



# Enriching SCFG Rules Directly from Efficient Bilingual Chart Parsing

Martin Čmejrek, Bowen Zhou, Bing Xiang  
{martin.cmejrek,bzhou,bxiang}@us.ibm.com  
IBM, T. J. Watson Research Center, Yorktown Heights

Europarl: German-English, hierarchical MT

die herausforderung besteht darin diese systeme zu den besten der welt zu machen

the challenge is to make the system the very best

r1:  $X \xrightarrow{\Omega}$  <machen, make>

r2:  $X \xrightarrow{\Omega}$  <die herausforderung, the challenge>

r3:  $X \xrightarrow{\Omega}$  <diese systeme, the system>

r4:  $X \xrightarrow{\Omega}$  <zu den besten der welt, the very best>

r5:  $X \xrightarrow{\Omega}$  <besteht darin  $X_1$  zu  $X_2$ , is to  $X_2$   $X_1$ >

r6:  $X \xrightarrow{\Omega}$  <besteht darin, is>

r7:  $X \xrightarrow{\Omega}$  < $X_1$  zu  $X_2$ , to  $X_2$   $X_1$ >

glue:  $X \xrightarrow{\Omega}$  < $X_1$   $X_2$ ,  $X_1$   $X_2$ >

Why is the rule r5 missing from the model?

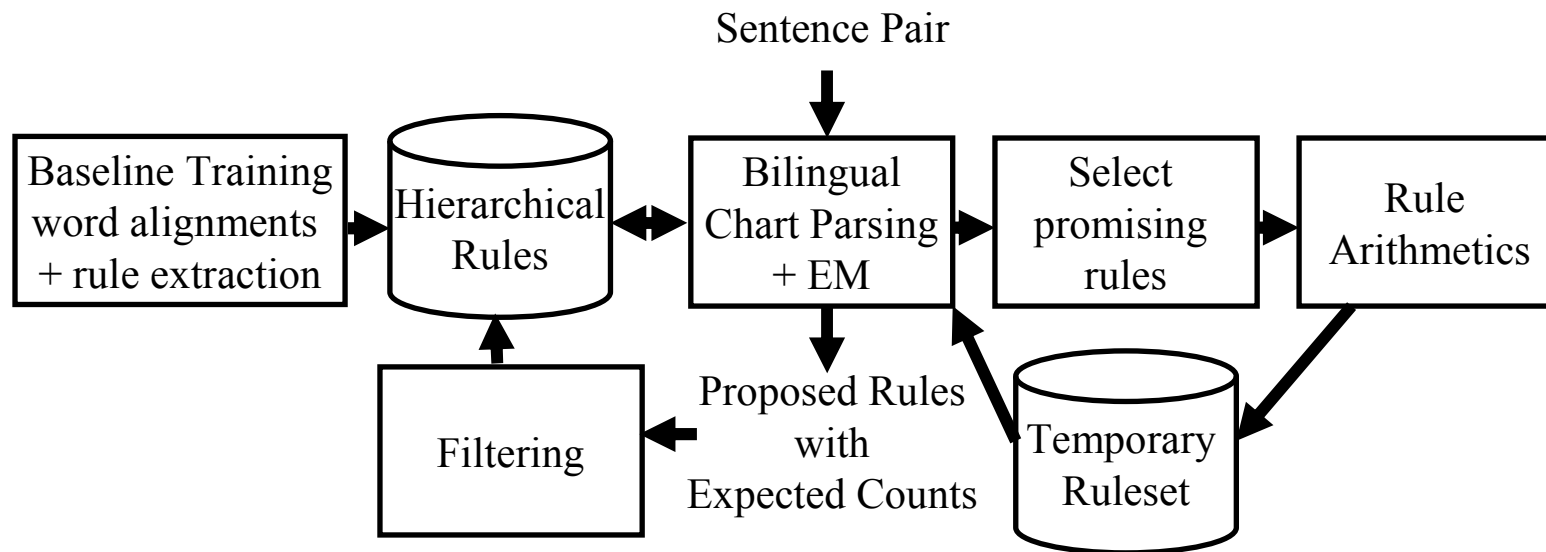
- Heuristic approaches to rule extraction  
(max phrase pair length, min span of non-terminals, etc.)

