A Combination of Hierarchical Systems with Forced Alignments from Phrase-Based Systems

IWSLT 2010, Paris, France

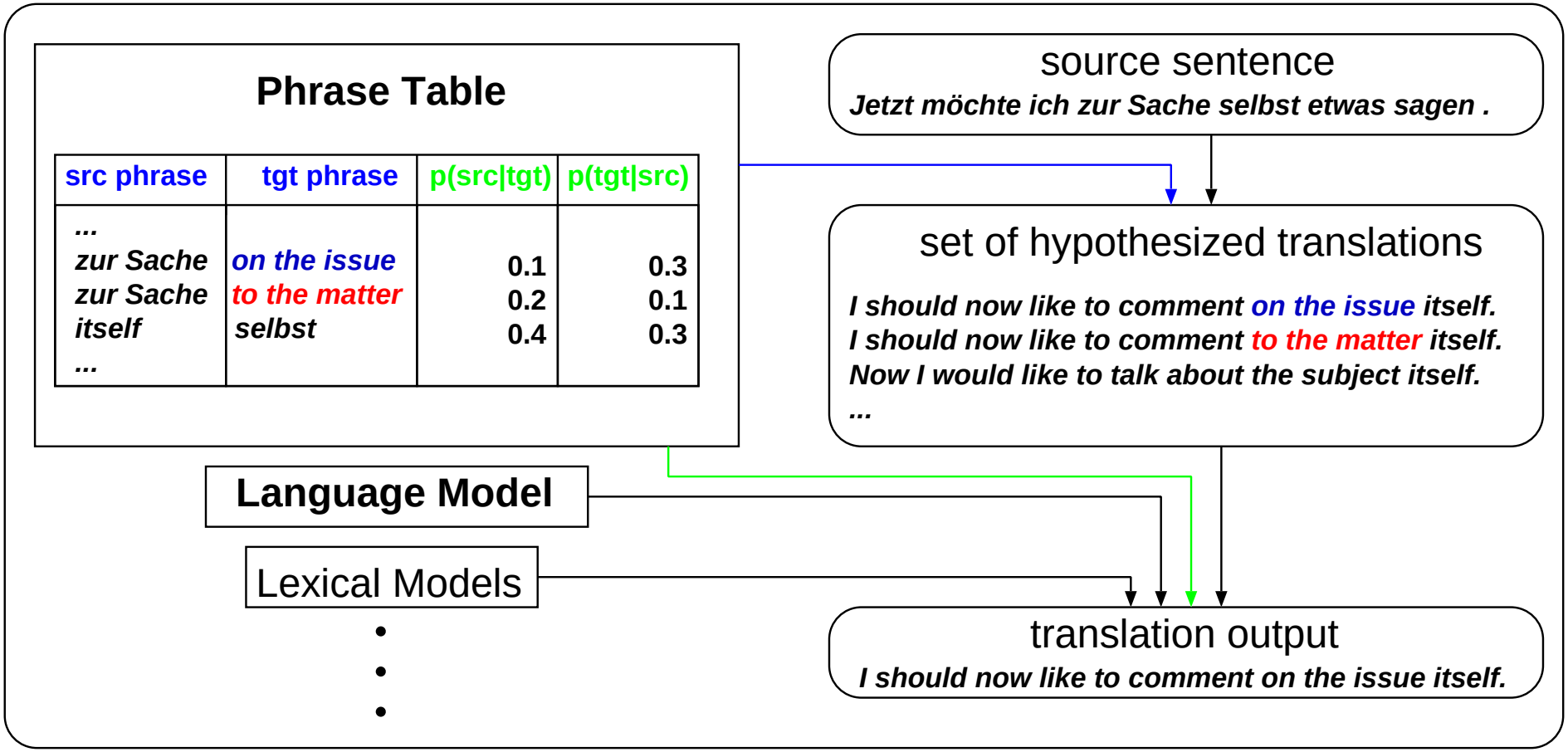
Carmen Heger, Joern Wuebker, David Vilar, Hermann Ney
<surname>@cs.rwth-aachen.de

December 3, 2010

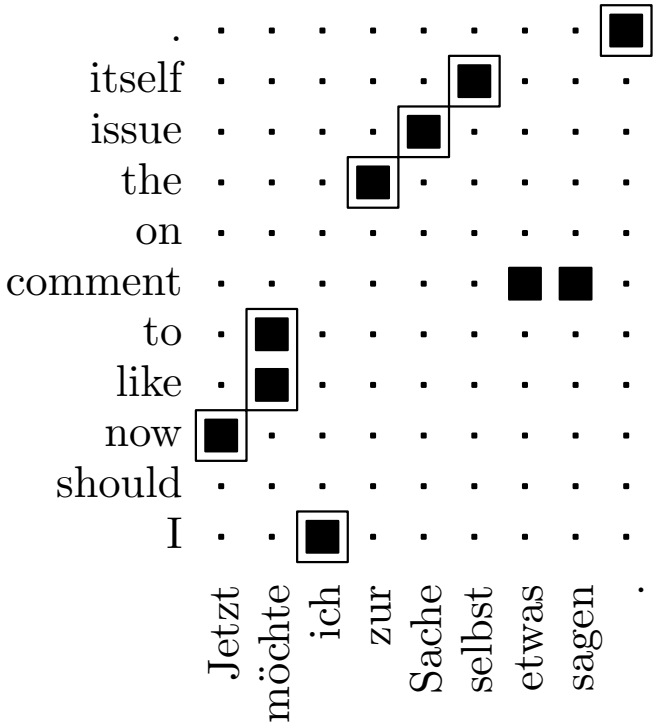
Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Motivation

