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

IWSLT '09, Tokyo

Dec 1, 2009

© 2009 IBM Corporation

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 ≙ <machen, make>

r2: X ≙ <die herausforderung, the challenge>

r3: X ≙ <diese systeme, the system>

r4: X ≙ <zu den besten der welt, the very best>

r5: X ≙ <besteht darin X_1 zu X_2 ,is to X_2 X_1>

r6: X ≙ <besteht darin, is>

r7: X ≙ <X_1 zu X_2 , to X_2 X_1>

glue: X ≙ <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 aritmetic** and propose r5 := r6 + r7*

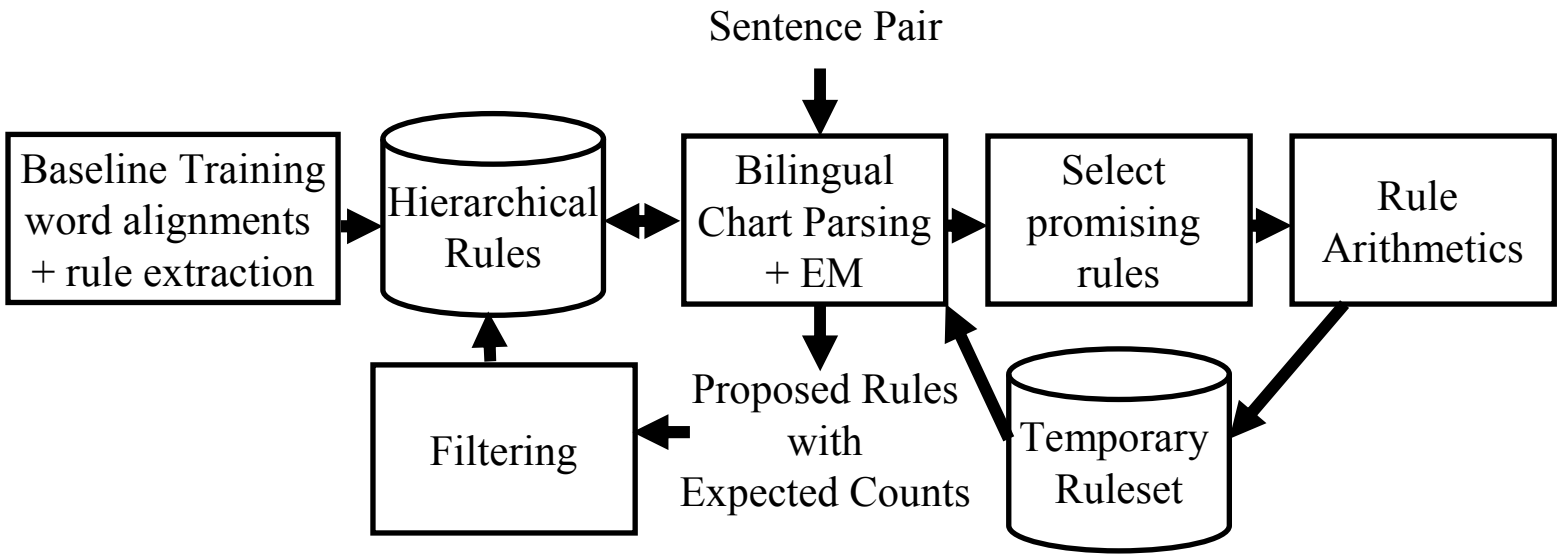
Rule r7 is too general. We like r5 better for its contextual restriction

(Ich bin zu Hause – I am at home)