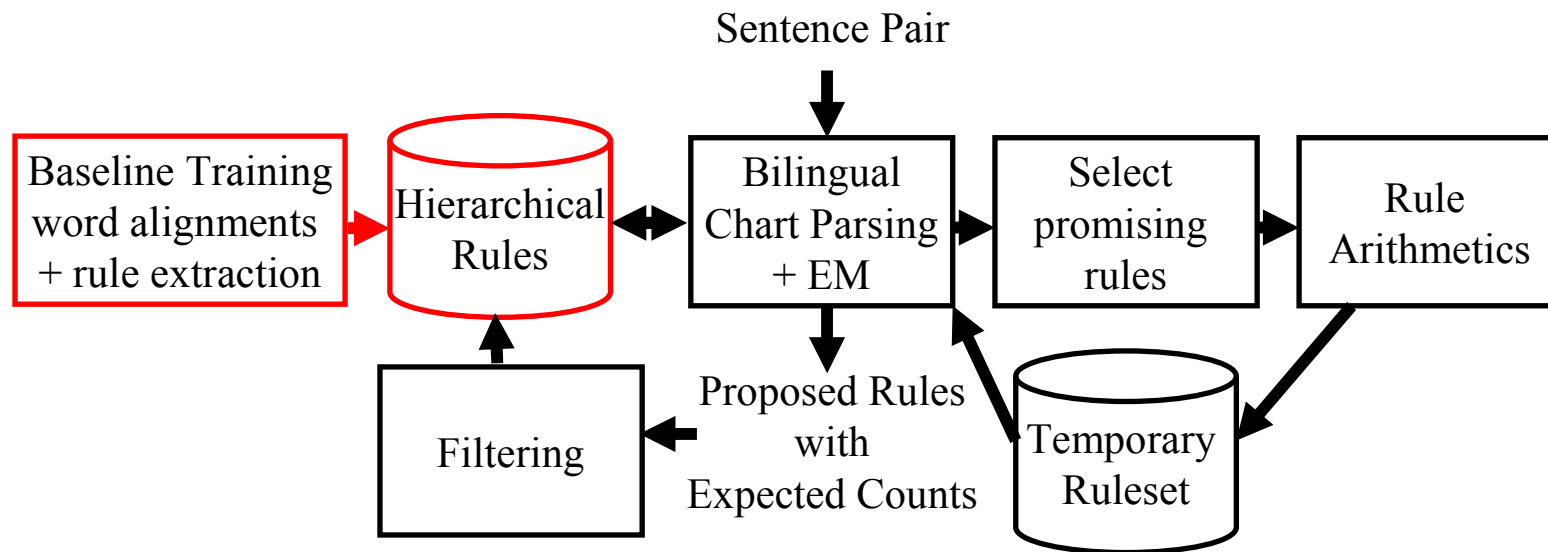
Let's define **rule arithmetic** and propose  $r5 := r6 + r7$

Rule r7 is too general. we like r5 better for its contextual restriction  
(Ich bin zu Hause - I am at home)

## Outline

- **Baseline: Formally syntax-based system (ForSyn)**
- **Bilingual chart parsing with SCFG**
- **Estimating rule probabilities from the parse forest using EM**
- **Improving grammar coverage**
- **Proposing new rules with rule arithmetic**
- **Experiments and results**





# Formally syntax-based system

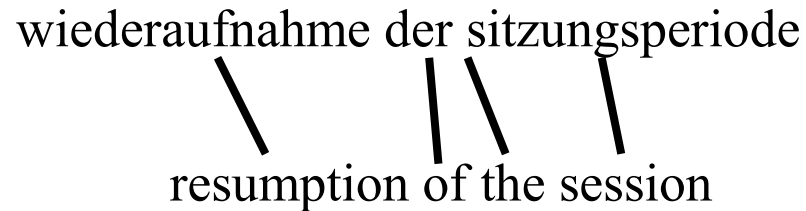
- **Synchronous Context-Free Grammar (SCFG)**

A synchronous rewriting system generating source and target pairs simultaneously  
 $X \rightarrow \langle \gamma, \alpha, \sim \rangle$  only one nterm.  $X, \gamma, \alpha$  are term./nterm. strings,  $\sim$  is coindexation.

- (David Chiang. 2007. Hierarchical phrase-based translation)
- Decoder: ForSyn (Bowen Zhou et al., 2008. Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels)

Parallel corpus

Alignments



**Phrase pairs**

- $X \stackrel{\Omega}{\sim} \langle \text{wiederaufnahme, resumption} \rangle$
- $X \stackrel{\Omega}{\sim} \langle \text{der, of the} \rangle$
- $X \stackrel{\Omega}{\sim} \langle \text{sitzungsperiode, session} \rangle$
- ...

**Abstract rules**

- $X \stackrel{\Omega}{\sim} \langle \text{wiederaufnahme } X_1, \text{resumption } X_1 \rangle$
- $X \stackrel{\Omega}{\sim} \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle$
- ...
- Glue**
- $X \stackrel{\Omega}{\sim} \langle X_1 X_2, X_1 X_2 \rangle$

