Statistical Phrase-Based Translation

Philipp Koehn, Franz Och, Daniel Marcu

koehn@isi.edu, och@isi.edu, marcu@isi.edu

Information Sciences Institute

University of Southern California

Motivation

- Phrase-based translation is the **best way** to do statistical machine translation
  - best performance in recent DARPA evaluations
  - also fairly simple
  - tools are freely available
- How do I construct a phrase translation table?

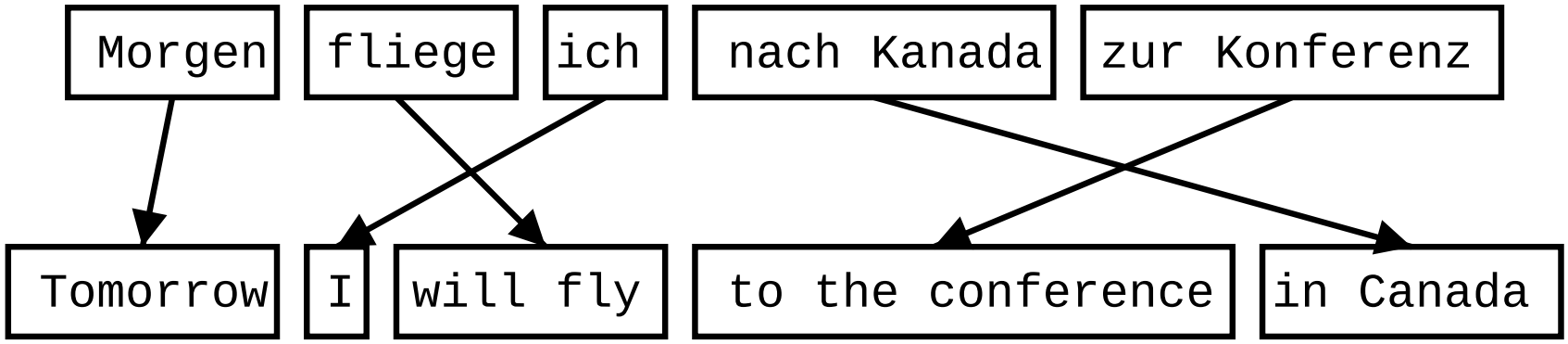
Goals

- Compare different approaches to learn phrases
- Examine properties of phrase-based translation
- Syntax and phrases

Overview

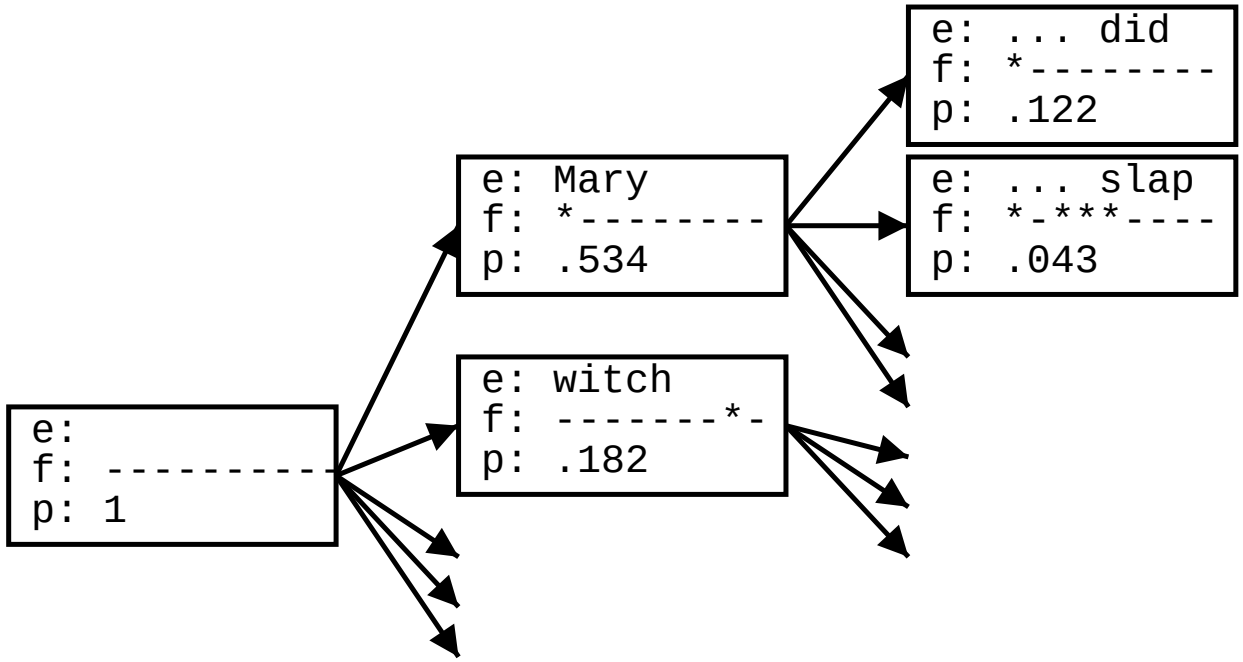
- Evaluation framework
  - unified model
  - decoder
  - corpus
- Three methods for learning phrases
  - word-alignment induced phrases
  - syntactic phrases
  - phrase-alignment
- Experiments