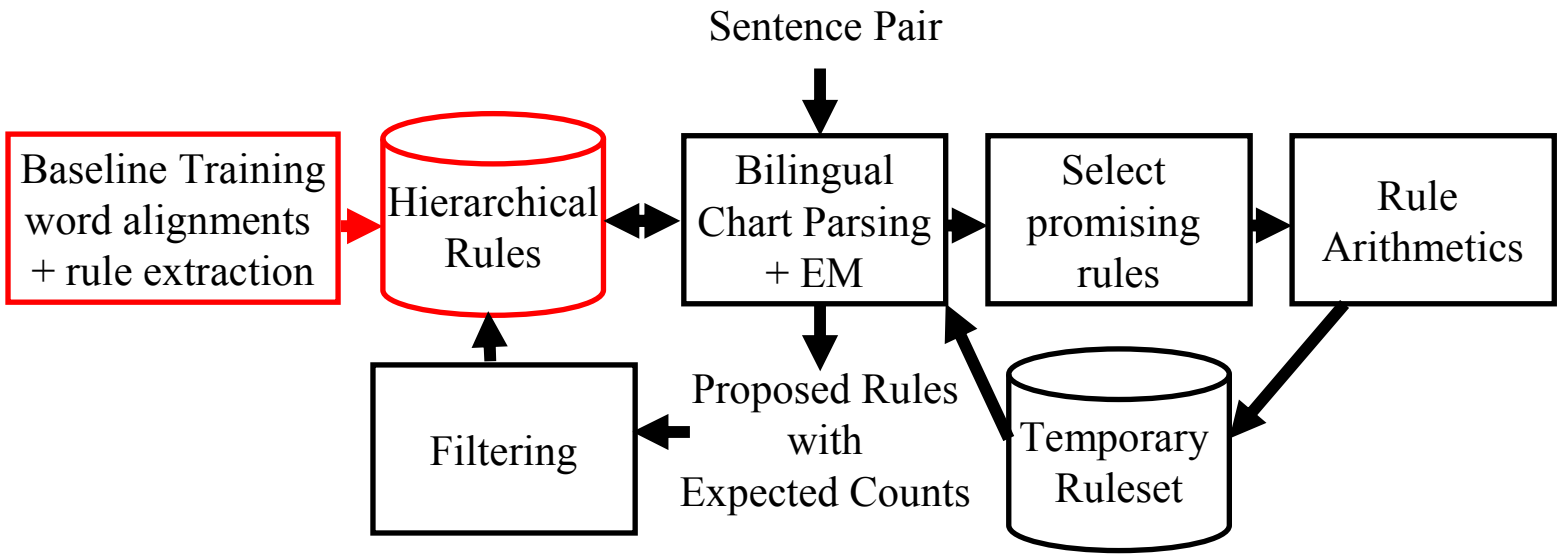
© 2009 IBM Corporation

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**



© 2009 IBM Corporation



© 2009 IBM Corporation

Formally syntax-based system

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

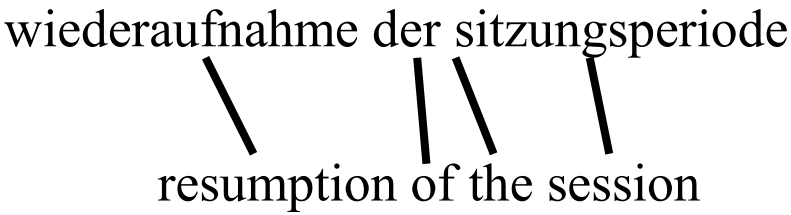
  A synchronous rewriting system generating source and target pairs simultaneously

  $X \rightarrow \langle \gamma, \alpha, \sim \rangle$ only one nterm. X, γ, α are term./nterm. strings, ~ 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)

