An Unsupervised Model for Joint Phrase Alignment and Extraction

Graham Neubig[1,2], Taro Watanabe[2], Eiichiro Sumita[2], Shinsuke Mori[1], Tatsuya Kawahara[1]

[1]Graduate School of Informatics, Kyoto University
[2]National Institute of Information and Communication Technology

Phrase Table Construction

The Phrase Table

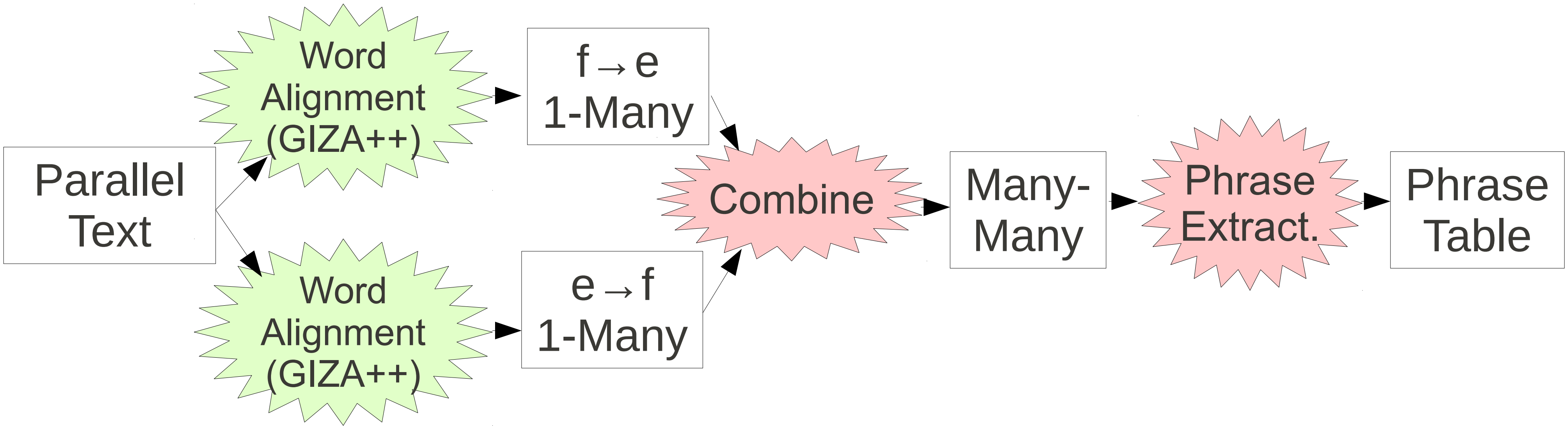
- The most important element of phrase-based SMT
  - Consists of scored bilingual phrase pairs

Source	Target	Scores
le	it	0.05 0.20 0.005 1
le admettre	admit it	1.0 1.0 1e-05 1
admettre	admit	0.4 0.5 0.02 1
	…	

- Usually learned from a parallel corpus aligned at the sentence level

→ Phrases must be aligned

Traditional Phrase Table Construction: 1-to-1 Alignment, Combination, Extraction



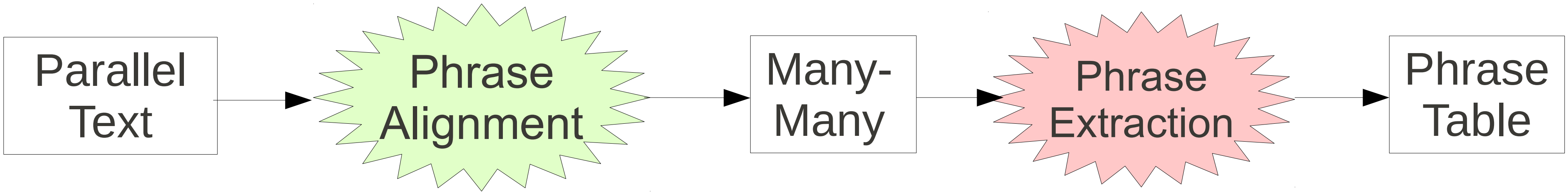
+ Generally quite effective, default for Moses

- Complicated, with lots of heuristics

- Does not directly acquire phrases, which are the final goal of alignment

- Phrase table is exhaustively extracted and thus large

Previous Work: Many-to-Many Alignment



- Significant recent research on many-to-many alignment [Zhang+ 08, DeNero+ 08, Blunsom+ 10]

  + Model is simplified, gains in accuracy

- Short phrases are aligned, then combined into longer phrases during the extraction step