## Modeling

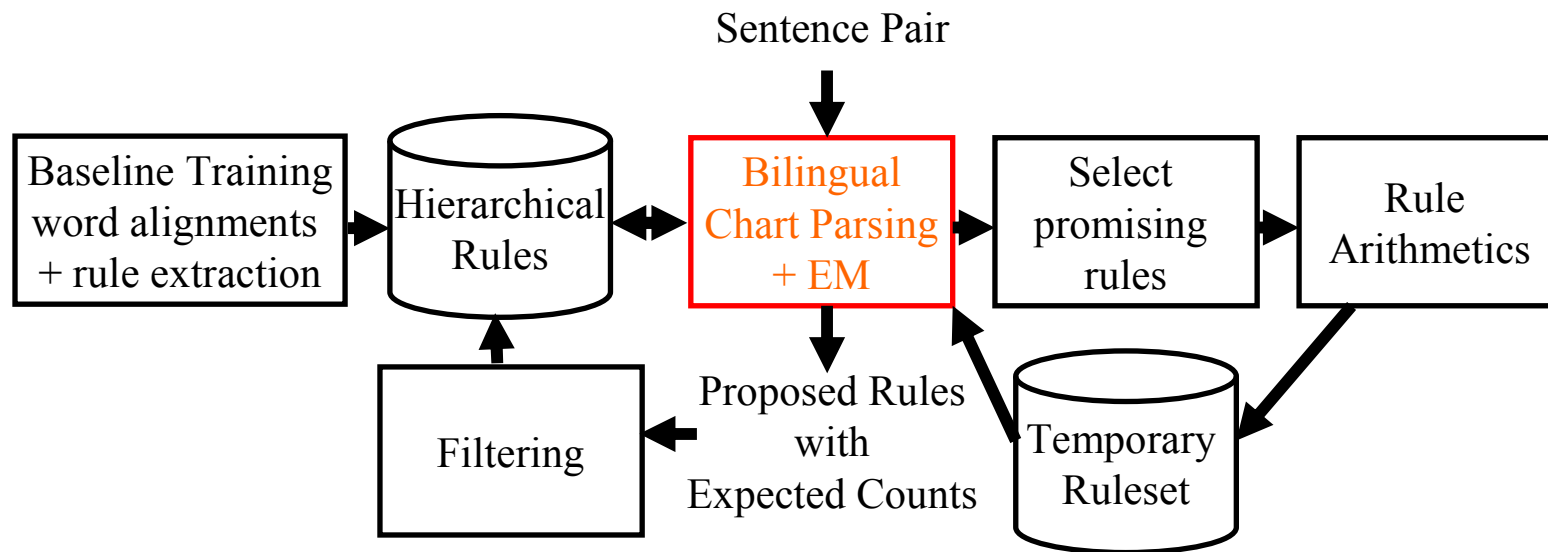
- **Baseline: a log-linear framework including 9 features**
- conditional rule probabilities in both directions:  $P(\alpha|\gamma)$  and  $P(\gamma|\alpha)$
- lexical weights in both directions:  $P_w(\alpha|\gamma)$  and  $P_w(\gamma|\alpha)$
- other 5 features: LM, word bonus, abstraction, rule, and glue penalties

Decoding implemented as searching for the optimal derivation

$$P(D) \propto P_{LM}(e)^{\lambda_{LM}} \times \prod_i \prod_{X \rightarrow \langle \gamma, \alpha \rangle \in D} \phi_i(X \rightarrow \langle \gamma, \alpha \rangle)^{\lambda_i},$$

## ... Modeling

- **Problem with  $P(\alpha|\gamma)$  and  $P(\gamma|\alpha)$** 
  - phrasal rules features can be estimated by maximum likelihood
  - abstract rules counts are not so reliable, eg:
    - $X \stackrel{\Omega}{\sim} \langle \text{wiederaufnahme X1, resumption X1} \rangle$   
was generated from both  
wiederaufnahme der sitzungsperiode – resumption of the session  
and  
wiederaufnahme der – resumption of the
- **Solution: Use EM to estimate SCFG rule probabilities!**
  - Use Bilingual Chart Parsing to obtain parse forest
  - Compute inside, outside probabilities, and expected counts
  - Re-estimate joint probabilities  $P(\alpha, \gamma)$ , then compute  $P(\alpha|\gamma)$  and  $P(\gamma|\alpha)$
  - S. Huang and B. Zhou, “An EM algorithm for SCFG...”, ICASSP’09





## Bilingual Chart Parsing (CKY)

1:wiederaufnahme 2:der 3:sitzungsperiode

1:resumption 2:of 3:the 4:session

chart cell            rule  
[i, j, k, l]

[1,1,1,1] X  $\stackrel{\Omega}{\Leftarrow}$  <wiederaufnahme, resumption>

[3,3,4,4] X  $\stackrel{\Omega}{\Leftarrow}$  <sitzungsperiode, session>

[2,2,2,3] X  $\stackrel{\Omega}{\Leftarrow}$  <der, of the>

[1,2,1,3] X  $\stackrel{\Omega}{\Leftarrow}$  <wiederaufnahme X<sub>1</sub>, resumption X<sub>1</sub>> [2,2,2,3]  
X  $\stackrel{\Omega}{\Leftarrow}$  <X<sub>1</sub> X<sub>2</sub>, X<sub>1</sub> X<sub>2</sub>> [[1,1,1,1],[2,2,2,3]]