Paralel corpus

Alignments



Phrase pairs

X ≙ <wiederaufnahme, resumption>

X ≙ <der, of the>

X ≙ <sitzungsperiode, session>

…

Abstract rules

X ≙ <wiederaufnahme X_1, resumption X_1>

X ≙ <X_1 der X_2 , X_1 of the X_2>

…

Glue

X ≙ <X_1 X_2 , X_1 X_2>

© 2009 IBM Corporation

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

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

© 2009 IBM Corporation

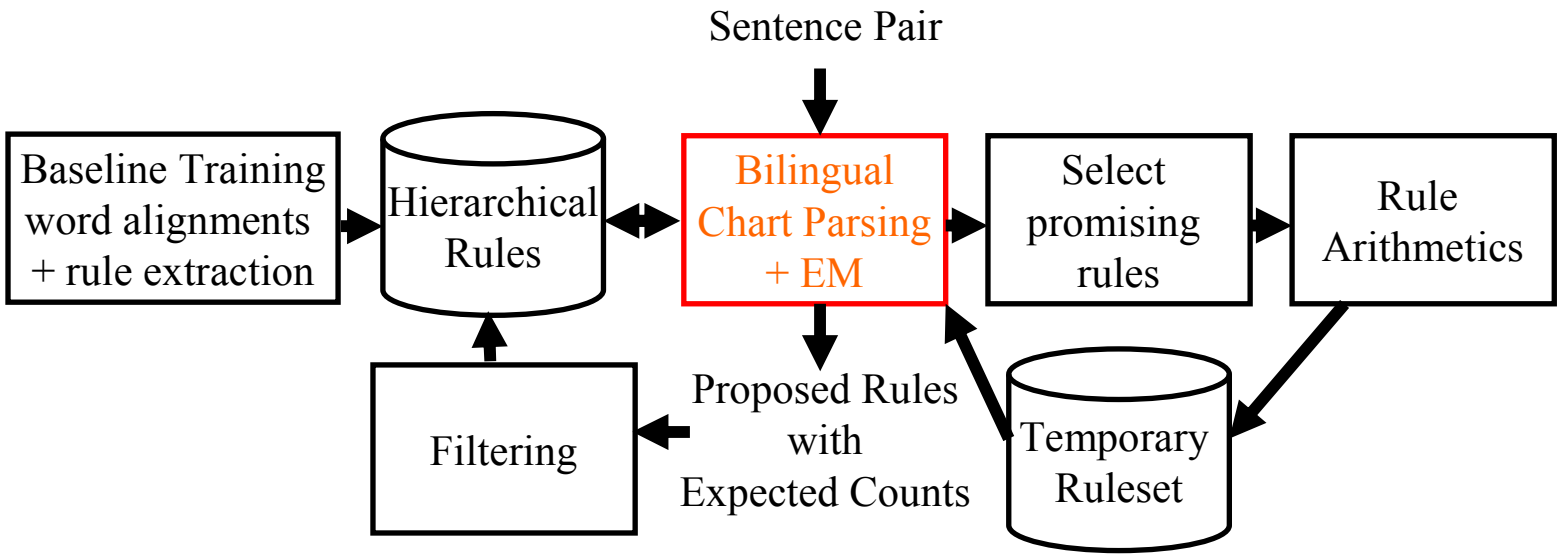
… Modeling

- **Problem with P(α|γ) and P(γ|α)**
  - phrasal rules features can be estimated by maximum likelihood
  - abstract rules counts are not so reliable, eg:

    X ≙ <wiederaufnahme X1, resumption X1>
    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(α,γ) , then compute P(α|γ) and P(γ|α)
  - S. Huang and B. Zhou, "An EM algorithm for SCFG… ", ICASSP'09

© 2009 IBM Corporation



© 2009 IBM Corporation

Bilingual Chart Parsing (CKY)

1:wiederaufnahme 2:der 3:sitzungsperiode
1:resumption 2:of 3:the 4:session

chart cell [i, j, k, l] — rule

[1,1,1,1] X ≙ <wiederaufnahme, resumption>

[3,3,4,4] X ≙ <sitzungsperiode, session>

[2,2,2,3] X ≙ <der, of the>

[1,2,1,3] X ≙ <wiederaufnahme X_1, resumption X_1> [2,2,2,3]
X ≙ <X_1 X_2 , X_1 X_2> [[1,1,1,1],[2,2,2,3]]

[2,3,2,4] X ≙ <X_1 X_2 , X_1 X_2> [[2,2,2,3],[3,3,4,4]]

[1,3,1,4] X ≙ <wiederaufnahme X_1, resumption X_1> [2,3,2,4]
X ≙ <X_1 der X_2 , X_1 of the X_2> [[1,1,1,1],[3,3,4,4]]
X ≙ <X_1 X_2 , X_1 X_2> [[1,1,1,1],[2,3,2,4]],[[1,2,1,3],[3,3,4,4]]

Success!