Model



- Bayes rule: $\text{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p(\mathbf{e})$
- Foreign sentence $\mathbf{f}$ is segmented into I phrases $\bar{f}_1^I$
- Each phrase is translated with $\phi(\bar{f}_i|\bar{e}_i)$
- Phrases are reordered with $d(\cdot)$
- Use of language model $p_{\text{LM}}(\mathbf{e})$ and word penalty $\omega^{|\mathbf{e}|}$

Decoder: Beam Search



- Build English by hypothesis expansion
  - from left to right
  - search space exponential with sentence length

  $\Rightarrow$ reduction by pruning weak hypothesis aided by future cost estimate

Evaluation on Europarl Corpus

- Collected from the European Parliament Proceedings
  - Available at `http://www.isi.edu/~koehn/`
  - 11 languages, 20 million words each
- Test set
  - German-English
  - 1755 sentence of length 5-15

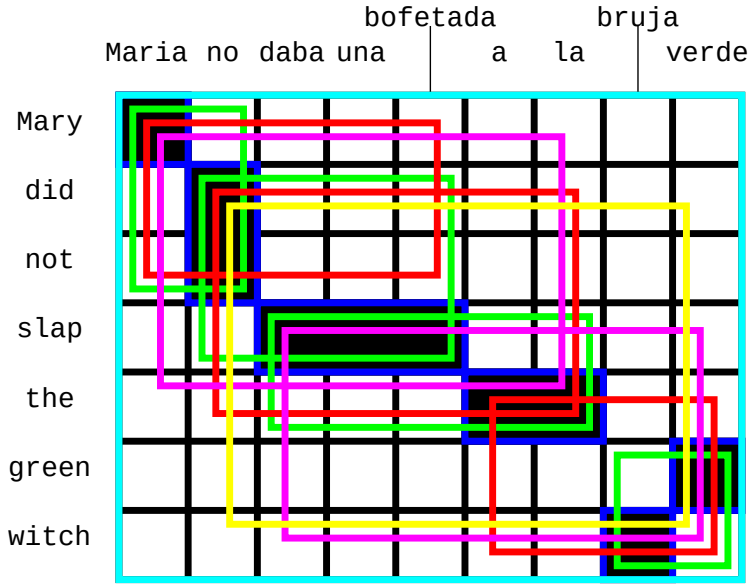
Three Methods for Learning Phrases

- Word-alignment induced phrases
  - similar to alignment templates [Och et al., 1999]
- Syntactic phrases
  - only syntactic phrases are learned
  - same restriction as in recently proposed syntactic transfer models
- Phrase-alignment
  - joint model [Marcu and Wong, 2002]

Word Alignment Induced Phrases

- Word alignment is generated using IBM Model 4
  - bidirectional alignments e$\rightarrow$f, f$\rightarrow$e
  - intersect alignments
  - grow additional alignment points with heuristics
- Collect phrase pairs consistent with word alignment
- This is alignment templates without word classes [Och et al., 1999]

Word Alignment Induced Phrases (2)