[2,3,2,4] X  $\stackrel{\Omega}{\Leftarrow}$  <X<sub>1</sub> X<sub>2</sub>, X<sub>1</sub> X<sub>2</sub>> [[2,2,2,3],[3,3,4,4]]

[1,3,1,4] X  $\stackrel{\Omega}{\Leftarrow}$  <wiederaufnahme X<sub>1</sub>, resumption X<sub>1</sub>> [2,3,2,4]

X  $\stackrel{\Omega}{\Leftarrow}$  <X<sub>1</sub> der X<sub>2</sub>, X<sub>1</sub> of the X<sub>2</sub>> [[1,1,1,1],[3,3,4,4]]

X  $\stackrel{\Omega}{\Leftarrow}$  <X<sub>1</sub> X<sub>2</sub>, X<sub>1</sub> X<sub>2</sub>> [[1,1,1,1],[2,3,2,4]],[[1,2,1,3],[3,3,4,4]]

$RSpans := precompute(R, e, f)$

**for**  $i, j, k, l$  in bottom-up order, such that

$1 \leq i \leq j \leq M,$

$1 \leq k \leq l \leq N$

**for**  $\rho \in RSpans(i, j, k, l)$

**switch**  $\rho.n$

**case 0:**

$t_{ijkl}.push(\rho)$

**case 1:**

**if** ( $filled(t_{\rho.bp1})$ )

$t_{ijkl}.push(\rho)$

**case 2:**

**if** ( $filled(t_{\rho.bp1}) \& filled(t_{\rho.bp2})$ )

$t_{ijkl}.push(\rho)$

**Success!**

## Computing inside probabilities

- Inside probability – probability of generating a parallel sequence from X

$$\beta_{ijkl}(X) = P(X \Rightarrow^* e_i^j; f_k^l)$$

- Can be defined recursively (there is only one non-terminal, so we can omit X)

$$\beta_{ijkl} = \sum_{\rho \in t_{ijkl}} P(\rho.r) \prod_{(i'j'k'l') \in \rho.bp} \beta_{i'j'k'l'}$$

- And computed dynamically while parsing...

$$\beta_{1,1,1,1} = P(X \stackrel{\Omega}{\Rightarrow} \langle \text{wiederaufnahme, resumption} \rangle)$$

...

$$\begin{aligned} \beta_{1,3,1,4} = & P(X \stackrel{\Omega}{\Rightarrow} \langle \text{wiederaufnahme } X_1, \text{ resumption } X_1 \rangle) \beta_{2,3,2,4} \\ & + P(X \stackrel{\Omega}{\Rightarrow} \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle) \beta_{1,1,1,1} \beta_{3,3,4,4} \\ & + P(X \stackrel{\Omega}{\Rightarrow} \langle X_1 X_2, X_1 X_2 \rangle) \beta_{1,1,1,1} \beta_{2,3,2,4} \\ & + P(X \stackrel{\Omega}{\Rightarrow} \langle X_1 X_2, X_1 X_2 \rangle) \beta_{1,2,1,3} \beta_{3,3,4,4} \end{aligned}$$

## Computing outside probabilities

- Outside probability – probability of generating the parallel sequence outside of X.

$$\alpha_{ijkl}(X) = P(S \Rightarrow^* e_1^{i-1}, X, e_{j+1}^M; f_1^{k-1}, X, f_{l+1}^N)$$

- Can be computed by iterating the chart in top-down ordering, starting from the root cell:

$$\alpha_{1,M,1,N} := 1$$

- And propagating the probability mass to backpointed cells for rules with 1 non-terminal:

$$\alpha_{\rho.bp1+} = P(\rho.r)\alpha_{ijkl}$$

$$\alpha_{\rho.bp1+} = P(\rho.r)\alpha_{ijkl}\beta_{\rho.bp2}$$

and 2 non-terminals:

$$\alpha_{\rho.bp2+} = P(\rho.r)\alpha_{ijkl}\beta_{\rho.bp1}$$

So that in the root cell [1,3,1,4] the updates would look like:

$$\alpha_{2,3,2,4} += P(X \stackrel{\Omega}{=} \langle \text{wiederaufnahme } X_1, \text{ resumption } X_1 \rangle) * 1$$

$$\alpha_{1,1,1,1} += P(X \stackrel{\Omega}{=} \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle) * 1 * \beta_{3,3,4,4}$$

$$\alpha_{3,3,4,4} += P(X \stackrel{\Omega}{=} \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle) * 1 * \beta_{1,1,1,1}$$

$$\alpha_{1,1,1,1} += P(X \stackrel{\Omega}{=} \langle X_1 X_2, X_1 X_2 \rangle) * 1 * \beta_{2,3,2,4}$$

$$\alpha_{2,3,2,4} += P(X \stackrel{\Omega}{=} \langle X_1 X_2, X_1 X_2 \rangle) * 1 * \beta_{1,1,1,1}$$

$$\alpha_{1,2,1,3} += P(X \stackrel{\Omega}{=} \langle X_1 X_2, X_1 X_2 \rangle) * 1 * \beta_{3,3,4,4}$$

$$\alpha_{3,3,4,4} += P(X \stackrel{\Omega}{=} \langle X_1 X_2, X_1 X_2 \rangle) * 1 * \beta_{1,2,1,3}$$

## Computing expected counts

- Contributions to rule expected counts can be computed by iterating the chart in any ordering

$$c(\rho.r)_+ = \frac{P(\rho.r)\alpha_{ijkl} \prod_{i=1}^{\rho.n} \beta_{\rho.bp_i}}{\beta_{1,M,1,N}},$$

- Collect counts from the whole training corpus.

