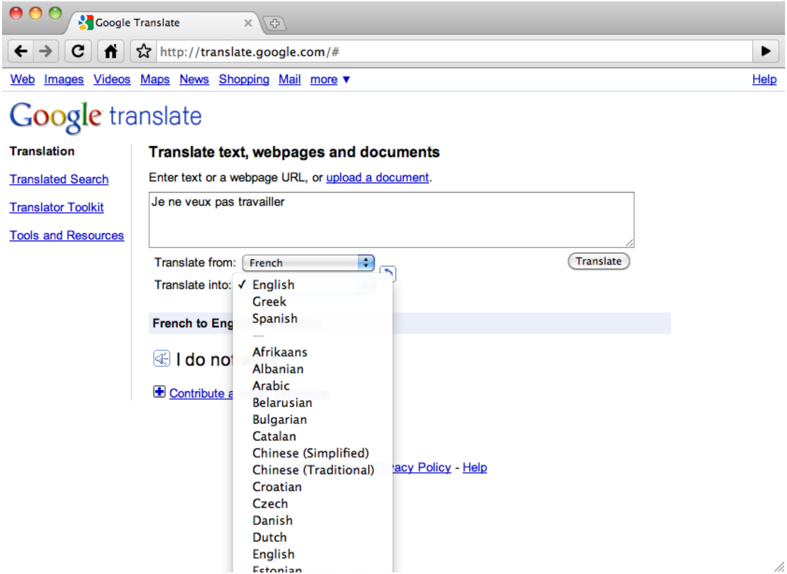
Urdu → English

اس حملہ کے بعد بڑی تعداد میں مقامی باشندوں نے علاقوں کو خالی کردیا ہے .

In this attack a large number of local residents has should vacate areas.

- Current state-of-the-art translation models struggle with language pairs which exhibit large differences in structure.

Statistical machine translation: successes



Statistical machine translation: limitations

Structural divergence between languages:

English	**Who wrote this letter?**
Arabic	من الذي كتب هذه الرسالة؟ (function-word) (who) (wrote) (this) (the-letter)
Chinese	这封信是谁写的？ (this) (letter) (be) (who) (write) (come-from) (function-word)

Statistical machine translation: limitations

Structural divergence between languages:

English	**Who wrote this letter?**
Arabic	من الذي كتب هذه الرسالة؟ (function-word) (who) (wrote) (this) (the-letter)
Chinese	这封 信 是 谁 写 的 ？ (this) (letter) (be) (who) (write) (come-from) (function-word)

Statistical machine translation: limitations

Structural divergence between languages:

English	**Who wrote this letter?**
Arabic	من الذي كتب هذه الرسالة؟ (function-word) (who) (wrote) (this) (the-letter)
Chinese	这封 信 是 谁 写 的 ？ (this) (letter) (be) (who) (write) (come-from) (function-word)

Statistical machine translation: limitations

Structural divergence between languages:

English	**Who wrote this letter?**
Arabic	من الذي كتب هذه الرسالة؟ (function-word) (who) (wrote) (this) (the-letter)
Chinese	这封 信 是 谁 写 的 ？ (this) (letter) (be) (who) (write) (come-from) (function-word)

- Phrasal translation equivalences
- Constituent reordering
- Morphology

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle \textit{Sie},\ \textit{She} \rangle$

$X \rightarrow \langle \textit{eine Tasse Kaffee},\ \textit{a cup of coffee} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$X \rightarrow \langle \textit{will},\ \textit{wants to} \rangle$

$X \rightarrow \langle \textit{trinken},\ \textit{drink} \rangle$

Example Derivation

$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \quad \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$

$\Rightarrow \langle \textit{Sie}\ X_{\boxed{3}},\ \textit{She}\ X_{\boxed{3}} \rangle \quad \Rightarrow \langle \textit{Sie}\ X_{\boxed{4}}\ X_{\boxed{5}},\ \textit{She}\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$

$\Rightarrow \langle \textit{Sie will}\ X_{\boxed{5}},\ \textit{She wants to}\ X_{\boxed{5}} \rangle \quad \Rightarrow \langle \textit{Sie will}\ X_{\boxed{6}} X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}} X_{\boxed{6}} \rangle$

$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee}\ X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}}\ \textit{a cup of coffee} \rangle$

$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee trinken},\ \textit{She wants to drink a cup of coffee} \rangle$

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle \mathit{Sie},\ \mathit{She} \rangle$

$X \rightarrow \langle \mathit{eine\ Tasse\ Kaffee},\ \mathit{a\ cup\ of\ coffee} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$X \rightarrow \langle \mathit{will},\ \mathit{wants\ to} \rangle$

$X \rightarrow \langle \mathit{trinken},\ \mathit{drink} \rangle$

Example Derivation

$$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \quad \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle \mathit{Sie}\ X_{\boxed{3}},\ \mathit{She}\ X_{\boxed{3}} \rangle \quad \Rightarrow \langle \mathit{Sie}\ X_{\boxed{4}}\ X_{\boxed{5}},\ \mathit{She}\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$$

$$\Rightarrow \langle \mathit{Sie\ will}\ X_{\boxed{5}},\ \mathit{She\ wants\ to}\ X_{\boxed{5}} \rangle \quad \Rightarrow \langle \mathit{Sie\ will}\ X_{\boxed{6}} X_{\boxed{7}},\ \mathit{She\ wants\ to}\ X_{\boxed{7}} X_{\boxed{6}} \rangle$$

$$\Rightarrow \langle \mathit{Sie\ will\ eine\ Tasse\ Kaffee}\ X_{\boxed{7}},\ \mathit{She\ wants\ to}\ X_{\boxed{7}}\ \mathit{a\ cup\ of\ coffee} \rangle$$

$$\Rightarrow \langle \mathit{Sie\ will\ eine\ Tasse\ Kaffee\ trinken},\ \mathit{She\ wants\ to\ drink\ a\ cup\ of\ coffee} \rangle$$

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle \textit{Sie},\ \textit{She} \rangle$

$X \rightarrow \langle \textit{will},\ \textit{wants to} \rangle$

$X \rightarrow \langle \textit{eine Tasse Kaffee},\ \textit{a cup of coffee} \rangle$

$X \rightarrow \langle \textit{trinken},\ \textit{drink} \rangle$

Example Derivation

$$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \quad \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle \textit{Sie}\ X_{\boxed{3}},\ \textit{She}\ X_{\boxed{3}} \rangle \quad \Rightarrow \langle \textit{Sie}\ X_{\boxed{4}}\ X_{\boxed{5}},\ \textit{She}\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$$

$$\Rightarrow \langle \textit{Sie will}\ X_{\boxed{5}},\ \textit{She wants to}\ X_{\boxed{5}} \rangle \quad \Rightarrow \langle \textit{Sie will}\ X_{\boxed{6}} X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}} X_{\boxed{6}} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee}\ X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}}\ \textit{a cup of coffee} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee trinken},\ \textit{She wants to drink a cup of coffee} \rangle$$

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle \mathit{Sie},\ \mathit{She} \rangle$

$X \rightarrow \langle \mathit{eine\ Tasse\ Kaffee},\ \mathit{a\ cup\ of\ coffee} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$X \rightarrow \langle \mathit{will},\ \mathit{wants\ to} \rangle$

$X \rightarrow \langle \mathit{trinken},\ \mathit{drink} \rangle$

Example Derivation

$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$

$\Rightarrow \langle \mathit{Sie}\ X_{\boxed{3}},\ \mathit{She}\ X_{\boxed{3}} \rangle \Rightarrow \langle \mathit{Sie}\ X_{\boxed{4}}\ X_{\boxed{5}},\ \mathit{She}\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$

$\Rightarrow \langle \mathit{Sie\ will}\ X_{\boxed{5}},\ \mathit{She\ wants\ to}\ X_{\boxed{5}} \rangle \Rightarrow \langle \mathit{Sie\ will}\ X_{\boxed{6}} X_{\boxed{7}},\ \mathit{She\ wants\ to}\ X_{\boxed{7}} X_{\boxed{6}} \rangle$

$\Rightarrow \langle \mathit{Sie\ will\ eine\ Tasse\ Kaffee}\ X_{\boxed{7}},\ \mathit{She\ wants\ to}\ X_{\boxed{7}}\ \mathit{a\ cup\ of\ coffee} \rangle$

$\Rightarrow \langle \mathit{Sie\ will\ eine\ Tasse\ Kaffee\ trinken},\ \mathit{She\ wants\ to\ drink\ a\ cup\ of\ coffee} \rangle$

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle Sie,\ She \rangle$

$X \rightarrow \langle eine\ Tasse\ Kaffee,\ a\ cup\ of\ coffee \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$X \rightarrow \langle will,\ wants\ to \rangle$

$X \rightarrow \langle trinken,\ drink \rangle$

Example Derivation

$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \quad \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$

$\Rightarrow \langle Sie\ X_{\boxed{3}},\ She\ X_{\boxed{3}} \rangle \quad \Rightarrow \langle Sie\ X_{\boxed{4}}\ X_{\boxed{5}},\ She\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$

$\Rightarrow \langle Sie\ will\ X_{\boxed{5}},\ She\ wants\ to\ X_{\boxed{5}} \rangle \quad \Rightarrow \langle Sie\ will\ X_{\boxed{6}} X_{\boxed{7}},\ She\ wants\ to\ X_{\boxed{7}} X_{\boxed{6}} \rangle$

$\Rightarrow \langle Sie\ will\ eine\ Tasse\ Kaffee\ X_{\boxed{7}},\ She\ wants\ to\ X_{\boxed{7}}\ a\ cup\ of\ coffee \rangle$

$\Rightarrow \langle Sie\ will\ eine\ Tasse\ Kaffee\ trinken,\ She\ wants\ to\ drink\ a\ cup\ of\ coffee \rangle$

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle \textit{Sie},\ \textit{She} \rangle$

$X \rightarrow \langle \textit{eine Tasse Kaffee},\ \textit{a cup of coffee} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$X \rightarrow \langle \textit{will},\ \textit{wants to} \rangle$

$X \rightarrow \langle \textit{trinken},\ \textit{drink} \rangle$

Example Derivation

$$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \quad \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle \textit{Sie}\ X_{\boxed{3}},\ \textit{She}\ X_{\boxed{3}} \rangle \quad \Rightarrow \langle \textit{Sie}\ X_{\boxed{4}}\ X_{\boxed{5}},\ \textit{She}\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$$

$$\Rightarrow \langle \textit{Sie will}\ X_{\boxed{5}},\ \textit{She wants to}\ X_{\boxed{5}} \rangle \quad \Rightarrow \langle \textit{Sie will}\ X_{\boxed{6}} X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}} X_{\boxed{6}} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee}\ X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}}\ \textit{a cup of coffee} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee trinken},\ \textit{She wants to drink a cup of coffee} \rangle$$

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle \textit{Sie},\ \textit{She} \rangle$

$X \rightarrow \langle \textit{eine Tasse Kaffee},\ \textit{a cup of coffee} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$X \rightarrow \langle \textit{will},\ \textit{wants to} \rangle$

$X \rightarrow \langle \textit{trinken},\ \textit{drink} \rangle$

Example Derivation

$$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \quad \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle \textit{Sie}\ X_{\boxed{3}},\ \textit{She}\ X_{\boxed{3}} \rangle \quad \Rightarrow \langle \textit{Sie}\ X_{\boxed{4}}\ X_{\boxed{5}},\ \textit{She}\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$$

$$\Rightarrow \langle \textit{Sie will}\ X_{\boxed{5}},\ \textit{She wants to}\ X_{\boxed{5}} \rangle \quad \Rightarrow \langle \textit{Sie will}\ X_{\boxed{6}} X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}} X_{\boxed{6}} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee}\ X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}}\ \textit{a cup of coffee} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee trinken},\ \textit{She wants to drink a cup of coffee} \rangle$$

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle \textit{Sie},\ \textit{She} \rangle$$

$$X \rightarrow \langle \textit{eine Tasse Kaffee},\ \textit{a cup of coffee} \rangle$$

$$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$$

$$X \rightarrow \langle \textit{will},\ \textit{wants to} \rangle$$

$$X \rightarrow \langle \textit{trinken},\ \textit{drink} \rangle$$

Example Derivation

$$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \quad \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle \textit{Sie}\ X_{\boxed{3}},\ \textit{She}\ X_{\boxed{3}} \rangle \quad \Rightarrow \langle \textit{Sie}\ X_{\boxed{4}}\ X_{\boxed{5}},\ \textit{She}\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$$

$$\Rightarrow \langle \textit{Sie will}\ X_{\boxed{5}},\ \textit{She wants to}\ X_{\boxed{5}} \rangle \quad \Rightarrow \langle \textit{Sie will}\ X_{\boxed{6}} X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}} X_{\boxed{6}} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee}\ X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}}\ \textit{a cup of coffee} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee trinken},\ \textit{She wants to drink a cup of coffee} \rangle$$

Using syntax in Machine Translation:

Synchronous Context Free Grammar (SCFG)

$S \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{2}}\ X_{\boxed{1}} \rangle$

$X \rightarrow \langle \textit{Sie},\ \textit{She} \rangle$

$X \rightarrow \langle \textit{eine Tasse Kaffee},\ \textit{a cup of coffee} \rangle$

$X \rightarrow \langle X_{\boxed{1}}\ X_{\boxed{2}},\ X_{\boxed{1}}\ X_{\boxed{2}} \rangle$

$X \rightarrow \langle \textit{will},\ \textit{wants to} \rangle$

$X \rightarrow \langle \textit{trinken},\ \textit{drink} \rangle$

Example Derivation

$$S \Rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \quad \Rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}},\ X_{\boxed{2}}\ X_{\boxed{3}} \rangle$$

$$\Rightarrow \langle \textit{Sie}\ X_{\boxed{3}},\ \textit{She}\ X_{\boxed{3}} \rangle \quad \Rightarrow \langle \textit{Sie}\ X_{\boxed{4}}\ X_{\boxed{5}},\ \textit{She}\ X_{\boxed{4}}\ X_{\boxed{5}} \rangle$$

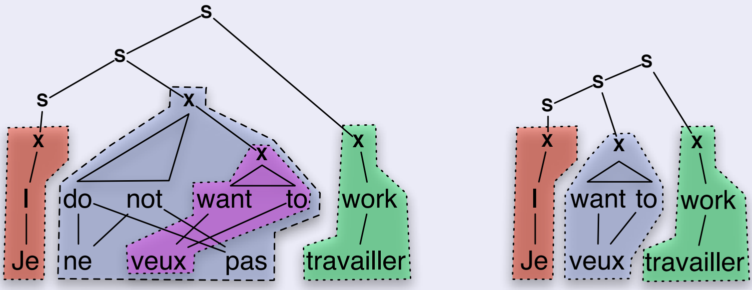
$$\Rightarrow \langle \textit{Sie will}\ X_{\boxed{5}},\ \textit{She wants to}\ X_{\boxed{5}} \rangle \quad \Rightarrow \langle \textit{Sie will}\ X_{\boxed{6}} X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}} X_{\boxed{6}} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee}\ X_{\boxed{7}},\ \textit{She wants to}\ X_{\boxed{7}}\ \textit{a cup of coffee} \rangle$$

$$\Rightarrow \langle \textit{Sie will eine Tasse Kaffee trinken},\ \textit{She wants to drink a cup of coffee} \rangle$$

Models of translation

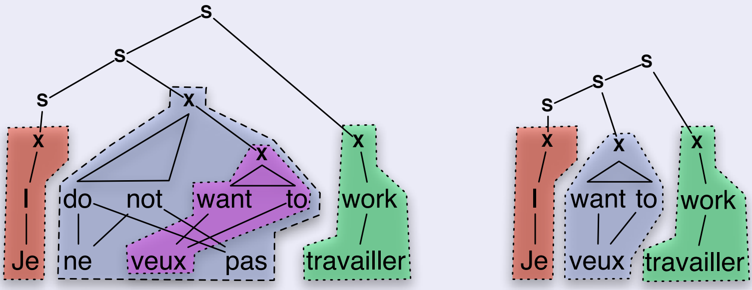
Unlabelled SCFG: Hiero



- S -> $\langle$ $X_{\boxed{1}}$, $X_{\boxed{1}}$ $\rangle$,
  S -> $\langle$ $S_{\boxed{1}}$ $X_{\boxed{2}}$, $S_{\boxed{1}}$ $X_{\boxed{2}}$ $\rangle$
- X -> $\langle$ Je, I $\rangle$, X -> $\langle$ ne $X_{\boxed{1}}$ pas, do not $X_{\boxed{1}}$ $\rangle$,
  X -> $\langle$ veux, want to$\rangle$, X -> $\langle$ travailler, work $\rangle$

Models of translation

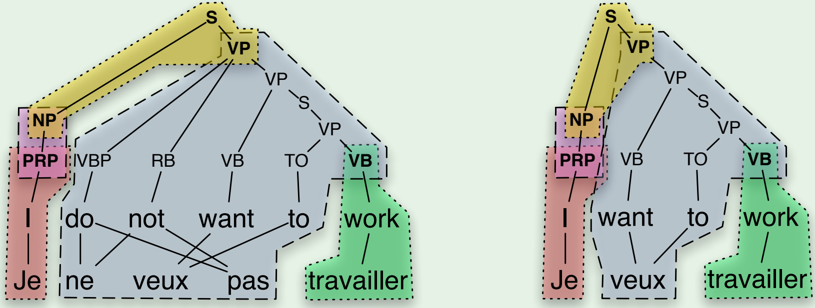
Unlabelled SCFG: Hiero



- Only requires the parallel corpus.
- But weak model of sentence structure.

Models of translation

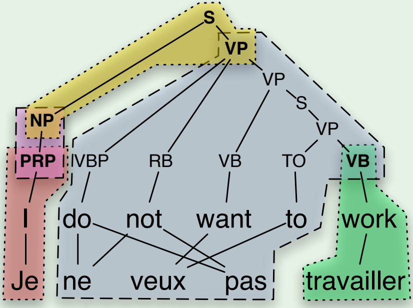
Supervised SCFG: Syntactic Tree-to-String



- S -> ⟨ $\text{NP}_{\boxed{1}}$ $\text{VP}_{\boxed{2}}$, $\text{NP}_{\boxed{1}}$ $\text{VP}_{\boxed{2}}$ ⟩,
  NP -> ⟨ $\text{PRP}_{\boxed{1}}$, $\text{PRP}_{\boxed{1}}$ ⟩
- PRP -> ⟨ Je, I ⟩, VB -> ⟨ travailler, work ⟩
  VP -> ⟨ ne veux pas $\text{VB}_{\boxed{1}}$, do not want to $\text{VB}_{\boxed{1}}$ ⟩

Models of translation

Supervised SCFG: Syntactic Tree-to-String



- Strong model of sentence structure.
- Reliant on a treebank to train the parser.

Impact

Language	Words	Domain
English	4.5M	Financial news
Chinese	0.5M	Broadcasting news
Arabic	300K (1M planned)	News
Korean	54K	Military

Table: Major treebanks: data size and domain

Impact

Parallel corpora far exceed treebanks (millions of words):

	7	90	83	55	40	50	55	28	29	12	12	8	10	8	7	21	6	6
	90	7	34	24	29	12	10	11	11	9	11	7	6	6	7	4	5	5
	83	34	7	17	16	12	10	12	11	9	10	8	6	6	7	6	6	5
	52	24	17	6	14	12	9	9	10	9	10	7	5	5	6	3	5	5
	39	29	16	14	6	9	10	7	8	8	10	8	6	6	6	3	5	5
	48	12	12	12	9	3	25	5	5	22	6	2	3	2	3	3	3	3
	55	10	10	9	10	26	2	2	2	8	5	2	2	2	2	2	2	2
	26	11	12	9	7	5	2	7	12	3	4	6	5	4	7	3	5	5
	29	11	11	10	8	5	2	12	6	3	4	6	6	5	6	3	5	5
	12	9	9	9	8	23	8	3	3	2	6	1	2	2	2	2	2	2
	11	11	10	10	10	6	5	4	4	6	4	5	3	3	4	1	3	3
	8	7	8	7	8	2	2	6	6	1	5	5	4	4	5	2	4	4

Models of translation

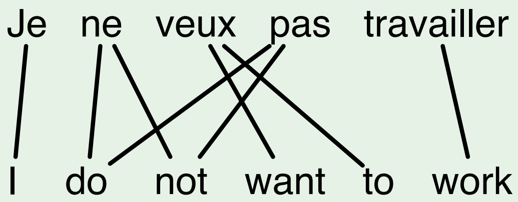
Phrase extraction:

Je ne veux pas travailler

I do not want to work

Models of translation

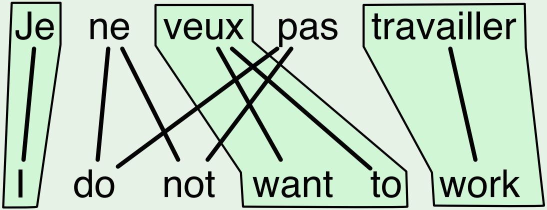
Phrase extraction:



- Use a word-based translation model to annotate the parallel corpus with word-alignments

Models of translation

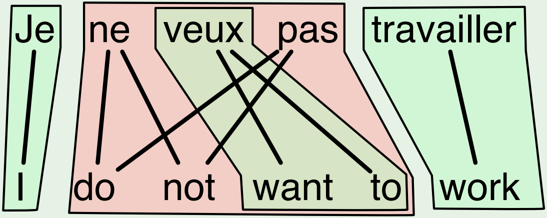
Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩

Models of translation

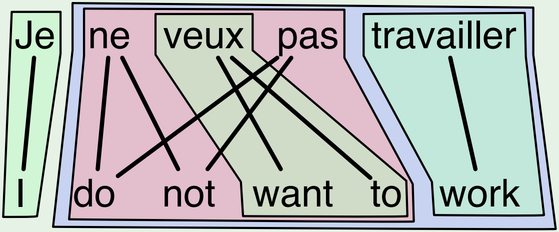
Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩, ⟨ ne veux pas, do not want to ⟩

Models of translation

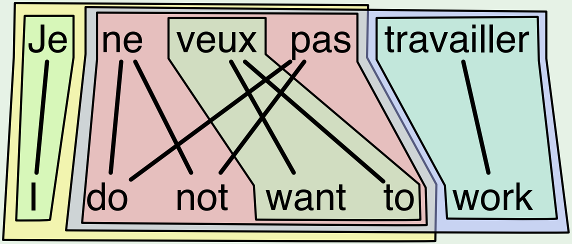
Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩, ⟨ ne veux pas, do not want to ⟩, ⟨ ne veux pas travailler, do not want to work ⟩

Models of translation

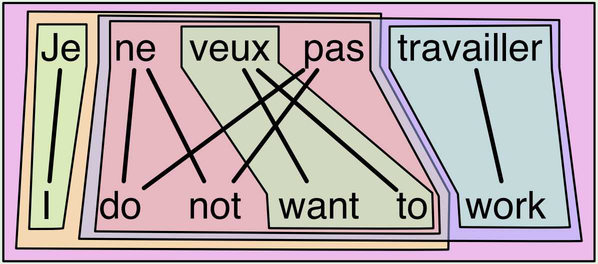
Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩, ⟨ ne veux pas, do not want to ⟩, ⟨ ne veux pas travailler, do not want to work ⟩, ⟨ Je ne veux pas, I do not want to ⟩

Models of translation

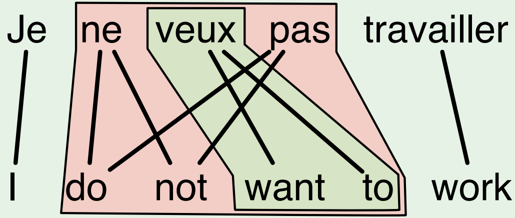
Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩, ⟨ ne veux pas, do not want to ⟩, ⟨ ne veux pas travailler, do not want to work ⟩, ⟨ Je ne veux pas, I do not want to ⟩, ⟨ Je ne veux pas travailler, I do not want to work ⟩

Models of translation

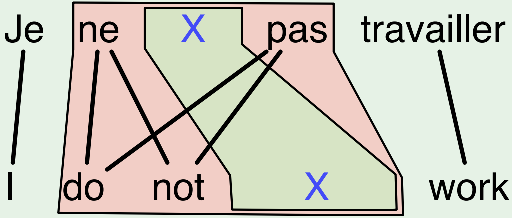
SCFG Rule extraction:



- X -> ⟨ ne veux pas, do not want to ⟩

Models of translation

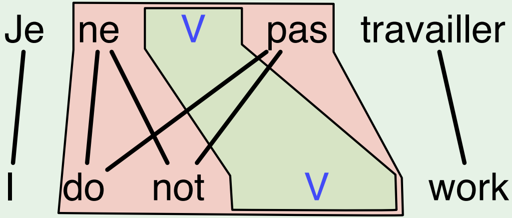
SCFG Rule extraction:



- X -> ⟨ ne veux pas, do not want to ⟩,
- X -> ⟨ ne $X_{\boxed{1}}$ pas, do not $X_{\boxed{1}}$ ⟩

Models of translation

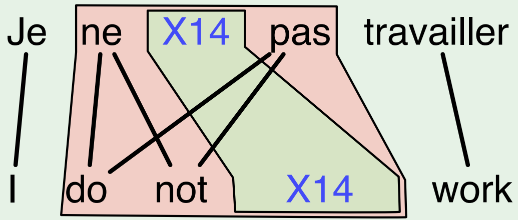
SCFG Rule extraction:



- VP/NN -> $\langle$ ne veux pas, do not want to $\rangle$,
- VP/NN -> $\langle$ ne $V_{\boxed{1}}$ pas, do not $V_{\boxed{1}}$ $\rangle$

Models of translation

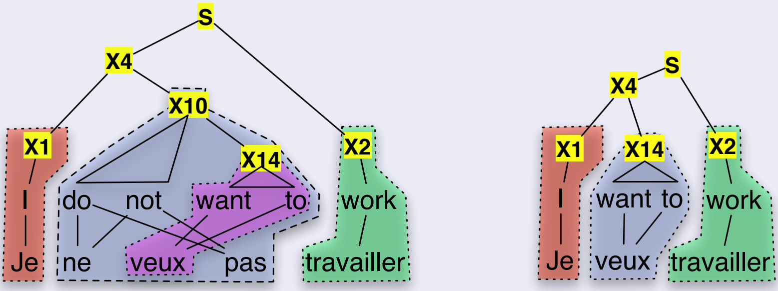
SCFG Rule extraction:



- X10 -> $\langle$ ne veux pas, do not want to $\rangle$,
- X10 -> $\langle$ ne $X14_{\boxed{1}}$ pas, do not $X14_{\boxed{1}}$ $\rangle$

Models of translation

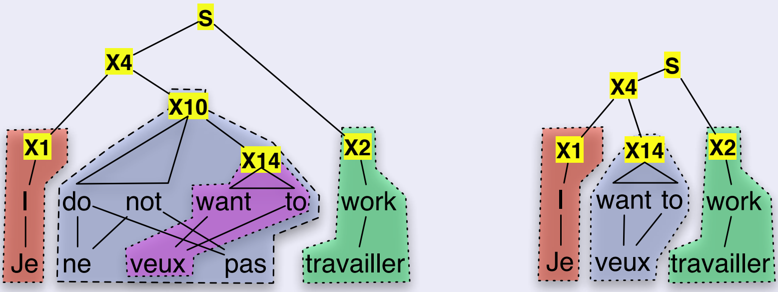
This workshop



- S -> ⟨ $X4_{\boxed{1}}$ $X2_{\boxed{2}}$, $X4_{\boxed{1}}$ $X2_{\boxed{2}}$ ⟩, X4 -> ⟨ $X1_{\boxed{1}}$ $X10_{\boxed{2}}$, $X1_{\boxed{1}}$ $X10_{\boxed{2}}$ ⟩
- X1 -> ⟨ Je, I ⟩, X10 -> ⟨ ne $X14_{\boxed{1}}$ pas, do not $X14_{\boxed{1}}$ ⟩, X14 -> ⟨ veux, want to⟩, X10 -> ⟨ travailler, work ⟩

Models of translation

This workshop



- Only requires the parallel corpus.
- But also gives a strong model of sentence structure.

Workshop overview

Input:

- Existing procedures for unlabelled synchronous grammar extraction

Output:

- New unsupervised models for large scale synchronous grammar extraction,
- A comparison and analysis of the existing and proposed models,
- Extended decoders (cdec/Joshua) capable of working efficiently with these models.

Workshop Streams

1. Implement scalable labelled SCFG grammar induction algorithms:
   - by clustering translation phrases which occur in the same context we can learn which phrases are substituteable,
   - we have implemented both parametric and non-parametric Bayesian clustering algorithms.
2. Improve SCFG decoders to efficiently handle the grammars produced:
   - translation complexity scales quadratically as we add more categories,
   - in order to decode efficiently with the grammars we've induced we have created faster search algorithms tuned for syntactic grammars.
3. Investigate discriminative training regimes to leverage features extracted from these grammars:
   - to make the most of our induced grammars we need discriminative training algorithms that learn from more than a handful of features,
   - we've implemented two large scale discriminative algorithms for training our models.

Extrinsic evaluation: Bleu

Ngram overlap metrics:

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

Extrinsic evaluation: Bleu

Ngram overlap metrics: 1-gram precision $p_1 = \frac{11}{14}$

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

Extrinsic evaluation: Bleu

Ngram overlap metrics: 2-gram precision $p_2 = \frac{5}{13}$

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

Extrinsic evaluation: Bleu

Ngram overlap metrics: 3-gram precision $p_3 = \frac{2}{12}$

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

Extrinsic evaluation: Bleu

Ngram overlap metrics: 4-gram precision $p_4 = \frac{1}{11}$

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

Extrinsic evaluation: Bleu

BLEU

$$BLEU_n = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

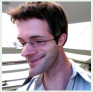
$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{R'}{C'}\right) & \text{if } c <= r \end{cases}$$

- BP is the *Brevity Penalty*, w_n is the ngram length weights (usually $\frac{1}{n}$), p_n is precision of ngram predictions, R' is the total length of all references and C' is the sum of the best matching candidates.
- statistics are calculate over the whole *document*, i.e. all the sentences.