  - Some issues still remain

  - Large phrase table, heuristics, no direct modeling of extracted phrases

Proposed Model for Joint Phrase Alignment and Extraction

- Phrases of multiple granularities directly modeled

  + No mismatch between alignment goal and final goal

  + Completely probabilistic model, no heuristics

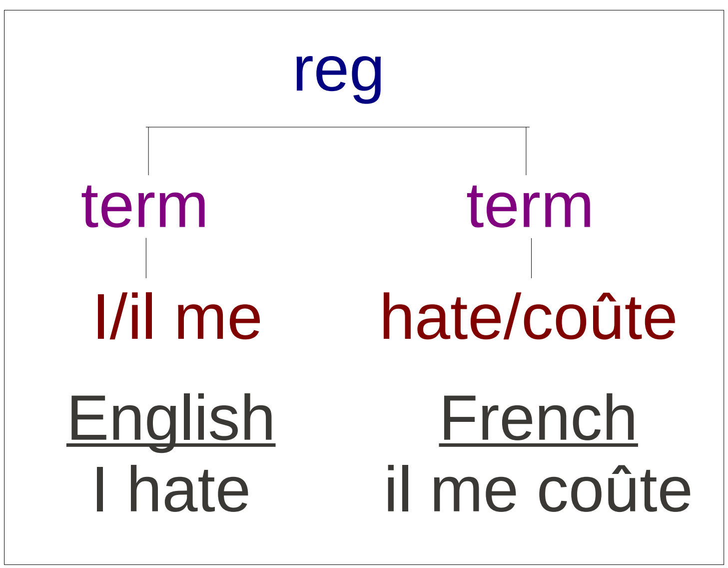
  + Competitive accuracy, smaller phrase table

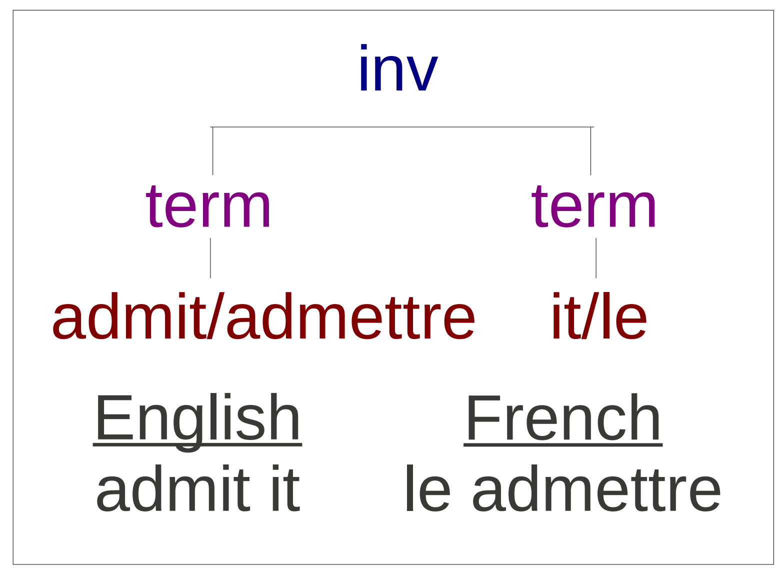
- Uses a hierarchical model for Inversion Transduction Grammars (ITG)

Phrasal Inversion Transduction Grammars (Previous Work)

Inversion Transduction Grammar (ITG)