## M-step

- Finally, joint probabilities are re-estimated as

$$P(r) = \frac{c(r)}{\sum_{r' \in R: L(r')=L(r)} c(r')}$$

L(r) means the left-hand side of the rule r. (Always X, trivial)

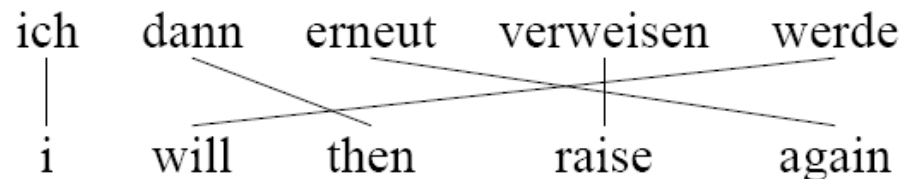
- Conditional probabilities P( $\alpha$ | $\gamma$ ) and P( $\gamma$ | $\alpha$ ) are computed as normalized probabilities of rules with the same source and target side, respectively.

## P(r) for scoring

- The joint probability of a rule is normalized for length by  $c_{size}(s)^{s-1}$
- Combination of all rule features is used (lexical weights are important too)

## Improving the grammar coverage

- **Many sentences cannot be parsed**
  - On Europarl data, 70% for *union*, 20% for *grow-diag-final*
  - Structural complexity



Every parse needs one of the following rules:

$X \stackrel{\Omega}{\Leftarrow} \langle \text{erneut } X_1, X_1 \text{ again} \rangle$

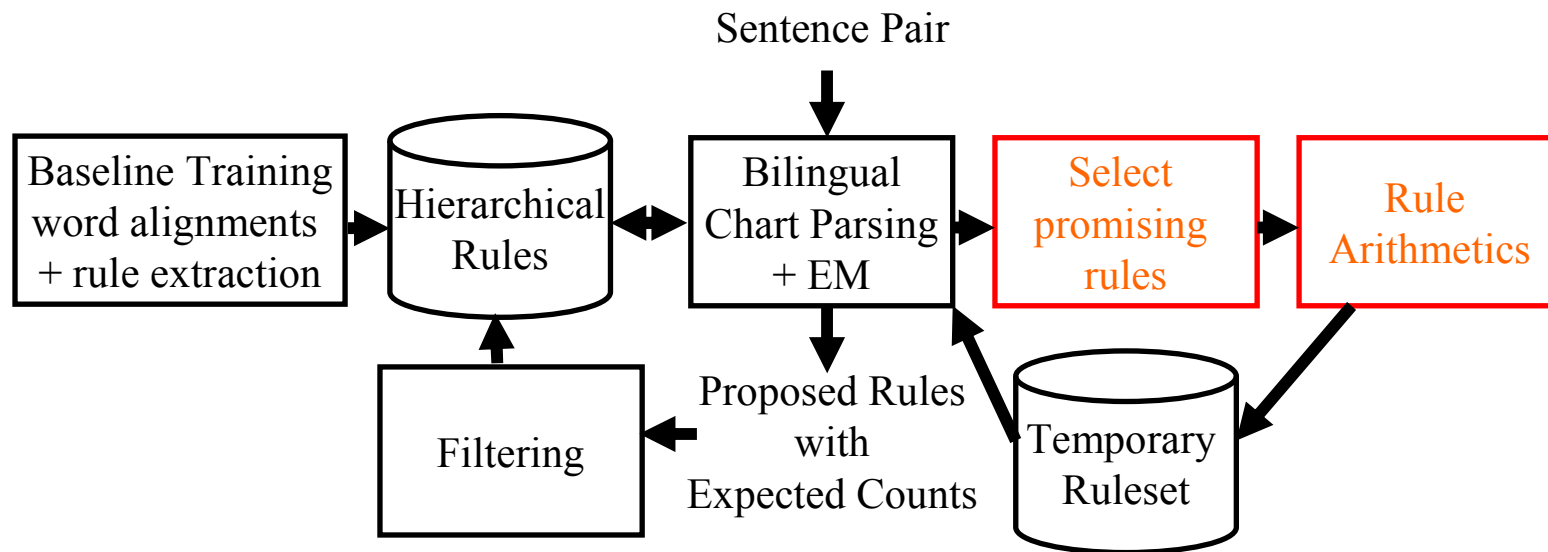
$X \stackrel{\Omega}{\Leftarrow} \langle X_1 \text{ verweisen}, \text{ raise } X_1 \rangle$

- Low frequency words and asymmetric translation pairs

**nach monatelangen und** weltweiten konsultationen wird nun im donaldson ...  
**and after consulting world wide for many months** the donaldson report ...