Phrase-based Statistical Machine Translation System



State of the art: Heuristic Phrase Extraction

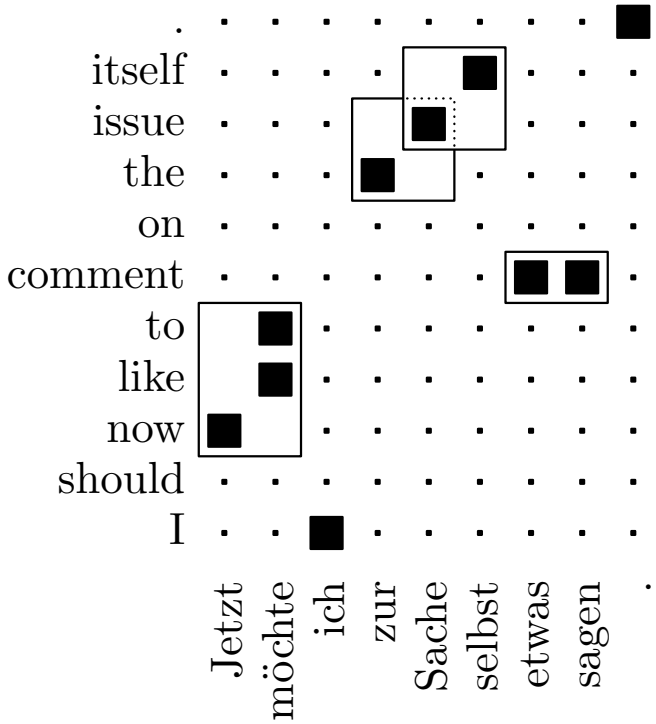


- heuristically extract phrases consistent with word alignment
- use relative frequencies to estimate phrase translation probabilities:

$$p(\tilde{f}|\tilde{e}) = \frac{C(\tilde{f}, \tilde{e})}{C(\tilde{e})}$$

for source phrase $\tilde{f}$ and target phrase $\tilde{e}$

State of the art: Heuristic Phrase Extraction

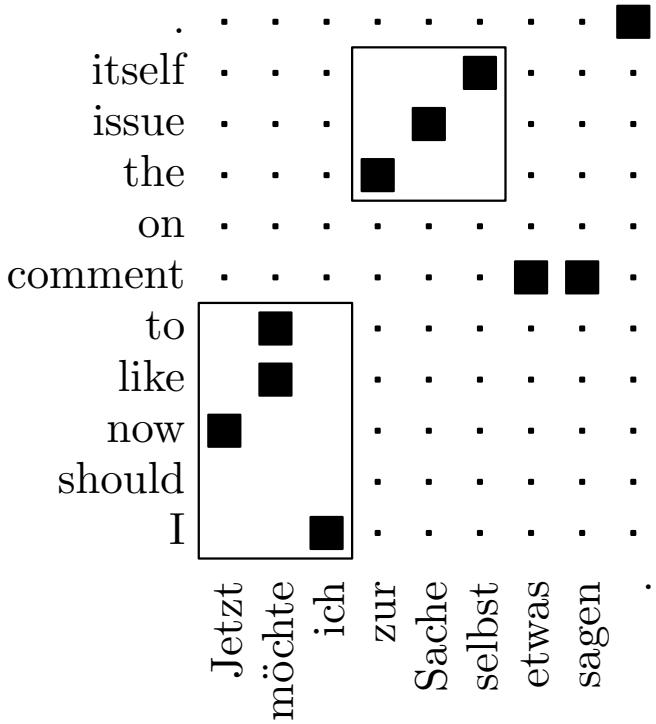


- heuristically extract phrases consistent with word alignment
- use relative frequencies to estimate phrase translation probabilities:

$$p(\tilde{f}|\tilde{e}) = \frac{C(\tilde{f}, \tilde{e})}{C(\tilde{e})}$$

for source phrase $\tilde{f}$ and target phrase $\tilde{e}$

State of the art: Heuristic Phrase Extraction

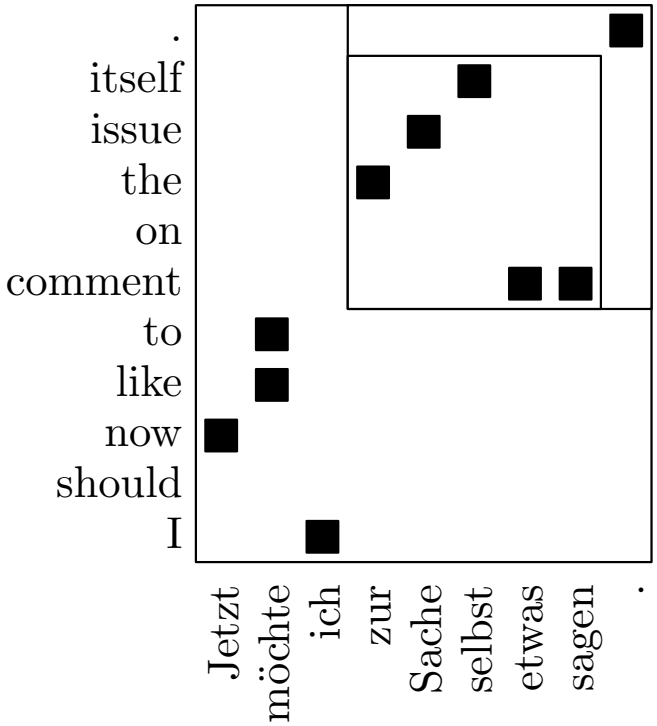


- heuristically extract phrases consistent with word alignment
- use relative frequencies to estimate phrase translation probabilities:

$$p(\tilde{f}|\tilde{e}) = \frac{C(\tilde{f}, \tilde{e})}{C(\tilde{e})}$$

for source phrase $\tilde{f}$ and target phrase $\tilde{e}$

State of the art: Heuristic Phrase Extraction



- heuristically extract phrases consistent with word alignment
- use relative frequencies to estimate phrase translation probabilities:

$$p(\tilde{f}|\tilde{e}) = \frac{C(\tilde{f}, \tilde{e})}{C(\tilde{e})}$$

for source phrase $\tilde{f}$ and target phrase $\tilde{e}$

Motivation

Combine two approaches:

- **hierarchical translation system:**
  - **rule table is constructed heuristically**
  - **inconsistency between training and translation**
- **phrase-based translation system:**
  - **phrase table trained with EM/Viterbi**
  - **training consistent with translation**

Goal:

- **use hierarchical paradigm for translation**
- **exploit trained phrase table from phrase-based system**

Results:

- **up to 0.7 BLEU and 1.0 TER improvement on IWSLT Ar-En**

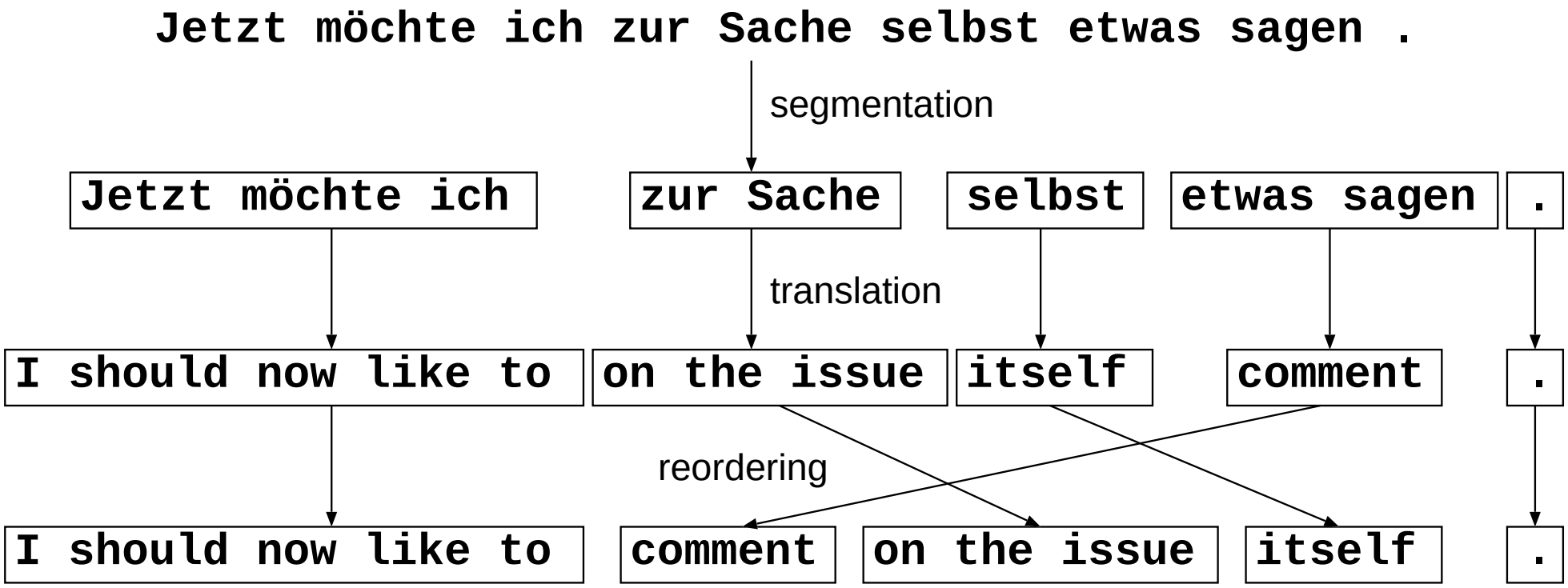
Outline

- Related Work
- Phrase-based vs. Hierarchical Translation
- Forced Alignment Phrase Training
- Experimental Setup
- Combining the Rule/Phrase Tables
- Results
- Conclusion