$RSpans := precompute(R, \mathbf{e}, \mathbf{f})$
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 $(filled(t_{\rho.bp1})\&filled(t_{\rho.bp2}))$
$t_{ijkl}.push(\rho)$

© 2009 IBM Corporation

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 ≙ <wiederaufnahme, resumption>)

…

$\beta_{1,3,1,4}$= P(X ≙ <wiederaufnahme X_1, resumption X_1>) $\beta_{2,3,2,4}$
+ P(X ≙ <X_1 der X_2 , X_1 of the X_2>) $\beta_{1,1,1,1}$ $\beta_{3,3,4,4}$
+ P(X ≙ <X_1 X_2 , X_1 X_2>) $\beta_{1,1,1,1}$ $\beta_{2,3,2,4}$
+ P(X ≙ <X_1 X2 , X_1 X_2>) $\beta_{1,2,1,3}$ $\beta_{3,3,4,4}$

© 2009 IBM Corporation

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

  and 2 non-terminals:
$$\alpha_{\rho.bp1} += P(\rho.r)\alpha_{ijkl}\beta_{\rho.bp2}$$
$$\alpha_{\rho.bp2} += P(\rho.r)\alpha_{ijkl}\beta_{\rho.bp1}$$

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

  $\alpha_{2,3,2,4}$ += P(X ≙ <wiederaufnahme X_1, resumption X_1>) * 1

  $\alpha_{1,1,1,1}$ += P(X ≙ <X_1 der X_2 , X_1 of the X_2>) * 1 * $\beta_{3,3,4,4}$
  $\alpha_{3,3,4,4}$ += P(X ≙ <X_1 der X_2 , X_1 of the X_2>) * 1 * $\beta_{1,1,1,1}$

  $\alpha_{1,1,1,1}$ += P(X ≙ <X_1 X_2 , X_1 X_2>) * 1 * $\beta_{2,3,2,4}$
  $\alpha_{2,3,2,4}$ += P(X ≙ <X_1 X_2 , X_1 X_2>) * 1 * $\beta_{1,1,1,1}$

  $\alpha_{1,2,1,3}$ += P(X ≙ <X_1 X_2 , X_1 X_2>) * 1 * $\beta_{3,3,4,4}$
  $\alpha_{3,3,4,4}$ += P(X ≙ <X_1 X_2 , X_1 X_2>) * 1 * $\beta_{1,2,1,3}$

© 2009 IBM Corporation

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(α|γ) and P(γ|α) are computed as normalized probabilities of rules with the same source and target side, respectively.

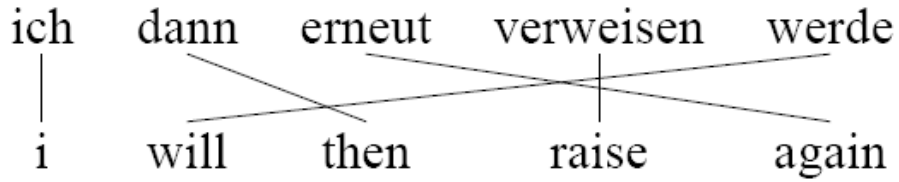
P(r) for scoring

- 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)

© 2009 IBM Corporation

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:
$X \triangleq \langle \text{erneut } X_1 , X_1 \text{ again}$
$X \triangleq \langle X_1 \text{ verweisen} , \text{raise } X_1 \rangle$