As a consequence, either the whole sentence pair cannot be parsed (losing expected counts for other words), or another asymmetric rules (trying to fix weltweiten konsultationen) are boosted

- Add Swap Glue to increase parsability:  $X \stackrel{\Omega}{\Leftarrow} \langle X_1 X_2, X_2, X_1 \rangle$
- Insertion and deletion rules (ITG - like)  
 $X \stackrel{\Omega}{\Leftarrow} \langle X_1, X_1 e \rangle, X \stackrel{\Omega}{\Leftarrow} \langle X_1, e X_1 \rangle, X \stackrel{\Omega}{\Leftarrow} \langle X_1 f, X_1 \rangle, X \stackrel{\Omega}{\Leftarrow} \langle f X_1, X_1 \rangle$



## Proposing new rules

- **Parse the sentence pair, estimate expected counts**
  
- **Select the “most promising” rule usages**
  - Currently selected as productions  $p$  with the highest contributions to expected counts
  
- **Use Rule arithmetic to combine rules**
  - 1) create span projections – for both rules
  - 2) merge span projections
  - 3) collect rules
  
- **Parse again, this time also using new proposed rules, estimate expected counts**

die herausforderung besteht darin diese systeme zu den besten der welt zu machen

the challenge is to make the system the very best

$$\begin{aligned}
 & X \stackrel{\Omega}{\sim} \langle \text{diese, the} \rangle \\
 + & X \stackrel{\Omega}{\sim} \langle \text{systeme, system} \rangle \\
 = & X \stackrel{\Omega}{\sim} \langle \text{diese systeme, the system} \rangle
 \end{aligned}$$

“Rule arithmetic Hello world”  
Combining phrasal rules



die herausforderung besteht darin diese **systeme** zu den besten der welt zu machen

the challenge is to make the **system** the very best

.....

.....

$X \stackrel{\Omega}{\Rightarrow} \langle \text{die, the} \rangle$

+  $X \stackrel{\Omega}{\Rightarrow} \langle \text{systeme, system} \rangle$

= *undefined*

Combining phrasal rules  
(contiguous projections required)

die herausforderung besteht darin diese systeme zu den besten der welt zu machen

the challenge is to make the system the very best

$X \stackrel{\Omega}{\Rightarrow} \langle \text{zu machen, to make} \rangle$

+  $X \stackrel{\Omega}{\Rightarrow} \langle X_1 X_2, X_2 X_1 \rangle$

=  $X \stackrel{\Omega}{\Rightarrow} \langle X_1 \text{ zu machen, to make } X_1 \rangle$

Combining phrasal rule  
and swap glue.

die herausforderung besteht darin diese systeme zu den besten der welt zu machen

the challenge is to make the system the very best

$X \xrightarrow{\Omega} \langle \text{besteht darin, is} \rangle$

+  $X \xrightarrow{\Omega} \langle X_1 \text{ zu } X_2, \text{to } X_2 X_1 \rangle$

=  $X \xrightarrow{\Omega} \langle \text{besteht darin } X_1 \text{ zu } X_2, \text{is to } X_2 X_1 \rangle$

Combining phrasal rule  
and abstract rule with  
2 non-terminals.

die herausforderung besteht darin diese systeme zu den besten der welt zu machen

the challenge is to make the system the very best

$X \stackrel{\Omega}{\Leftarrow} \langle \text{diese } X_1, \text{the } X_1 \rangle$

+  $X \stackrel{\Omega}{\Leftarrow} \langle X_1 \text{ zu } X_2, \text{to } X_2 X_1 \rangle$

=  $X \stackrel{\Omega}{\Leftarrow} \langle \text{diese } X_1 \text{ zu } X_2, \text{to } X_2 \text{ the } X_1 \rangle$

