Theoretical and Methodological Issues in MT (TMI),
Skövde, Sweden, Sep. 7-9, 2007

Statistical MT from TMI-1988 to TMI-2007:
What has happened?

Hermann Ney
E. Matusov, A. Mauser, D. Vilar, R. Zens

Human Language Technology and Pattern Recognition
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany

Contents

1 History

use of statistics has been controversial in NLP:

- Chomsky 1969:
  ... the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term.
- was considered to be true by most experts in NLP and AI

Statistics and NLP: Myths and Dogmas

History: Statistical Translation

short (and simplified) history:

- **1949 Shannon/Weaver: statistical (=information theoretic) approach**
- **1950–1970 empirical/statistical approaches to NLP ('empiricism')**
- **1969 Chomsky: ban on statistics in NLP**
- **1970–? hype of AI and rule-based approaches**
- **1988 TMI: Brown presents IBM's statistical approach**
- **1988–1995 statistical translation at IBM Research:**
  - corpus: Canadian Hansards: English/French parliamentary debates
  - DARPA evaluation in 1994:
    comparable to 'conventional' approaches (Systran)
- **1992 TMI: *Empiricist vs. Rationalist Methods in MT***
  controversial panel discussion (?)

After IBM: 1995 – ...

limited domain:

- **speech translation:**
  travelling, appointment scheduling,...
- **projects:**
  – **Verbmobil (German)**
  – **EU projects: Eutrans, PF-Star**

’unlimited’ domain:

- **DARPA TIDES 2001-04: written text (newswire):**
  Arabic/Chinese to English
- **EU TC-Star 2004-07: speech-to-speech translation**
- **DARPA GALE 2005-07+:**
  – **Arabic/Chinese to English**
  – **speech and text**
  – **ASR, MT and information extraction**
  – **measure: HTER (= human translation error rate)**

Verbmobil 1993-2000

German national project:
– general effort in 1993-2000: about 100 scientists per year
– statistical MT in 1996-2000: 5 scientists per year

task:

- input: SPOKEN language for restricted domain:
  appointment scheduling, travelling,
  tourism information, ...
- vocabulary size:
  about 10 000 words (=full forms)
- competing approaches and systems
  – end-to-end evaluation
    in June 2000 (U Hamburg)
  – human evaluation (blind):
    is sentence approx. correct: yes/no?

Translation Method	Error [%]
Semantic Transfer	62
Dialog Act Based	60
Example Based	51
Statistical	29

- overall result: statistical MT highly competitive

similar results for European projects:
Eutrans (1998-2000) and PF-Star (2001-2004)

ingredients of the statistical approach:

- **Bayes decision rule:**
  - **minimizes the decision errors**
  - **consistent and holistic criterion**
- **probabilistic dependencies:**
  - **toolbox of statistics**
  - **problem-specific models (in lieu of 'big tables')**
- **learning from examples:**
  - **statistical estimation and machine learning**
  - **suitable training criteria**

approach:

$$\textbf{statistical MT} = \textbf{structural (linguistic?) modelling}$$
$$+ \textbf{ statistical decision/estimation}$$

Analogy: ASR and Statistical MT

Klatt in 1980 about the principles of DRAGON and HARPY (1976);
p. 261/2 in 'Lea, W. (1980): Trends in Speech Recognition':

> **"...the application of simple structured models to speech recognition. It might seem to someone versed in the intricacies of phonology and the acoustic-phonetic characteristics of speech that a search of a graph of expected acoustic segments is a naive and foolish technique to use to decode a sentence. In fact such a graph and search strategy (and probably a number of other simple models) can be constructed and made to work very well indeed if the proper acoustic-phonetic details are embodied in the structure".**

my adaption to statistical MT:

> **"...the application of simple structured models to machine translation. It might seem to someone versed in the intricacies of morphology and the syntactic-semantic characteristics of language that a search of a graph of expected sentence fragments is a naive and foolish technique to use to translate a sentence. In fact such a graph and search strategy (and probably a number of other simple models) can be constructed and made to work very well indeed if the proper syntactic-semantic details are embodied in the structure".**

2 EU Project TC-Star (2004-2007)

March 2007: state-of-the-art for speech/language translation

domain: speeches given in the European Parliament

- **work on a real-life task:**
  - **’unlimited’ domain**
  - **large vocabulary**
- **speech input:**
  - **cope with disfluencies**
  - **handle recognition errors**
- **sentence segmentation**
- **reasonable performance**

Speech-to-Speech Translation

speech in source language

ASR: automatic speech recognition

text in source language

SLT: spoken language translation

text in target language

TTS: text-to-speech synthesis

speech in target language

characteristic features of TC-Star:

- **full chain of core technologies:**
  ASR, SLT(=MT), TTS and their interactions
- **unlimited domain and real-life world task:**
  primary domain: speeches in European Parliament
- **periodic evaluations of all core technologies**