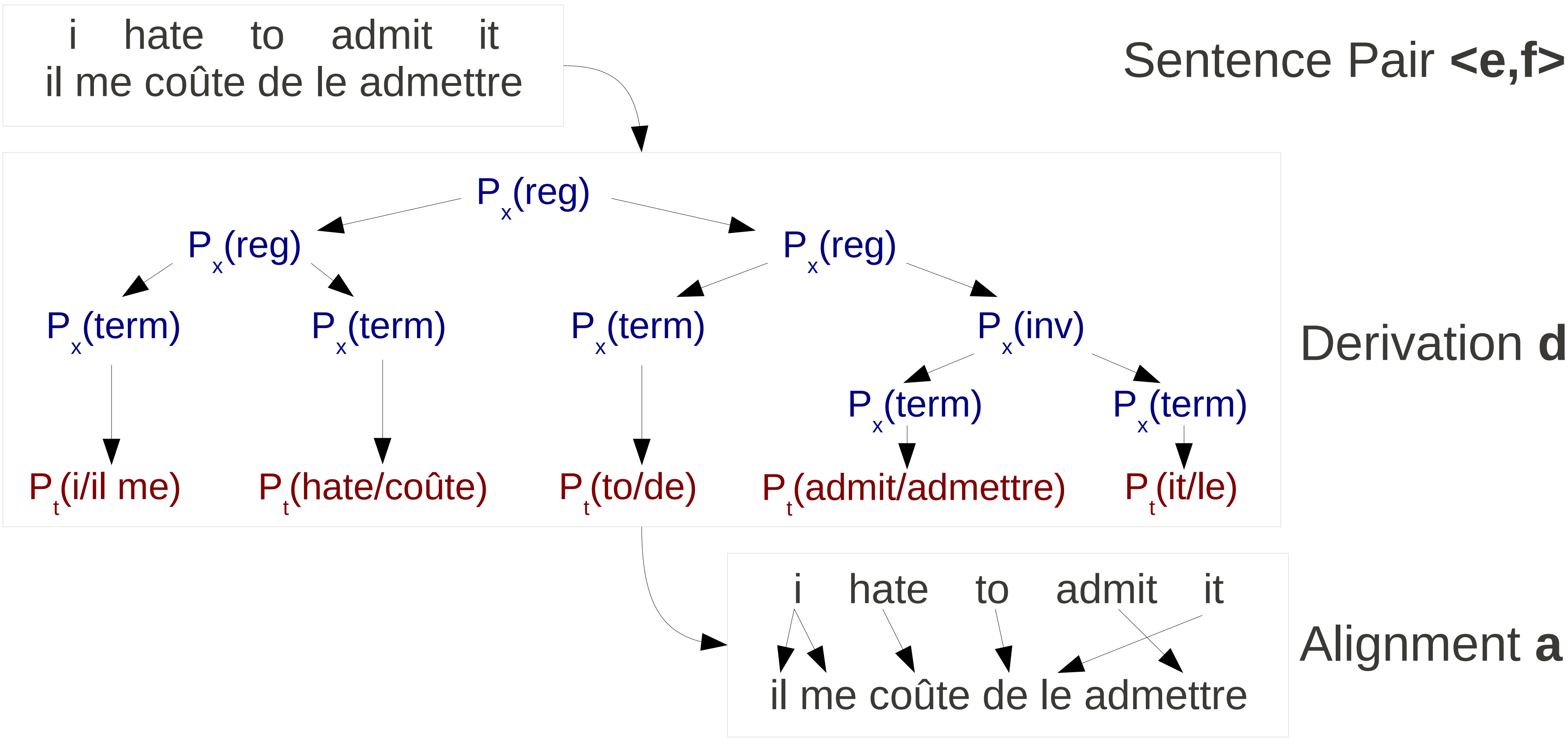
- Like a CFG over two languages
  - Have non-terminals for regular and inverted productions
  - One pre-terminal
  - Terminals specifying phrase pairs