Related Work

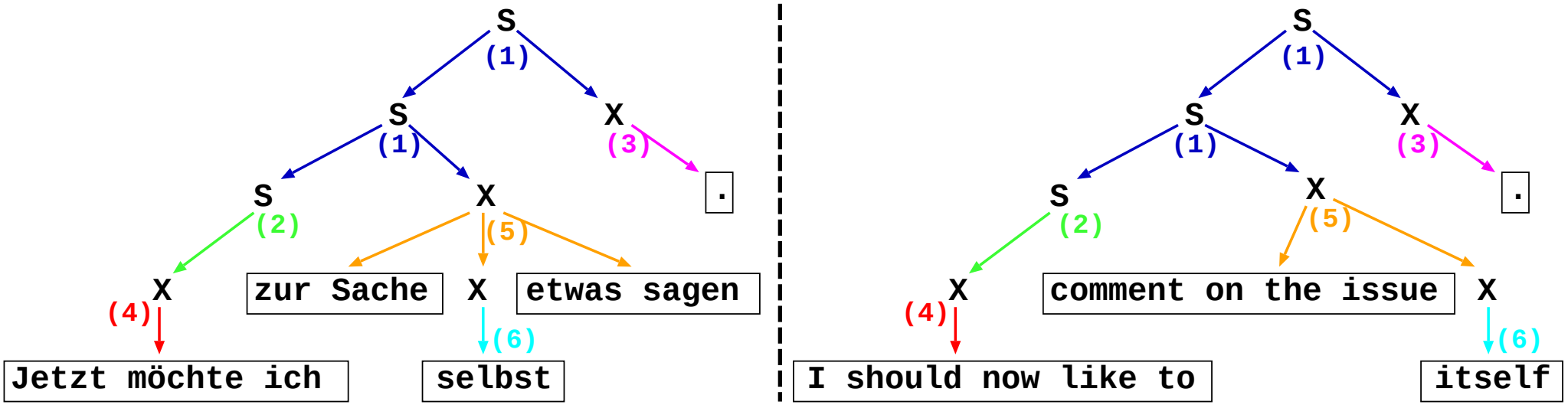
- **Blunsom, Cohn, Osborne: A Discriminative Latent Variable Model, 2008**
  - **discriminative phrase model training for hierarchical translation**
  - **parameter regularization to prevent over-fitting**
- **Čmejrek, Zhou, Xiang: Enriching SCFG Rules Directly From Efficient Bilingual Chart Parsing, 2009**
  - **EM training for hierarchical rules**
- **Wuebker, Mauser, Ney: Training Phrase Translation Models with Leaving-One-Out, 2010**
  - **EM/Viterbi training for phrase-based translation**
  - **leaving-one-out to prevent over-fitting**

Phrase-based Translation



Hierarchical Translation

Synchronous context free grammar



(1) S → <SX,SX>
(2) S → <X,X>
(3) X → <.,.>
(4) X → <Jetzt möchte ich,I should now like to>
(5) X → <zur Sache X etwas sagen,comment on the issue X>
(6) X → <selbst,itself>

Log-linear Model Combination

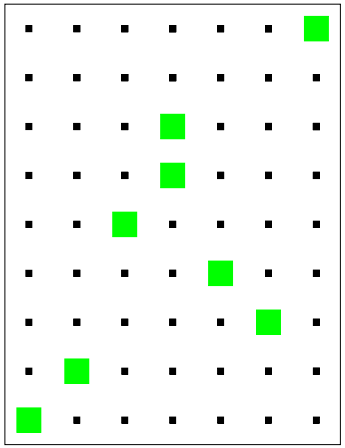
- source sentence $f_1^J = f_1, \ldots, f_J$
- target sentence $e_1^I = e_1, \ldots, e_I$
- weighted log-linear combination of M models
- best translation $\hat{e}_1^{\hat{I}}$ is defined as:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I, s}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, s, f_1^J) \right\}$$

- hidden structure s
  - phrase-based translation: phrase segmentation and alignment
  - hierarchical translation: parse tree

Phrase Training

State of the art:



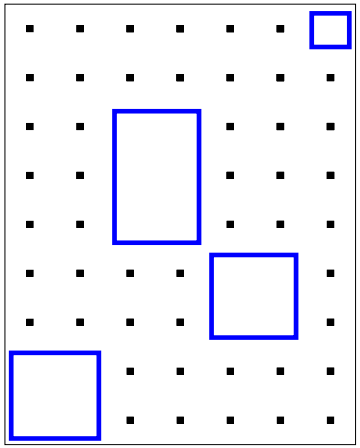
Viterbi Word Alignment

word translation models
trained by EM Algorithm

heuristic phrase counts

Phrase Translation Table

Forced Alignment:



Phrase Alignment

phrase translation models
trained by EM/Viterbi

phrase translation probabilities

Phrase Translation Table

Training (Forced Alignment) - Concept

constrain translation system to reference translation

- **given:**
  - **source sentence** $f_1^J = f_1, \ldots, f_J$
  - **target sentence** $e_1^I = e_1, \ldots, e_I$
  - $(M-1)$ **models: no language model necessary**
- **unknown: phrase alignment** s