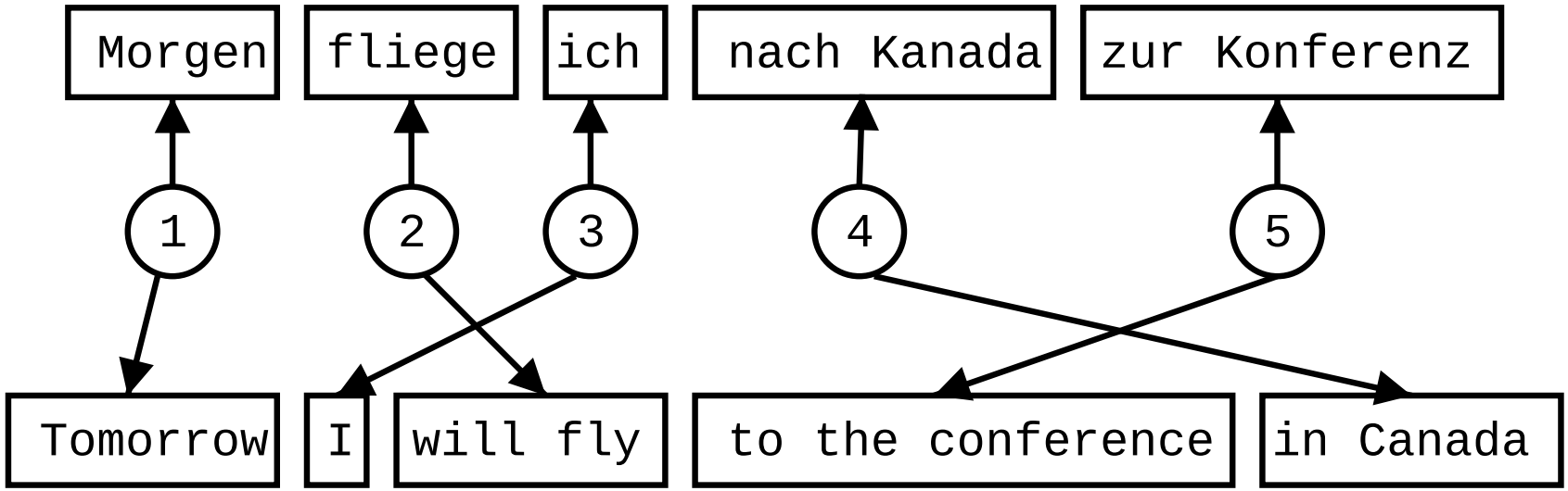


(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch), (Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch), (Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Syntactic Phrases

- Syntactic phrases span whole constituents in parse tree
- Motivation
  - only these phrases used syntactic transfer models,
    e.g., [Yamada and Knight, 2002]
  - does syntax help or hurt?
- Extract syntactic phrase pairs
  - parse both sides (with statistical parsers)
  - use word alignment as before
  - limit to phrases to syntactic constituents in parse tree