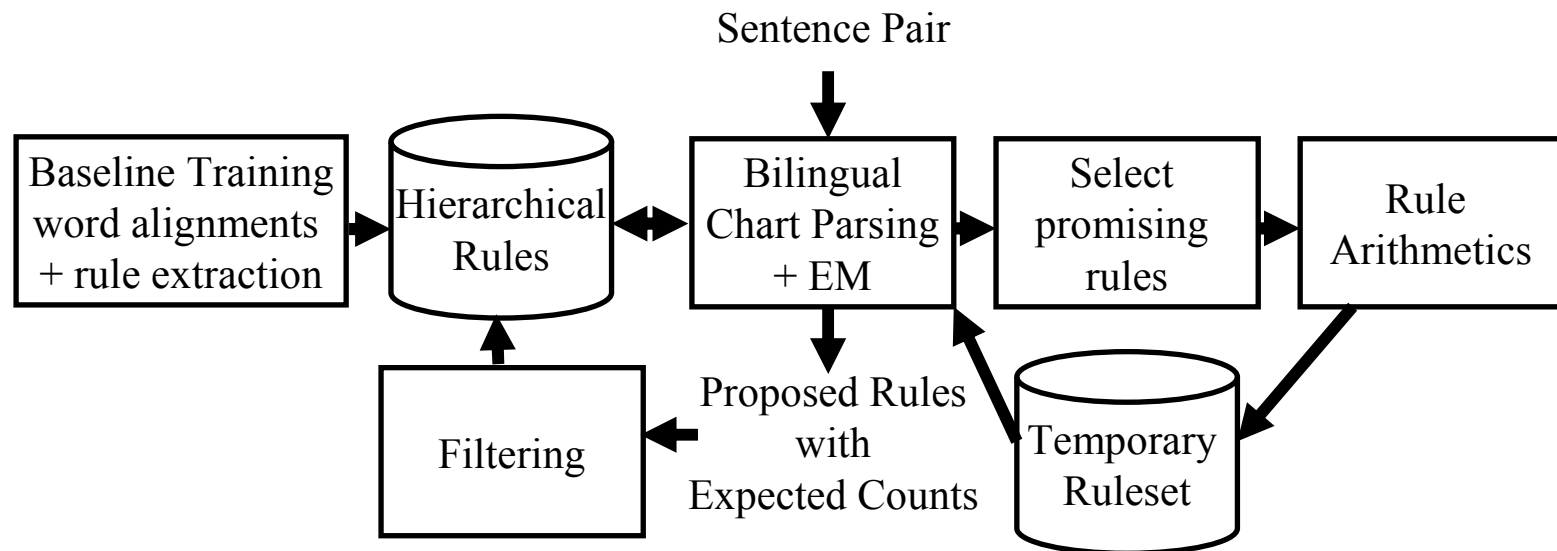
Combining 2 abstract rules

Note that the span  $X_1$  of the second rule was shortened

## Experiment

### ■ Experiment

- Presenting 3 sets of results (BLEU)
  - 11 Iterations of EM training
  - Proposing new rules after the 1<sup>st</sup> iteration (adding new rules to the baseline)
    - filtering (using only rules proposed at least from 2 sentence pairs)
  - + 1 iteration of EM with new rules



## Results (BLEU)

### German-English data from the Europarl

- 297k sentence pairs, lowercase, punctuation removed
- 1k dev set, 1 reference
- 1k test set, 1 reference
- 13M baseline rules
- 100k proposed rules

### Farsi-English conversational data

- 1.4k dev set, 1 reference
- 417 test set, 4 references
- 11.3M baseline rules
- 120k proposed rules

	Ger-En dev	Ger-En test	Farsi-En dev	Farsi-En test
<b>baseline</b>	23.852	25.447	41.095	38.248
<b>EM i0</b>	24.394	<b>26.122</b>	40.764	<b>39.114</b>
<b>i1</b>	24.365	25.826	41.295	38.509
<b>i2</b>	24.433	25.936	41.424	38.238
<b>i3</b>	24.375	26.047	41.339	39.339
<b>i4</b>	24.409	25.936	41.563	39.594
<b>i5</b>	24.300	26.259	41.557	39.242
<b>i6</b>	24.339	26.197	41.726	39.209
<b>i7</b>	23.985	25.827	41.542	39.445
<b>i8</b>	24.129	<b>26.305</b>	41.634	39.326
<b>i9</b>	23.988	25.940	41.422	39.583
<b>i10</b>	24.079	26.226	41.287	<b>39.650</b>
<b>proposed i0</b>	24.418	<b>26.122</b>	40.729	38.382
<b>proposed i0 + EM</b>	24.837	<b>26.408</b>	41.836	<b>40.246</b>

## Conclusion

- Introduced algorithms for bilingual parsing, and estimation of rule probabilities
- Presented a new method for synthesizing new rules from the most confident rules within the parse forest
- The method is independent of bilingual word alignments and complementary to heuristic alignment-based approaches
- Presented results on conversational data from two different language pairs: +1 BLEU on German-English translations of Europarl data, +2.00 BLEU on Farsi.