- **best phrase alignment** $\hat{s}$ **is defined as:**

$$\hat{s} = \underset{K,s}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M-1} \lambda_m h_m(e_1^I, s, f_1^J) \right\}$$

- **details:**

  [Wuebker, Mauser, Ney: Training Phrase Translation Models with Leaving-One-Out, ACL 2010]

Phrase Model

- phrase translation probabilities are estimated as relative frequencies:

$$p_{FA}(\tilde{f}|\tilde{e}) = \frac{C_{FA}(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} C_{FA}(\tilde{f}', \tilde{e})}$$

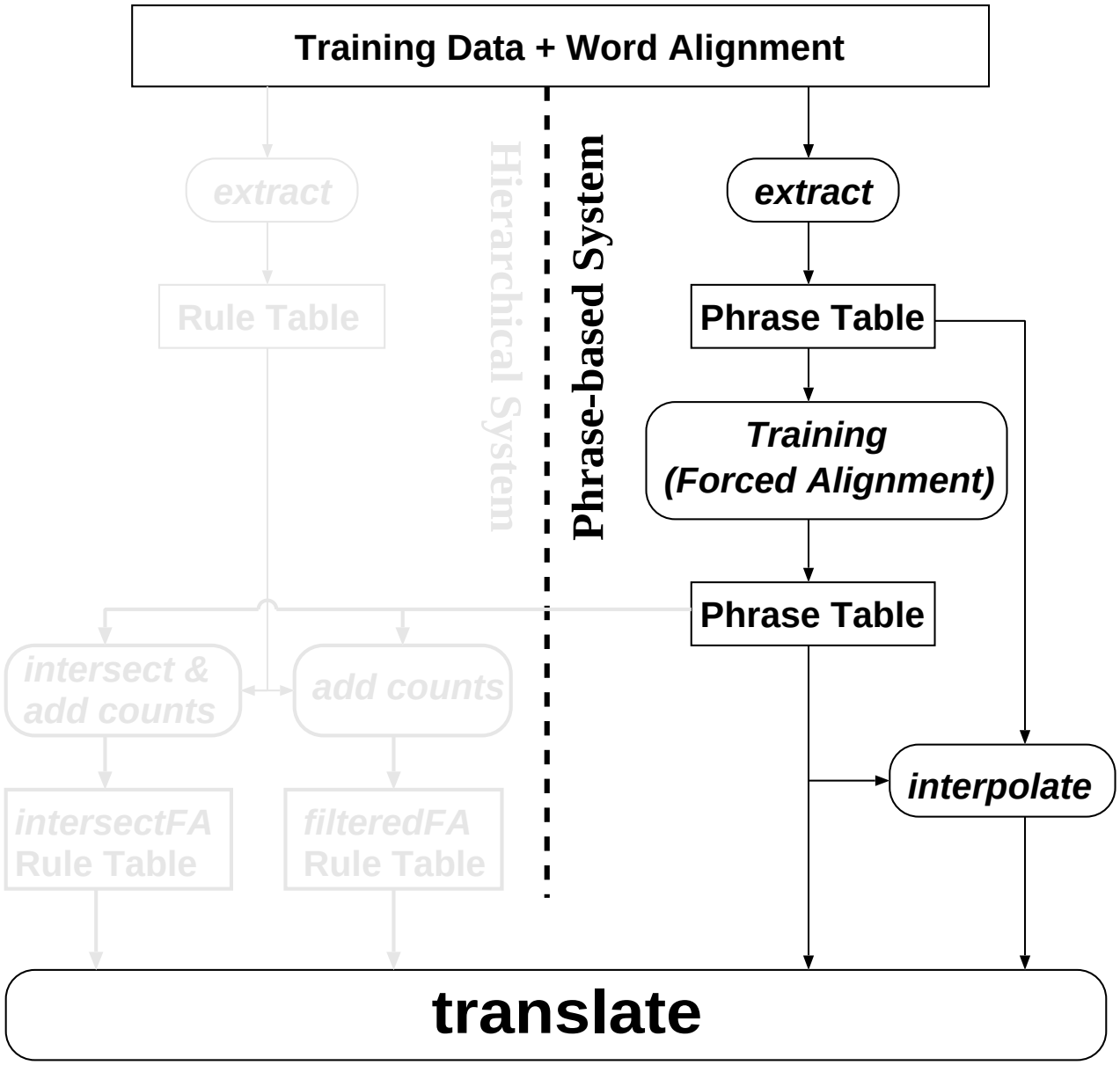
- the (weighted) counts $C_{FA}(\tilde{f}, \tilde{e})$ can be computed from
  - the Viterbi phrase alignment
  - count model:
    the n-best alignments, weighted equally