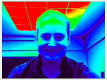
Language pairs

- BTEC Chinese-English:
  - 44k sentence pairs, short sentences
  - Widely reported 'prototyping' corpus
  - Hiero baseline score: 57.0 (16 references)
- NIST Urdu-English:
  - 50k sentence pairs
  - Hiero baseline score: 21.1 (4 references)
  - Major challenges: major long-range reordering, SOV word order
- Europarl Dutch-French:
  - 100k sentence pairs, standard Europarl test sets
  - Hiero baseline score: Europarl 2008 - 15.75 (1 reference)
  - Major challenges: V2 / V-final word order, morphology

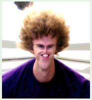
Outline



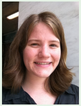
Trevor Cohn



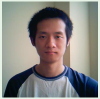
Chris Dyer



Jan Botha



Olivia Buzek



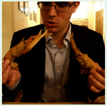
Desai Chen

- 1:55pm Grammar induction and evaluation. Trevor
- 2:10pm Non-parametric models of category induction. Chris
- 2:25pm Inducing categories for morphology. Jan
- 2:35pm Smoothing, backoff and hierarchical grammars. Olivia
- 2:45pm Parametric models: posterior regularisation. Desai
- 3:00pm Break.

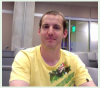
Outline

Vlad Eidelman

Ziyuan Wang



Adam Lopez



Jon Graehl

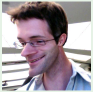
ThuyLinh Nguyen

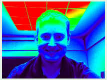
- 3:15pm Training models with rich features spaces. Vlad
- 3:30pm Decoding with complex grammars. Adam
- 4:00pm Closing remarks. Phil
- 4:05pm Finish.

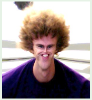
Remember:

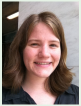
- Idea: Learn synchronous grammar labels which encode substituteability; phrases which occur in the same context should receive the same label.
- Result: Better models of translation structure, morphology and improved decoding algorithms.

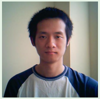
Outline


Trevor Cohn


Chris Dyer


Jan Botha


Olivia Buzek


Desai Chen

- 1:55pm Grammar induction and evaluation. Trevor
- 2:10pm Non-parametric models of category induction. Chris
- 2:25pm Inducing categories for morphology. Jan
- 2:35pm Smoothing, backoff and hierarchical grammars. Olivia
- 2:45pm Parametric models: posterior regularisation. Desai
- 3:00pm Break.

Grammar Induction

Trevor Cohn

- Problem recap
- Clustering hypothesis
- Evaluation

Baseline (Chiang, 2007):

ne $\boxed{\textit{veux}}$ *pas* , *does not* $\boxed{\textit{want}}$

Baseline (Chiang, 2007):

ne $\boxed{}$ *pas* , *does not* $\boxed{}$

$$X \rightarrow \langle \textit{ne}\ X\ \textit{pas}, \textit{does not}\ X \rangle$$

Baseline (Chiang, 2007):

$$\textit{ne}\ \boxed{}\ \textit{pas}\ ,\ \textit{does not}\ \boxed{}$$

$$X \rightarrow \langle \textit{ne}\ X\ \textit{pas}, \textit{does not}\ X \rangle$$

Problem: over-generation

Baseline (Chiang, 2007):

$$\textit{ne}\ \boxed{}\ \textit{pas}\ ,\ \textit{does not}\ \boxed{}$$

$$\mathrm{X} \rightarrow \langle\, \textit{ne}\ \mathrm{X}\ \ \textit{pas}\,,\textit{does not}\ \mathrm{X} \rangle$$

Problem: over-generation

$$\mathrm{X} \rightarrow \langle\, \textit{chat}\ ,\ \textit{cat}\ \rangle$$

Baseline (Chiang, 2007):

$$\textit{ne} \ \boxed{} \ \textit{pas} \ , \ \textit{does not} \ \boxed{}$$

$$\mathrm{X} \rightarrow \langle \textit{ne} \ \mathrm{X} \ \textit{pas} , \textit{does not} \ \mathrm{X} \rangle$$

Problem: over-generation

$$\mathrm{X} \rightarrow \langle \textit{chat} , \textit{cat} \rangle$$

licences

$$\mathrm{X} \Rightarrow \langle \textit{ne chat pas} , \textit{does not cat} \rangle$$

A solution

Use categories which encode the *syntactic role* of the phrase(pair)

ne $\boxed{6}$ *pas* , *does not* $\boxed{6}$

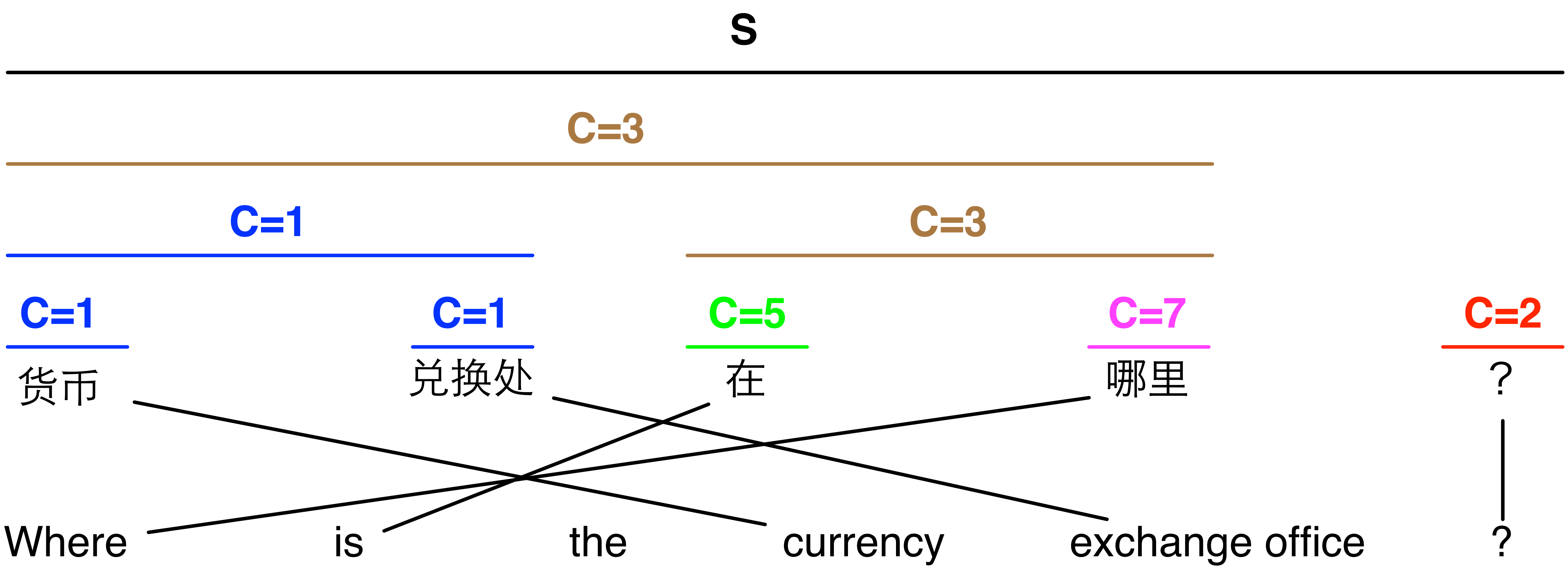
$$X_{18} \rightarrow \langle \textit{ne}\ X_6\ \textit{pas}, \textit{does not}\ X_6 \rangle$$

18 comes from classification of

ne veux pas ,*does not want*

6 comes from classification of

veux , *want*



Clustering must label every n-gram ‘phrase’

constituents: the currency exchange office

and distituents: where is

“A word is known by the company it keeps.”

“Words that occur in the same contexts tend to have similar meanings.”
(Harris, 1954)

Find instances of phrases in context

and teaching them how to	***sing***	with the correct pronuncia
a Koshetz , she went on to	***sing***	with the New York metrop
old your friend's hand and	***sing***	along with teacher ... "
eed with us but on how to	***deal***	with the threat.
and the sailors move on to	***deal***	with the next emergency.
What a	***deal***	!
"It was a	***deal***	," the broadcaster quoted E
“What a	***disgrace***	!
“It was a	***disgrace***	," Clinton said bitterly.
lived in excess and died in	***disgrace***	.

Cluster based on neighbouring words

and teaching them how to	***sing***	with the correct pronuncia
a Koshetz , she went on to	***sing***	with the New York metrop
old your friend's hand and	***sing***	along with teacher ... "
eed with us but on how to	***deal***	with the threat.
and the sailors move on to	***deal***	with the next emergency.
What a	***deal***	!
"It was a	***deal***	," the broadcaster quoted E
“What a	***disgrace***	!
“It was a	***disgrace***	," Clinton said bitterly.
lived in excess and died in	***disgrace***	.

Cluster based on neighbouring words

and teaching them how to **_sing_** with the (Verbs ronunciat

a Koshetz , she went on to **_sing_** with the New York metrop

old your friend's hand and **_sing_** along with teacher ... "

eed with us but on how to **_deal_** with the threat.

and the sailors move on to **_deal_** with the next emergency.

What a **_deal_** !

"It was a **_deal_** ," the broadcaster quoted B

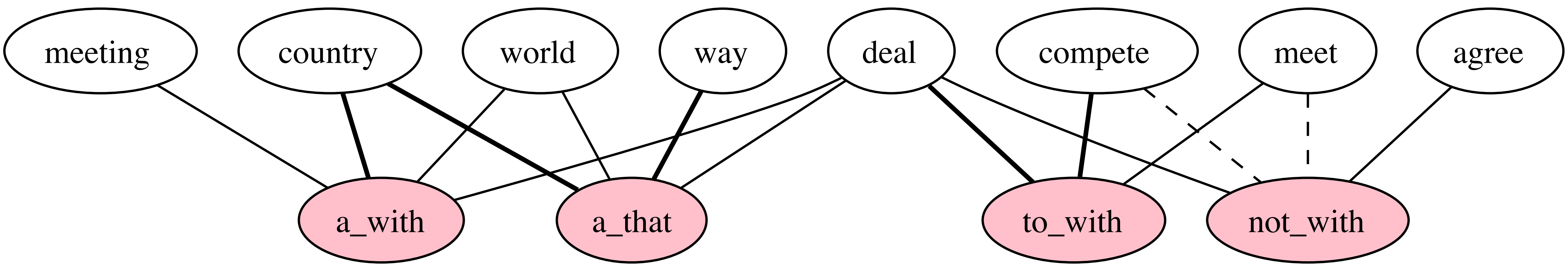
“What a **_disgrace_** !

“It was a **_disgrace_** ," Clinton said bitterly.

lived in excess and died in **_disgrace_** .

Nouns

Phrase-Context Graph



- Desiderata:
  - Edges from a phrase have few category labels
  - Edges from a context have few category labels
  - Similar phrases and contexts share labels

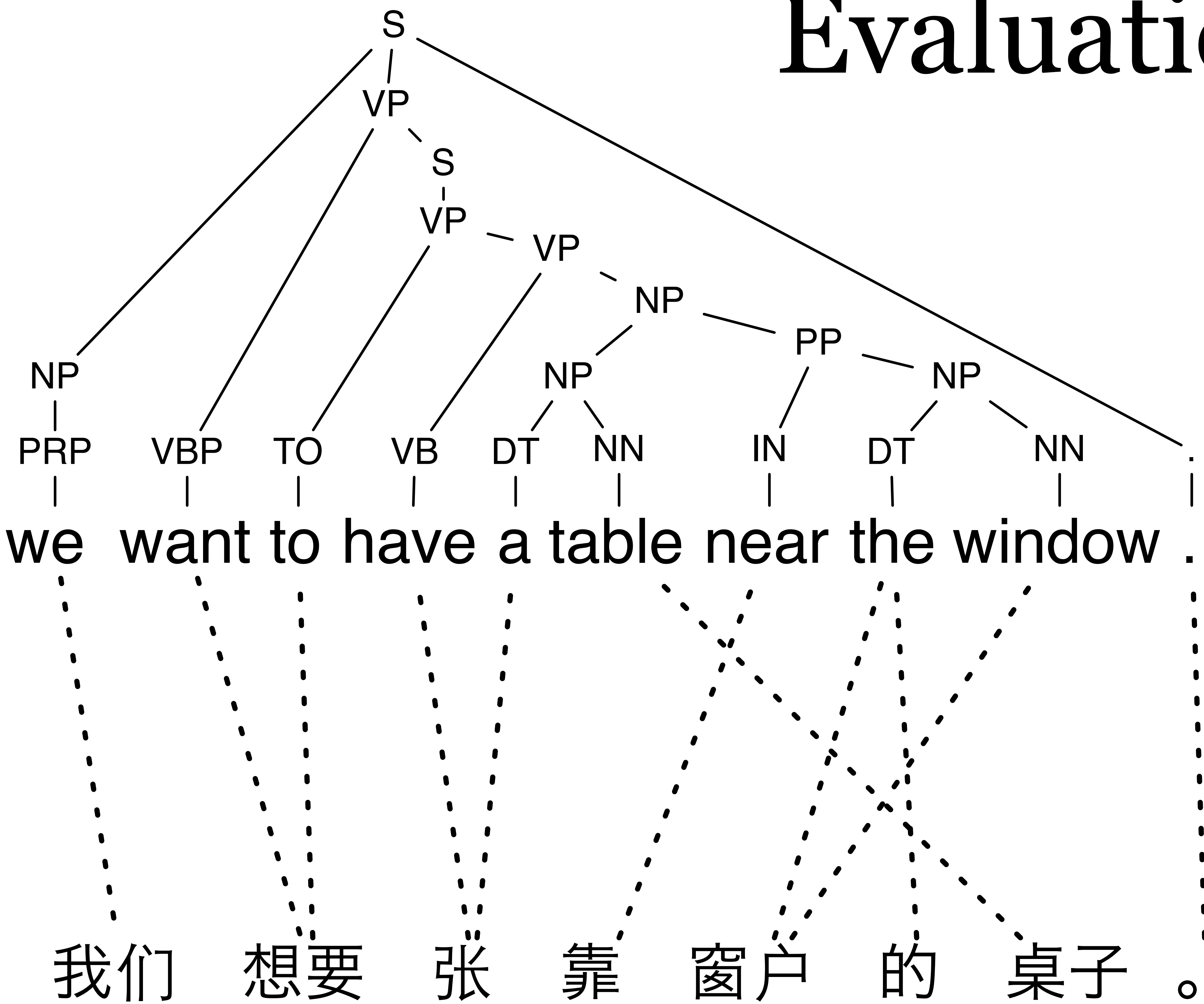
Chris: target parameter sparsity using a hierarchical Pitman-Yor process prior

Desai: find models with sparse posterior distributions using Posterior Regularisation

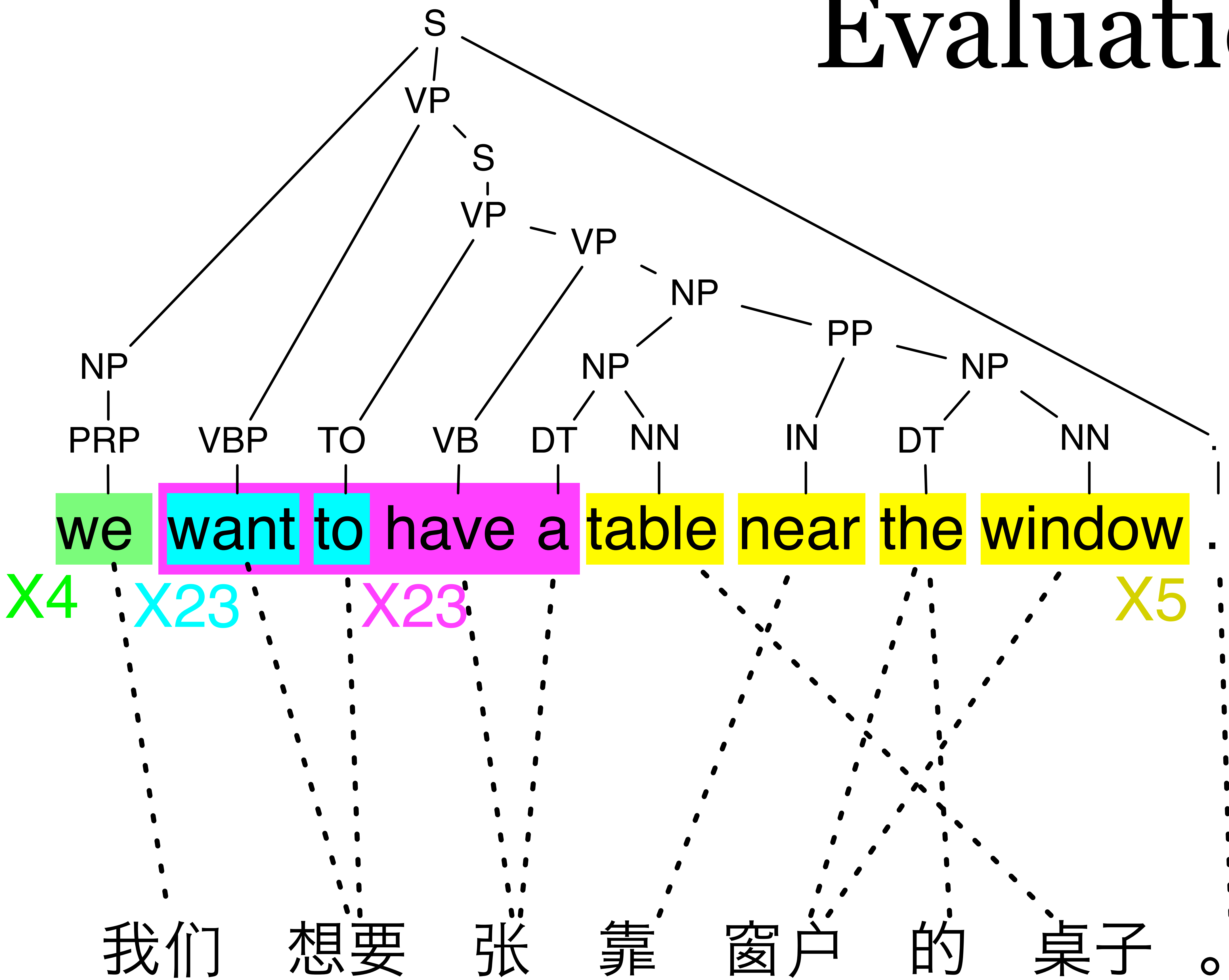
Evaluation

- Primary
  - translation quality (BLEU)
- Secondary
  - *intrinsic* evaluation against treebank parsers
  - compare induced categories to syntactic constituent labels

Evaluation

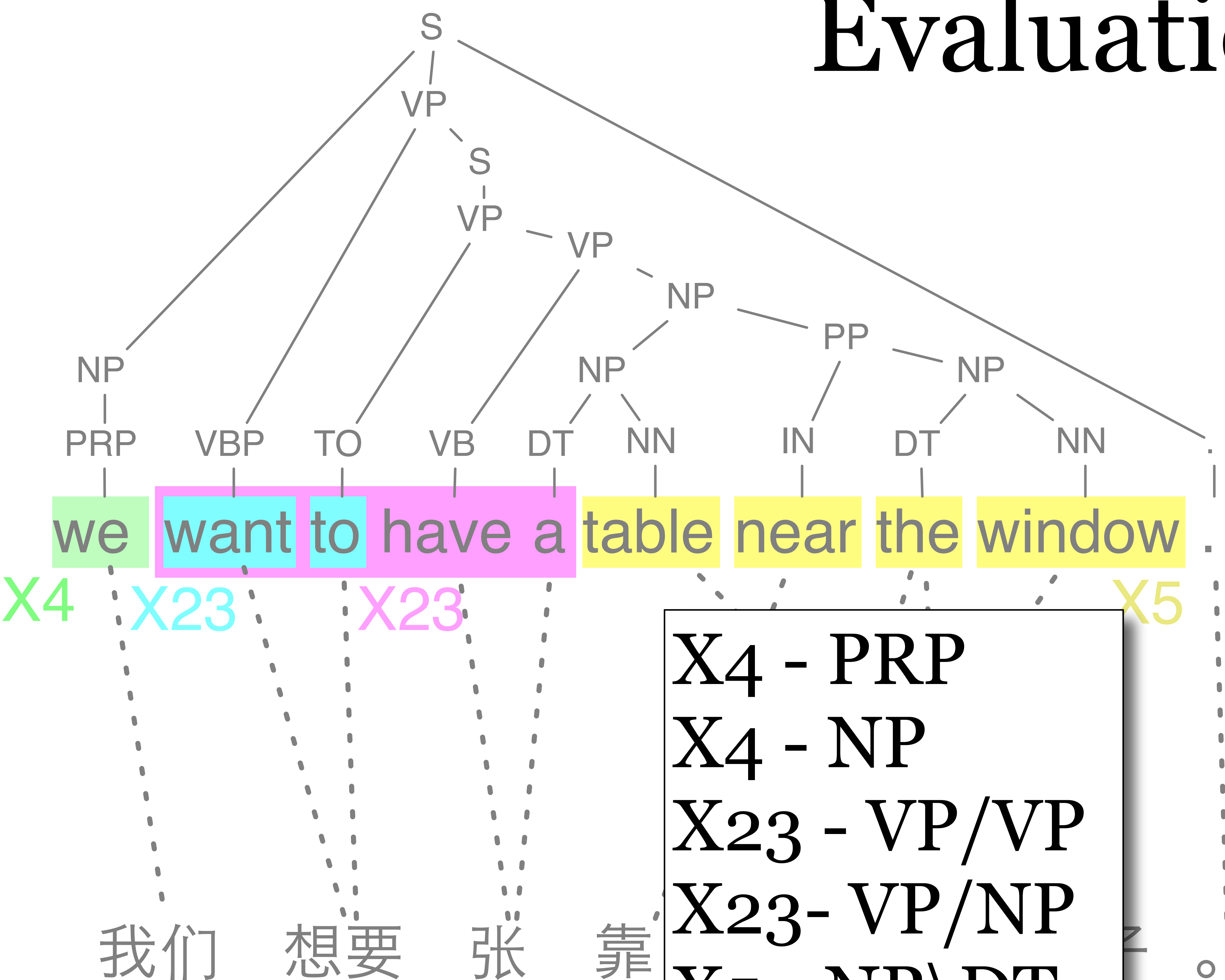

S
VP
S
VP
VP
NP
NP
PP
NP
NP
PRP VBP TO VB DT NN IN DT NN .
we want to have a table near the window .
我们 想要 张 靠 窗户 的 桌子 。

Evaluation



S
VP
S
VP
VP
NP
NP
PP
NP
NP
PRP
VBP
TO
VB
DT
NN
IN
DT
NN
.
we want to have a table near the window .
X4
X23
X23
X5
我们 想要 张 靠 窗户 的 桌子 。

Evaluation

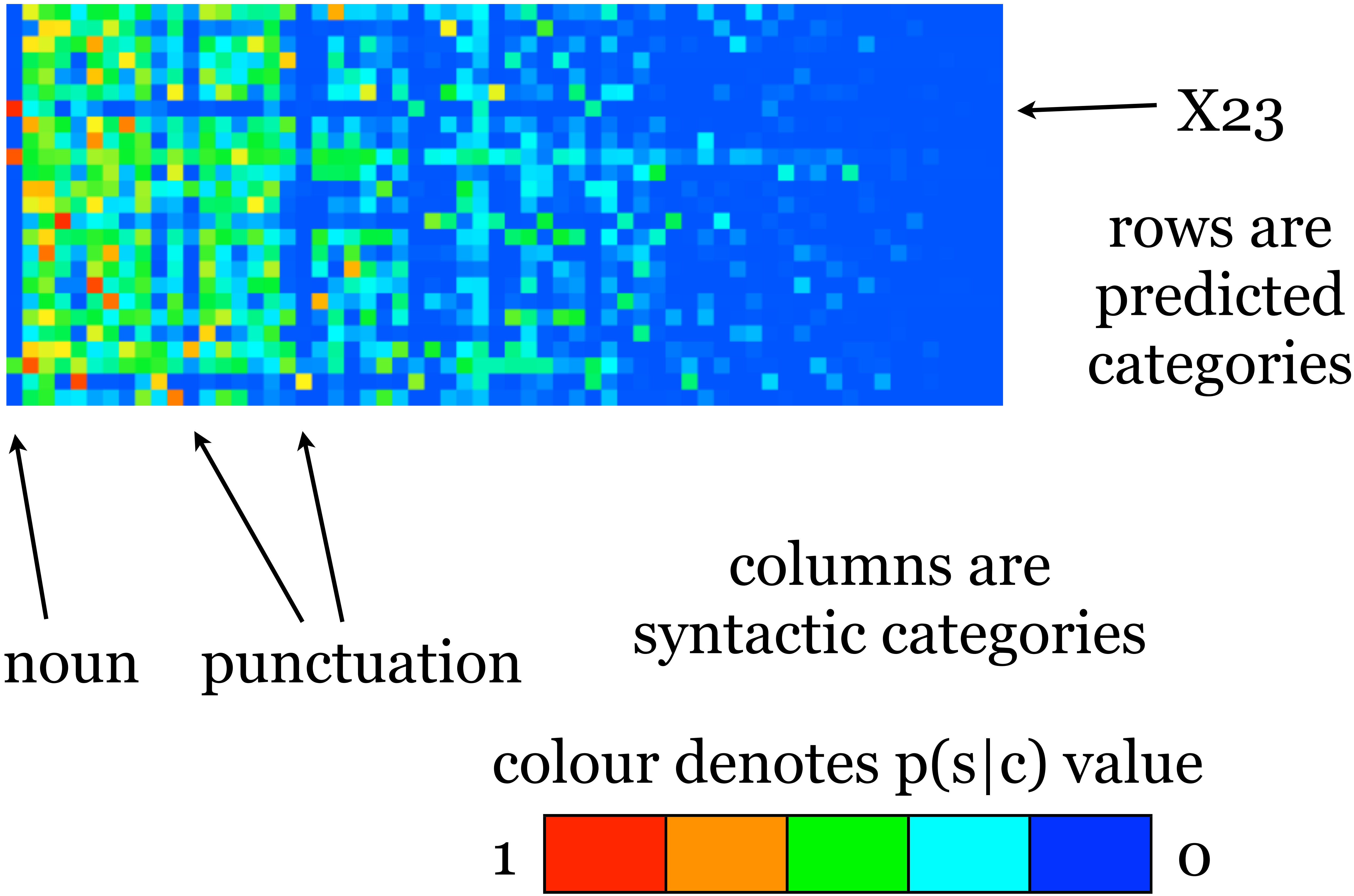


X4 - PRP
X4 - NP
X23 - VP/VP
X23- VP/NP
X5 - NP\DT

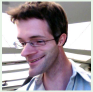
Conditional Entropy

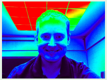
$$H(S|C) = \sum_{s,c} p(s,c) \log \frac{p(c)}{p(s,c)}$$

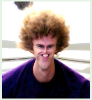
- quantifies the 'surprise' at seeing the syntactic category, *s*, given the predicted category, *c*
- p(s,c) and p(c) are simple frequency estimates

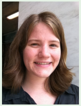

X23
rows are predicted categories
noun
punctuation
columns are syntactic categories
colour denotes p(s|c) value
1
0

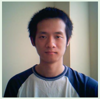
Outline


Trevor Cohn


Chris Dyer


Jan Botha


Olivia Buzek


Desai Chen

- 1:55pm Grammar induction and evaluation. Trevor
- 2:10pm Non-parametric models of category induction. Chris
- 2:25pm Inducing categories for morphology. Jan
- 2:35pm Smoothing, backoff and hierarchical grammars. Olivia
- 2:45pm Parametric models: posterior regularisation. Desai
- 3:00pm Break.