Thank you



## Example of rules proposed by rule arithmetic

	1 ...	
	um $X_1$	in order $X_1$
	natuerlich $X_1$	of course $X_1$
	deshalb $X_1$	this is why $X_1$
	$X_1$ zu koennen	to $X_1$
	$X_1$ ist	it is $X_1$
nach der tagesordnung folgt die	$X_1$	the next item is the $X_1$
herr $X_1$ herr kommissar	$X_2$	mr $X_1$ commissioner $X_2$
die $X_1$ der	$X_2$	$X_1$ the $X_2$
im gegenteil	$X_1$	on the contrary $X_1$
nach der tagesordnung folgt	$X_1$	the next item is $X_1$
$X_1$ die	$X_2$	the $X_1$ the $X_2$
die $X_1$ die		the $X_1$
ausserdem	$X_1$	in addition $X_1$
daher	$X_1$	that is why $X_1$
wir	$X_1$ nicht $X_2$	we $X_1$ not $X_2$
die $X_1$ der	$X_2$	the $X_2$ $X_1$
deshalb	$X_1$	for this reason $X_1$
um $X_1$ zu	$X_2$	to $X_2$ $X_1$
$X_1$ nicht $X_2$ werden		$X_1$ not be $X_2$

## ... Example of rules proposed by rule arithmetic

	<b>... 50001</b>	
nach der tagesordnung folgt die $X_1$ ueber		the next item is $X_1$ on
das $X_1$		$X_1$ that for
wird $X_1$ zur $X_2$		$X_1$ to $X_2$
das parlament nimmt den entwurf einer legislativen		parliament adopted the draft legislative
sehr		it is very
$X_1$ und herren abgeordneten $X_2$		$X_1$ and gentlemen $X_2$
ich habe nicht $X_1$		i do not have $X_1$
es $X_1$		there $X_1$ the
die $X_1$ von lissabon		the lisbon $X_1$
$X_1$ dabei		in this $X_1$
frau kommissarin $X_1$ moechte $X_2$		commissioner $X_1$ would like to $X_2$
	<b>... 99991</b>	
$X_1$ genehmigt		$X_1$ approved the
$X_1$ koennen nicht		$X_1$ cannot have
diese		for them
auch		also needs
dass sich		believe
vorgeschlagen		have proposed
sie $X_1$ ein		you $X_1$
es ist		they are
diese vorschlaege gestimmt		voted in favour of these proposals
$X_1$ rechnungshof		$X_1$ auditors

## Results – German 2 English

### ▪ Data

- German-English data from the Europarl corpus.
  - 297k sentence pairs,
  - 1k dev set, 1 reference
  - 1k test set, 1 reference
  - 13M baseline rules
  - 100k proposed rules

	Ger-En dev	Ger-En test
<b>baseline</b>	23.852	25.447
<b>i0</b>	24.394	<b>26.122</b>
<b>i1</b>	24.365	25.826
<b>i2</b>	24.433	25.936
<b>i3</b>	24.375	26.047
<b>i4</b>	24.409	25.936
<b>i5</b>	24.300	26.259
<b>i6</b>	24.339	26.197
<b>i7</b>	23.985	25.827
<b>i8</b>	24.129	<b>26.305</b>
<b>i9</b>	23.988	25.940
<b>i10</b>	24.079	26.226
<b>proposed i0</b>	24.418	<b>26.122</b>
<b>proposed i0 + EM</b>	24.837	<b>26.408</b>

## Results - Farsi

- **Data**
  - Farsi-English conversational data
    - 1439 dev set, 1 reference
    - 417 test set, 4 references

	Farsi-En dev	Farsi-En test
<b>baseline</b>	41.095	38.248
<b>i0</b>	40.764	<b>39.114</b>
<b>i1</b>	41.295	38.509
<b>i2</b>	41.424	38.238
<b>i3</b>	41.339	39.339
<b>i4</b>	41.563	39.594
<b>i5</b>	41.557	39.242
<b>i6</b>	41.726	39.209
<b>i7</b>	41.542	39.445
<b>i8</b>	41.634	39.326
<b>i9</b>	41.422	39.583
<b>i10</b>	41.287	<b>39.650</b>
<b>proposed i0</b>	40.729	38.382
<b>proposed i0 + EM</b>	41.836	<b>40.246</b>

## Parsing efficiency

- **The efficient implementation of  $RSpans(i,j,k,l)$  is very important.**
- **Create 2 prefix trees to encode all rule sides relevant to the sentence pair. And to store their spans, and spans of their non-terminals.**

$[1,2,1,3] X \stackrel{\Omega}{\Leftarrow} \langle \text{wiederaufnahme } X_1, \text{resumption } X_1 \rangle [2,2,2,3]$

$[1,3,1,4] X \stackrel{\Omega}{\Leftarrow} \langle \text{wiederaufnahme } X_1, \text{resumption } X_1 \rangle [2,3,2,4]$

- **The most time consuming is the glue rule, since there is no non-terminal between  $X_1$  and  $X_2$**

## Discussion, future directions

- **Speed**
  - ~5s/sentence
  - Pruning?
  - In literature, usually, parsing is done once, then 50 – 100 EM iterations, remembering the parse forest.  
Can we update the forest just for proposed rules?
  
- **Additional features, more linguistic information**
  
- **In monolingual parsing, rules are extracted from a treebank, then EM runs on a different data...**
  
- **Operation of subtraction?**

## Acknowledgements

- **Bowen Zhu**
  - discussions of the BCP and rule arithmetics, ForSyn decoder
- **Bing Xiang**
  - discussions over the BCP
- **Songfang Huang**
  - initial EM implementation
  
- **and to the whole S2S team!**