Phrase Alignment



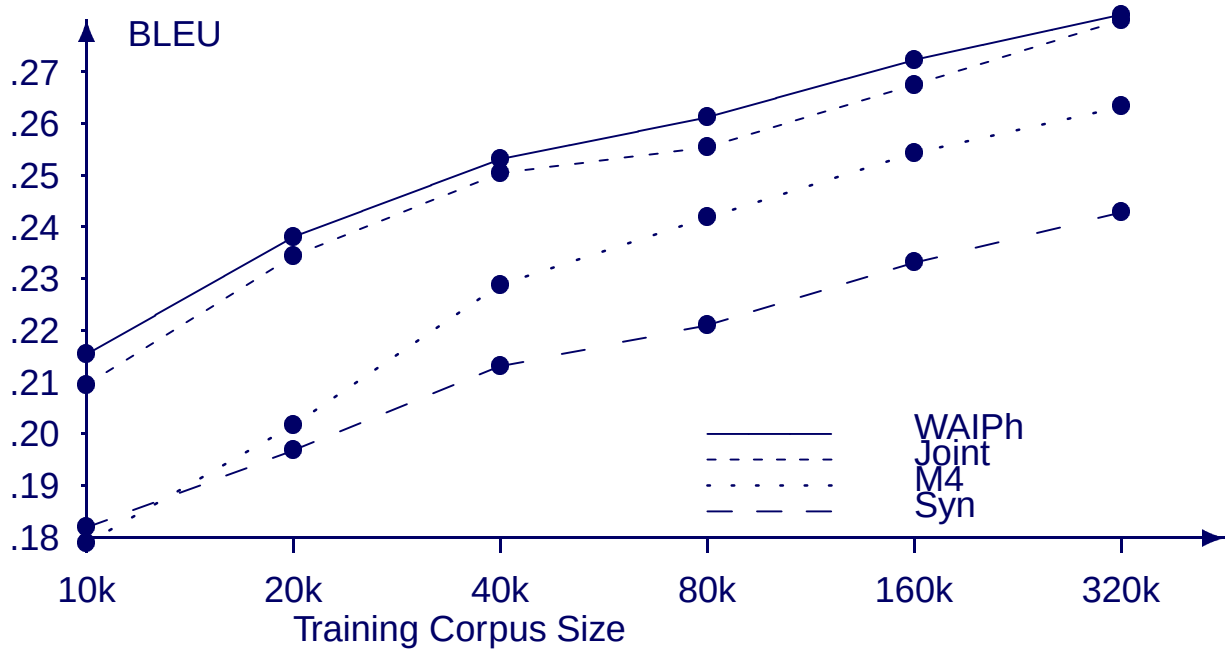
- Direct Phrase Alignment of Parallel Corpus
  [Marcu and Wong, 2002]
- Generative Story
  - a number of concepts are created
  - each concept generates a foreign and English phrase

Experiments

- Comparison of core methods
- Maximum phrase length
- Lexical weighting
- Phrase extraction heuristics
- Simpler word alignment models
- Other language pairs

Comparison of Core Methods

- Same decoder, same training data, same language model
  - except for IBM Model 4: uses greedy decoder [Germann et al., 2001]
- WAIPh best, syntactic phrases very bad



- All following experiments on WAIPh only

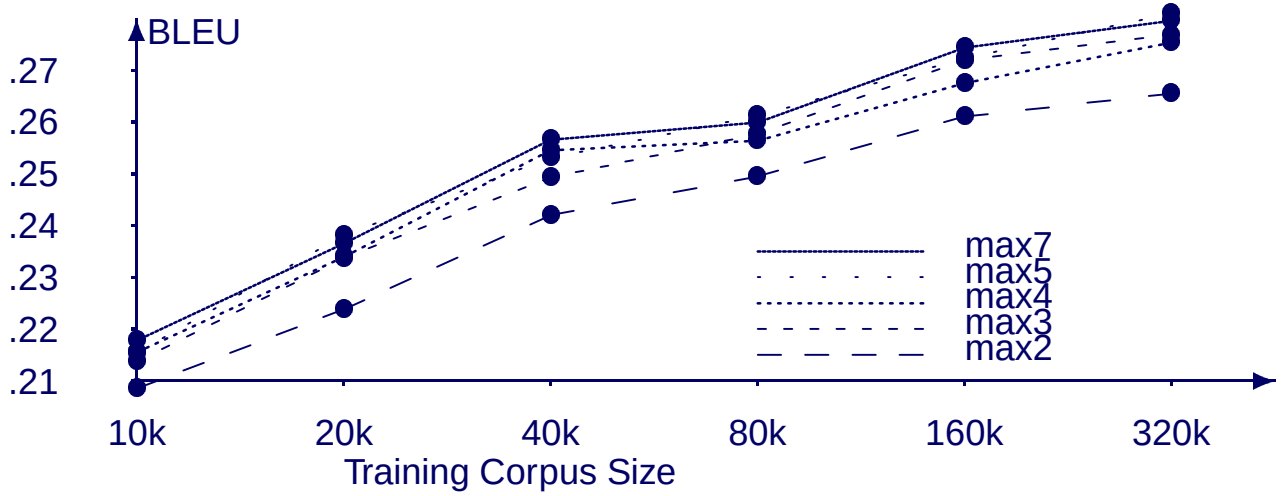
Maximum Phrase Length

- Maximum limit on length of phrases
  - higher limit $\rightarrow$ larger phrase translation table
  - all tables still fit into memory of modern machines