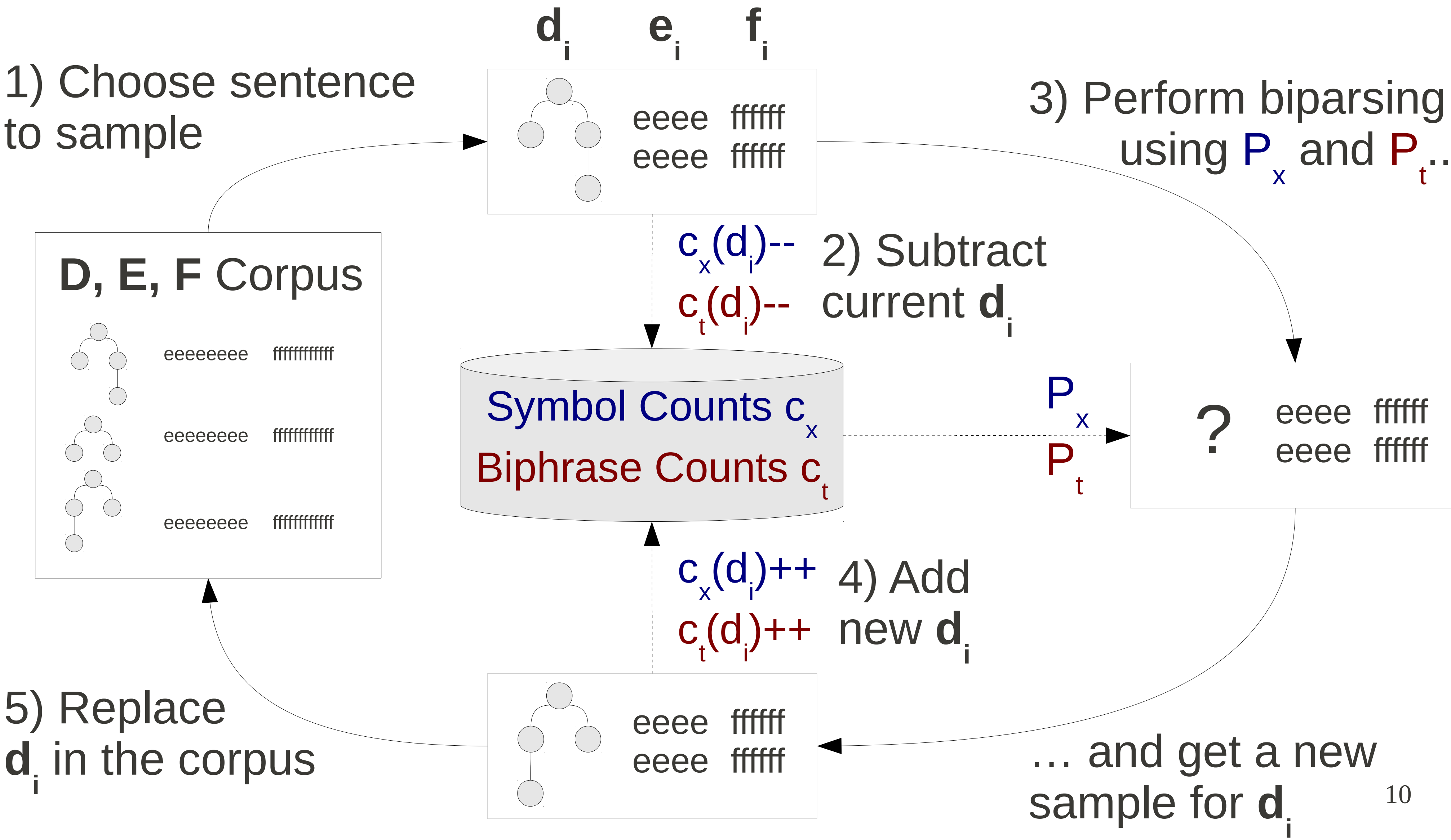
Biparsing-based Alignment with ITGs

- Non/pre-terminal distribution P_x, and phrase distribution P_t



- Viterbi parsing and sampling both possible in $O(n^6)$

Learning Phrasal ITGs with Blocked Gibbs Sampling [Blunsom+ 10]



Calculating Probabilities given Counts

c_t(it/le)=12	c_t(I/il me)=3	c_t(hate/coûte)=0	…
c_x(reg)=415	c_x(inv)=43	c_x(term)=312	

- Adapt Bayesian approach, assume that probabilities were generated from Pitman-Yor process, Dirichlet distribution

$$P_t \sim PY(d, \theta, P_{base})$$
$$P_x \sim Dirichlet(\alpha = 1, 1/3)$$

- Marginal probabilities can be calculated (in example, ignoring *d* for the PY process)

$$P_t(f, e) = \frac{c_t(f, e) + \theta_t P_{base}(f, e)}{\sum_{f,e} c_t(f, e) + \theta_t} \qquad P_x(x) = \frac{c_x(x) + \alpha_x / 3}{\sum_x c_x(x) + \alpha_x}$$

Base Measure

$$P_t(f,e)=\frac{c_t(f,e)+\theta_t P_{base}(f,e)}{\sum_{f,e} c_t(f,e)+\theta_t}$$