Phrase Table Size

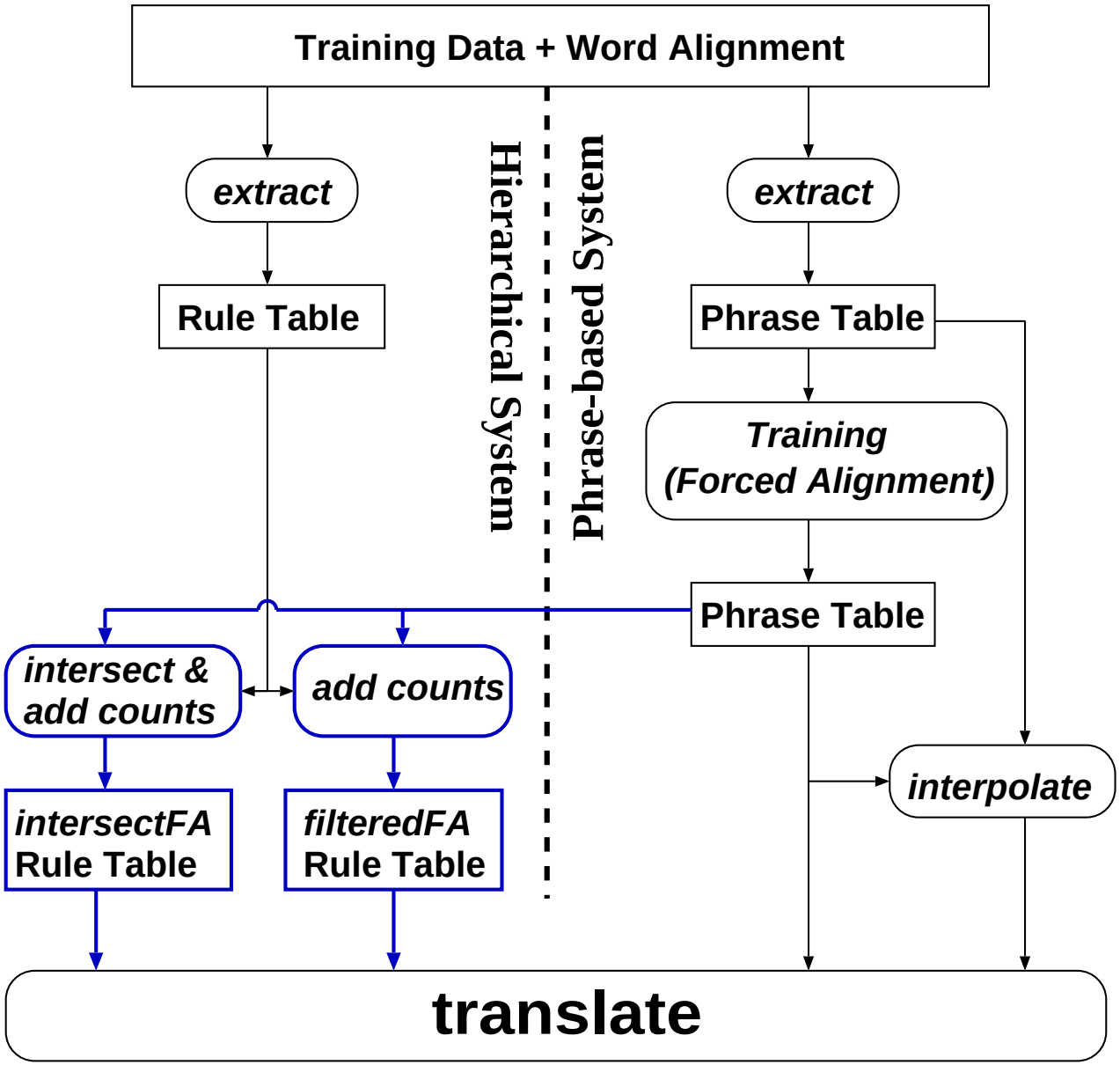
IWSLT 2010, Ar-En		
	# phrases	% of full table
FA, count model	348K	36.7
heuristic	947K	100

Quaero 2010, En-De		
	# phrases	% of full table
FA, count model	17.8M	14.7
heuristic	121M	100

Experimental Setup



Experimental Setup



Combining the Rule/Phrase Tables

- leave hierarchical rules untouched
- re-estimate non-hierarchical phrase translation probabilities:
  - compute new counts $C_{comb}(\tilde{f}, \tilde{e})$ and $C_{comb}(\tilde{e})$:

$$C_{comb}(\tilde{f}, \tilde{e}) = C_{FA}(\tilde{f}, \tilde{e}) + C_{H}(\tilde{f}, \tilde{e})$$
$$C_{comb}(\tilde{e}) = C_{FA}(\tilde{e}) + C_{H}(\tilde{e})$$

$C_{FA}(\tilde{f}, \tilde{e})$ count in forced alignment
$C_{H}(\tilde{f}, \tilde{e})$ count in original hierarchical extraction

Combining the Rule/Phrase Tables

- leave hierarchical rules untouched
- re-estimate non-hierarchical phrase translation probabilities:
  - compute new counts $C_{comb}(\tilde{f}, \tilde{e})$ and $C_{comb}(\tilde{e})$:

$$C_{comb}(\tilde{f}, \tilde{e}) = C_{FA}(\tilde{f}, \tilde{e}) + C_{H}(\tilde{f}, \tilde{e})$$
$$C_{comb}(\tilde{e}) = C_{FA}(\tilde{e}) + C_{H}(\tilde{e})$$