Max. Length	Training corpus size					
	10k	20k	40k	80k	160k	320k
2	37k	70k	135k	250k	474k	882k
3	63k	128k	261k	509k	1028k	1996k
4	84k	176k	370k	736k	1536k	3152k
5	101k	215k	459k	925k	1968k	4119k
7	130k	278k	605k	1217k	2657k	5663k

Maximum Phrase Length (2)

- Impact of limit on translation quality
  - not much improvement if maximum length is extended beyond 3
  - independent of training corpus size



Lexical Weighting

- Augment phrase translation probability $\phi(\bar{f}|\bar{e})$ with lexical translation probabilities $w(f|e)$

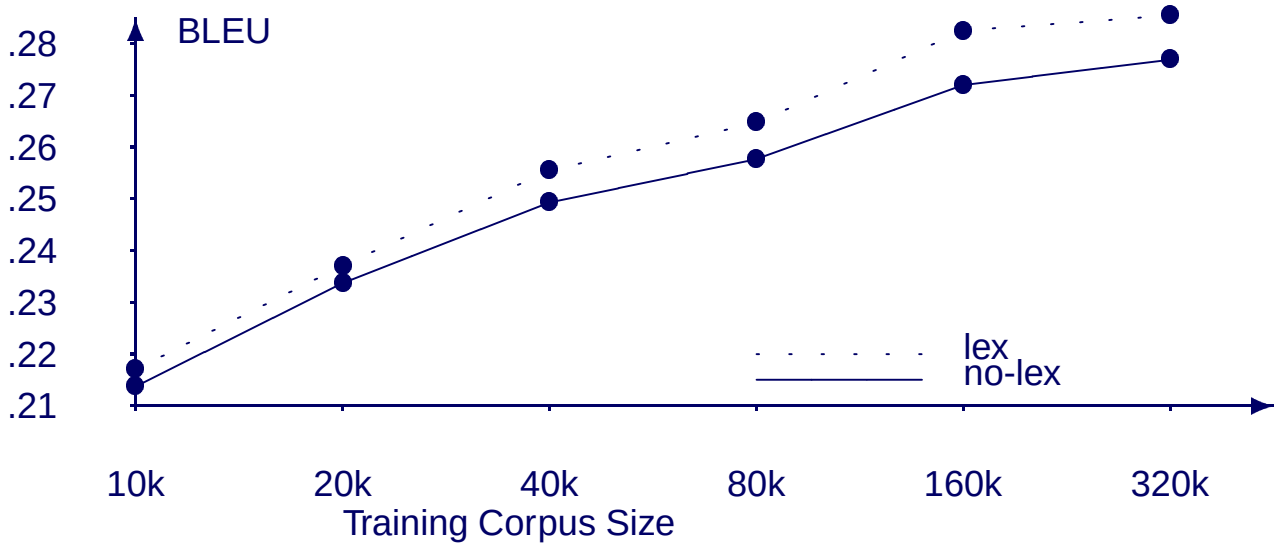
	la	bruja	verde
the	###	---	---
green	---	---	###
witch	---	###	---

- Lexical weight:

$$p_w = w(\text{la}|\text{the}) \times w(\text{bruja}|\text{witch}) \times w(\text{verde}|\text{green})$$