Nonparametric Clustering for Category Induction

Blunsom & Dyer

Clustering with Nonparametrics

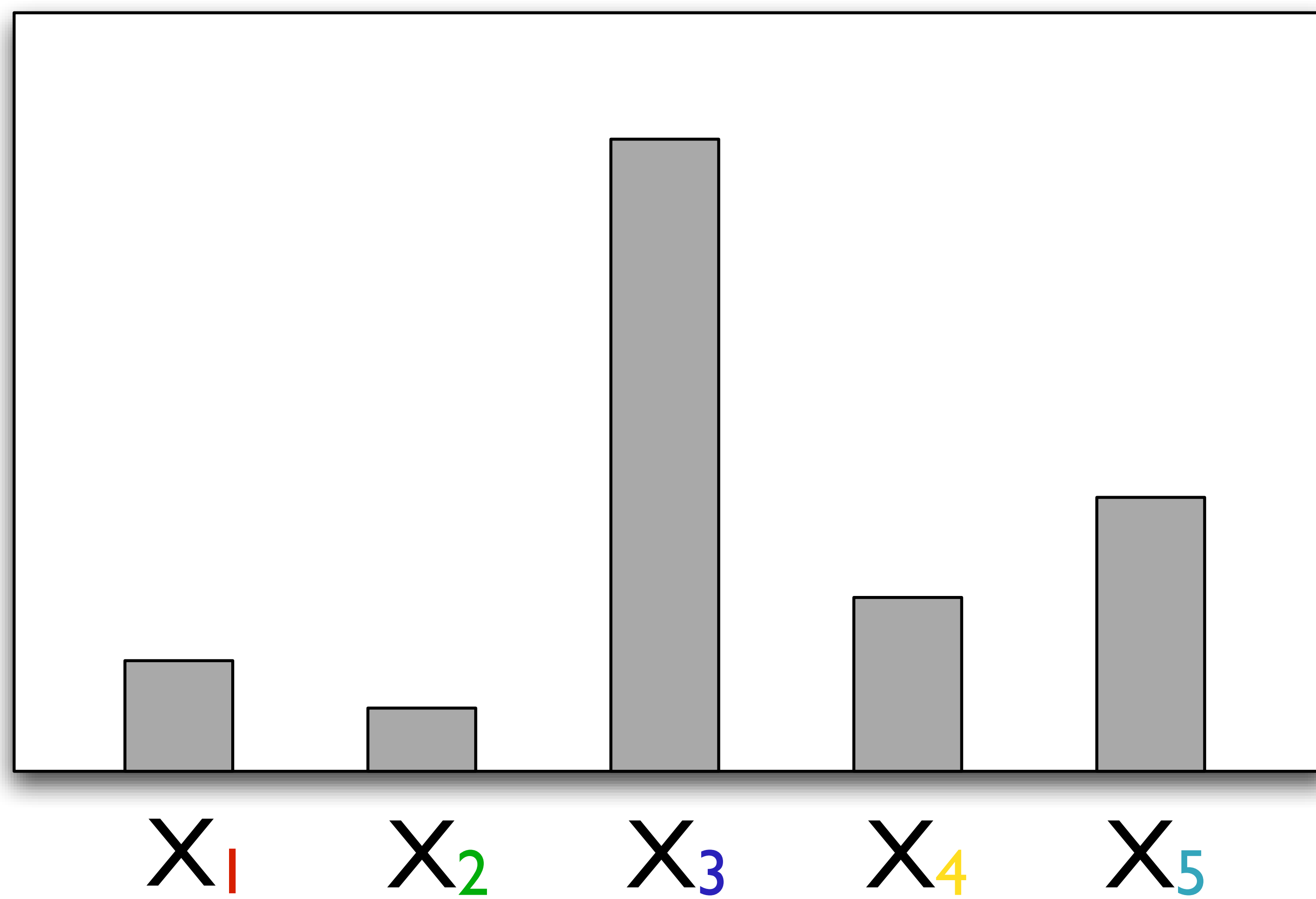
- Generalization of LDA model (Blei, 2001)
- Corpus consists of **phrases**, each of which occurs in one or more **contexts**
- Generative model
  - Each **phrase** is mixture of **categories**
  - **Categories** generate **contexts**

The Model I

Every **phrase** is characterized as a **mixture of categories** (X_1, X_2, X_3, ...):

phrase =

"Prime Minister"



The Model II

Each **category** generates **contexts** with some probability

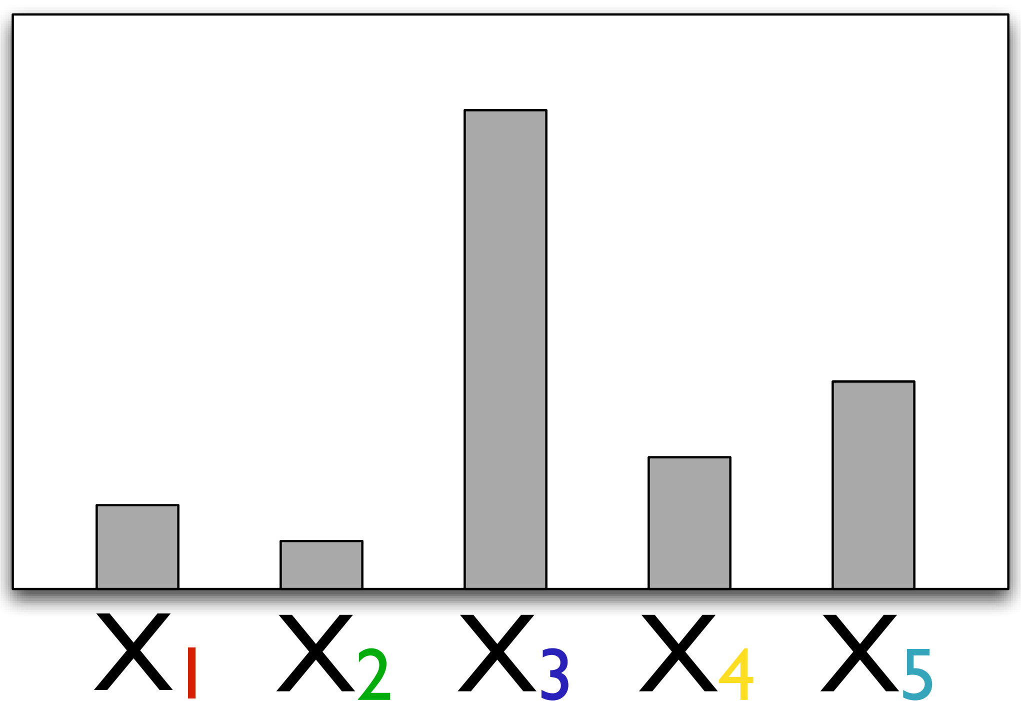
category = X_3

The Model III: Priors

- Use **priors** to impose beliefs about the solutions we would like to find
  - Each category should generate a **small number** of contexts
  - Each phrase should be a mixture of a **few categories**

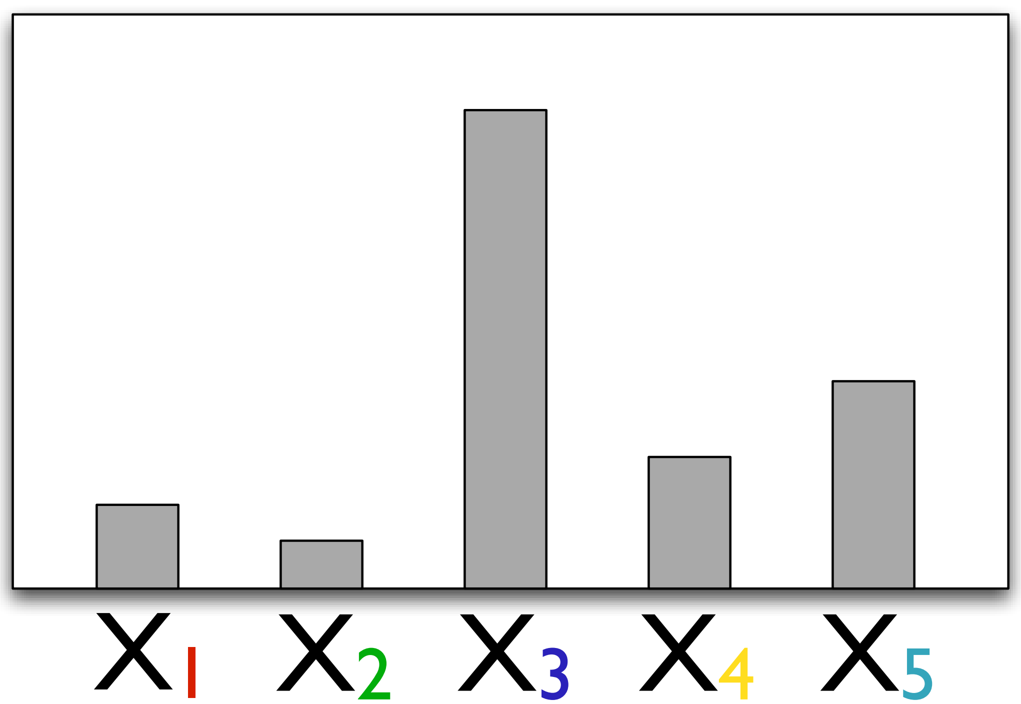
Our prior beliefs

Hypothesis 1

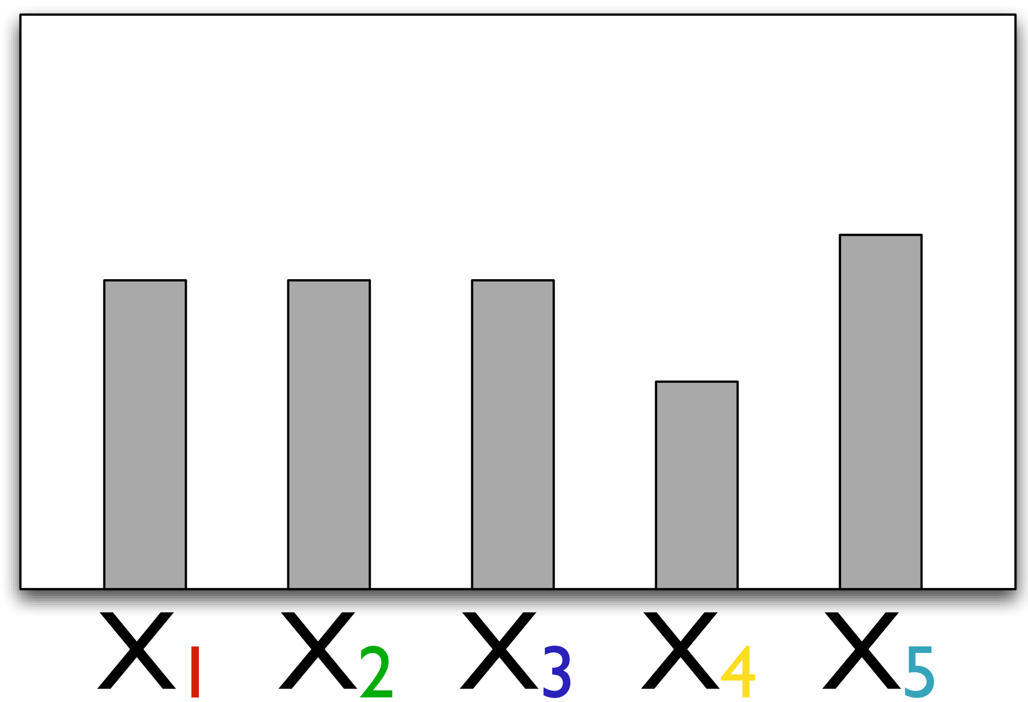


Our prior beliefs

Hypothesis 1

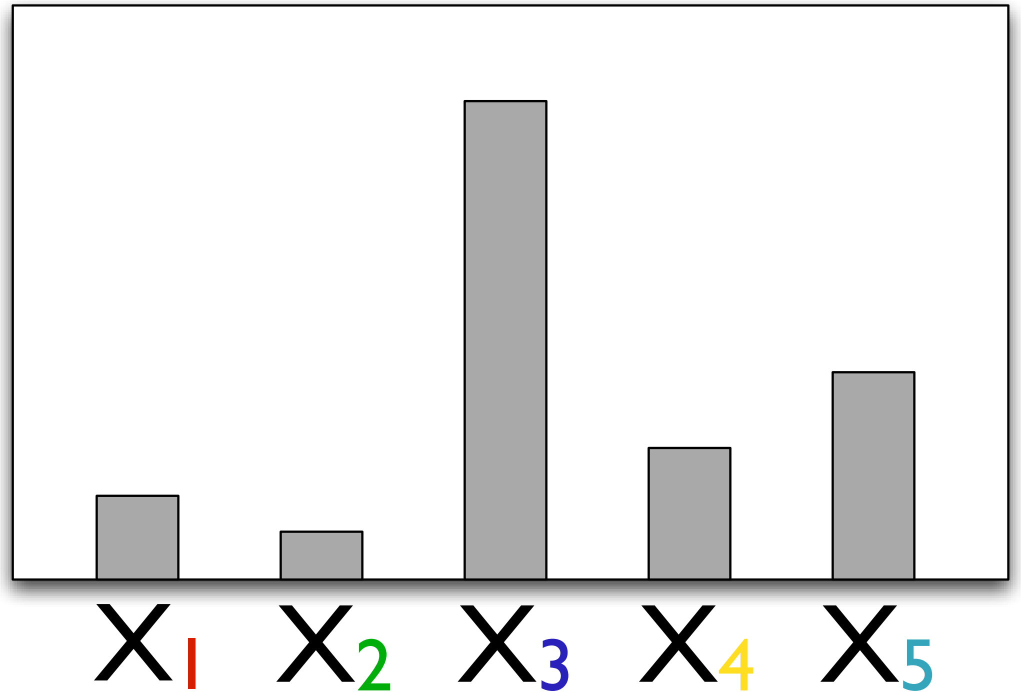


Hypothesis 2



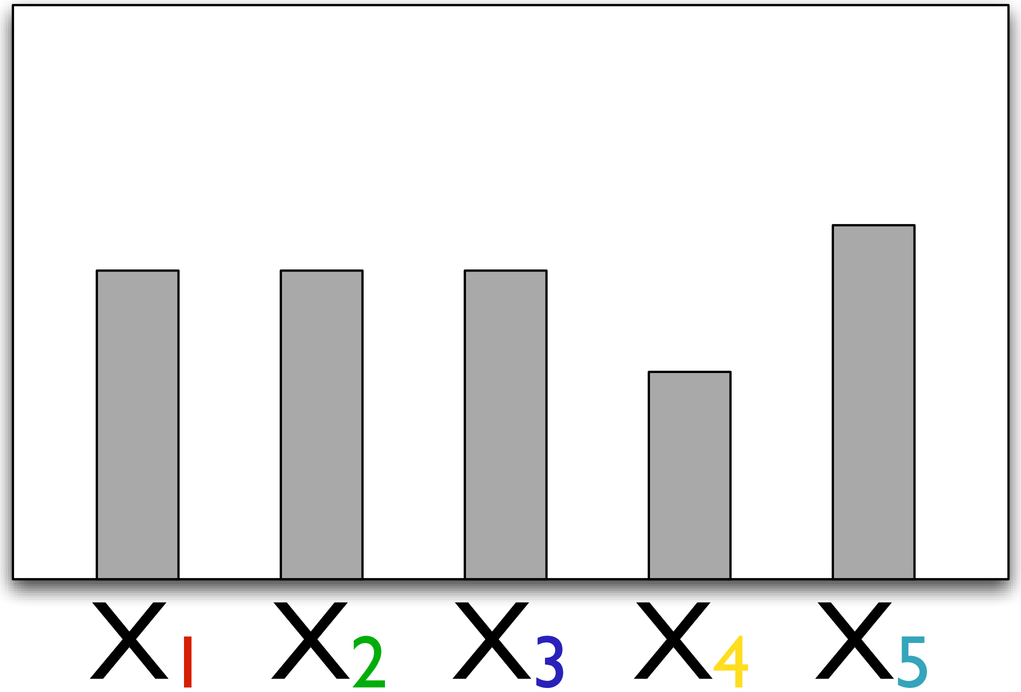
Our prior beliefs

Hypothesis 1



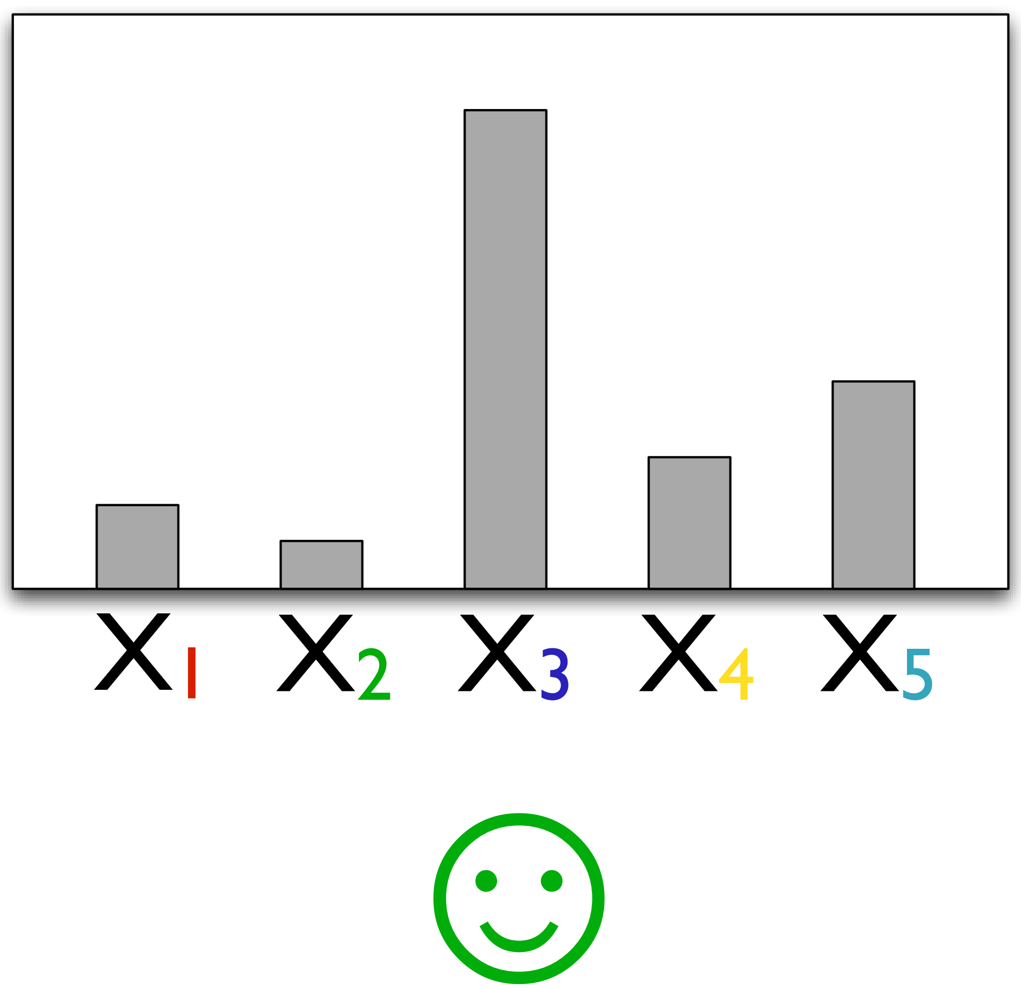
☺

Hypothesis 2

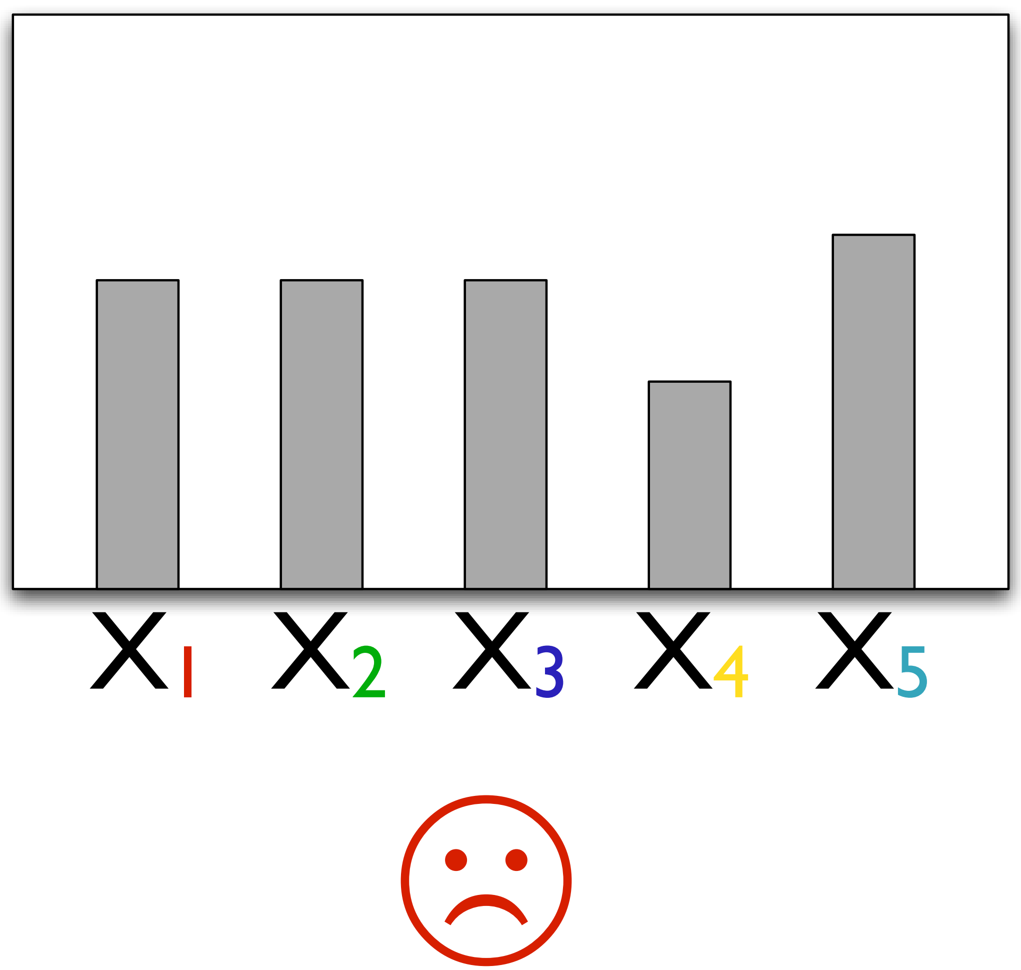


Our prior beliefs

Hypothesis 1



Hypothesis 2



Modeling contexts

$X_{16} \rightarrow$

that **the** __ **of** Great

Modeling contexts

$X_{16} \rightarrow$

that **the** __ **of** Great

$X_{16} \rightarrow$

is **the** __ **of** the

Modeling contexts

$X_{16} \rightarrow$ that **the** __ **of** Great

$X_{16} \rightarrow$ is **the** __ **of** the

the __ **of**

Modeling contexts

$X_{16} \rightarrow$ that **the** __ **of** Great

$X_{16} \rightarrow$ is **the** __ **of** the

the __ **of**

How we do it...

$\mathbf{c}$	$\sim H_{X_n}$	$\mathbf{c} = c_l\ c_0\ c_r$
$H_{X_n} \mid a_1, b_1$	$\sim PYP(a_1, b_1, G_{X_n}(c_0) \times U)$	$U = (1/V)^2\ ,\ \forall\ c_0$
$G_{X_n} \mid a_0, b_0, P_0$	$\sim PYP(a_0, b_0, P_0 = U)$	

The Chinese Restaurant Process

- We use Pitman-Yor Processes to
  - enforce sparsity in the distribution over contexts for each **category**
  - enforce sparsity in the distribution over categories for each **phrase**
- Values of **hyperparameters** (concentration, discount) have priors as

Remarks

- Caveats
  - Prior beliefs are about **parameters** (i.e., not posterior distributions)
  - No global consistency constraints on grammars
    - Independence assumptions (i.e., "bag of contexts") enable fast inference.

Inference

- Given the **data** (phrases and their contexts)
- And given the **priors** infer what **categories** generated what **contexts**

Inference

- We use **collapsed Gibbs sampling**
  - We don't explicitly represent category-context parameters or category mixture proportions
  - **Only represent assignments of contexts to categories**!
  - Sample for *n* iterations
    - Reason about assignments in last sample
    - Reason about MAP category (given context) in last sample

Inference

Prime minister

the_of

<s>_Blair

the_of

a_is

British_David

traveled

representatives_to

has_to

has_to

has_long

reported

has_that

has_that

the_problem

Inference

Prime minister

the_of

<s>_Blair

the_of

a_is

British_David

traveled

representatives_to

has_to

has_to

has_long

reported

has_that

has_that

the_problem

Inference

Prime minister	**traveled**	**reported**
the_of	representatives_to	has_that
<s>_Blair	has_to	has_that
the_of	has_to	the_problem
a_is	has_long	
British_David		

Inference

Prime minister

the_of

<s>_Blair

the_of

a_is

British_David

traveled

representatives_to

has_to

has_to

has_long

reported

has_that

has_that

the_problem

Inference

Prime minister

- the_of
- <s>_Blair
- the_of
- a_is
- British_David

traveled

- representatives_to
- has_to
- has_to
- has_long

reported

- has_that
- has_that
- the_problem

Inference

Prime minister

the_of

<s>_Blair

the_of

a_is

British_David

traveled

representatives_to

has_to

has_to

has_long

reported

has_that

has_that

the_problem

Inference

Prime minister

the_of
<s>_Blair
the_of
a_is
British_David

traveled

representatives_to
has_to
has_to
has_long

reported

has_that
has_that
the_problem

Inference

Prime minister

the_of
<s>_Blair
the_of
a_is
British_David

traveled

representatives_to
has_to
has_to
has_long

reported

has_that
has_that
the_problem

Inference

Prime minister

the_of
<s>_Blair
the_of
a_is
British_David

traveled

representatives_to
has_to
has_to
has_long

reported

has_that
has_that
the_problem

Inference

Prime minister

- the_of
- <s>_Blair
- the_of
- a_is
- British_David

traveled

- representatives_to
- has_to
- has_to
- has_long

reported

- has_that
- has_that
- the_problem

Inference

Prime minister

- the_of
- <s>_Blair
- the_of
- a_is
- British_David

traveled

- representatives_to
- has_to
- has_to
- has_long

reported

- has_that
- has_that
- the_problem

Inference

Prime minister

the_of

<s>_Blair

the_of

a_is

British_David

traveled

representatives_to

has_to

has_to

has_long

reported

has_that

has_that

the_problem

Inference

Prime minister	**traveled**	**reported**
the_of	representatives_to	has_that
<s>_Blair	has_to	has_that
the_of	has_to	the_problem
a_is	has_long	
British_David		

Do this many 1000s of times, and it will converge!

Experiments

- Questions
  - **What should we cluster?**
    - Source or target?
    - Words, word clusters, POS tags?
    - Proper context size?
  - **How many classes?**

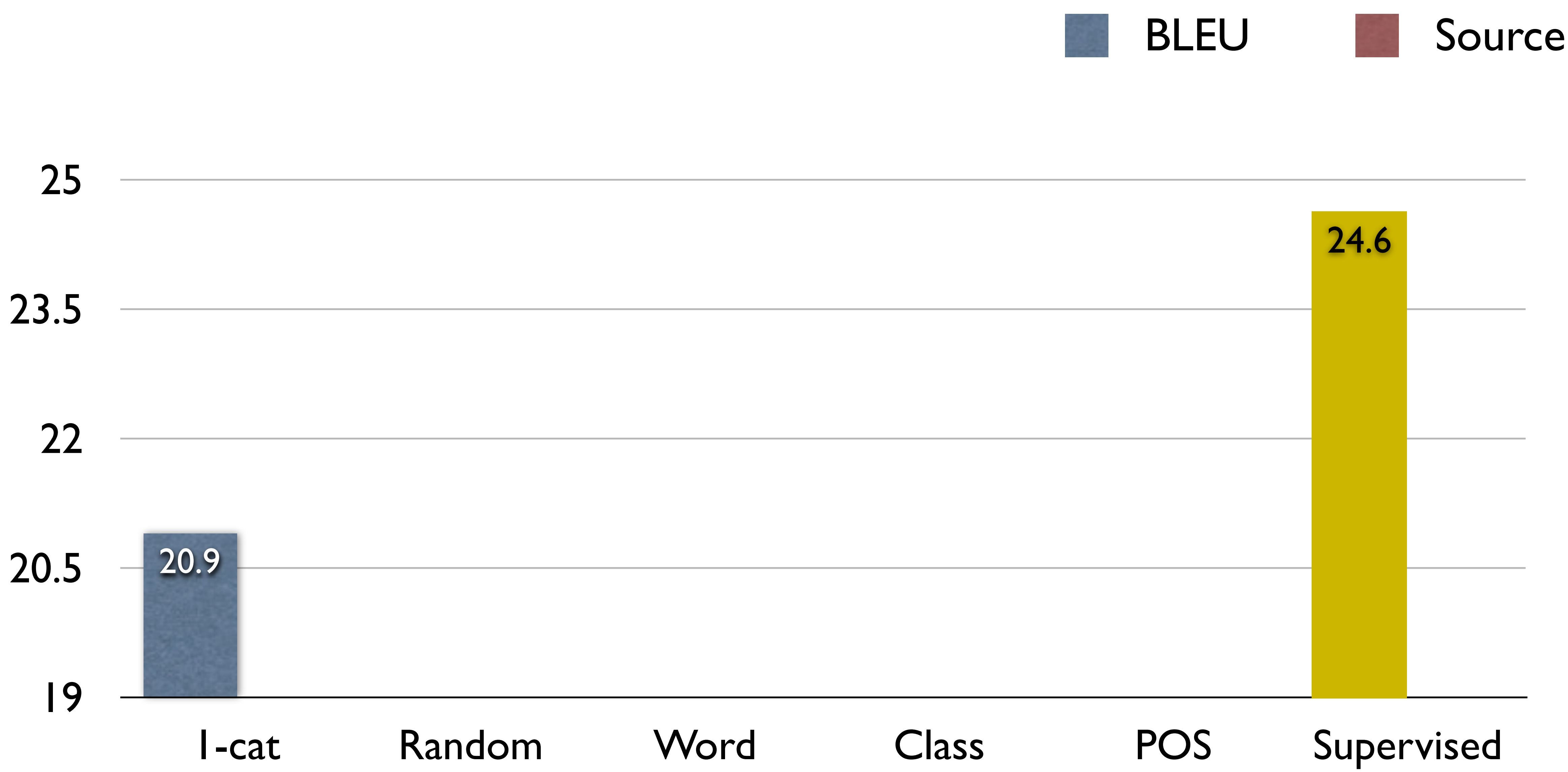
Evaluation

- Extrinsic evaluation
  - **BLEU score** (translation quality)
- Intrinsic evaluation
  - **conditional entropy** with respect to supervised baseline
- How well does the intrinsic metric correlate with extrinsic performance?