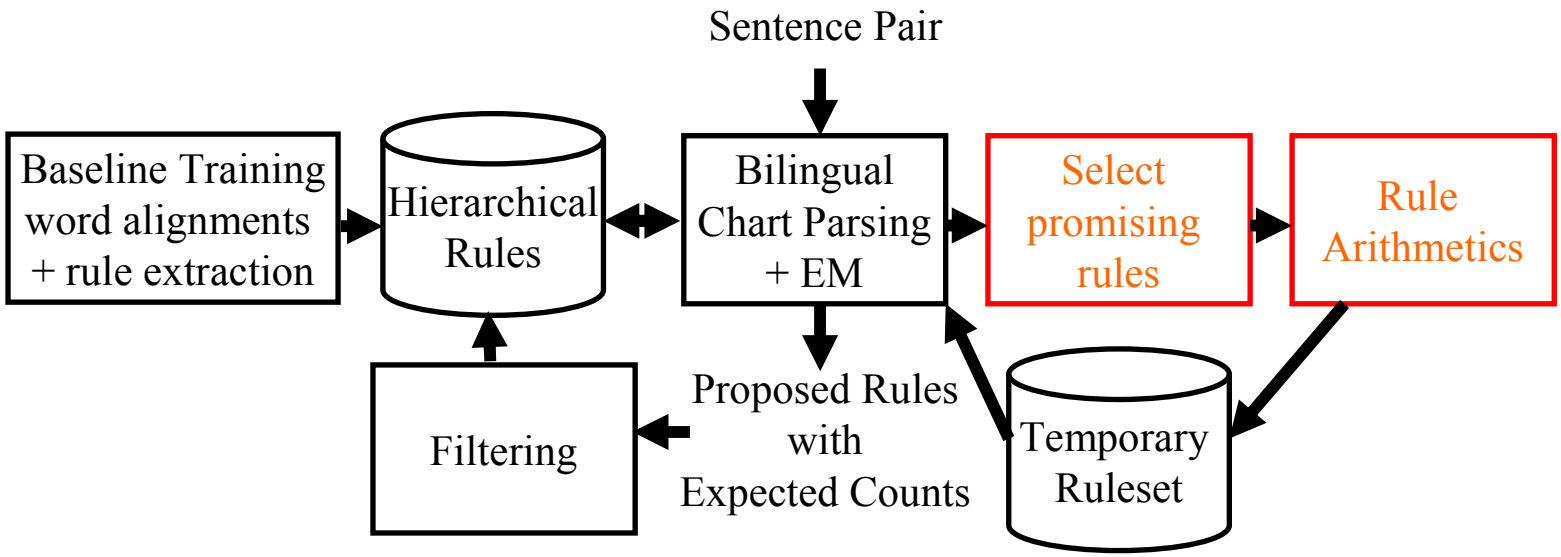
  - Low frequency words and asymmetric translation pairs

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 konsultationnen) are boosted

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

© 2009 IBM Corporation



© 2009 IBM Corporation

Proposing new rules

- **Parse the sentence pair, estimate expected counts**
- **Select the "most promising" rule usages**
  - Currently selected as productions ρ 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**