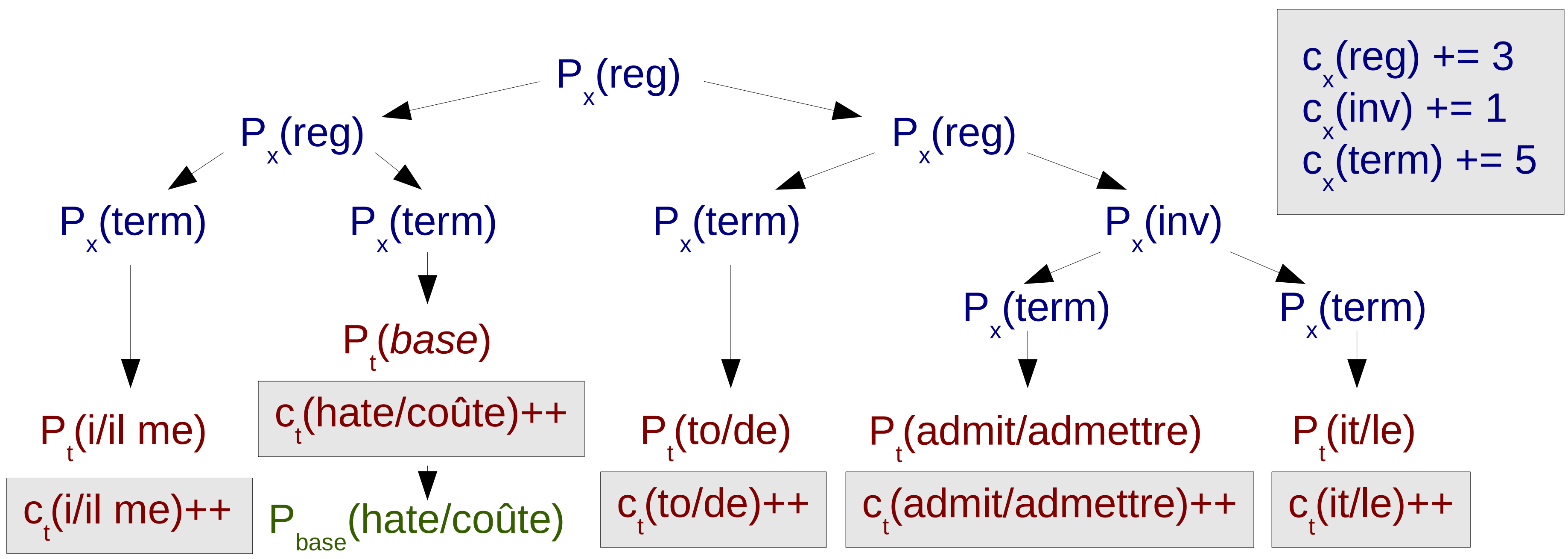
- P_{base} has an effect of smoothing probabilities
  - Particularly for low frequency pairs
- To bias towards good phrase pairs, use geometric mean of word-based Model 1 probabilities [DeNero+ 08]

$$P_{base}(e,f)=\left(P_{m1}(f|e)\,P_{uni}(e)\,P_{m1}(e|f)\,P_{uni}(f)\right)^{\frac{1}{2}}$$

- Good word match in both directions = good phrase match

Calculating Counts given Derivations

- Elements generated from each distribution P_x and P_t added to the counts used to calculate the probabilities