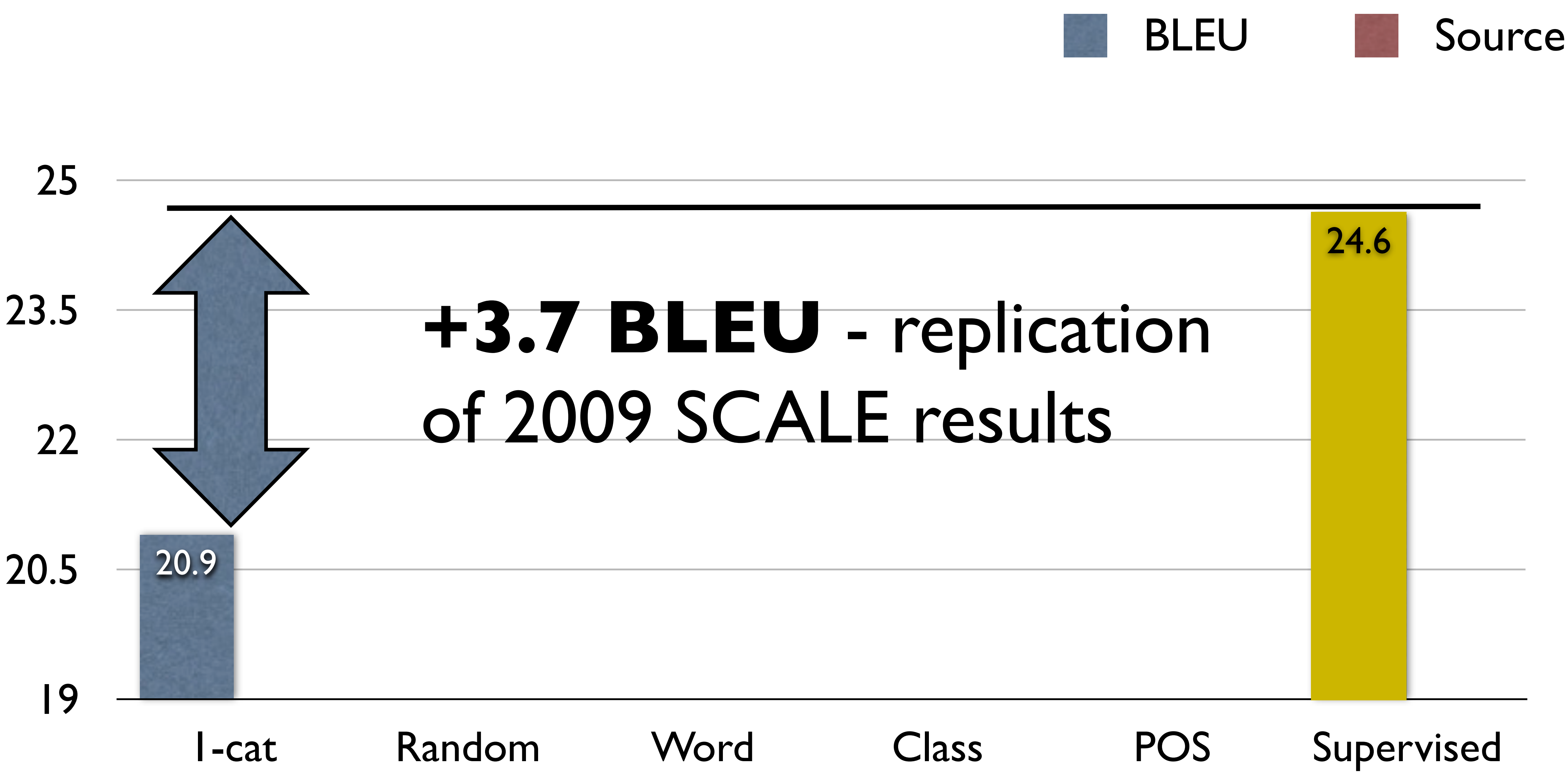
Predictions

- **Target language** clustering will be better for translation than source language
- **Larger contexts** (with sensible backoff) will improve clustering / translation

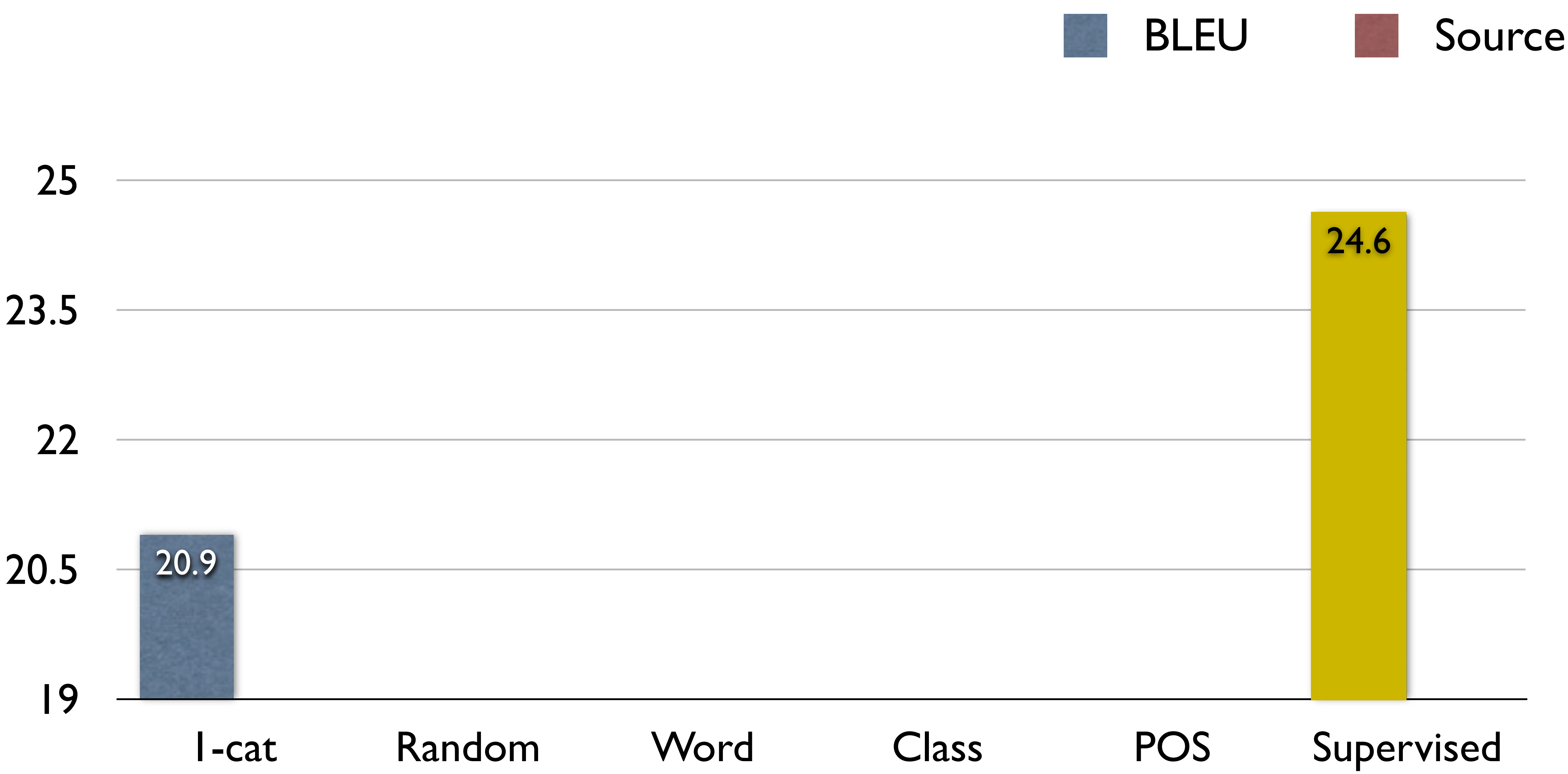
Urdu-English



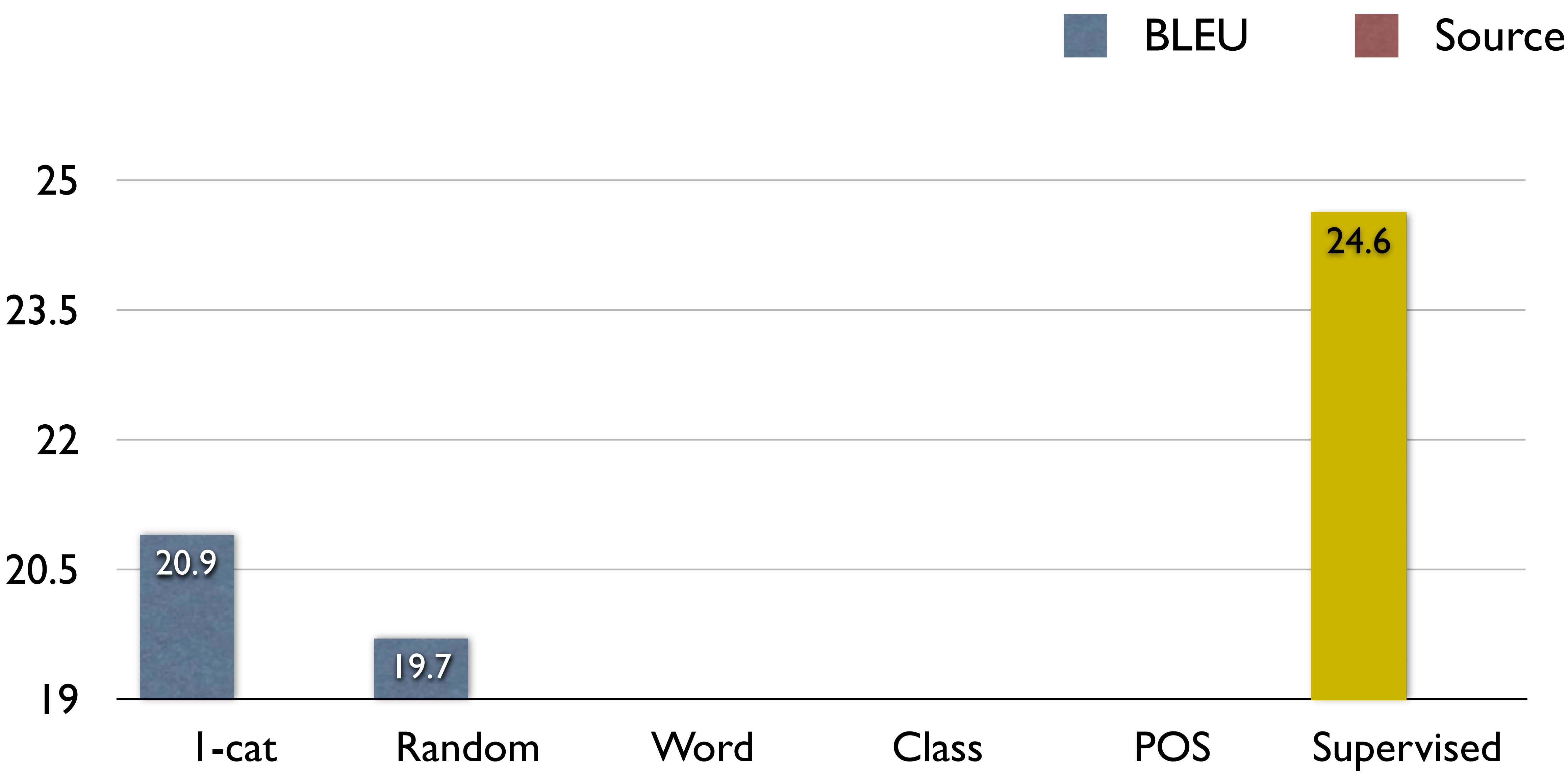
Urdu-English



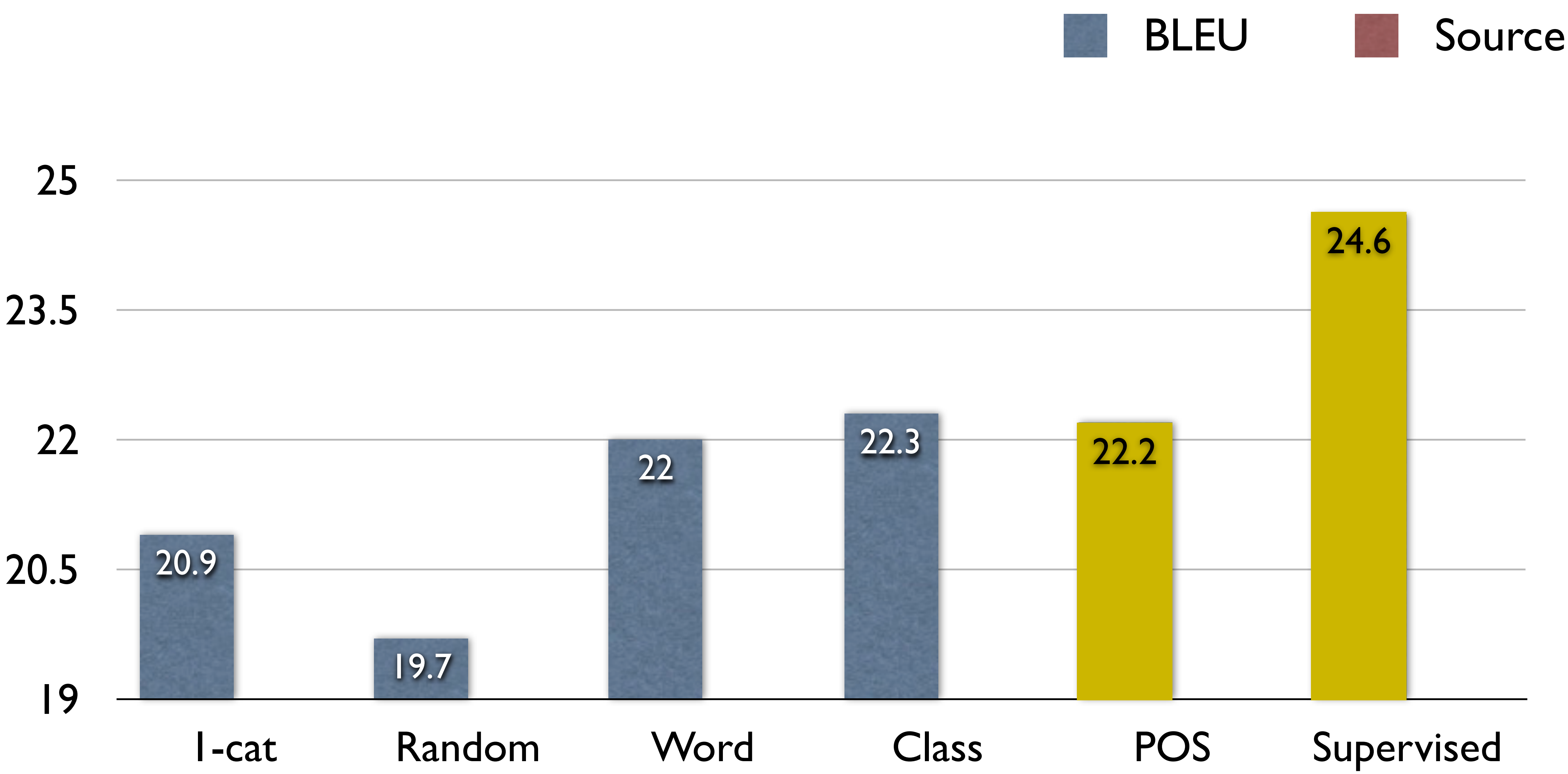
Urdu-English



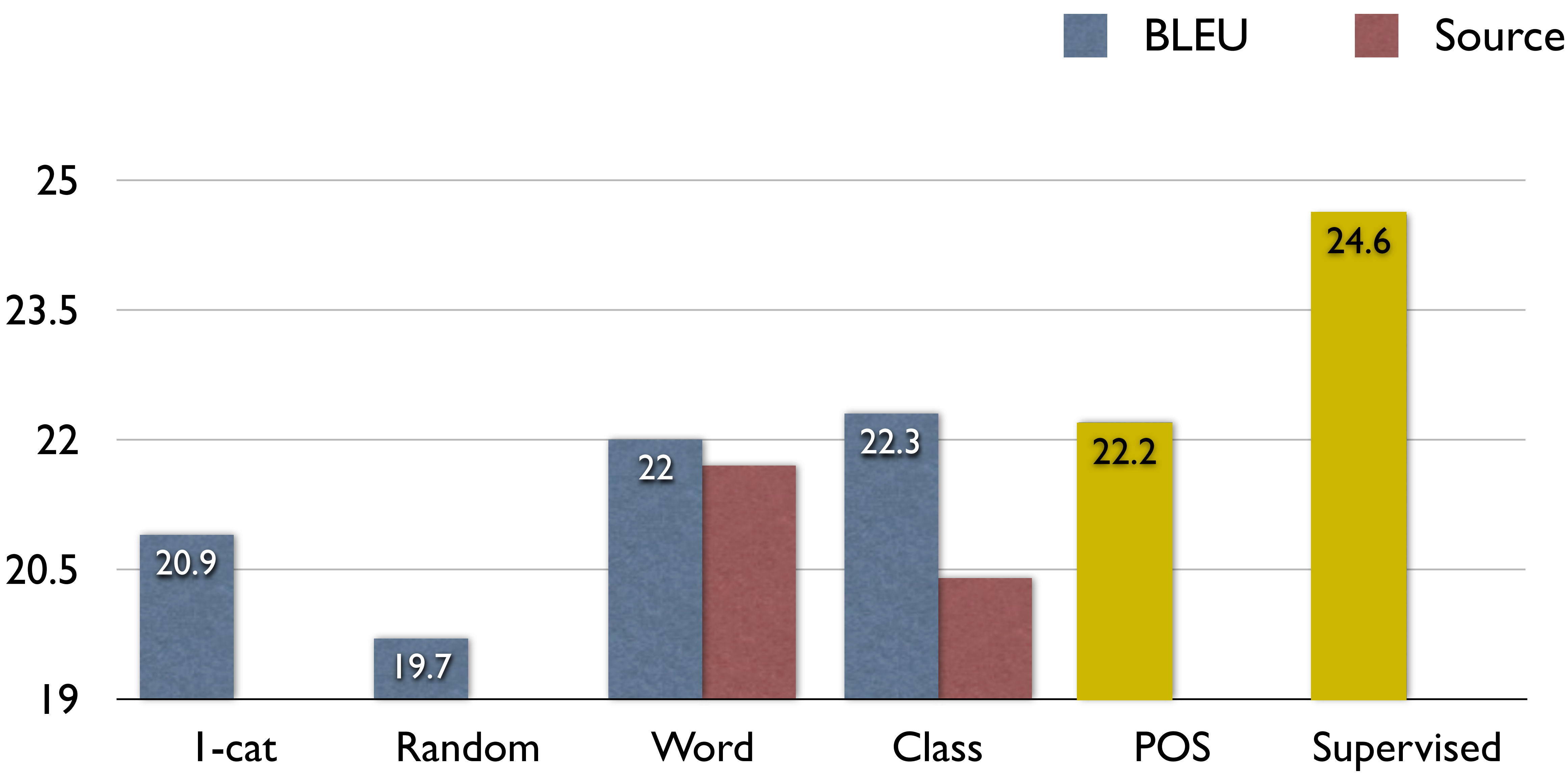
Urdu-English



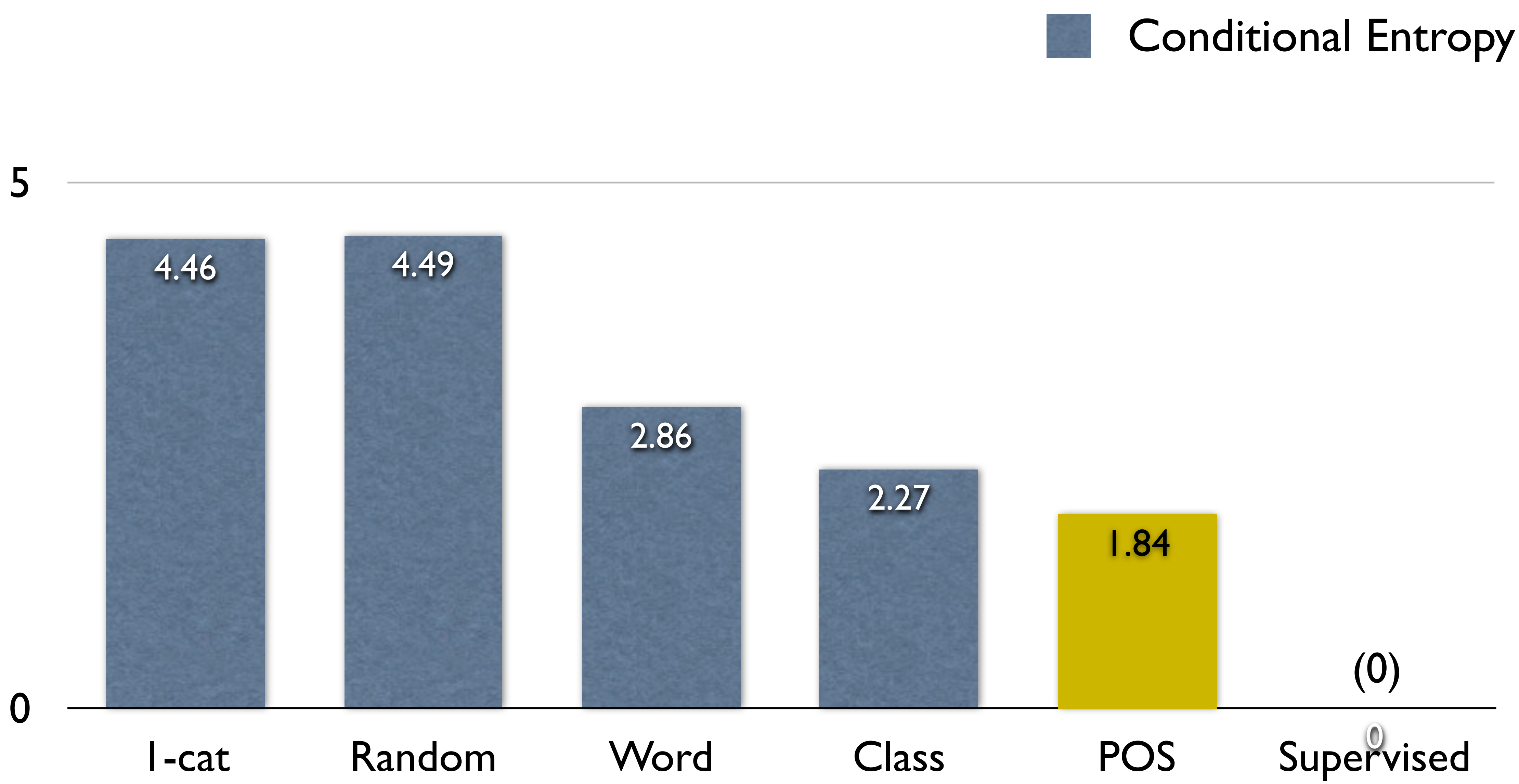
Urdu-English

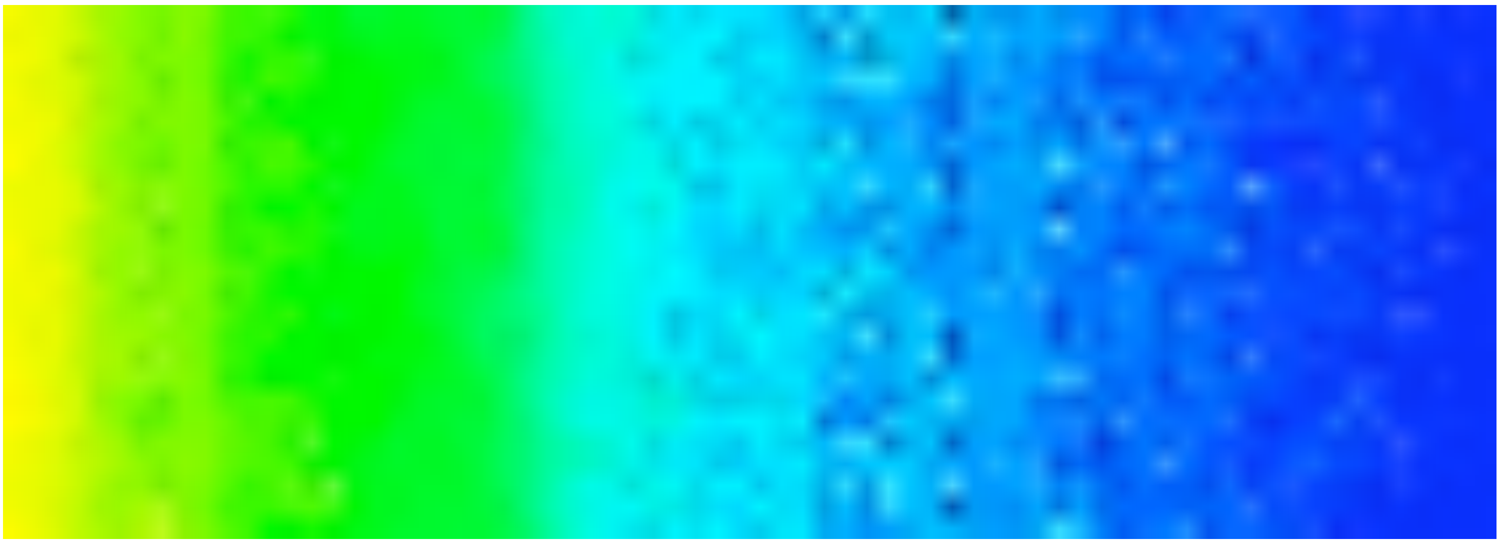


Urdu-English

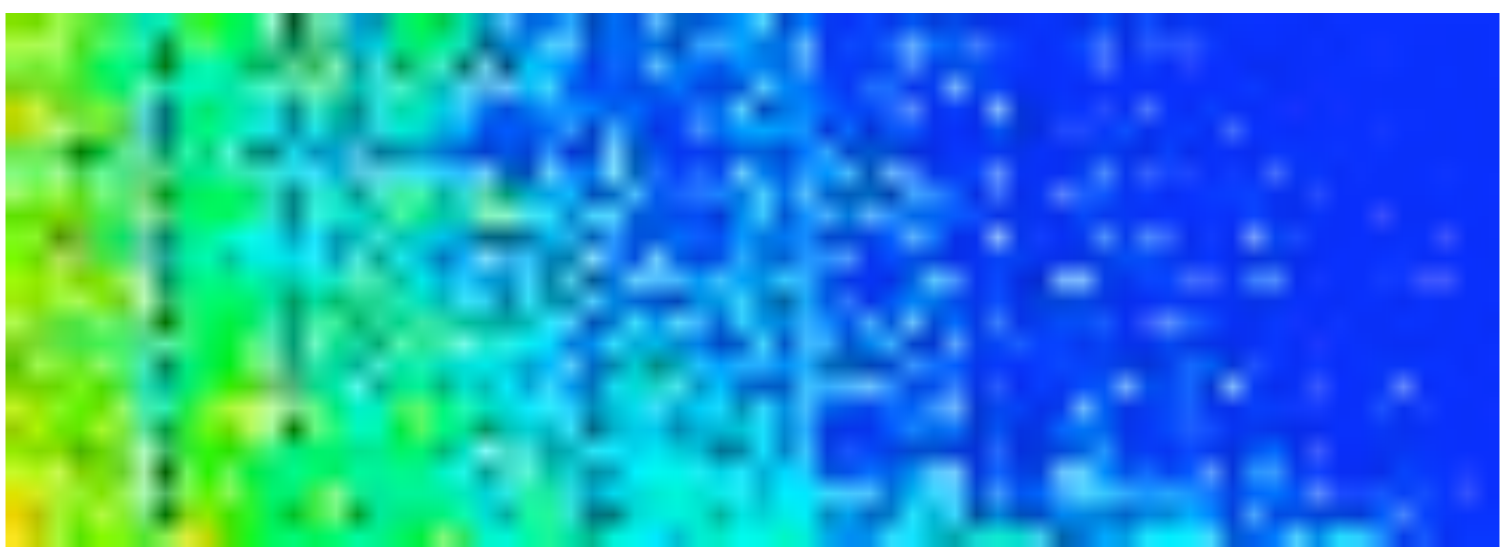


Intrinsic evaluation

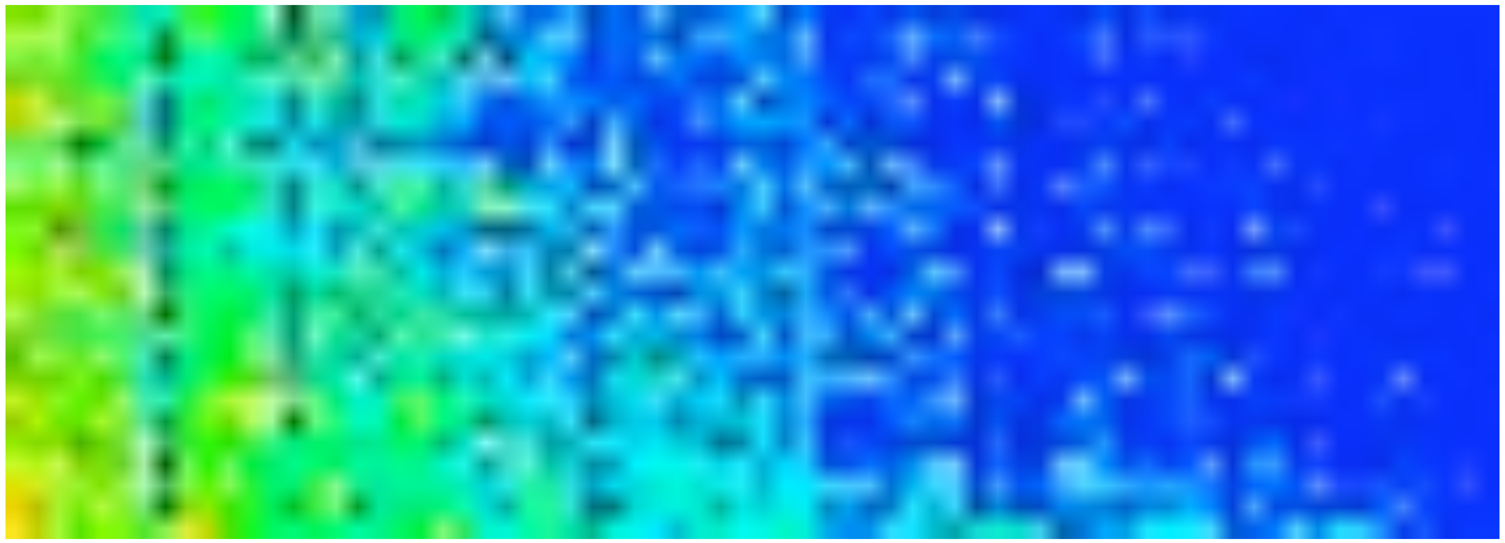




Random word
(Entropy=4.49)