© 2009 IBM Corporation

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

the challenge is to make the system the very best

X ≙ <diese, the>

\+ X ≙ <systeme, system>

= X ≙ <diese systeme, the system>

"Rule arithmetic Hello world"
Combining phrasal rules

© 2009 IBM Corporation

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

the challenge is to make the system the very best

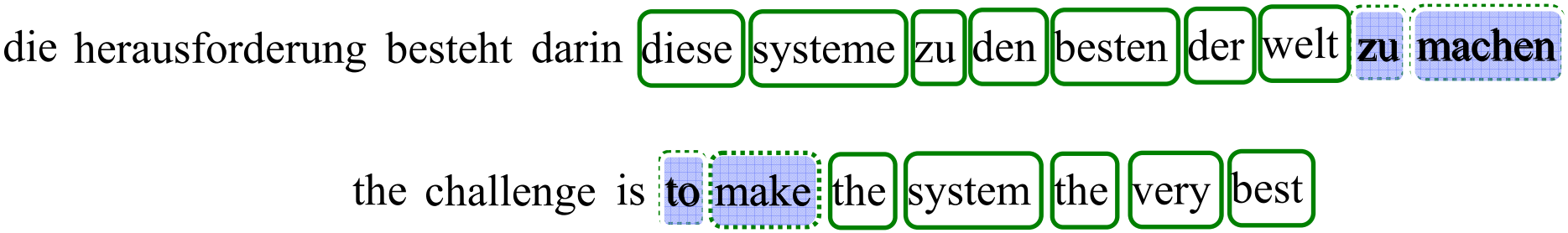
X ≙ <die, the>

\+ X ≙ <systeme, system>

= *undefined*

Combining phrasal rules
(contiguous projections required)

© 2009 IBM Corporation



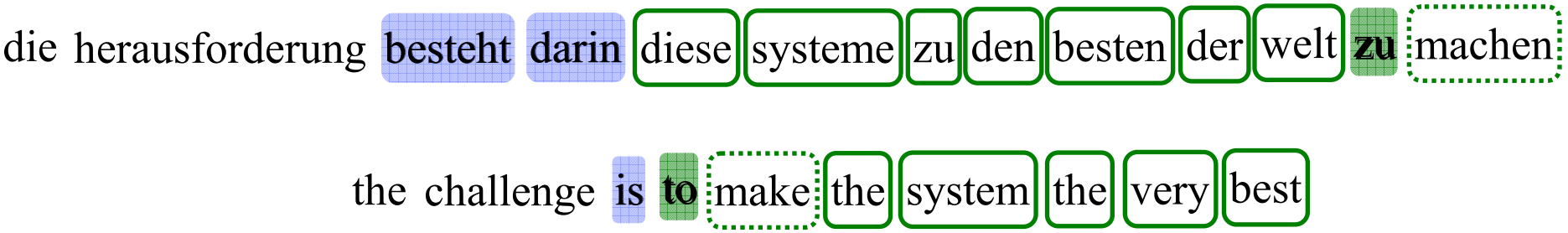
$X \triangleq \langle$zu machen, to make$\rangle$

$+ \quad X \triangleq \langle X_1 X_2, X_2 X_1 \rangle$

$= \quad X \triangleq \langle X_1$ zu machen, to make $X_1 \rangle$

Combining phrasal rule and swap glue.

© 2009 IBM Corporation



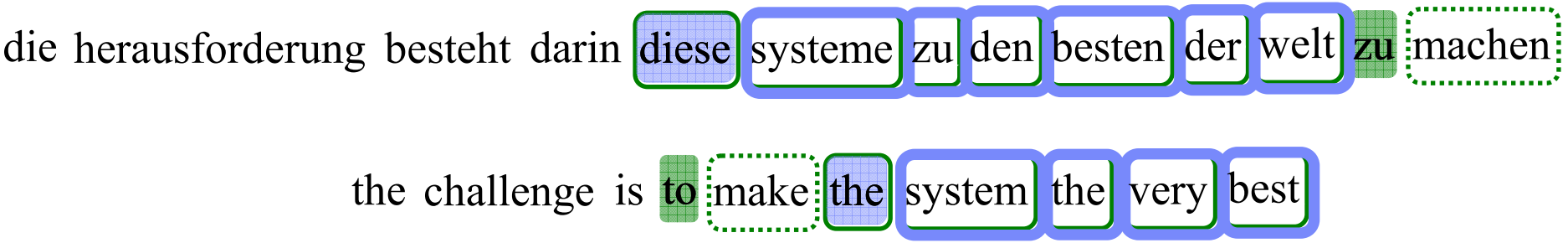
$X \triangleq \langle$besteht darin, is$\rangle$

\+ $X \triangleq \langle X_1$ zu X_2, to $X_2 X_1 \rangle$

= $X \triangleq \langle$besteht darin X_1 zu X_2, is to X_2 $X_1 \rangle$

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

© 2009 IBM Corporation



$X \triangleq \langle \text{diese } X_1, \text{ the } X_1 \rangle$

$+ \quad X \triangleq \langle X_1 \text{ zu } X_2, \text{ to } X_2 X_1 \rangle$

$= \quad X \triangleq \langle \text{diese } X_1 \text{ zu } X_2, \text{ to } X_2 \text{ the } X_1 \rangle$