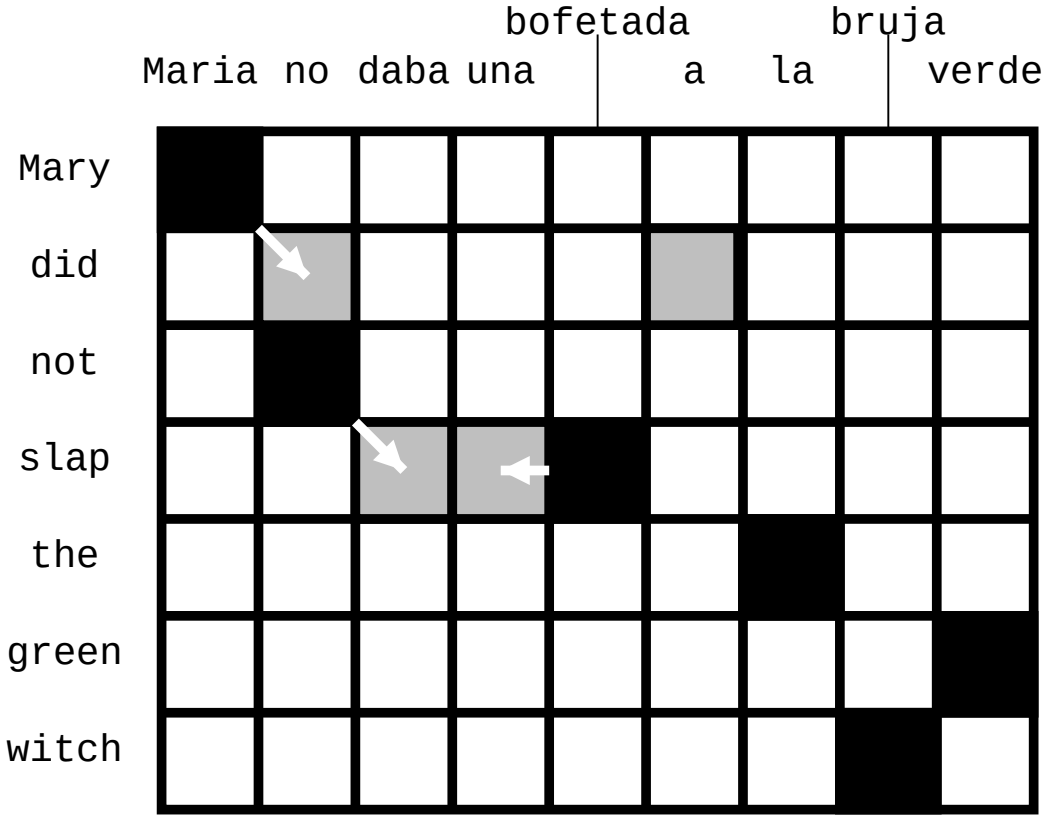
Lexical Weighting

- Improves translation quality



Phrase Extraction Heuristics

- Recall: word alignment based on intersection of bidirectional IBM Model 4 alignments + heuristics

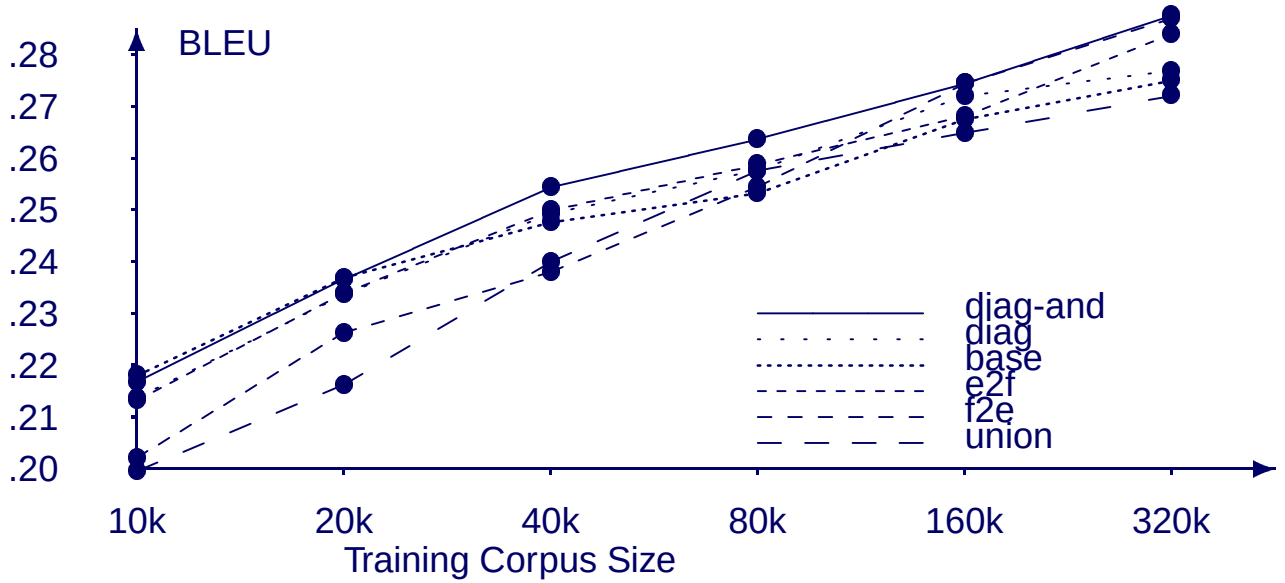


Phrase Extraction Heuristics (2)