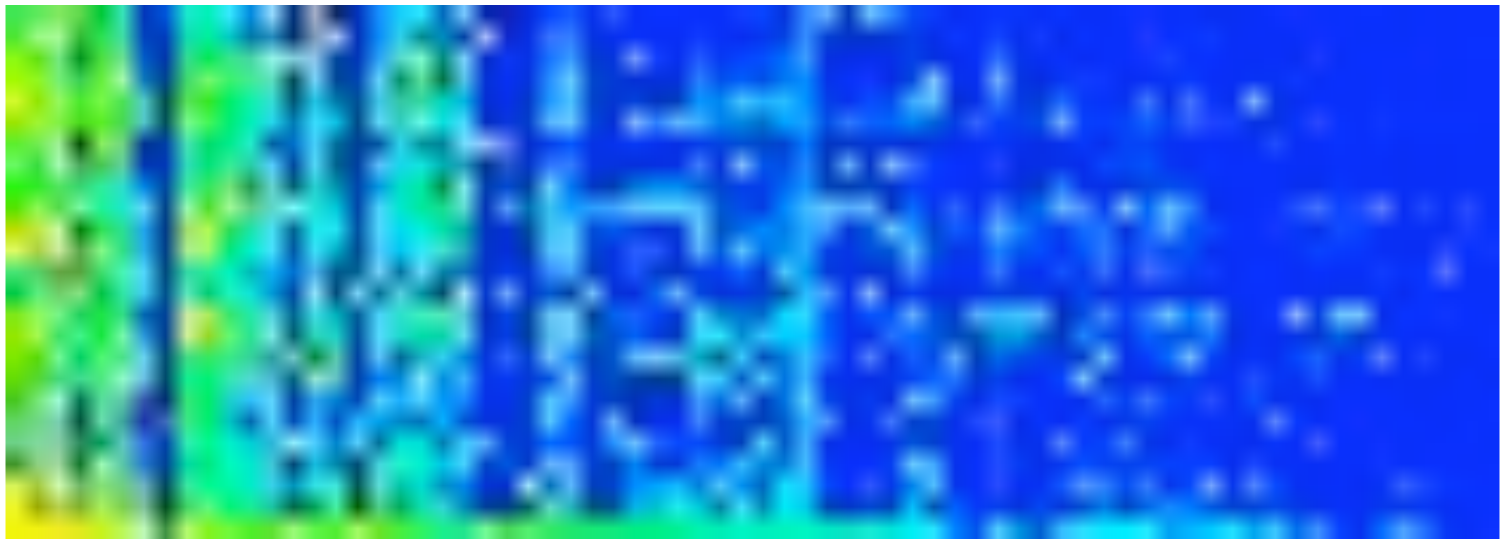


Source word
(Entropy=3.25)



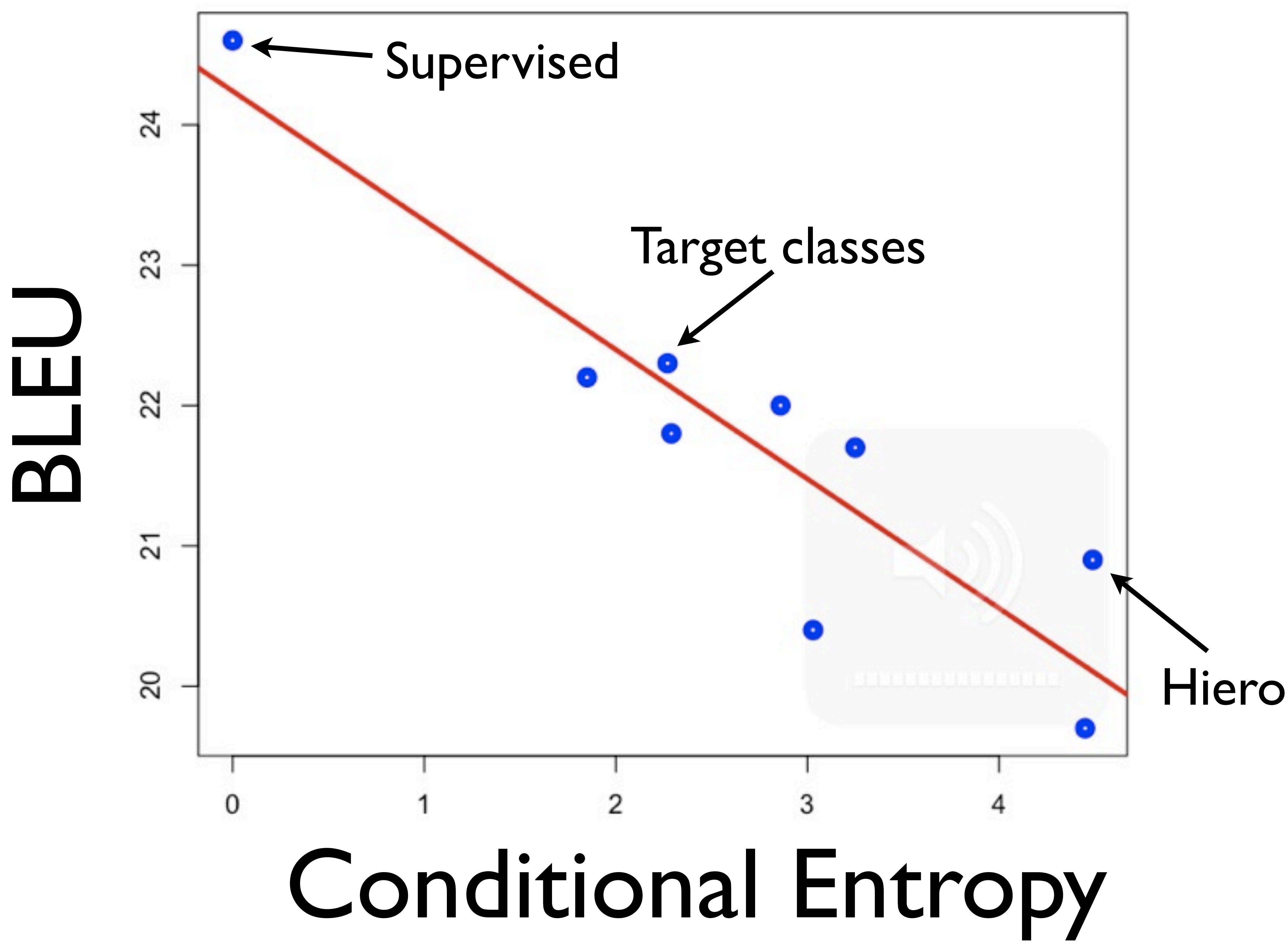
Source word
(Entropy=3.25)

Target word
(Entropy=2.86)

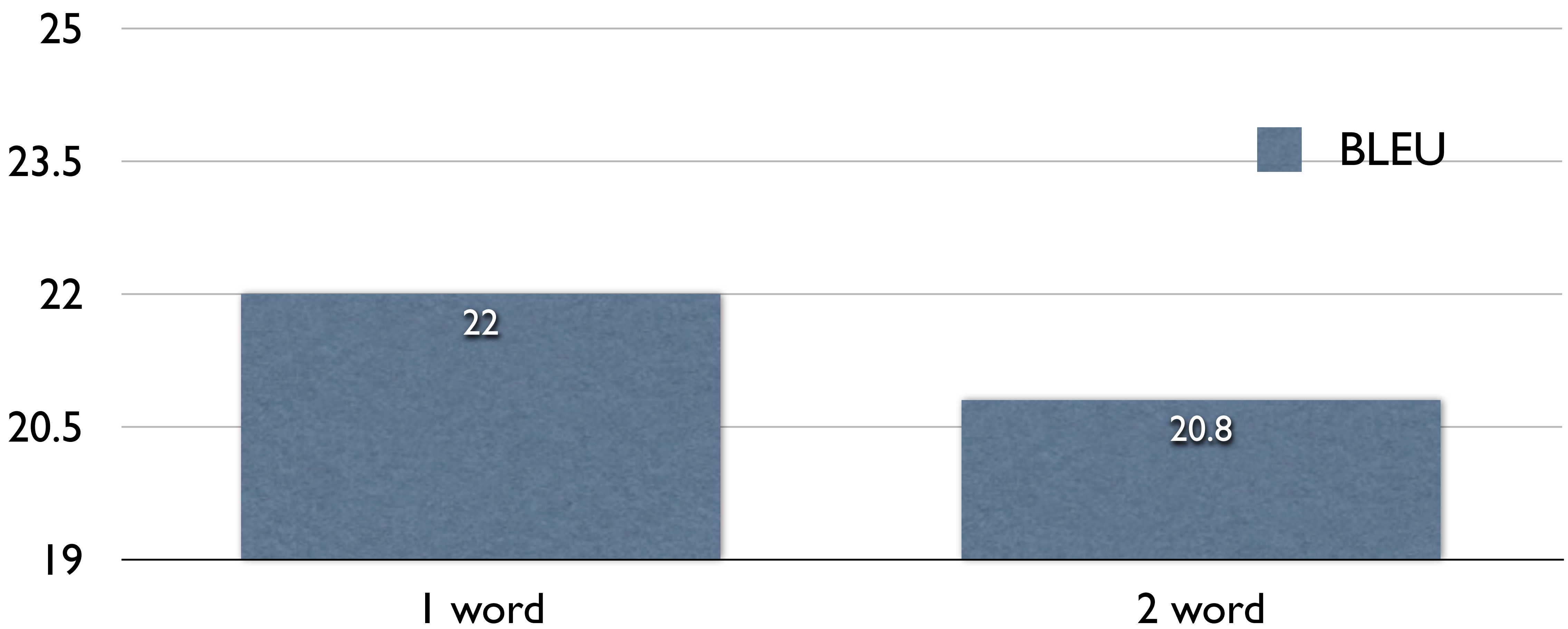


Target word
(Entropy=2.85)

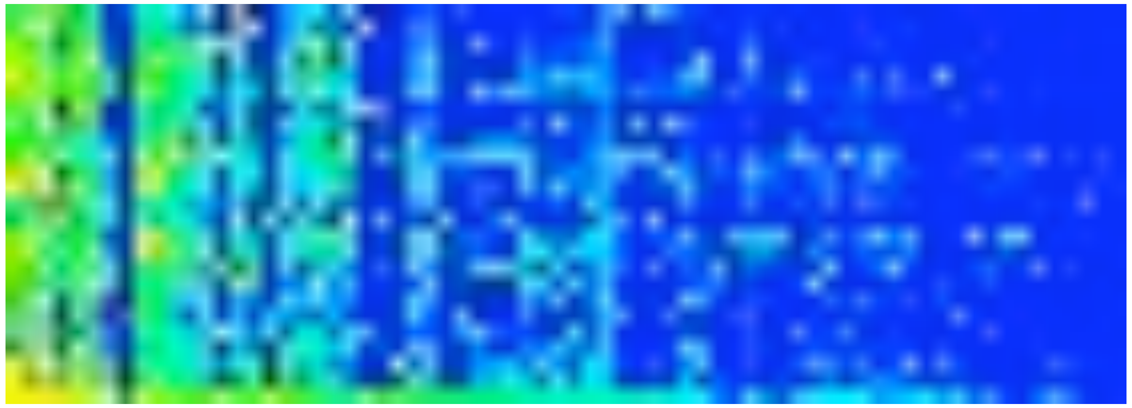
Target POS
(Entropy=1.85)


Supervised
Target classes
Hiero
BLEU
24
23
22
21
20
0
1
2
3
4
Conditional Entropy

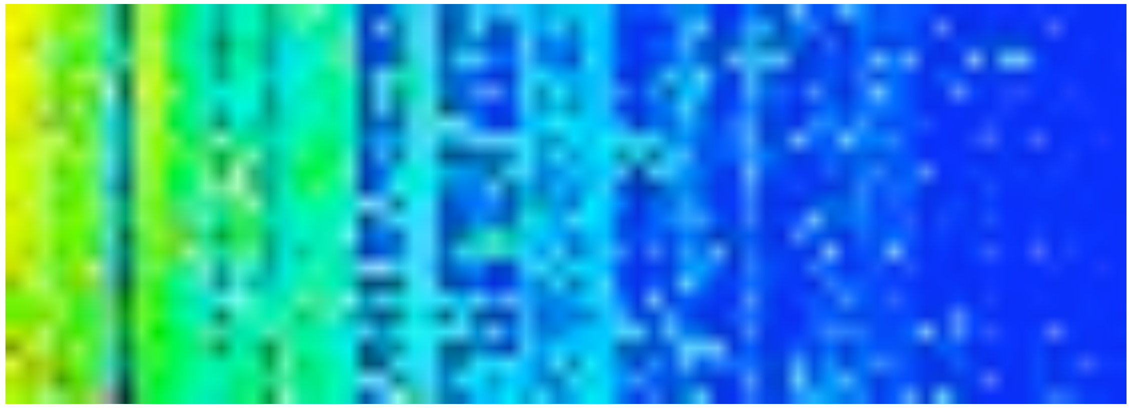
Context size?



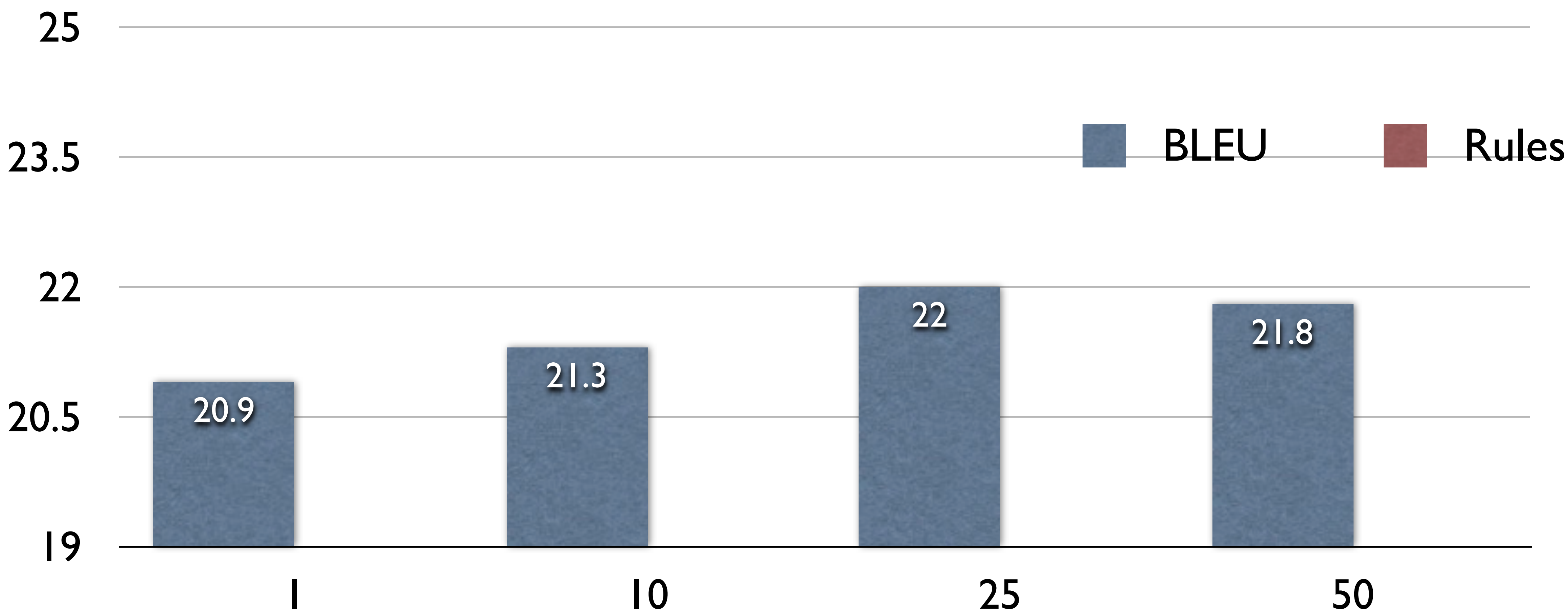
Entropy=2.86



Entropy=3.16



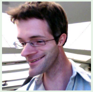
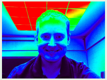
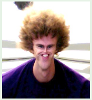
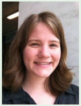
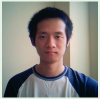
How many categories?



Summary

- Unsupervised syntax, induced using Pitman-Yor clustering from contextual information improves translation
- "Bag of contexts" assumption not unreasonable
- Context back-off (using hierarchical PYPs) needs more investigation

Outline


Trevor Cohn

Chris Dyer

Jan Botha

Olivia Buzek

Desai Chen

- 1:55pm Grammar induction and evaluation. Trevor
- 2:10pm Non-parametric models of category induction. Chris
- 2:25pm Inducing categories for morphology. Jan
- 2:35pm Smoothing, backoff and hierarchical grammars. Olivia
- 2:45pm Parametric models: posterior regularisation. Desai
- 3:00pm Break.

Inducing structured morphology

Can labelled SCFGs be used to model word formation in MT?

Jan Botha

Outline

- Why bother with morphology in MT?
- The case for using a labelled grammar
- Results:
  - Categories learnt
  - Effect on translation
- What goes wrong

Morphology + MT

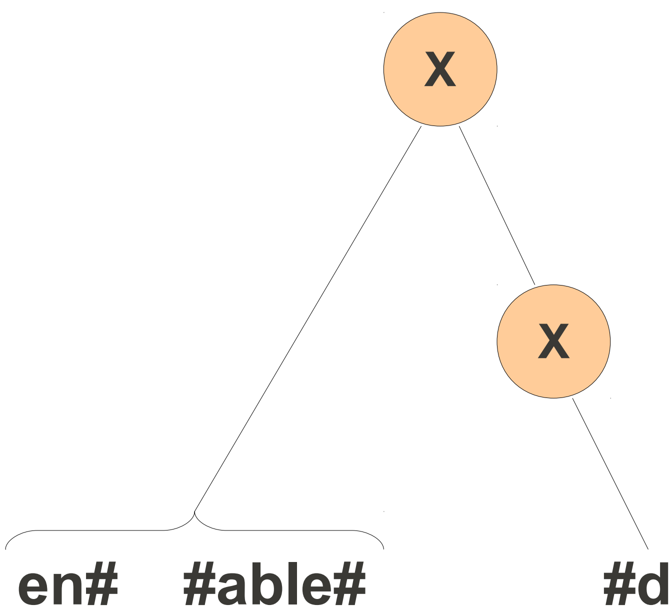
- Sparse data: will never observe all inflections
  - Observed:
    - j'entend**s** I hear
    - nous répond**ons** we reply
  - Not observed:
    - nous entend**ons** we hear
- Want to generate unobserved form using the observed *morphemes*
- Need rules for how morphemes combine
  - Induce rules instead of hand-crafting them

en# #able# #d

$\mathbf{X} \rightarrow$ **en# #able# X**

en# #able# #d

$X \rightarrow$ en# #able# X



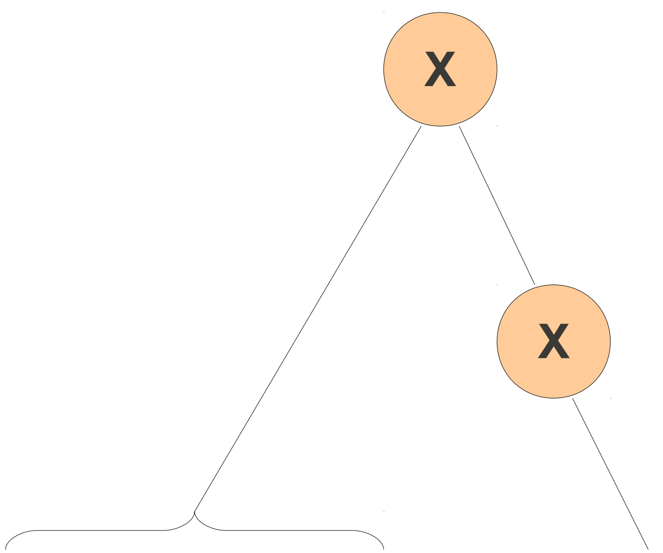
$X \rightarrow$ #d

#s — $X \rightarrow$ #s

#er — $X \rightarrow$ #er

Hiero's X = village bicycle

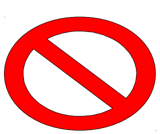
en# #able# #d **X → en# #able# X**



en# **#able#** **#d** **X → #d**

#ity to X → #ity to

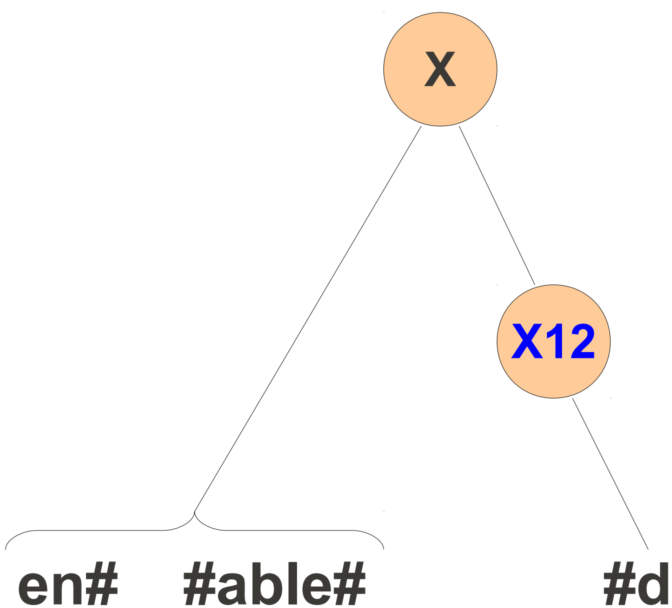
elephant cow X → elephant cow



Constraining with Labelled Categories

en# #able# #d

X → en# #able# X12



Chris just showed us a source for such labels

~~#ity to~~

~~elephant cow~~

X12 → #d

X → #ity to

X → elephant cow

Better

Overview of Strategy

- Segment text into morphemes
- Learn categories over morphemes/words using the grammar induction model
- Label spans in the training data
- Extract a SCFG as usual, but now
  - **labelled NTs** only deal with **word formation**
  - everything else is handled by the generic X NT
- Translate and hope BLEU goes up

What gets labelled

- les **modifi#** #cation# #s n' ont pas lieu d' être .

  X5

- les **justifi#** #cation# #s ...

  X5

Inside some Dutch Categories

1) 85% noun stems — mostly with plural endings
- **resolutie#** #s | **kilo# #meter#** #s

4) 99% verb stems — taking various prefixes
- ge# **#maakt** | ver# **#werpt** | samen# **#brengt**

6) 99% adjective stems — taking suffix #e
- **interessant#** #e | **etisch#** #e

10) 65% full words — mostly compound nouns
- **eind# #resultaat** | **drie# #jaren# #plan**

0) 75% concerns the joining infix #s#

* de **europe#** #s# | europe# **#s#** #e

Translation Results

		BLEU
	Without segmentation	15.75
baseline	**Unlabelled**	**15.60**
this attempt	**Labelled (source)**	**15.43**
	Labelled (target)	**15.34**

A previously unseen inflection was generated <u>correctly</u>:

Input:	het ivoriaanse model	*the (pertaining to Ivory Coast) model*
Reference:	du modèle **ivoirien**	
Baseline:	du modèle **ivoirïenne**	- adjective has wrong gender
Labelled (src):	du **ivoirien** modèle	- correct gender, wrong word order

Aligning Morphemes

Before

(Nothing about that may be changed.)

- *Dutch:* daaraan mag niets veranderd worden .
- *French:* les modifications n' ont pas lieu d' être .

After

- daar# #aan mag niet# #s veranderd word# #en .
- les modifi# #cation# #s n' ont pas lieu d' être .

Aligning Morphemes

Before

(Nothing about that may be changed.)

- *Dutch:* daaraan mag niets veranderd worden .
- *French:* les modifications n' ont pas lieu d' être .

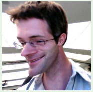
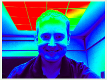
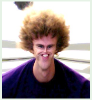
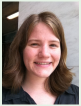
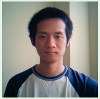
After

- daar# #aan mag niet# #s veranderd word# #en .
- les modifi# #cation# #s n' ont pas lieu d' être .

Summary

- A way of thinking about morphology
- Basic idea seems worthwhile
  - strong patterns in induced categories
- Further work
  - address problem of morpheme alignment

Outline


Trevor Cohn

Chris Dyer

Jan Botha

Olivia Buzek

Desai Chen

- 1:55pm Grammar induction and evaluation. Trevor
- 2:10pm Non-parametric models of category induction. Chris
- 2:25pm Inducing categories for morphology. Jan
- 2:35pm Smoothing, backoff and hierarchical grammars. Olivia
- 2:45pm Parametric models: posterior regularisation. Desai
- 3:00pm Break.