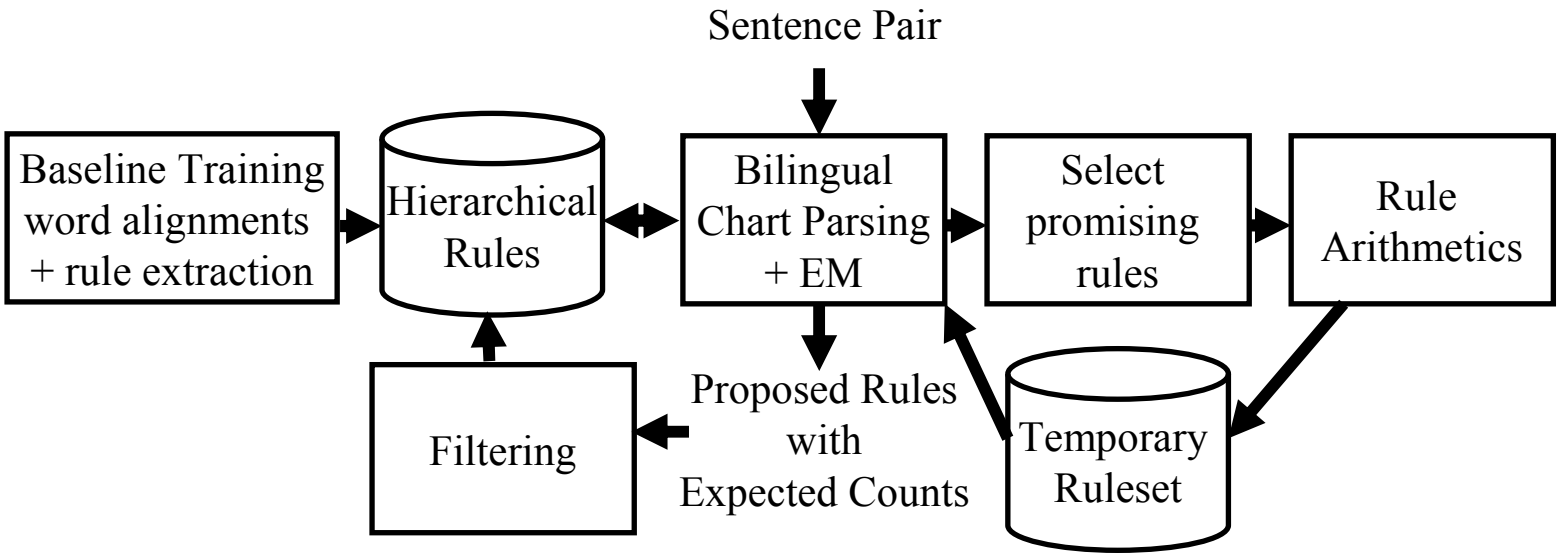
Combining 2 abstract rules

Note that the span X_1 of the second rule was shortened

© 2009 IBM Corporation

Experiment

- **Experiment**
  - Presenting 3 sets of results (BLEU)
    - 11 Iterations of EM training
    - Proposing new rules after the 1st 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



© 2009 IBM Corporation

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

© 2009 IBM Corporation

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.

© 2009 IBM Corporation

Thank you

 © 2009 IBM Corporation

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

© 2009 IBM Corporation

… Example of rules proposed by rule arithmetic

... 50001	
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
... 99991	
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

© 2009 IBM Corporation

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

© 2009 IBM Corporation

Results - Farsi

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

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

© 2009 IBM Corporation

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.**

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

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

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

© 2009 IBM Corporation

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?**

© 2009 IBM Corporation

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!**

© 2009 IBM Corporation