- Problem: only minimal phrases are added

→Must still heuristically combine into multiple granularities

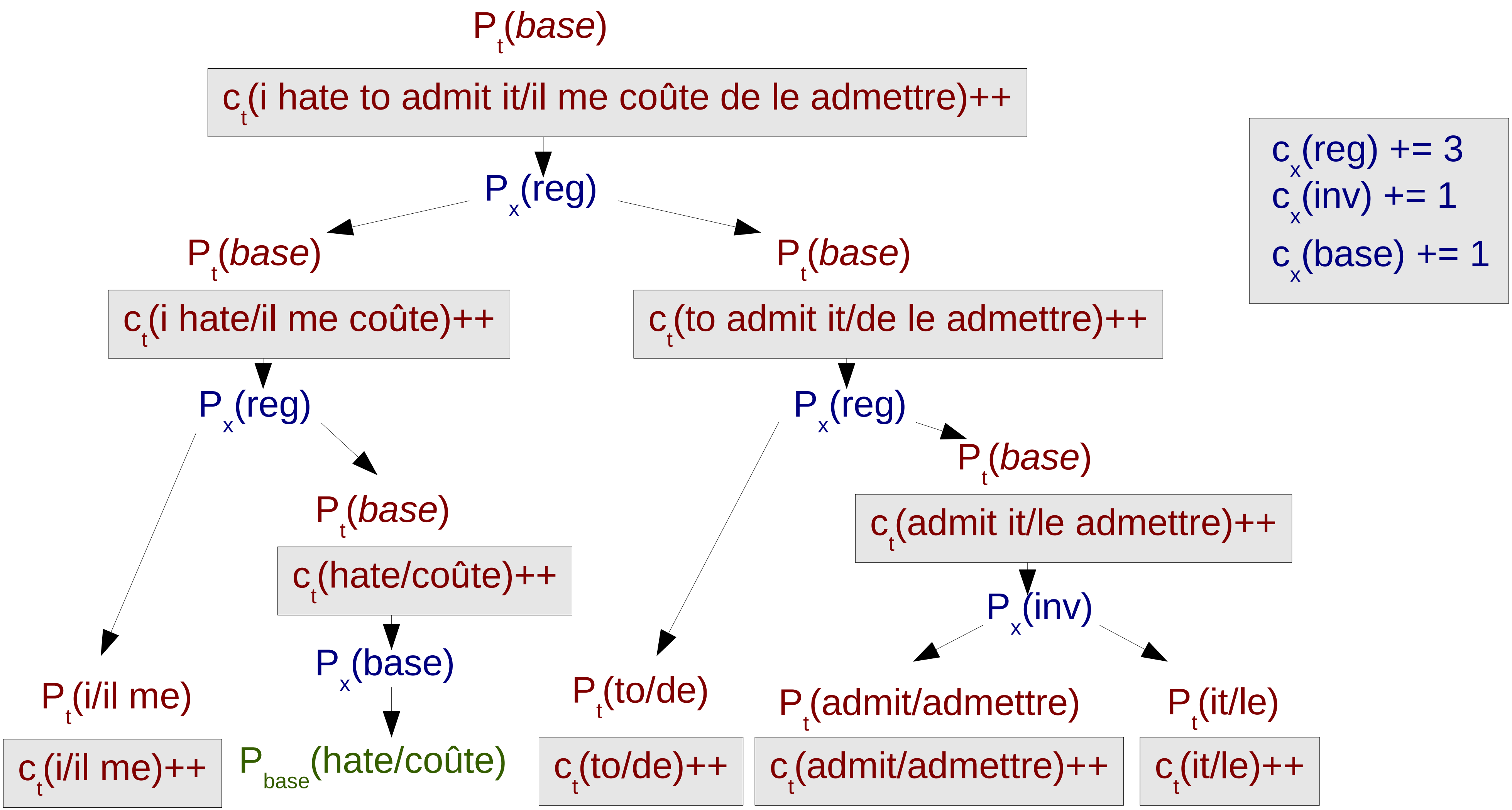
Joint Phrase Alignment and Extraction (Our Work)

Basic Idea

- Generative story in reverse order
- Traditional ITG Model:
  - Generate branches (reordering structure) from P_x
  - Generate leaves (phrase pairs) from P_t
- Proposed ITG Model:
  - From the top, try to generate phrase pair from P_t
  - Divide and conquer using P_x to handle sparsity

Derivation in the Proposed Model

- Phrases of many granularities generated from P_t, added to c_t



- No extraction needed, as multiple granularities are included!

Recursive Base Measure

- Previous work: high prob. words = high prob. phrases
- Proposed: Build new phrase pairs by combining existing phrase pairs in P_{dac} (“divide-and-conquer”)

P_t(I/il me) ← high

P_t(hate/coûte) ← high

P_{dac}(I hate/il me coûte) ← high

$$P_t(f,e) = \frac{c_t(f,e) + \alpha_t P_{dac}(f,e)}{\sum_{f,e} c_t(f,e) + \alpha_t}$$

- High probability sub-phrases → high probability phrases
- P_t is included in P_{dac}, P_{dac} is included in P_t

Details of P_{dac}

- Choose from P_x one of three patterns for P_{dac}, like ITG

Regular: P_x(reg) * P_t(I/il me) * P_t(hate/coûte) →
I hate/il me coûte

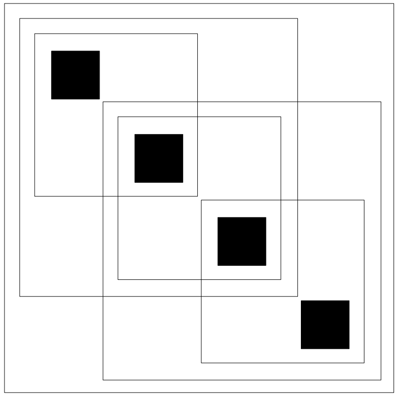
Inverted: P_x(inv) * P_t(admit/admettre) * P_t(it/le) →
admit it/le admettre

Base: P_x(base) * P_{base}(hate/coûte) →
hate/coûte

- P_{base} is the same as before

Phrase Extraction

- Traditional Heuristics: Exhaustively combine and count all neighboring phrases
  - $O(n^2)$ phrases per sent.



Phrase Table Scores

$P(e|f) = c(e,f) / c(f)$

$P(f|e) = c(e,f) / c(e)$

- Model Probabilities: Calculate phrase table from model probabilities where c(e,f) >= 1
  - O(n) phrases per sent.