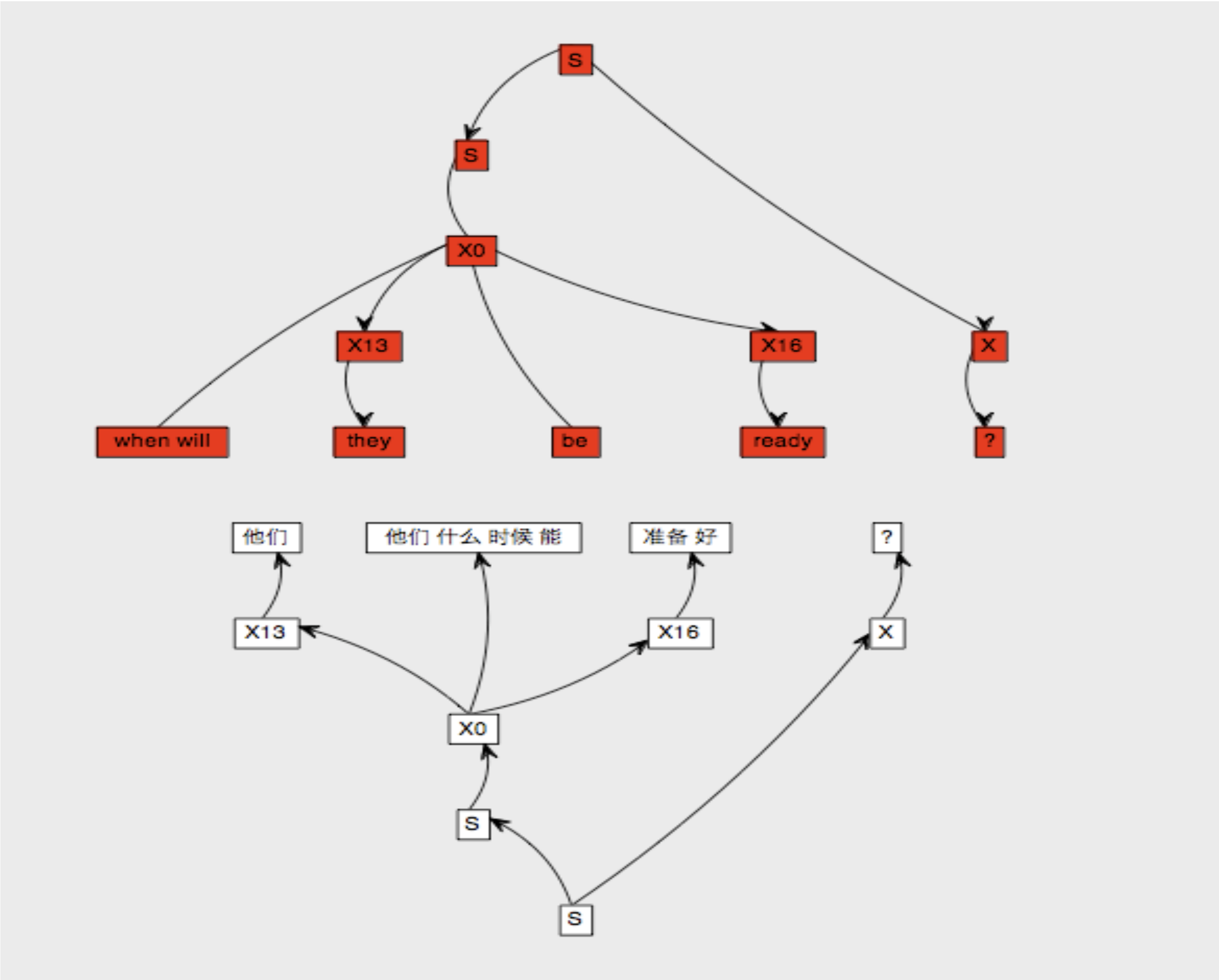
SMOOTHING WITH BACKOFF GRAMMARS

Olivia Buzek

Hierarchical Translation Overview

- Induce a synchronous CFG which simultaneously parses a sentence in both the source and target languages

Hierarchical Translation Overview



Hierarchical Translation Overview

- Induce a synchronous CFG which simultaneously parses a sentence in both the source and target languages
- Can result in problems where rules for certain constructions are absent
  - Natural language data is inherently sparse

Motivation

- Translations affected by data sparsity
  - Rules are too specific
- Backing off to more general categories allows handling of constructions not in the training data

Naïve Backoff Grammar

Rather than specific rules…

$$[X0] \rightarrow < \alpha \; [X1] \; \beta, \;\; \gamma \; [X1] \; \delta >$$

…we should be able to optionally move to any category, with a penalty:

$$[X0] \rightarrow < \alpha \; [X1_{\text{backoff}}] \; \beta, \;\; \gamma \; [X1_{\text{backoff}}] \; \delta >$$

$$[X1_{\text{backoff}}] \rightarrow \text{any category}$$

Naïve Backoff Grammar

- Based on 25-cat PYP-induced grammar

$$X_{\boxed{25}}$$

- Plus backoff rules

$$[X0] \rightarrow < \alpha \; [X1_{\text{backoff}}] \; \beta, \; \gamma \; [X1_{\text{backoff}}] \; \delta >$$
$$[X1_{\text{backoff}}] \rightarrow \text{any category}$$

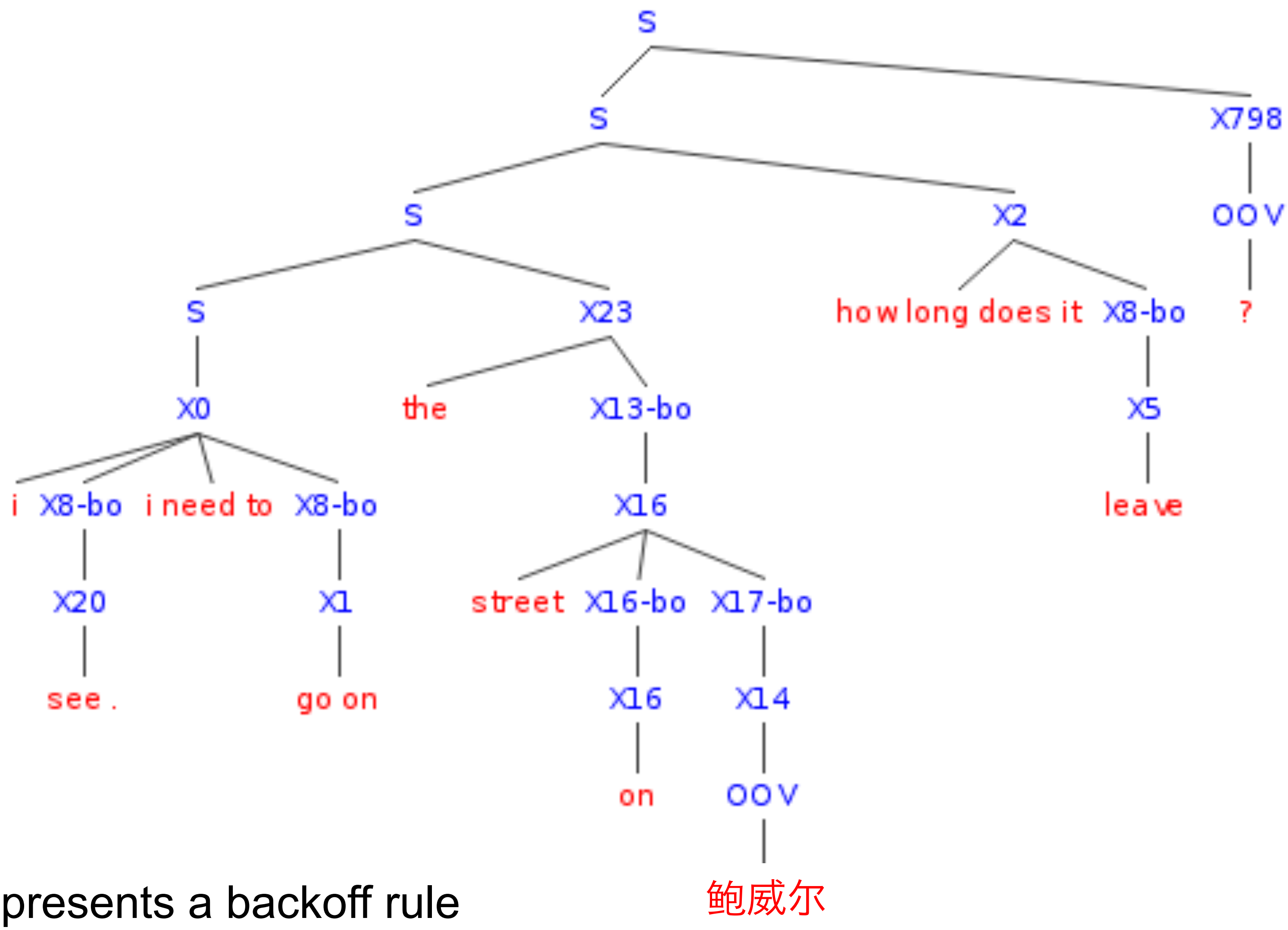
- BackoffRule feature weights
  - BR=0 when backing off to the same category

$$[X1_{\text{backoff}}] \rightarrow < [X1], [X1] > \text{BR}{=}0$$

  - BR=1 when backing off to a different category

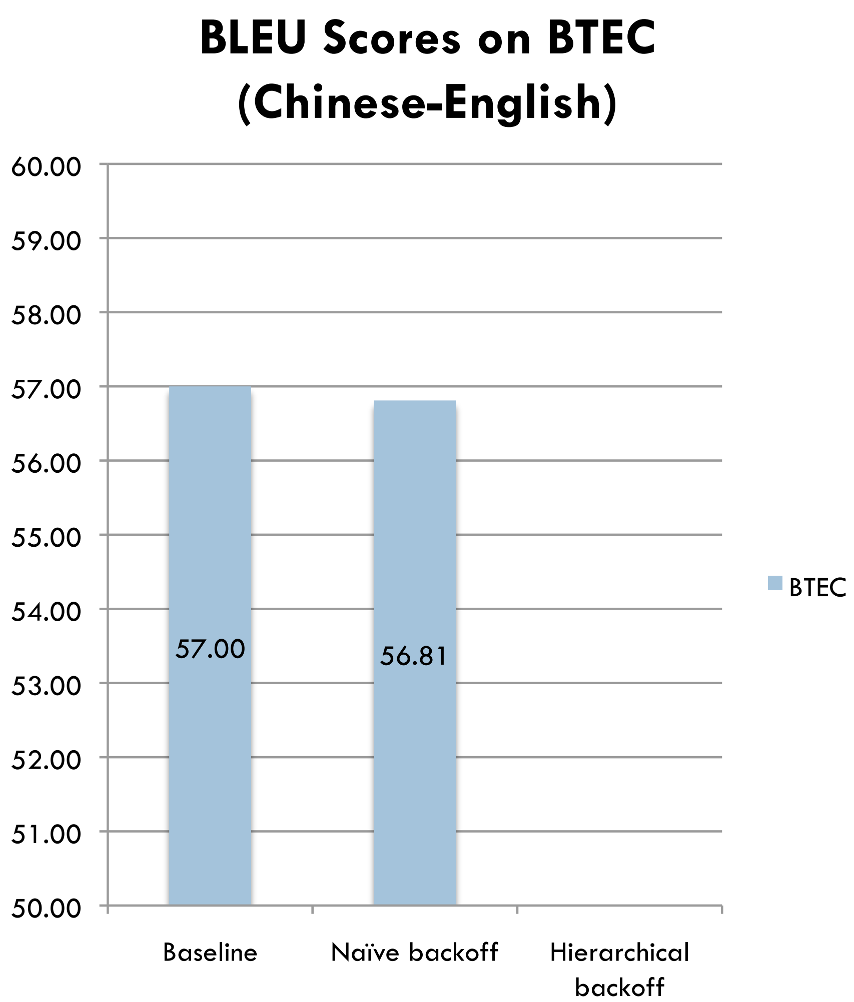
$$[X1_{\text{backoff}}] \rightarrow < [X3], [X3] > \text{BR}{=}1$$

Naïve Backoff Grammar



-bo represents a backoff rule

Naïve Backoff Grammar



- Results from Chinese-English corpus BTEC
- Didn't perform as well
  - Backing off with no preference performs poorly
  - Need to encode preference in structure or features

Hierarchical Backoff Grammar

- Can we encode a backoff hierarchy from our induced grammars?
  - Backoff categories could preferentially move to categories which are similar to the expected category
  - Linguistic motivation: subcategories of nouns behave similarly
- Inducing a strict hierarchy tricky, possibly unnecessary

Hierarchical Backoff Grammar

- Instead, we derive a rough hierarchy based on four induced grammars at varying levels of granularity

$$X_{\boxed{10}} \qquad X_{\boxed{30}}$$

$$X_{\boxed{15}} \qquad X_{\boxed{50}}$$

- Backoff rules allow redirecting from coarser categories to finer categories

$$[X_{\boxed{10}}] \rightarrow [\text{any } X_{\boxed{15}} \text{ category}]$$
$$[X_{\boxed{15}}] \rightarrow [\text{any } X_{\boxed{30}} \text{ category}]$$
$$[X_{\boxed{30}}] \rightarrow [\text{any } X_{\boxed{50}} \text{ category}]$$

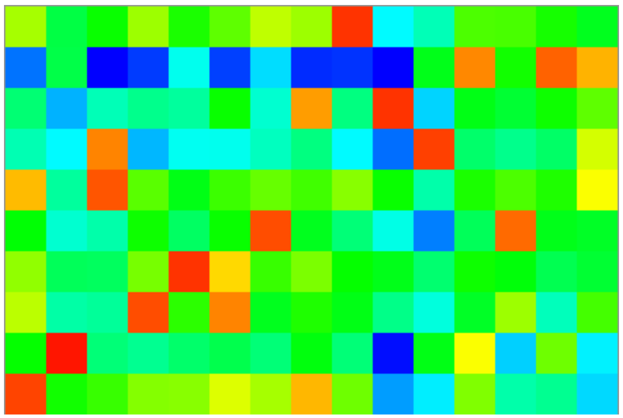
- BackoffRule feature:

$$BR = log_2\ P(X_{\boxed{15}}|X_{\boxed{10}})$$

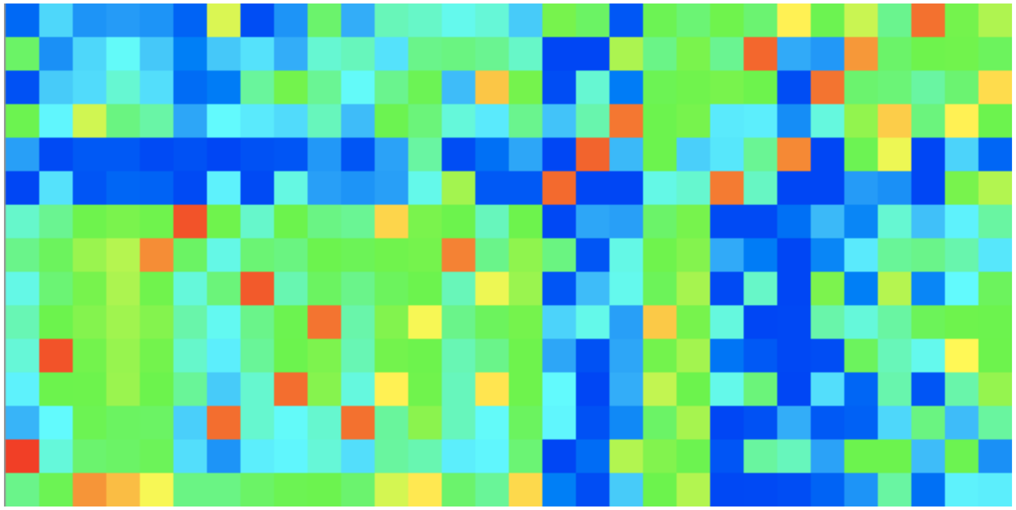
$$P(X_{\boxed{15}}|X_{\boxed{10}}) = \frac{\#\ (phrases\ assigned\ X_{\boxed{10}}\ category\ \cap phrases\ assigned\ X_{\boxed{15}}\ category\ )}{\#\ phrases\ assigned\ X_{\boxed{10}}\ category}$$

Hierarchical Backoff Grammar

- Not truly hierarchical, but does encode some similarities in the categories

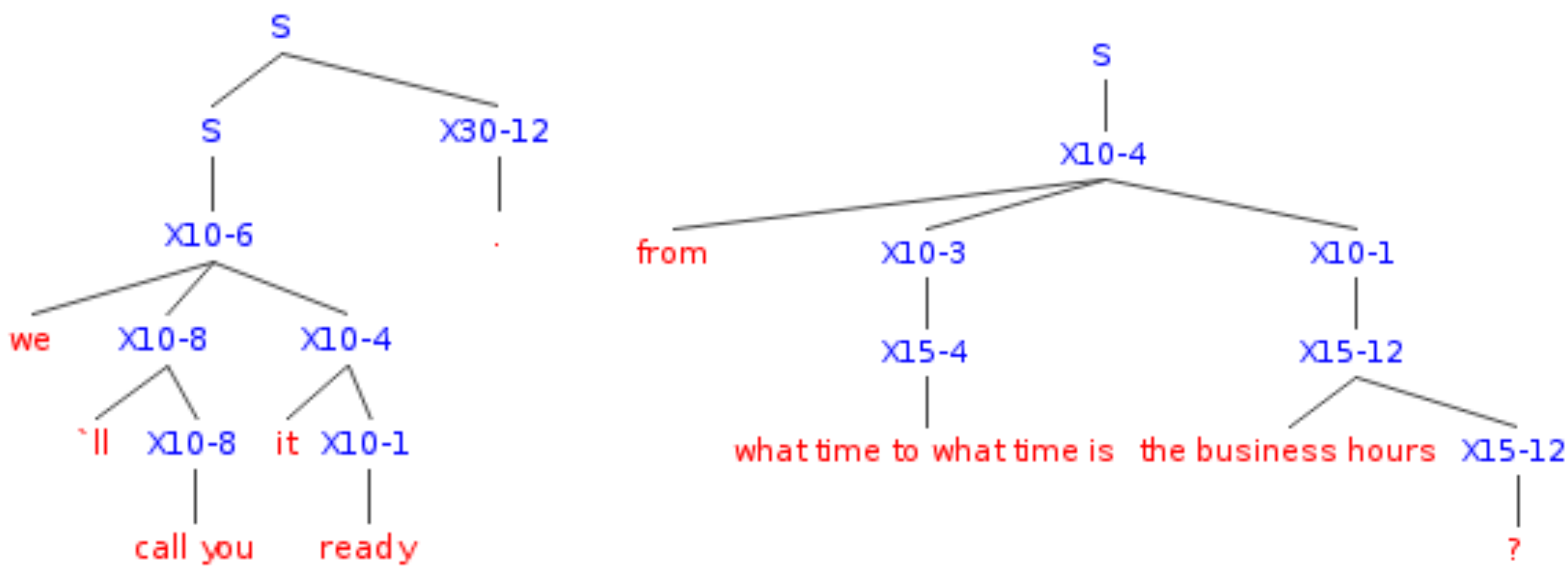


$$H(X_{\boxed{15}} \mid X_{\boxed{10}})$$



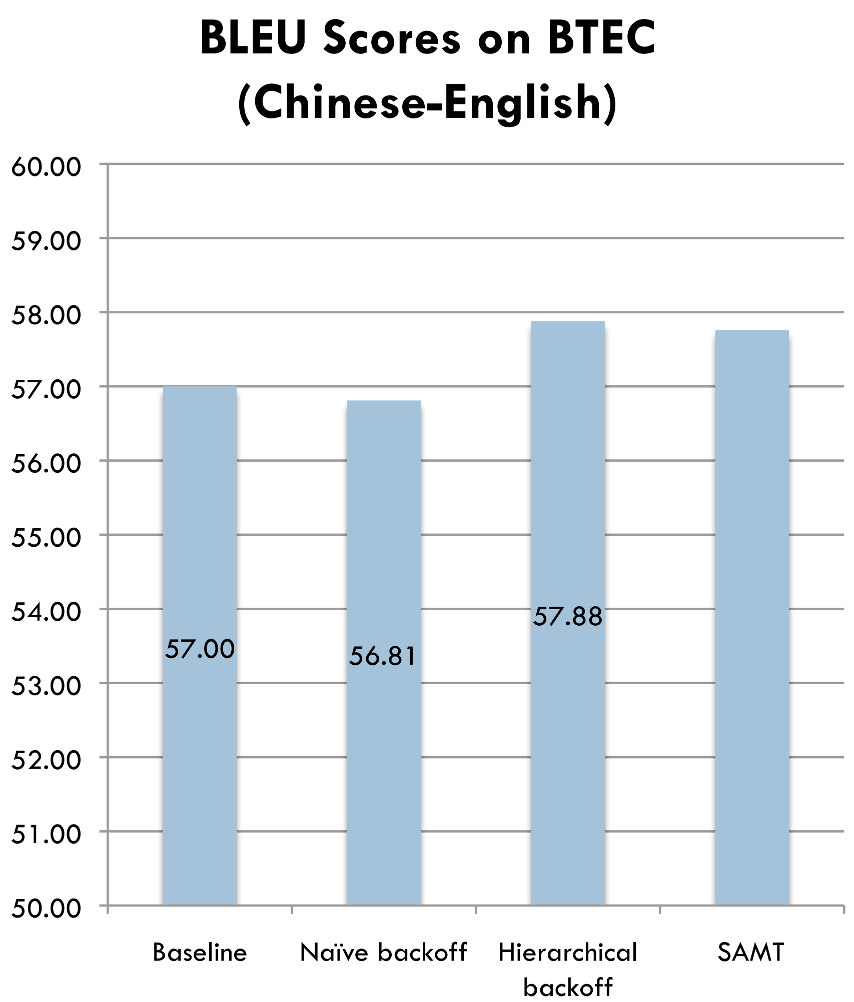
$$H(X_{\boxed{30}} \mid X_{\boxed{15}})$$

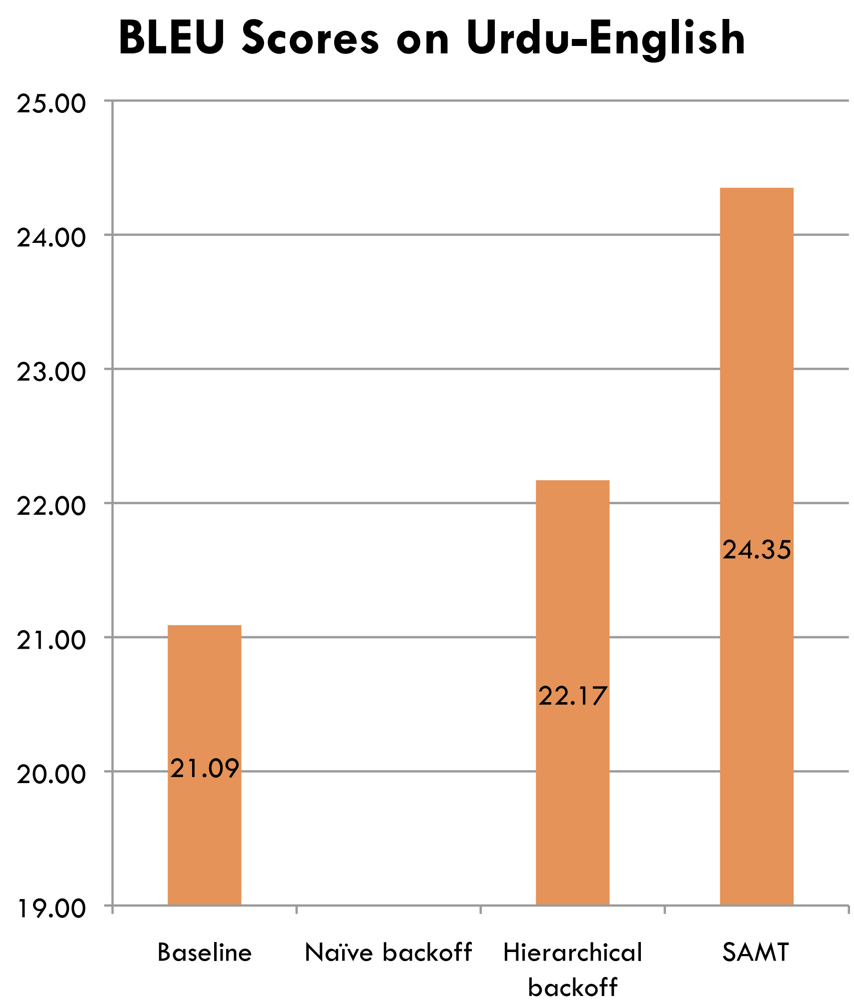
Hierarchical Backoff Grammar



X10-* represent coarse X_{10} categories
X15-* represent fine(-r) X_{15} categories

Results and Future Work



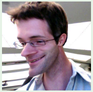


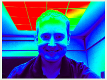
SAMT is Syntax-Augmented Machine Translation

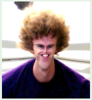
Results and Future Work

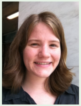
- Hierarchical backoff performs very well
  - Not quite the improvements of supervised syntax-based translation, but good for automated
- Possible improvements
  - Vary levels of granularity
  - More sophisticated feature weighting

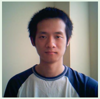
Outline


Trevor Cohn


Chris Dyer


Jan Botha


Olivia Buzek


Desai Chen

- 1:55pm Grammar induction and evaluation. Trevor
- 2:10pm Non-parametric models of category induction. Chris
- 2:25pm Inducing categories for morphology. Jan
- 2:35pm Smoothing, backoff and hierarchical grammars. Olivia
- 2:45pm Parametric models: posterior regularisation. Desai
- 3:00pm Break.

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

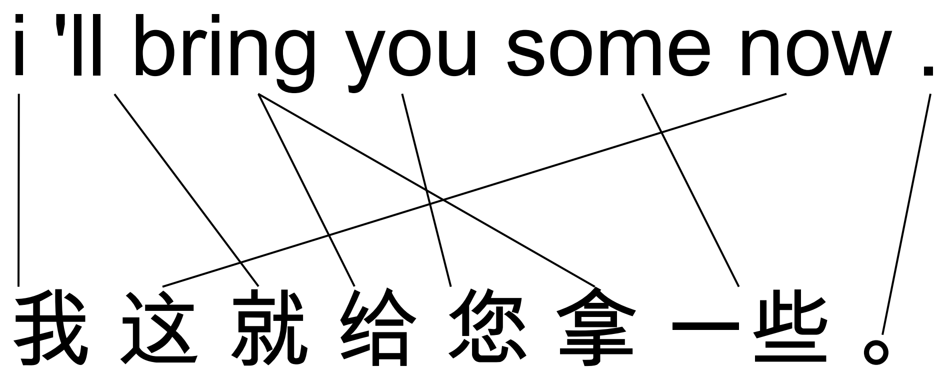
i 'll bring you some now .

我 这 就 给 您 拿 一些 。

Example from btec

Phrase clustering

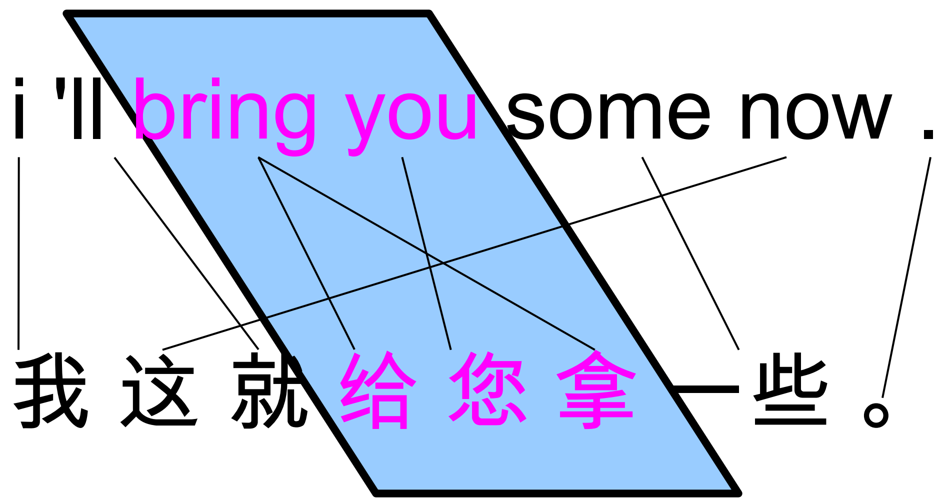
Phrases are defined as contiguous spans aligned with each other



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

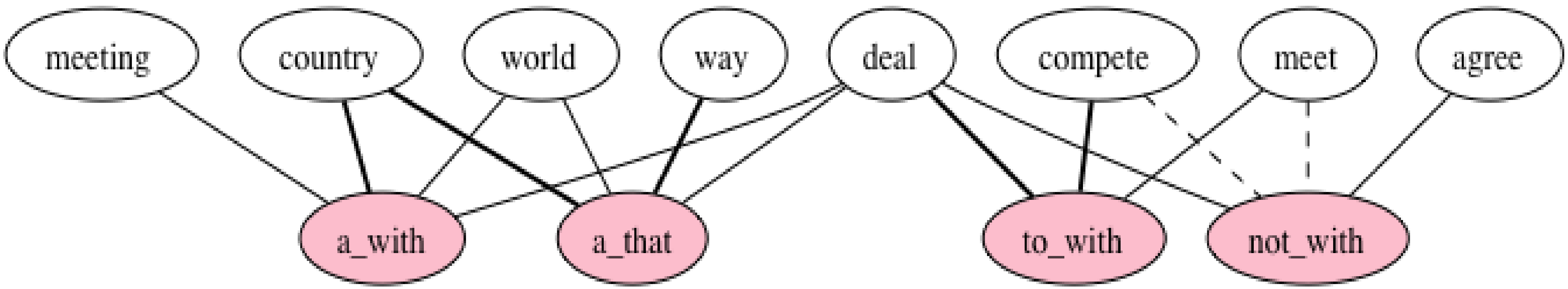
i 'll bring you some now .