- Different phrases are learned, if heuristic to create word alignment is changed.
- Variations in heuristics:
  - only to directly neighboring
  - also to diagonally neighboring
  - also to non-neighboring
  - prefer English-foreign or foreign-to-English
  - use lexical probabilities or frequencies
  - extend only to unaligned words
  - ...

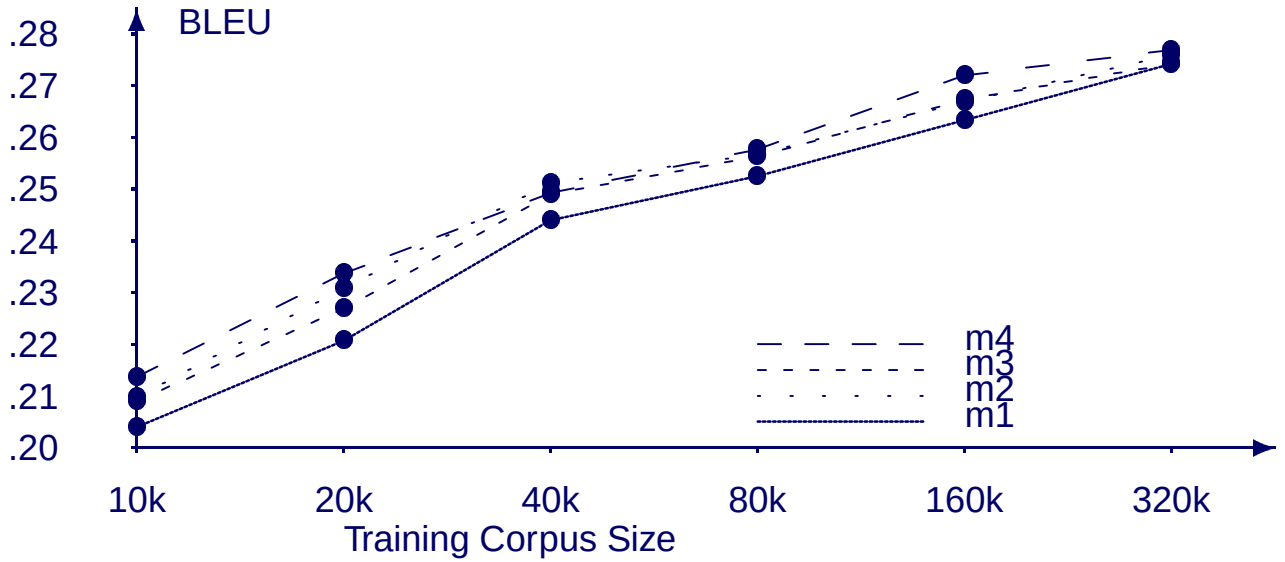
Phrase Extraction Heuristics (3)

- No clear advantage to any strategy
  - large differences, but ...
  - ... depending on corpus size
  - ... depending on language pair



Simpler Word Alignment Models

- Using simpler IBM Models for word alignment
  - not much impact, if simpler models used
  - simpler models computationally much cheaper



Other Language Pairs

- Finding hold for other language pairs, other corpora
  - Phrase translation better than IBM Model 4
  - Lexicalization helps (about +0.01 BLEU)

Language Pair	Model4	Phrase	Lex
English-German	0.2040	0.2361	0.2449
French-English	0.2787	0.3294	0.3389
English-French	0.2555	0.3145	0.3247
Finnish-English	0.2178	0.2742	0.2806
Swedish-English	0.3137	0.3459	0.3554
Chinese-English	0.1190	0.1395	0.1418

Conclusions

- Phrase-based translation better than word-based translation
- Limit to syntactic phrases hurts a lot
- Small phrases (up to 3 words) good enough
- Lexical weighting helpful
- Phrase extraction heuristics matter, but best heuristics vary on corpus size, language pair