  - renormalize:

$$p_{comb}(\tilde{f}|\tilde{e}) = \frac{C_{comb}(\tilde{f}, \tilde{e})}{C_{comb}(\tilde{e})}$$

- filteredFA: keep all phrases in rule table
- intersectFA: keep only phrases seen in forced alignment

Results on IWSLT 2010, Arabic-English

	test04		test05		test08	
	BLEU	TER	BLEU	TER	BLEU	TER
hier. baseline	58.8	27.6	57.1	29.6	62.5	25.4
hier. + filteredFA	59.3	26.6	57.5 (+.4)	28.6 (-1)	62.2 (-.3)	25.2 (-.2)
hier. + intersectFA	58.6	27.8	56.3 (-.8)	29.9 (+.3)	61.5 (-1)	26.0 (+.6)
hier. + syntax	59.0	26.7	57.5	28.6	61.4	25.9
hier. + syntax + filteredFA	60.6	26.2	57.5	28.6	62.1 (+.7)	25.5 (-.4)

- filteredFA: keep all phrases in rule table
- intersectFA: keep only phrases seen in forced alignment

Results on Quaero 2010, English-German News

- training data: Europarl + News (cf. WMT 2010)
- test data: News domain

	Dev		Test	
	BLEU	TER	BLEU	TER
hier. baseline	17.6	66.9	18.6	65.0
hier. + filteredFA	17.8	66.9	19.0 (+.4)	64.6 (-.4)
phrase-based baseline	17.1	67.8	17.8	65.7
phrase-based + interpolate	17.3	66.4	18.4	64.5

- filteredFA: keep all phrases in rule table
- intersectFA: keep only phrases seen in forced alignment

Translation Examples

source	This makes an increase in immigration unavoidable.
hierarchical	Dies ist ein Anstieg der Einwanderung unvermeidlich.
hierarchical+FA	Das macht eine zunehmende Einwanderung unvermeidlich.
reference	Schon das macht eine vermehrte Einwanderung unvermeidlich.
source	Saving Alstom by nationalizing the company is obviously wrong.
hierarchical	Die Rettung von Alstom durch das Unternehmen verstaatlicht ist offenkundig falsch.
hierarchical+FA	Die Rettung von Alstom durch Verstaatlichung des Unternehmens ist offenkundig falsch.
reference	Alstom durch die Nationalisierung des Unternehmens zu retten, ist offensichtlich falsch.

- $\tilde{f}$ = "nationalizing"
- $\tilde{e}$ = "Verstaatlichung"

$$p_H(\tilde{f}|\tilde{e}) = 0.14 \rightarrow p_{comb}(\tilde{f}|\tilde{e}) = 0.21$$

Conclusions

- first step in the direction of real training for hierarchical rule table
- results indicate that hierarchical systems can benefit from phrase training
  - up to 0.7 BLEU and 1.0 TER improvement on IWSLT Ar-En
  - up to 0.4 BLEU and 0.4 TER improvement on Quaero En-De
- new approach to combining hierarchical and phrase-based paradigms

Thank you for your attention

Carmen Heger
Joern Wuebker

<surname>@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

Corpus Statistics, Ar-En IWSLT 2010 (BTEC)

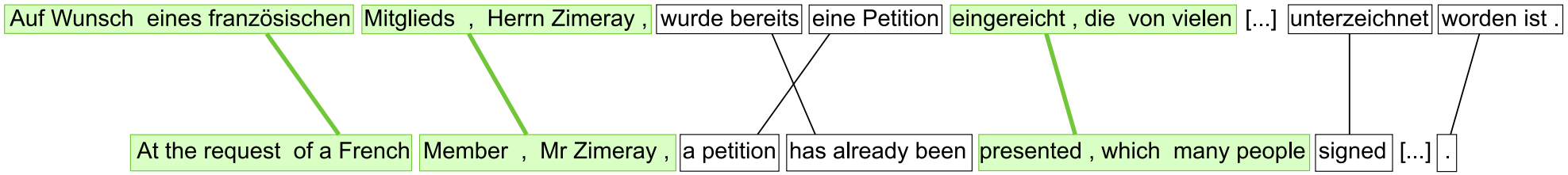
		Arabic	English
Train:	Sentences	23,940	
	Running Words	206,008	240,125
	Vocabulary	15,861	8,258
	Singletons	7,152	3,516
test04:	Sentences	500	
	Running Words	3,537	–
	Vocabulary	1,183	–
	OOV rate	3%	–
test05:	Sentences	506	
	Running Words	3,421	–
	Vocabulary	1,185	–
	OOV rate	3.5%	–
test08:	Sentences	507	
	Running Words	3,478	–
	Vocabulary	1,141	–
	OOV rate	3.9%	–

Corpus Statistics, En-De Quaero 2010

		English	German
Train:	Sentences	1,799,293	
	Running Words	46,708,144	44,626,470
	Vocabulary	121,857	369,859
	Singletons	45,648	176,657
Dev:	Sentences	2,121	
	Running Words	51,343	52,946
	Vocabulary	7,313	9,863
	OOV rate	0.6%	1.1%
Test:	Sentences	2,007	
	Running Words	49,763	51,119
	Vocabulary	7,119	9,680
	OOV rate	0.5%	1.1%