我 这 就 给 您 拿 一些 。

source side context

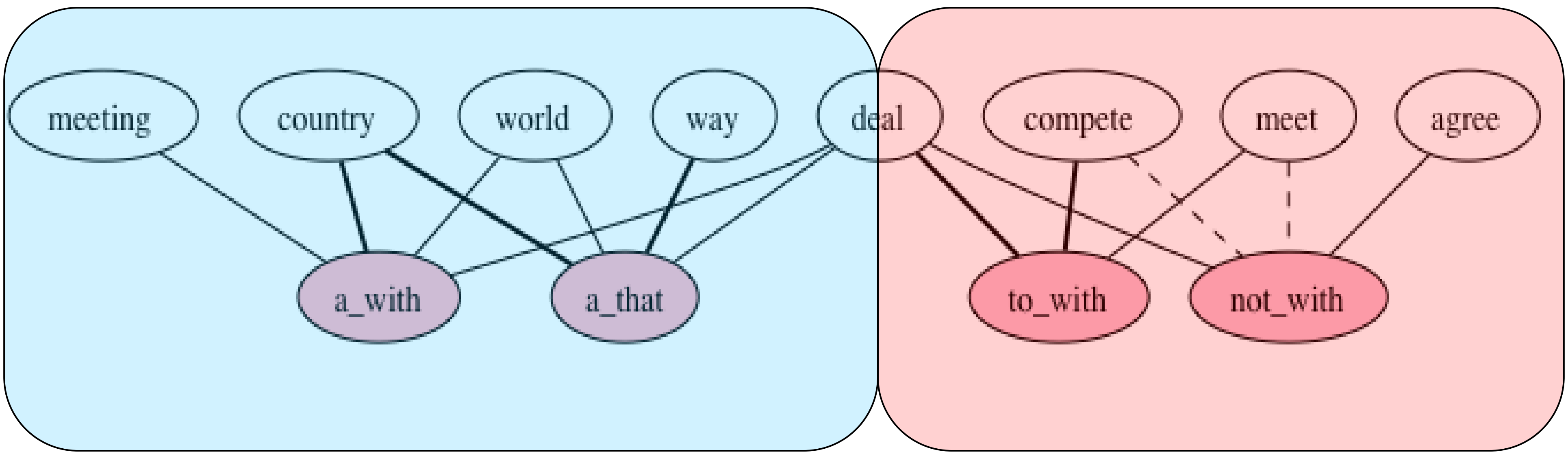
Objective

Put all phrase-context pairs into categories



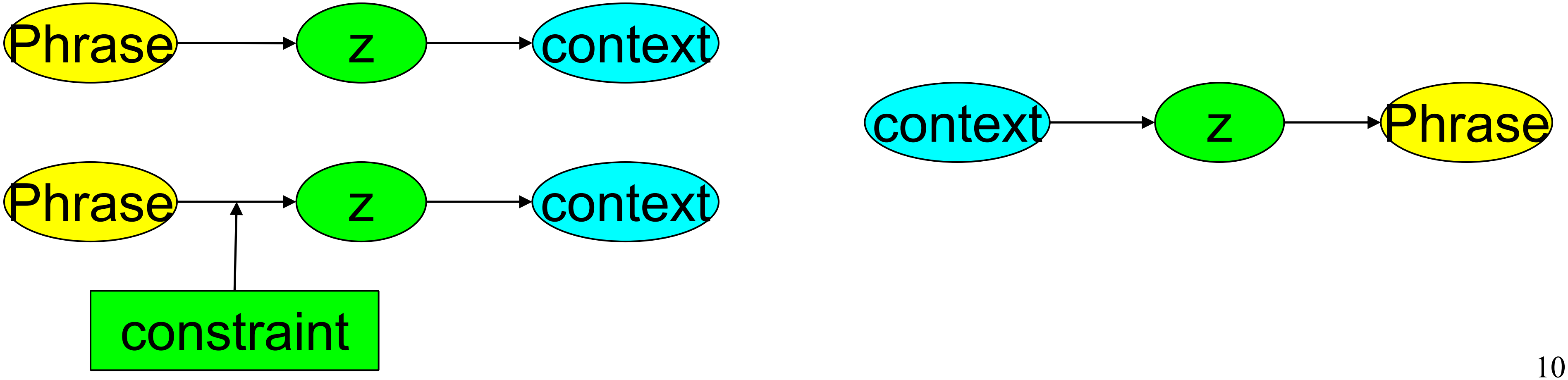
Objective

Put all phrase-context pairs into categories



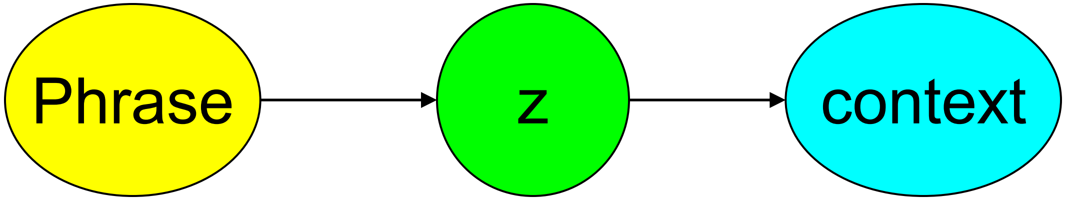
Outline

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



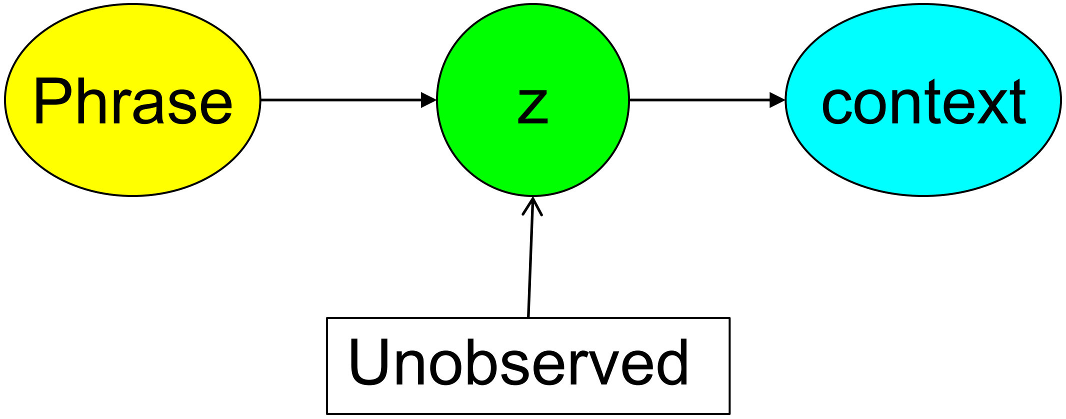
Expectation-Maximization

- naïve Bayes model for phrase labeling



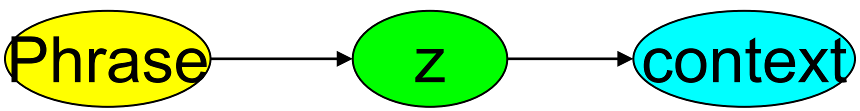
EM clustering

- naïve Bayes model for phrase labeling



EM clustering

- naïve Bayes model for phrase labeling

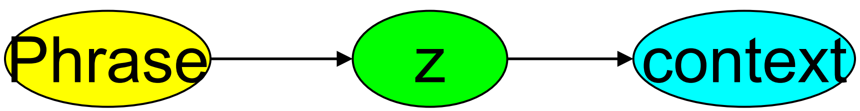


E-step

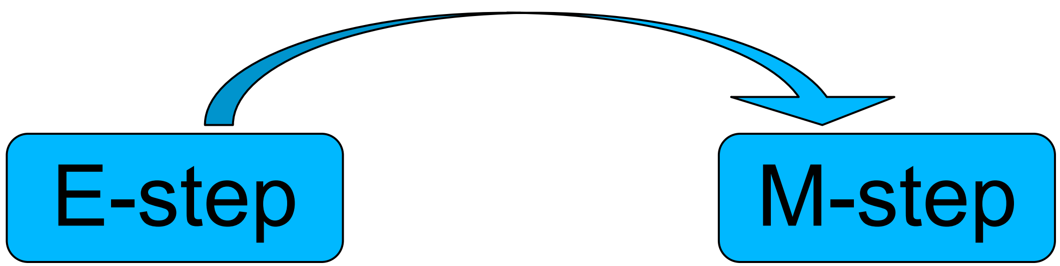
M-step

EM clustering

- naïve Bayes model for phrase labeling

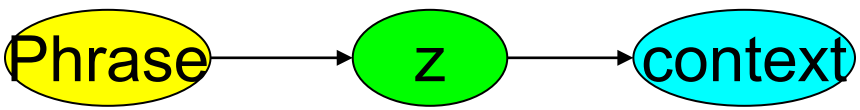


$$q(z|p,c) = P_\theta(z|p,c)$$

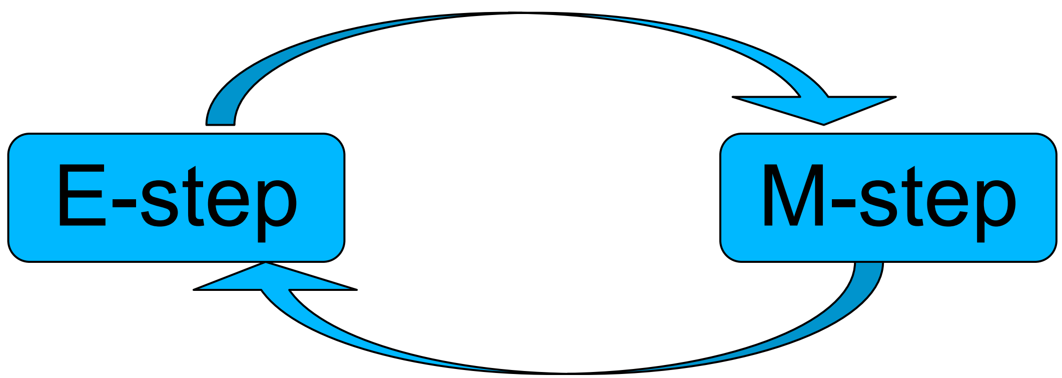


EM clustering

- naïve Bayes model for phrase labeling



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



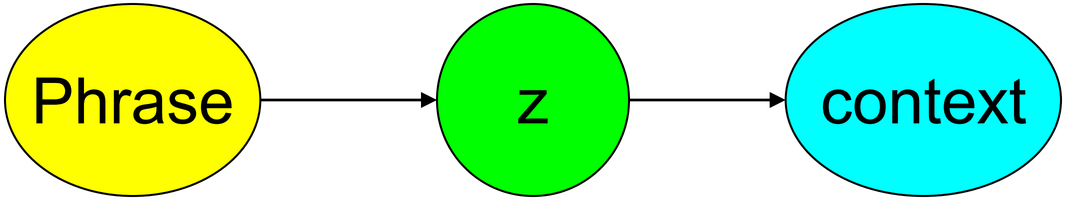
$$\theta = MLE\ q(z|p,c)$$

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.

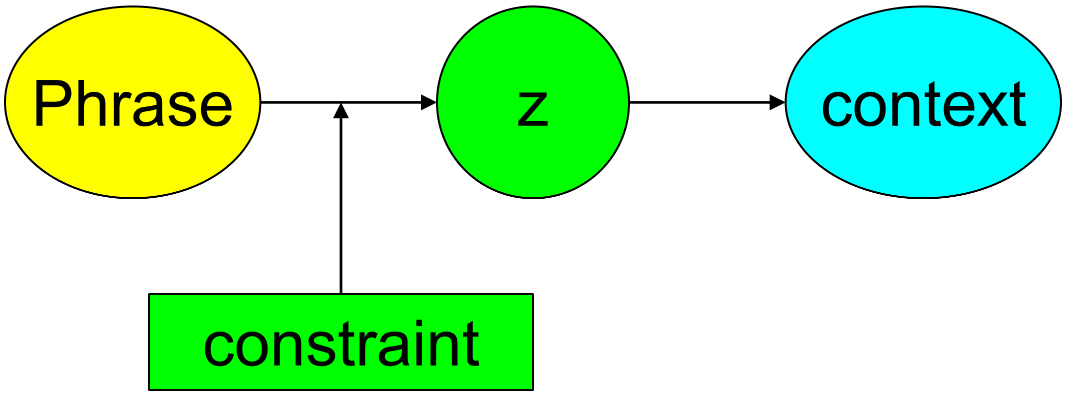
Sparsity constraints

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

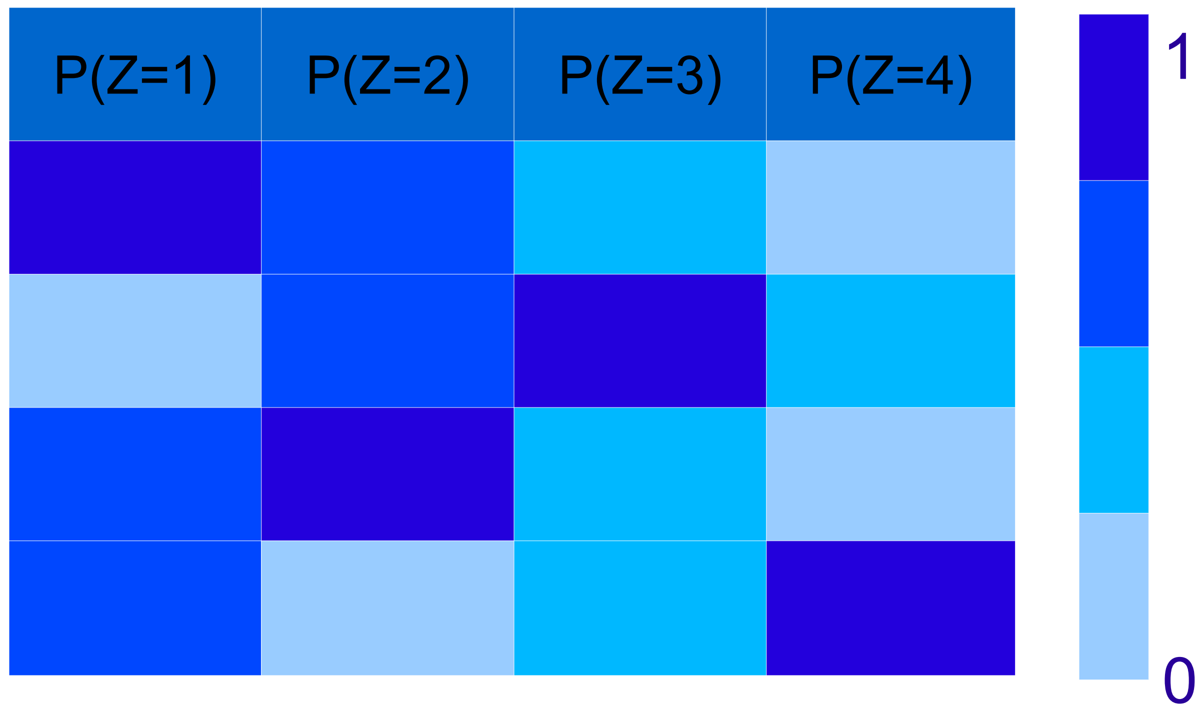
Phrase: there are

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis



Sparsity constraints

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

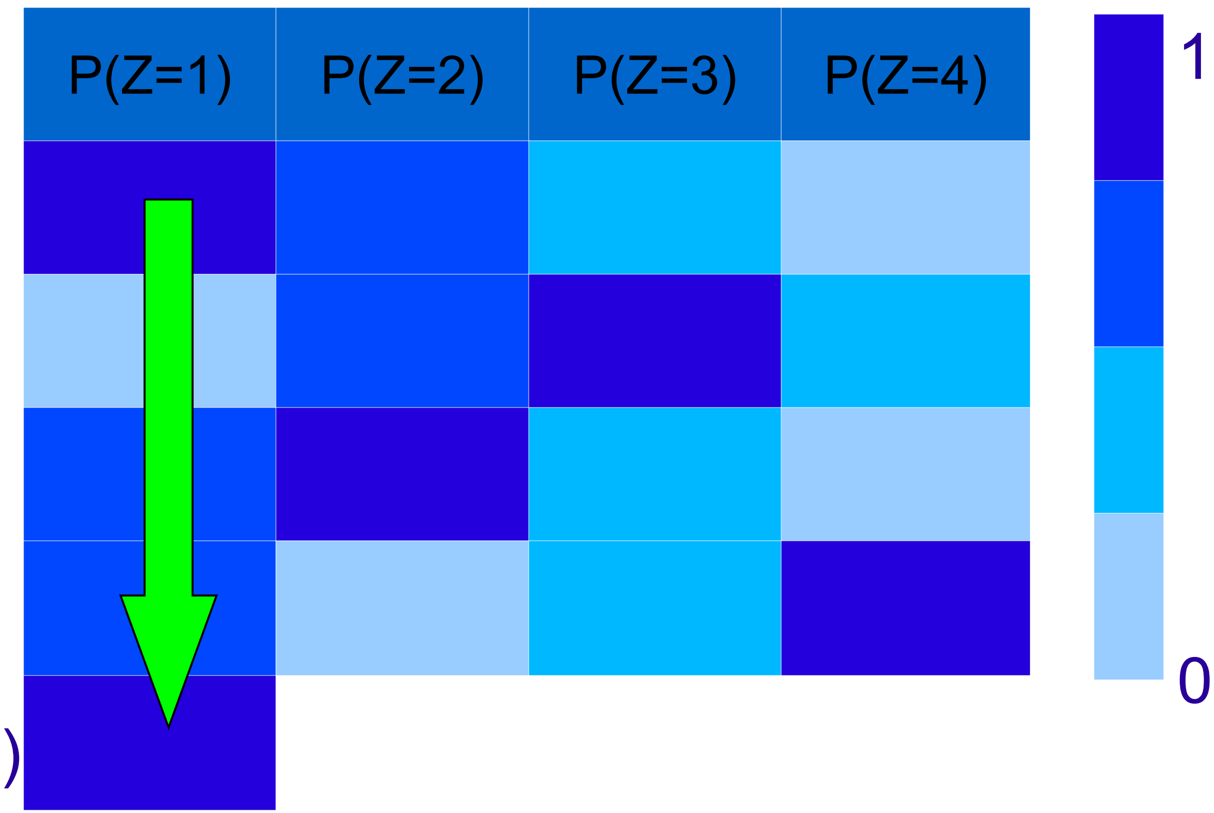
Phrase: there are

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

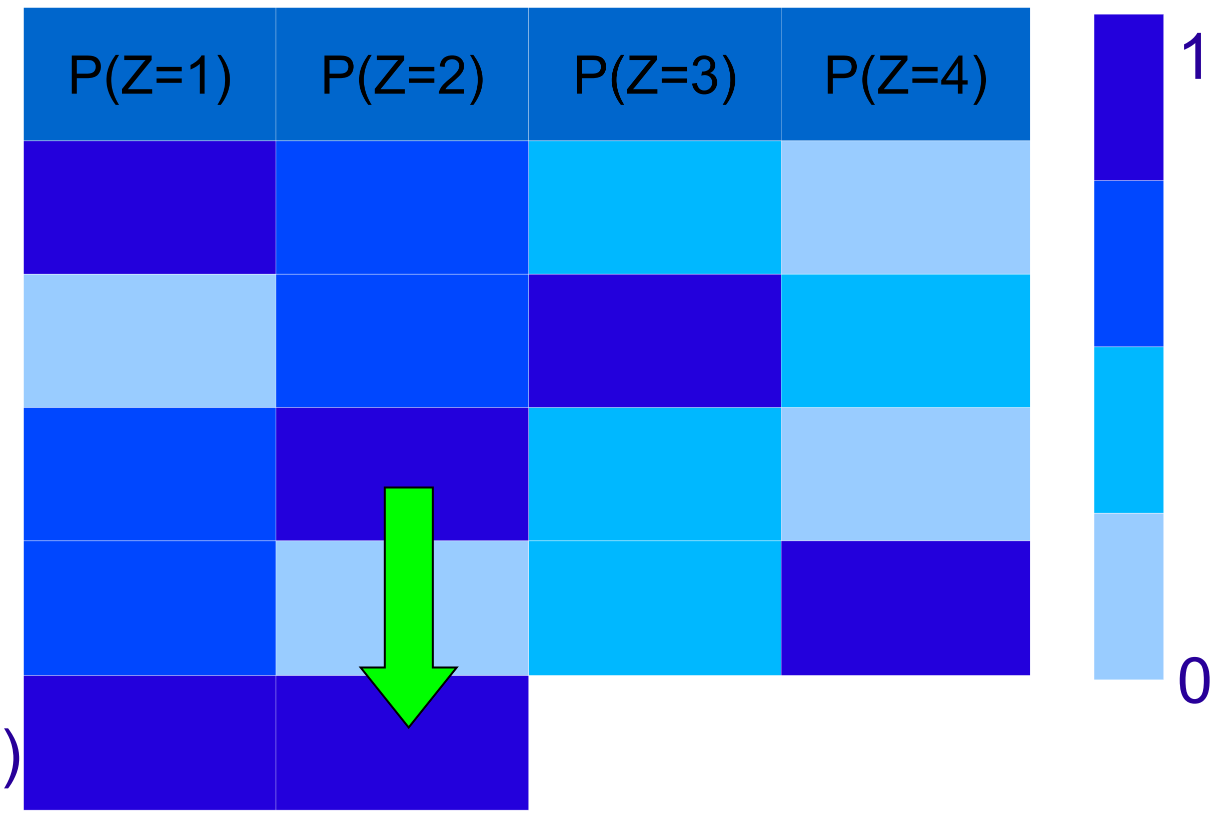
Contexts:
i understand _ some sightseeing

<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

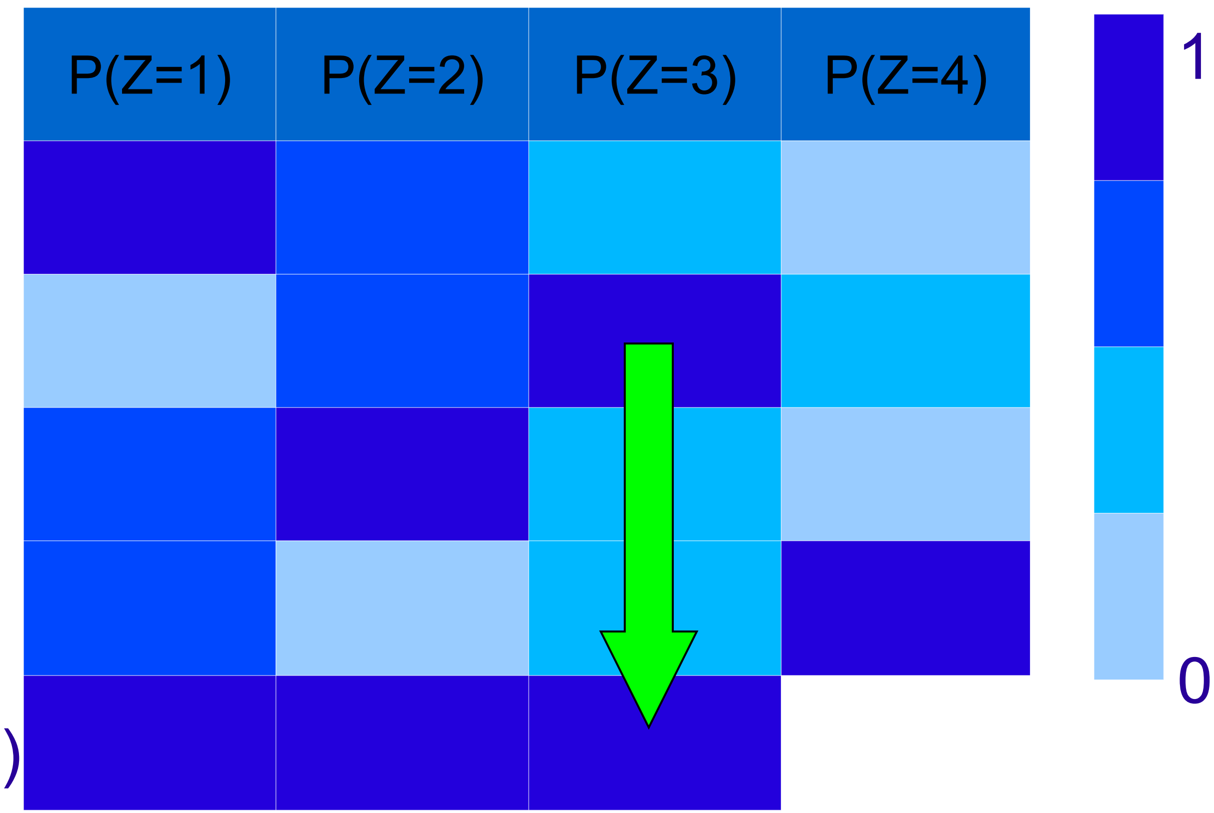
Contexts:
i understand _ some sightseeing

<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

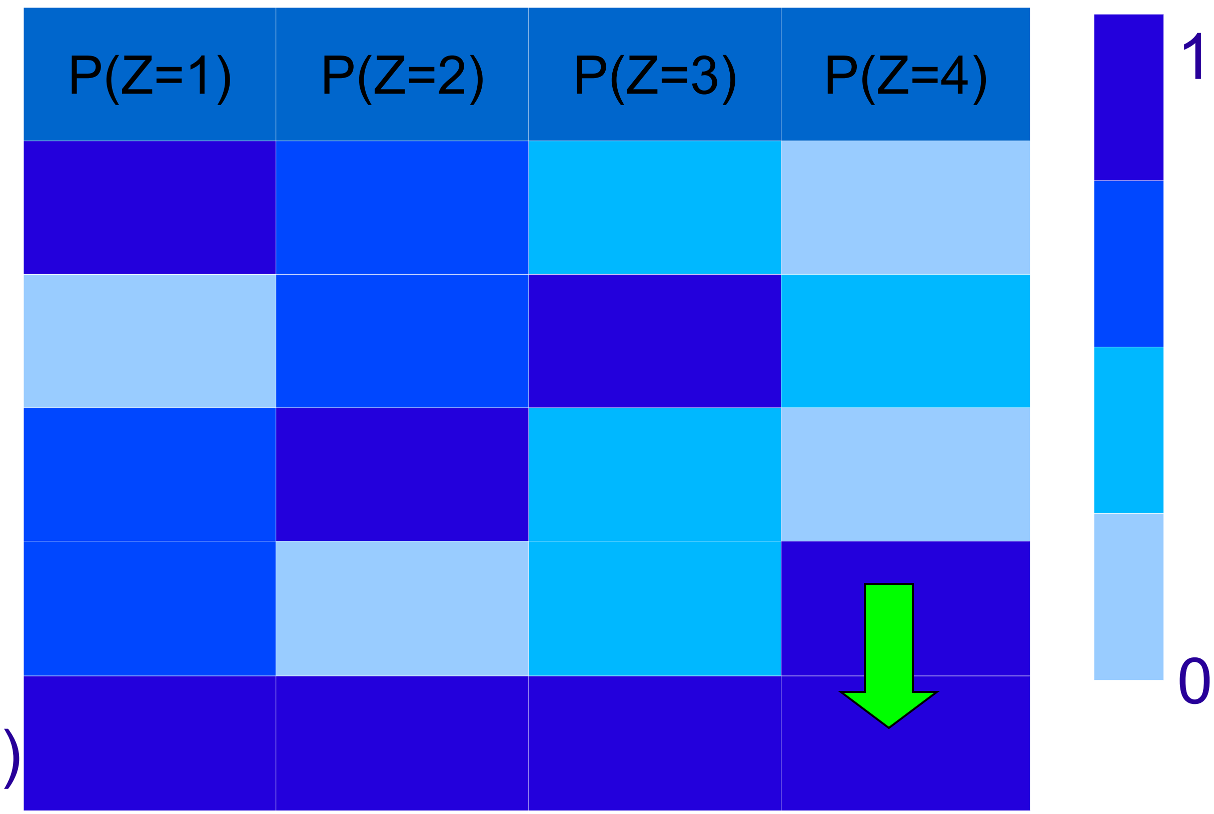
Contexts:
i understand _ some sightseeing

<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

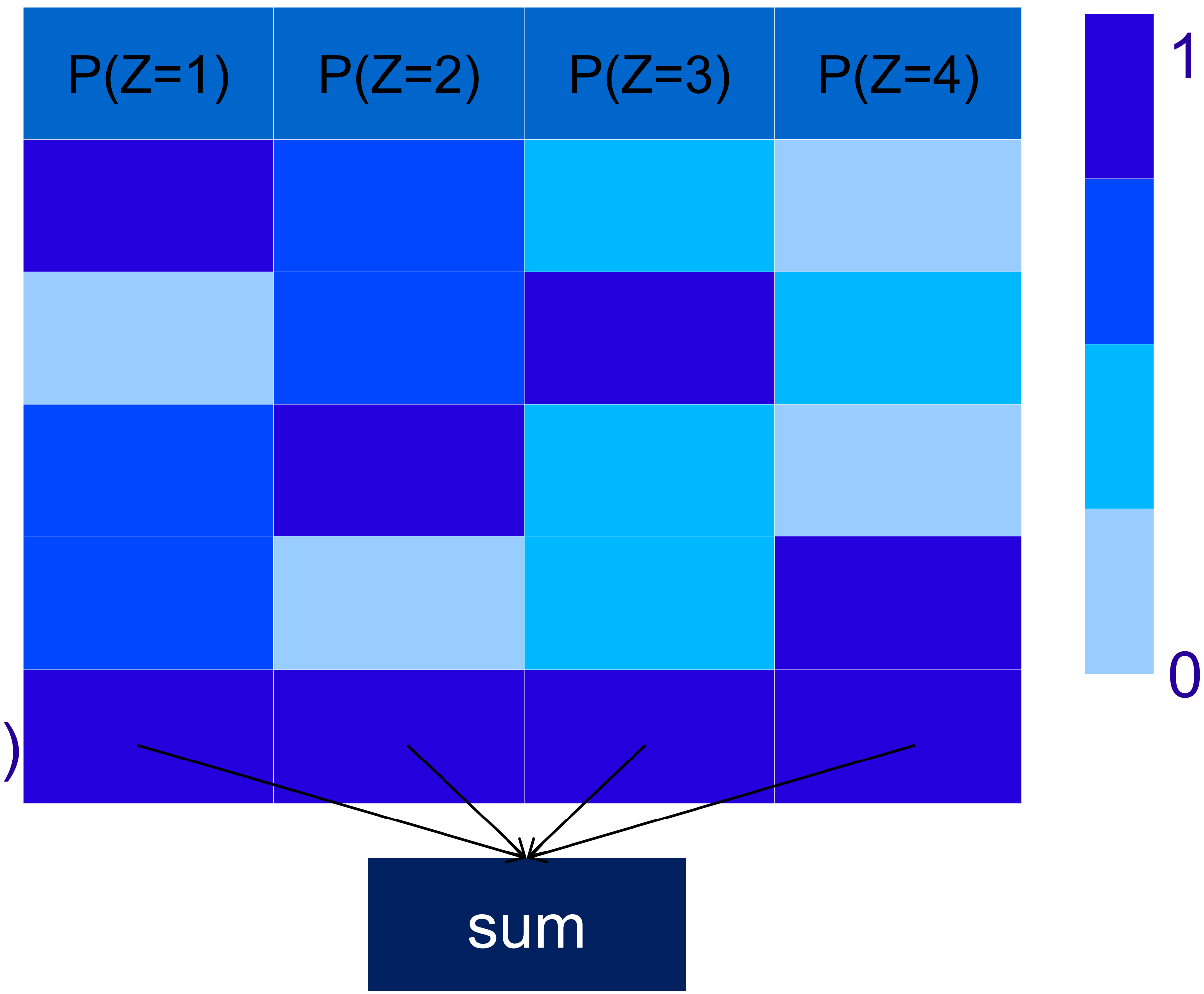
Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max P(tag|phrase)



Sparsity constraints

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

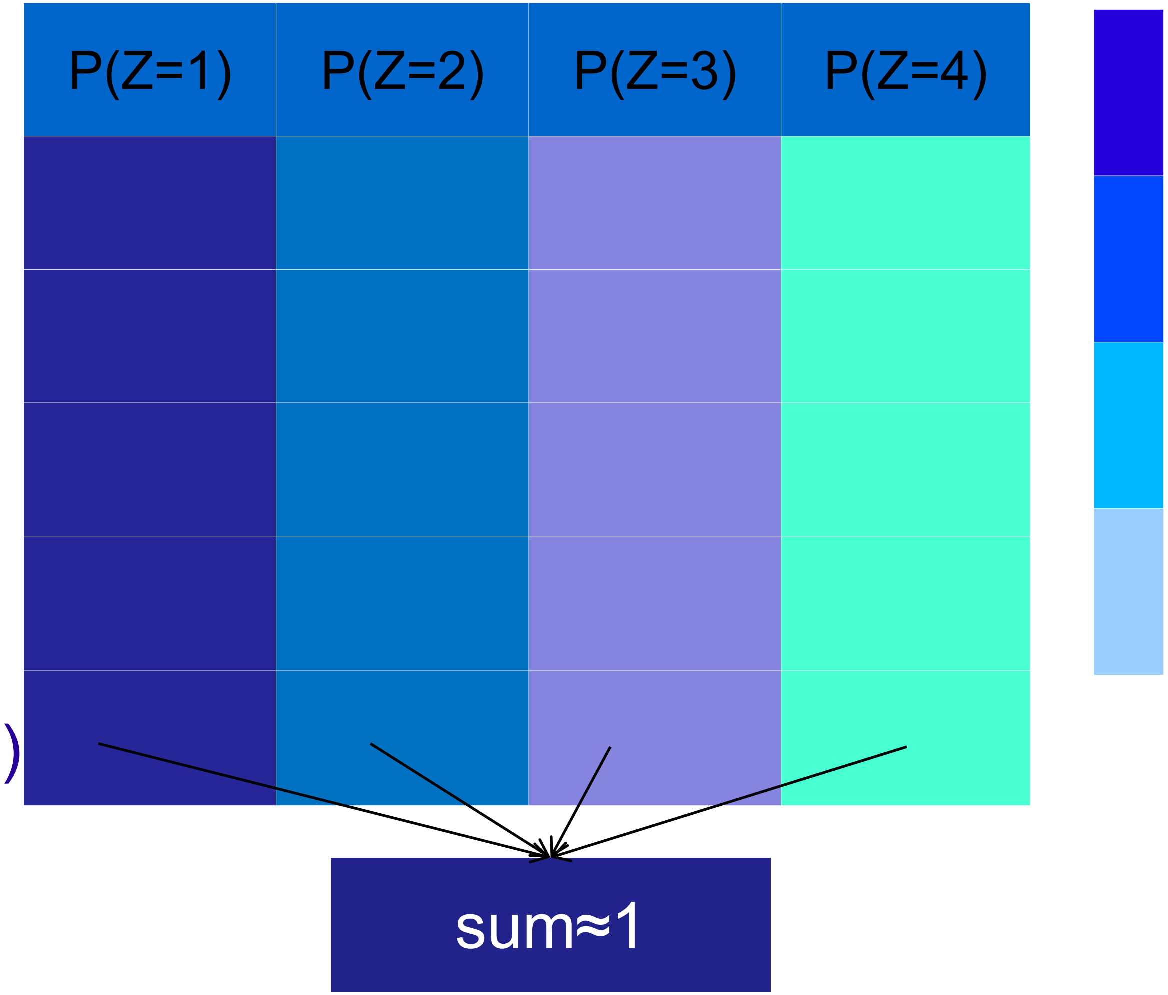
Phrase: there are

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants
brochure shows _ some tennis

max P(tag|phrase)