Phrase Table Scores

$P(e|f) = P_t(e,f) / P_t(f)$

$P(f|e) = P_t(e,f) / P_t(e)$

Experiments

Tasks/Data

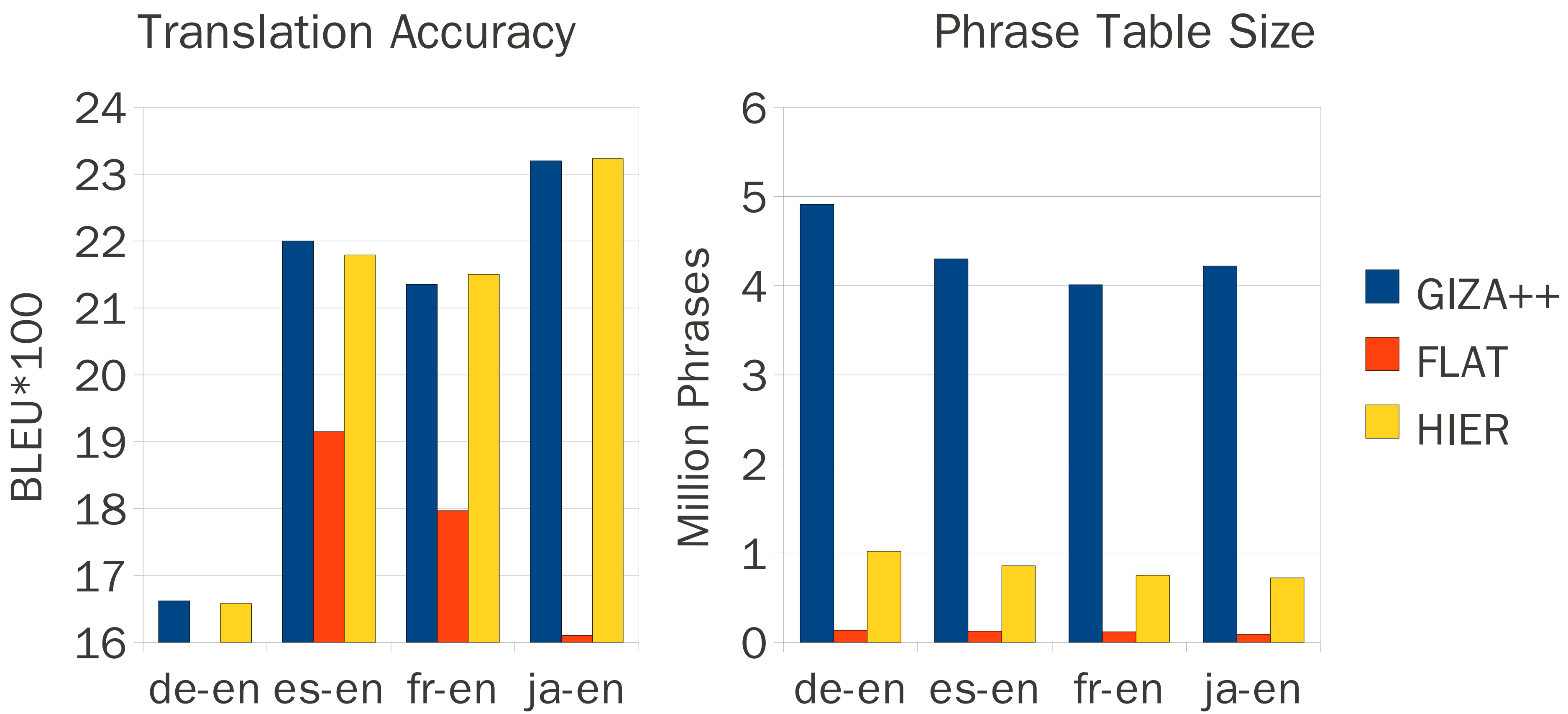
- 4 Languages, 2 tasks: es-en, de-en, fr-en, ja-en
  - de-en, es-en, fr-en: WMT10 news-commentary
  - ja-en: NTCIR08 patent translation
- Data was lowercased, tokenized, and sentences of length 40 and under were used

	WMT				NTCIR	
	de	es	fr	en	ja	en
TM	1.85M	1.82M	1.56M	1.80M/1.62M/1.35M	2.78M	2.38M
LM	-	-	-	52.7M	-	44.7M
Tune	47.2k	52.6k	55.4k	49.8k	80.4k	68.9k
Test	62.7k	68.1k	72.6k	65.6k	48.7k	40.4k