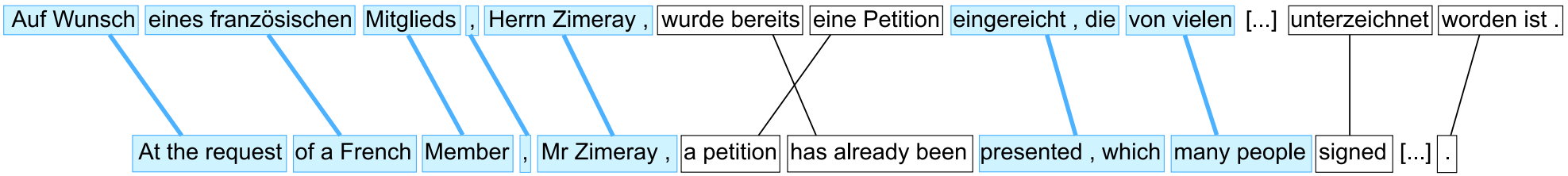
Leaving-One-Out - Motivation

- previous approaches have reported problems with over-fitting
- EM training learns unintuitive, long phrases



Leaving-One-Out - Motivation

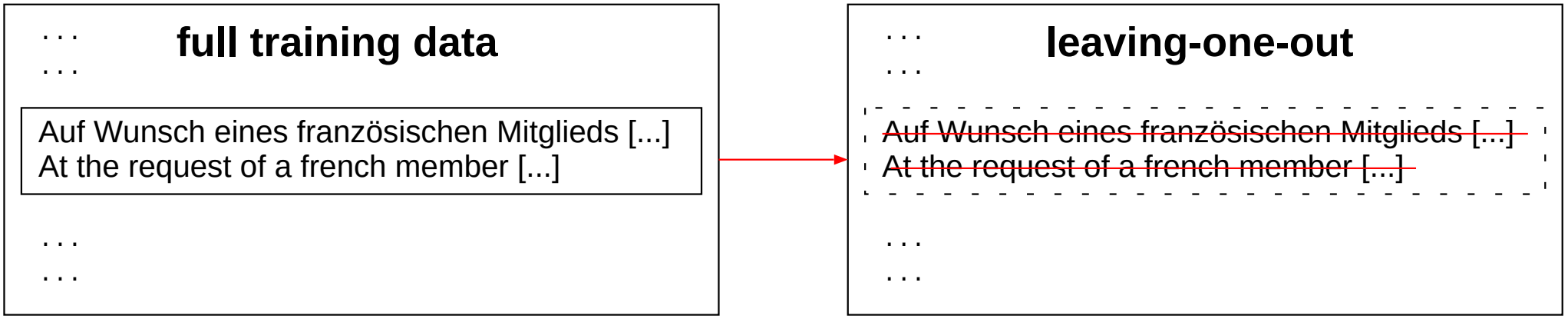
- previous approaches have reported problems with over-fitting
- EM training learns unintuitive, long phrases
- With leaving-one-out: shorter, more intuitive phrases



Leaving-One-Out - Example

current sentence:

Auf Wunsch eines französischen Mitglieds [...]
At the request of a french member [...]



$$C(\tilde{f},\tilde{e}) = 2$$
$$C(\tilde{e}) = 3$$
$$\Rightarrow p(\tilde{f}|\tilde{e}) = \frac{2}{3}$$

$$C_{l1o}(\tilde{f},\tilde{e}) = 2-1 = 1$$
$$C_{l1o}(\tilde{e}) = 3-1 = 2$$
$$\Rightarrow p_{l1o}(\tilde{f}|\tilde{e}) = \frac{1}{2}$$

$\tilde{f}$ = **"Auf Wunsch eines französischen"**
$\tilde{e}$ = **"At the request of a French"**

Leaving-One-Out - Formal Definition

- store phrase counts $C(\tilde{f}, \tilde{e})$ and marginals $C(\tilde{e})$ in phrase table
- for each sentence pair (f_n, e_n) in training:
  - perform phrase extraction on (f_n, e_n)
  - "local" phrase counts for (f_n, e_n): $C_n(\tilde{f}, \tilde{e})$, $C_n(\tilde{e})$
  - re-estimate phrase translation probabilities:

$$p_{l1o,n}(\tilde{f}|\tilde{e}) = \frac{C(\tilde{f}, \tilde{e}) - C_n(\tilde{f}, \tilde{e})}{C(\tilde{e}) - C_n(\tilde{e})}$$

  - perform search with re-estimated probabilities $p_{l1o,n}$

Leaving-One-Out - Singleton Phrases

- **singleton phrases are often necessary to find phrase alignment**
  - **retain with low probability** $p_{l1o,n}(\tilde{f}|\tilde{e})$
- **standard leaving-one-out:** $p_{l1o,n}(\tilde{f}|\tilde{e}) := \alpha$ **, with fixed** α
- **length based leaving-one-out:** $p_{l1o,n}(\tilde{f}|\tilde{e}) := \beta^{(|\tilde{f}|+|\tilde{e}|)}$**, with fixed** β