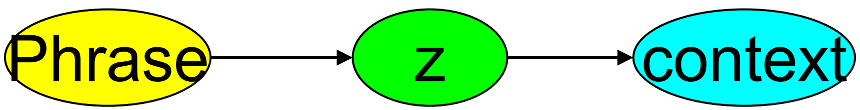


Posterior Regularization

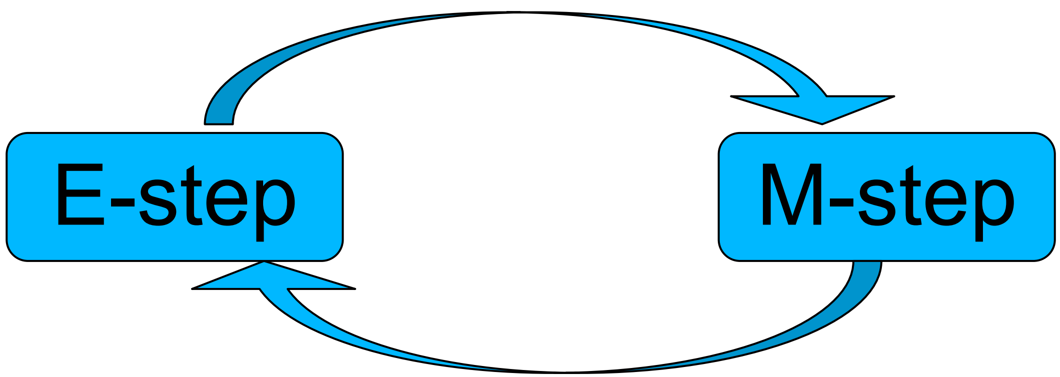
- Follows *Posterior Regularization for Structured Latent Variable Models,* Ganchev et al., 2009
- During E-step, impose constraints on the posterior q to guide the search

Posterior Regularization

- impose constraints on the posterior q



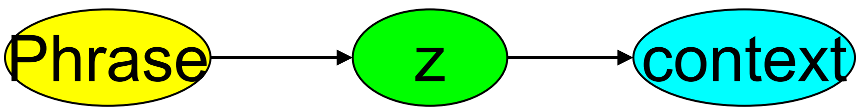
$$q(z|p,c) = \underset{q \in Q}{\arg\min}\, KL(q||P_\theta)$$

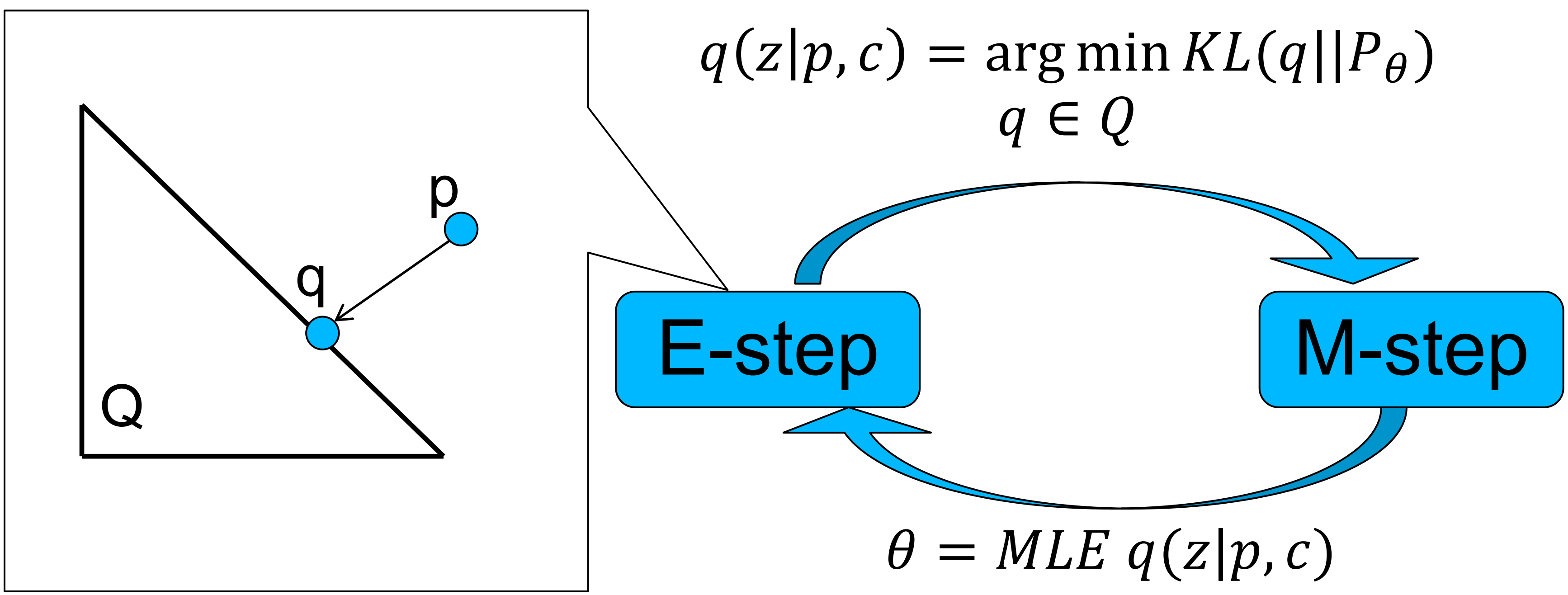


$$\theta = MLE\ q(z|p,c)$$

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:

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

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

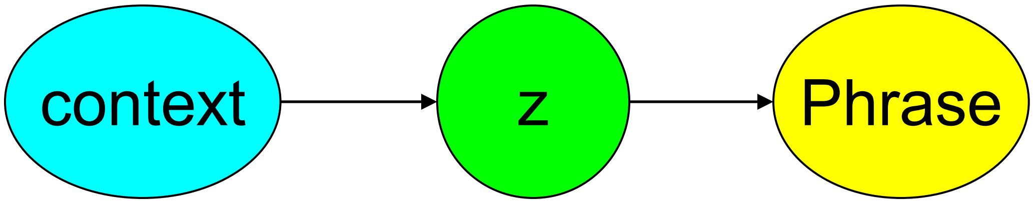
- Constrained model gives clustering that's more sparse
- Clustering for a few phrases with 25 tags on BTEC ZH-EN

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

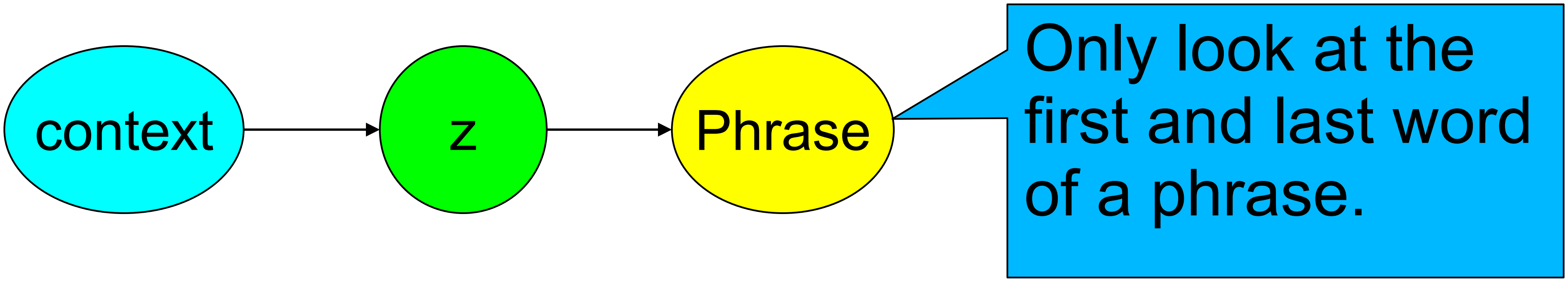
- agreement constraint: different “good” models should agree on posterior distribution
- what model to agree with: another naïve Bayes model in the reverse direction or in the other language.

Agreement model



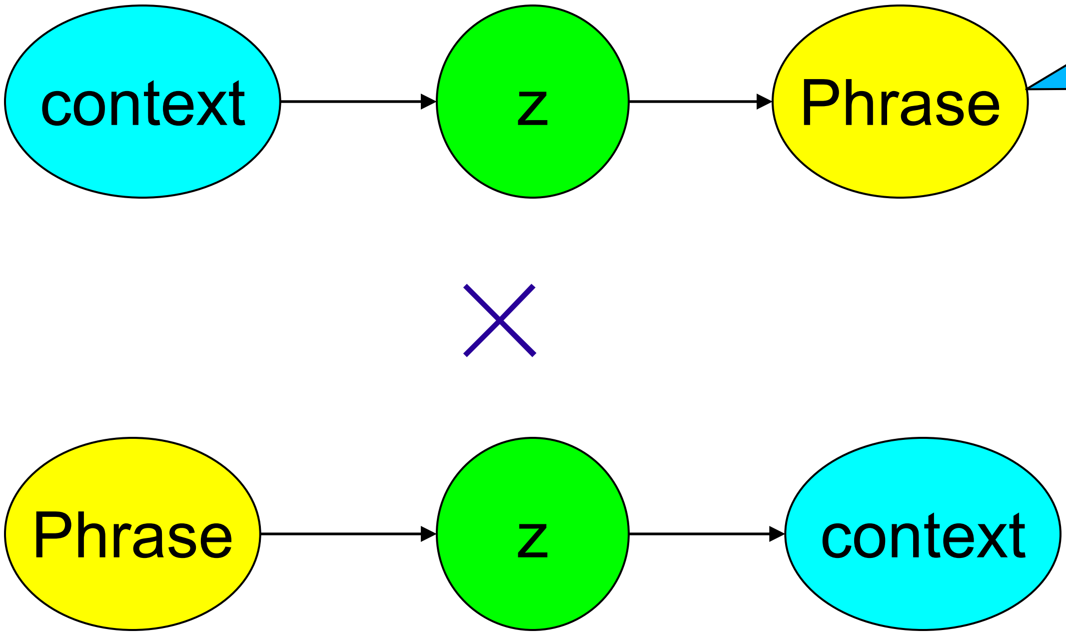
- implementation: multiply posteriors of two models together.

Agreement model



- implementation: multiply posteriors of two models together.

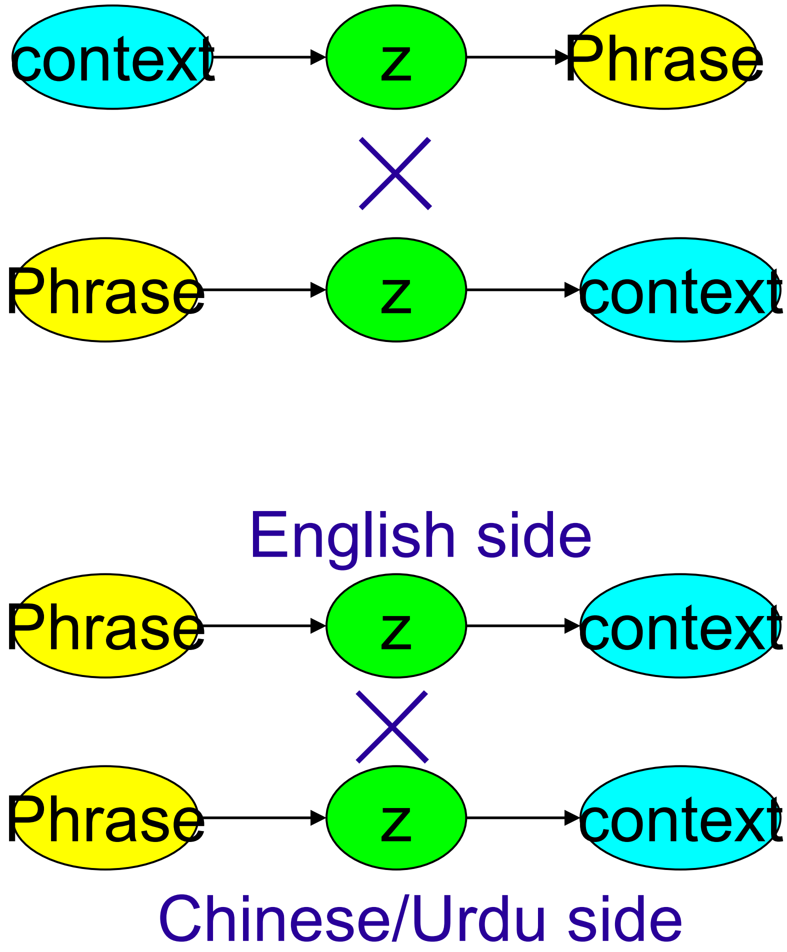
Agreement model



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

- implementation: multiply posteriors of two models together.

Agreement model

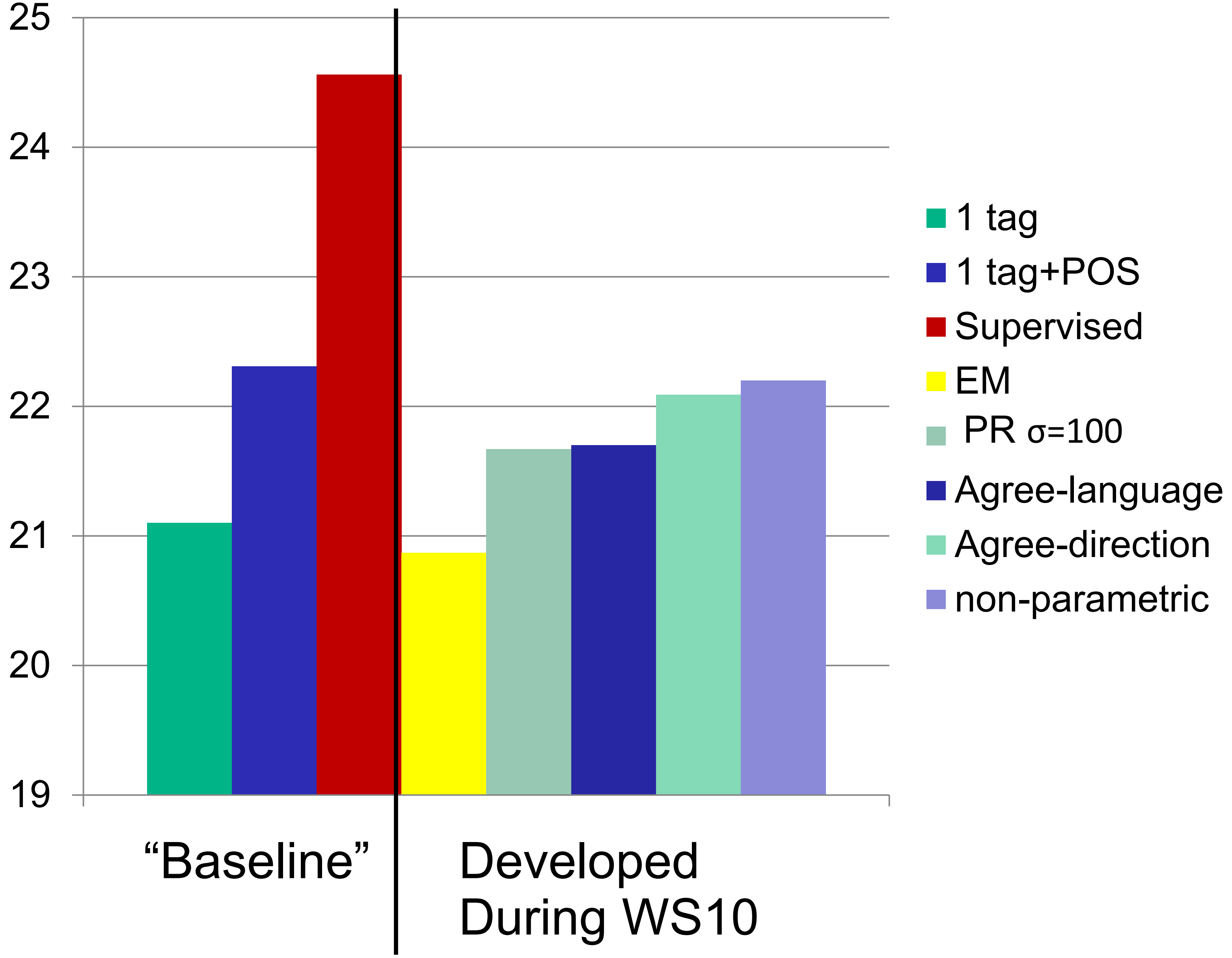


- implementation: multiply posteriors of two models together.

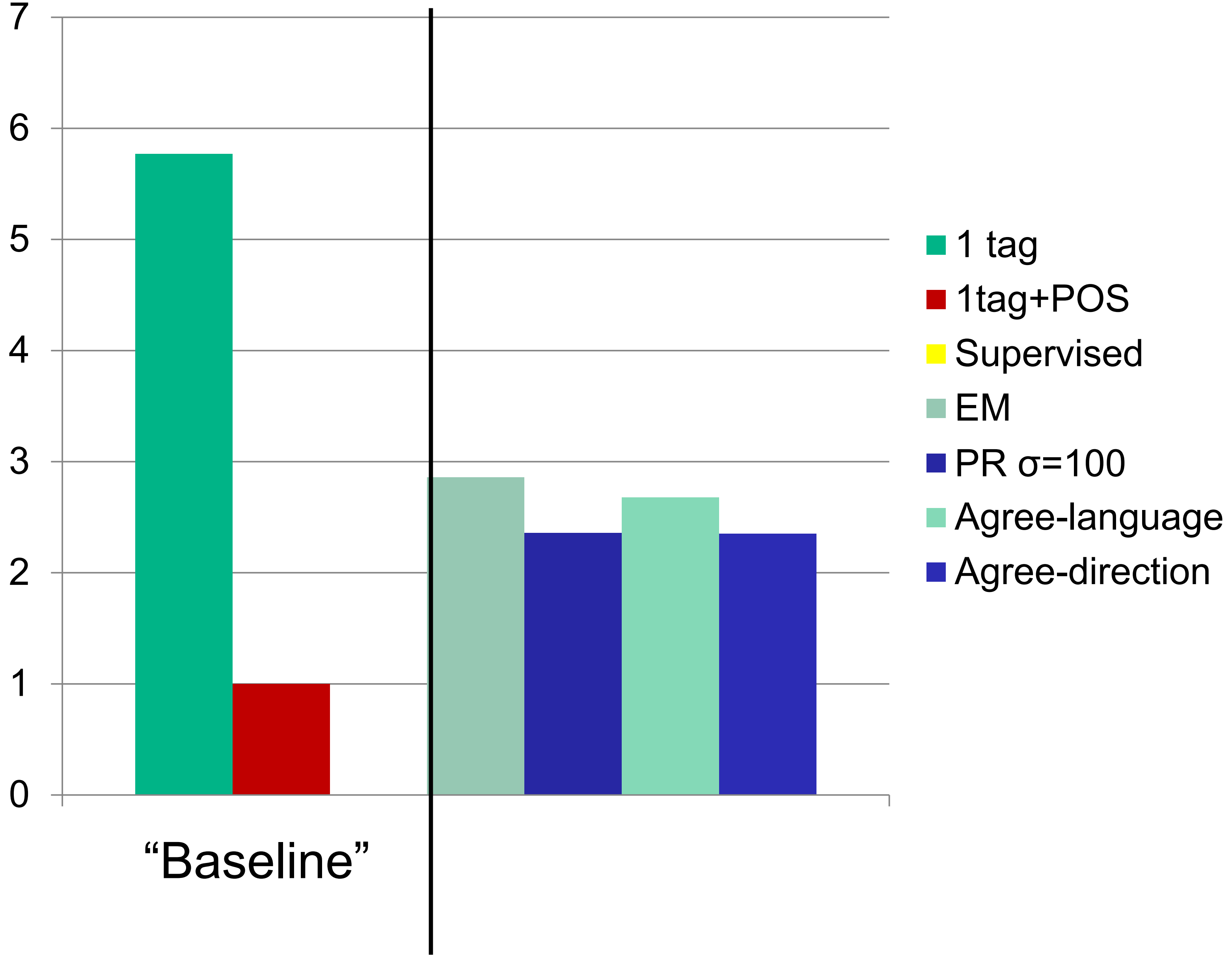
Outline

- 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

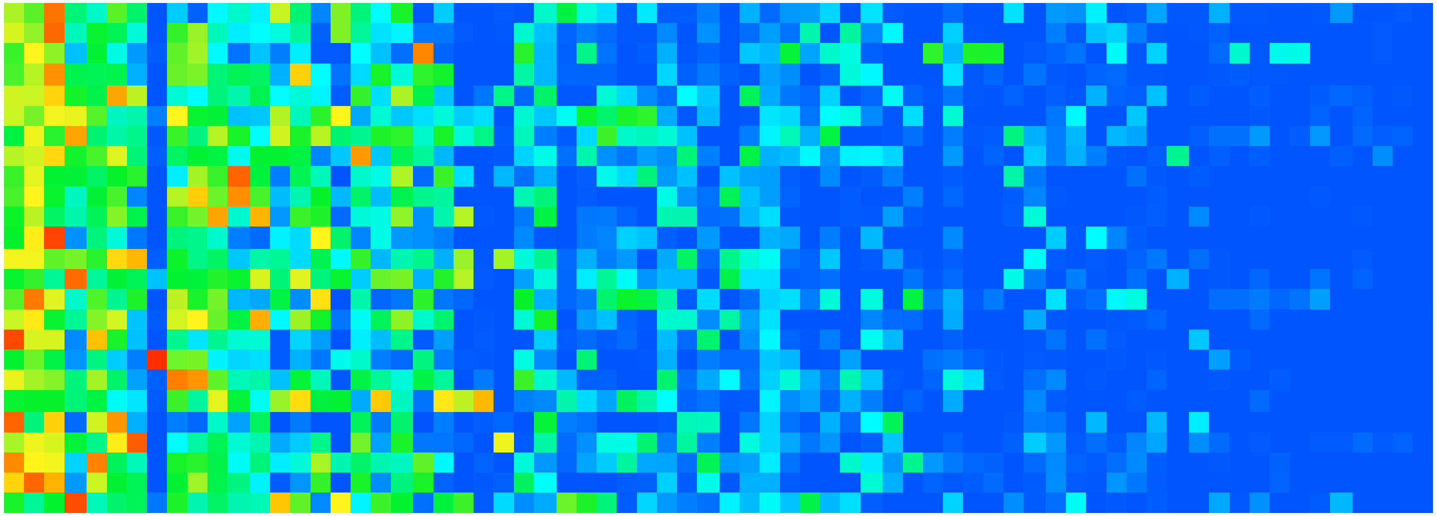


Evaluation against supervised grammar (Conditional Entropy, lower is better)

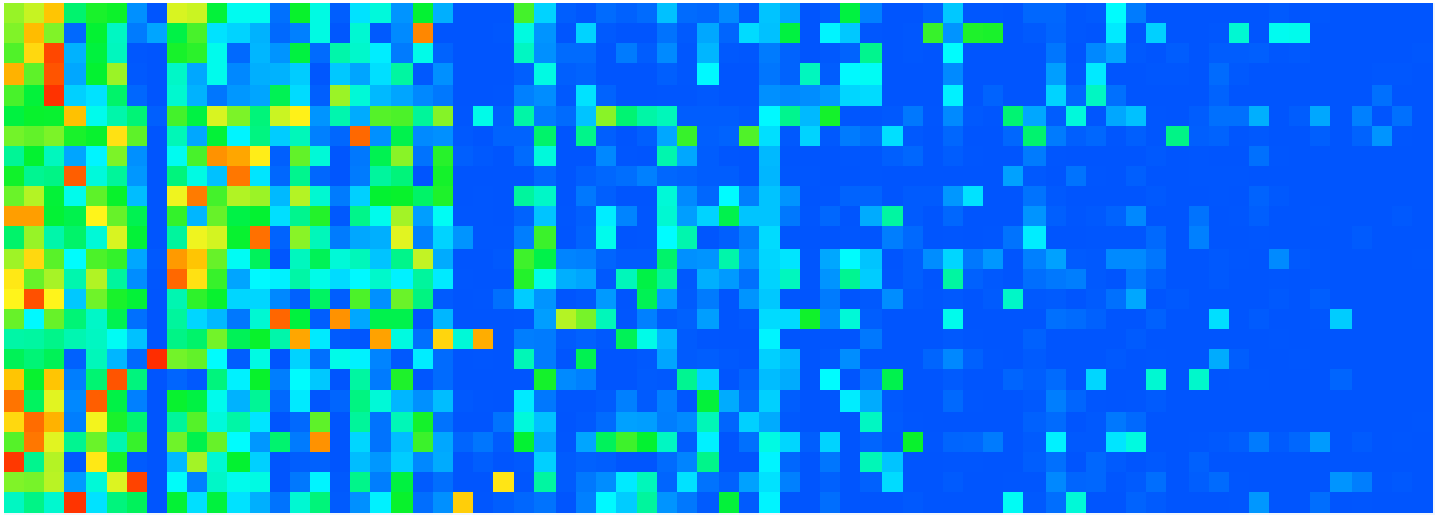


Confusion matrix against supervised labeling

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

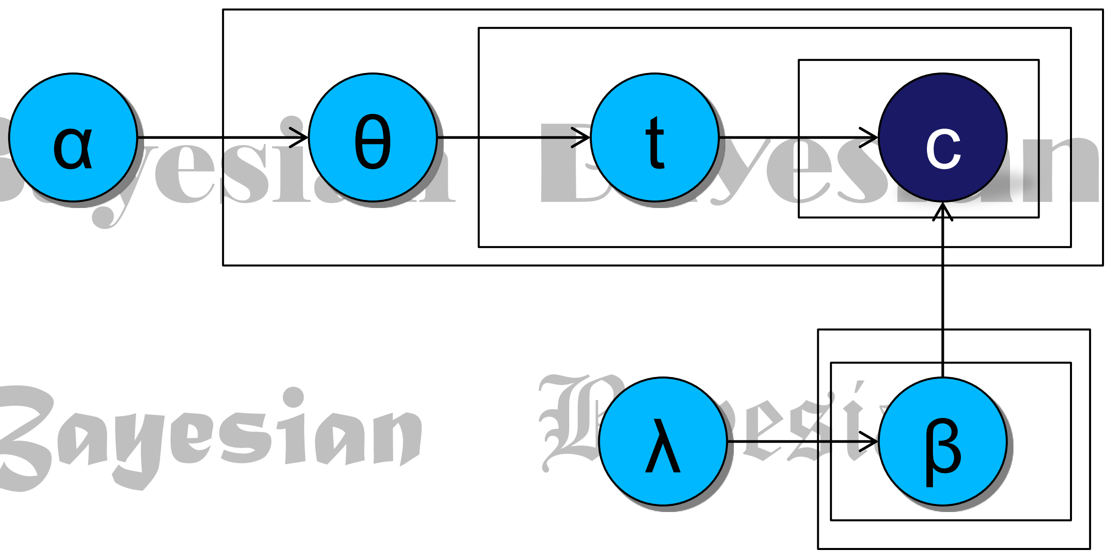
Bayesian Bayesian Bayesian

Things we didn't have time to get working

Bayesian Bayesian Bayesian

- variational Bayes inference

Bayesian Bayesian Bayesian



Bayesian Bayesian Bayesian

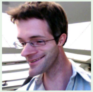
Bayesian Bayesian Bayesian

Outline

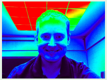
- Where do phrases come from?
- EM with posterior regularization
- results and future experiments

Thanks!

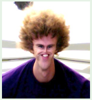
Outline



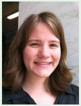
Trevor Cohn



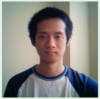
Chris Dyer



Jan Botha



Olivia Buzek



Desai Chen

- 1:55pm Grammar induction and evaluation. Trevor
- 2:10pm Non-parametric models of category induction. Chris
- 2:25pm Inducing categories for morphology. Jan
- 2:35pm Smoothing, backoff and hierarchical grammars. Olivia
- 2:45pm Parametric models: posterior regularisation. Desai
- 3:00pm Break.