Setting

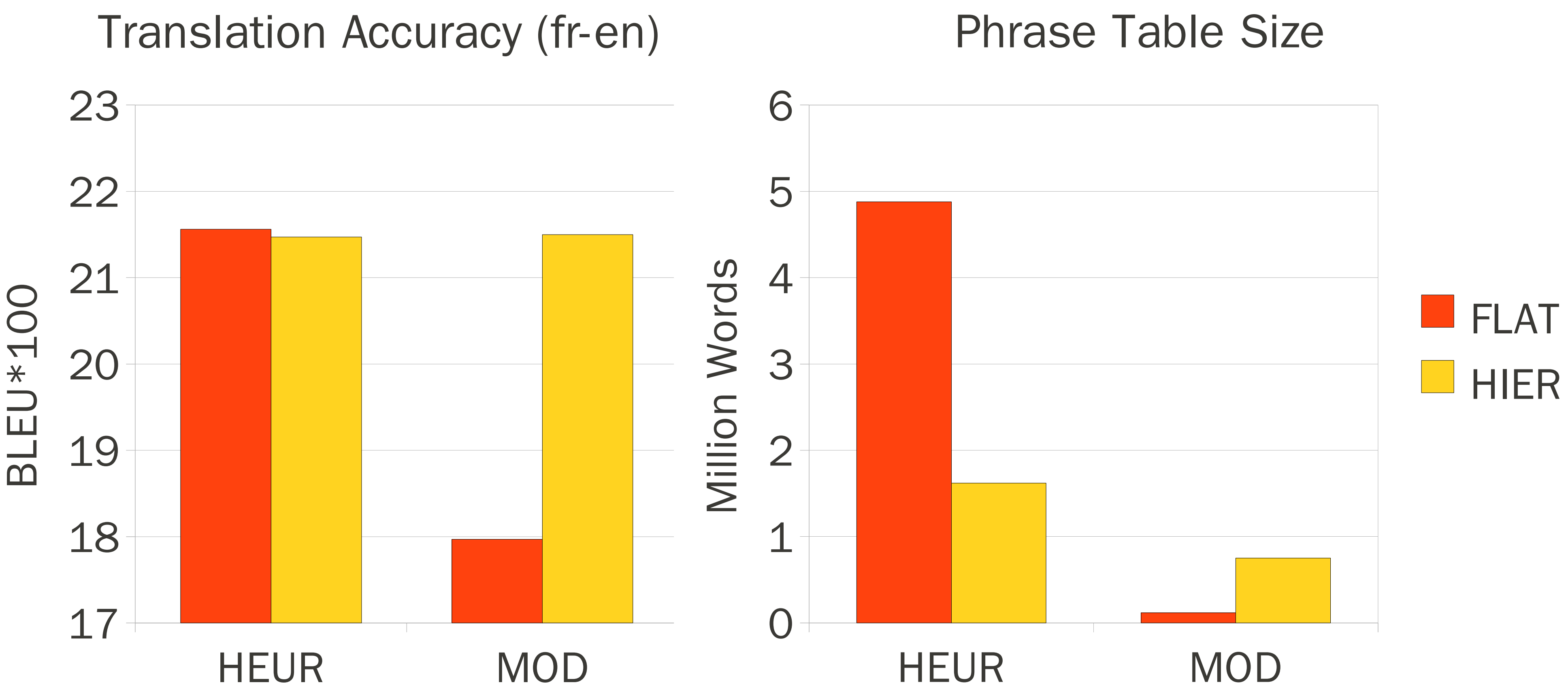
- Used Moses as a decoder
- Evaluated using BLEU score
- 3 Alignment Methods:
  - GIZA++ and *grow-diag-final-and* heuristic
  - Traditional ITG model (FLAT)
  - Proposed ITG model (HIER)
- 2 Phrase Extraction Methods:
  - Heuristic phrase extraction
  - Using the model probabilities P_t

Results



- GIZA++ uses heuristic extraction, others use model probabilities
- Same accuracy as GIZA++, phrase table smaller
- Higher accuracy than FLAT (when using model probs.)

Phrase Table: Heuristic Extraction vs. Model Probabilities



- HIER + Model Probabilities has competitive accuracy, smaller table size

Conclusion

- Used a hierarchical model to include phrases of multiple granularities in the alignment process
- Able to achieve competitive accuracy directly using model probabilities in the phrase table
- Future work:
  - Expansion to tree-based translation
  - Further refinement of modeling and search techniques
- Software is released open source:

pialign – Phrasal ITG Aligner
http://www.phontron.com/pialign

Thank You!