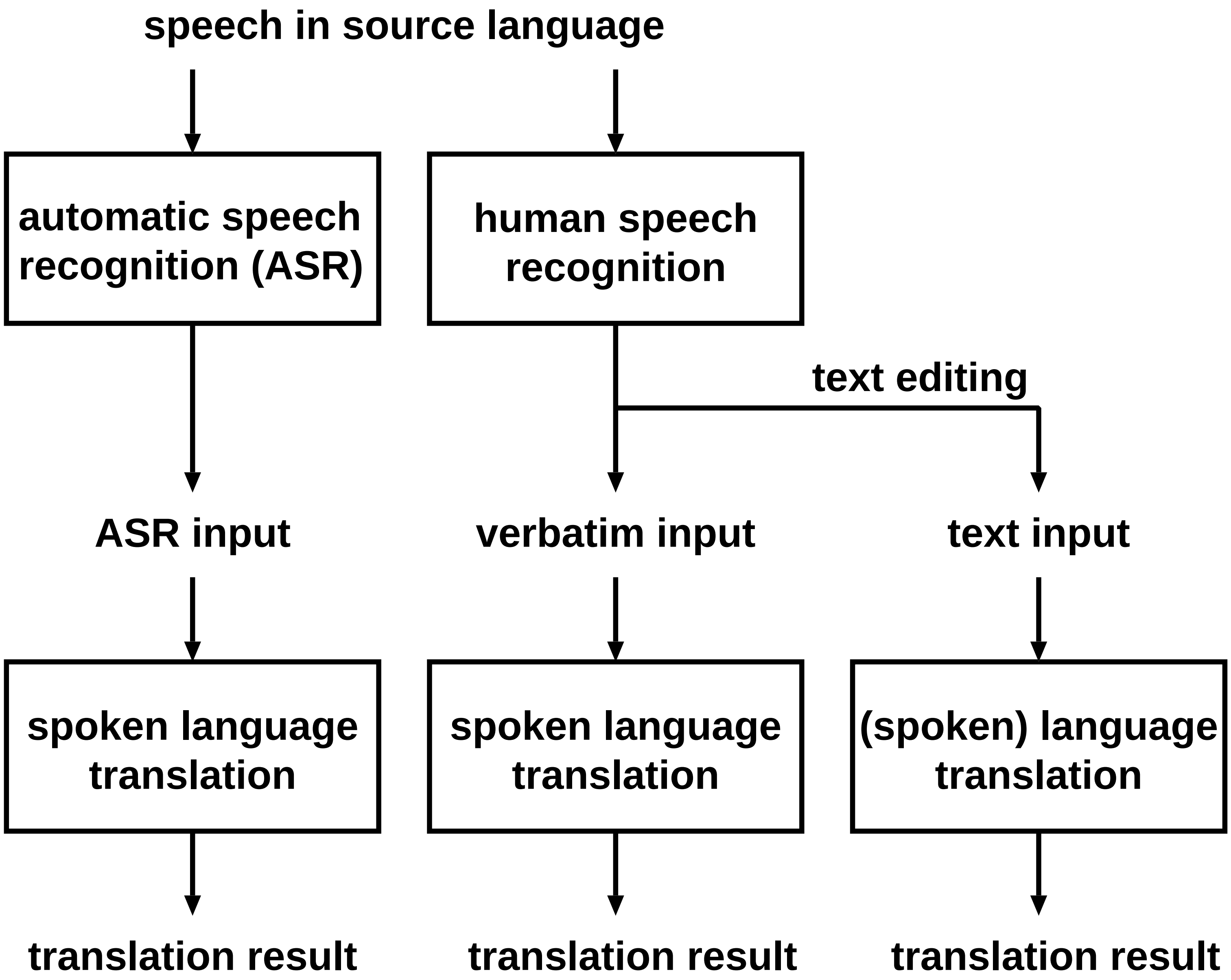
TC-Star: Approaches to MT (IBM, IRST, LIMSI, RWTH, UKA, UPC)

- phrase-based approaches and extensions
  – extraction of phrase pairs, weighted FST, ...
  – estimation of phrase table probabilities
- improved alignment methods
- log-linear combination of models
  (scoring of competing hypotheses)
- use of morphosyntax
  (verb forms, numerus, noun/adjective,...)
- language modelling
  (neural net, sentence level, ...)
- word and phrase re-ordering
  (local re-ordering, shallow parsing, MaxEnt for phrases)
- generation (search):
  efficiency is crucial

- system combination for MT
  - generate improved output from several MT engines
  - problem: word re-ordering

- interface ASR-MT:
  - effect of word recognition errors
  - pass on ambiguities of ASR
  - sentence segmentation

more details: webpage + papers


speech in source language
automatic speech recognition (ASR)
human speech recognition
text editing
ASR input
verbatim input
text input
spoken language translation
spoken language translation
(spoken) language translation
translation result
translation result
translation result

Evaluation 2007: Spanish → English

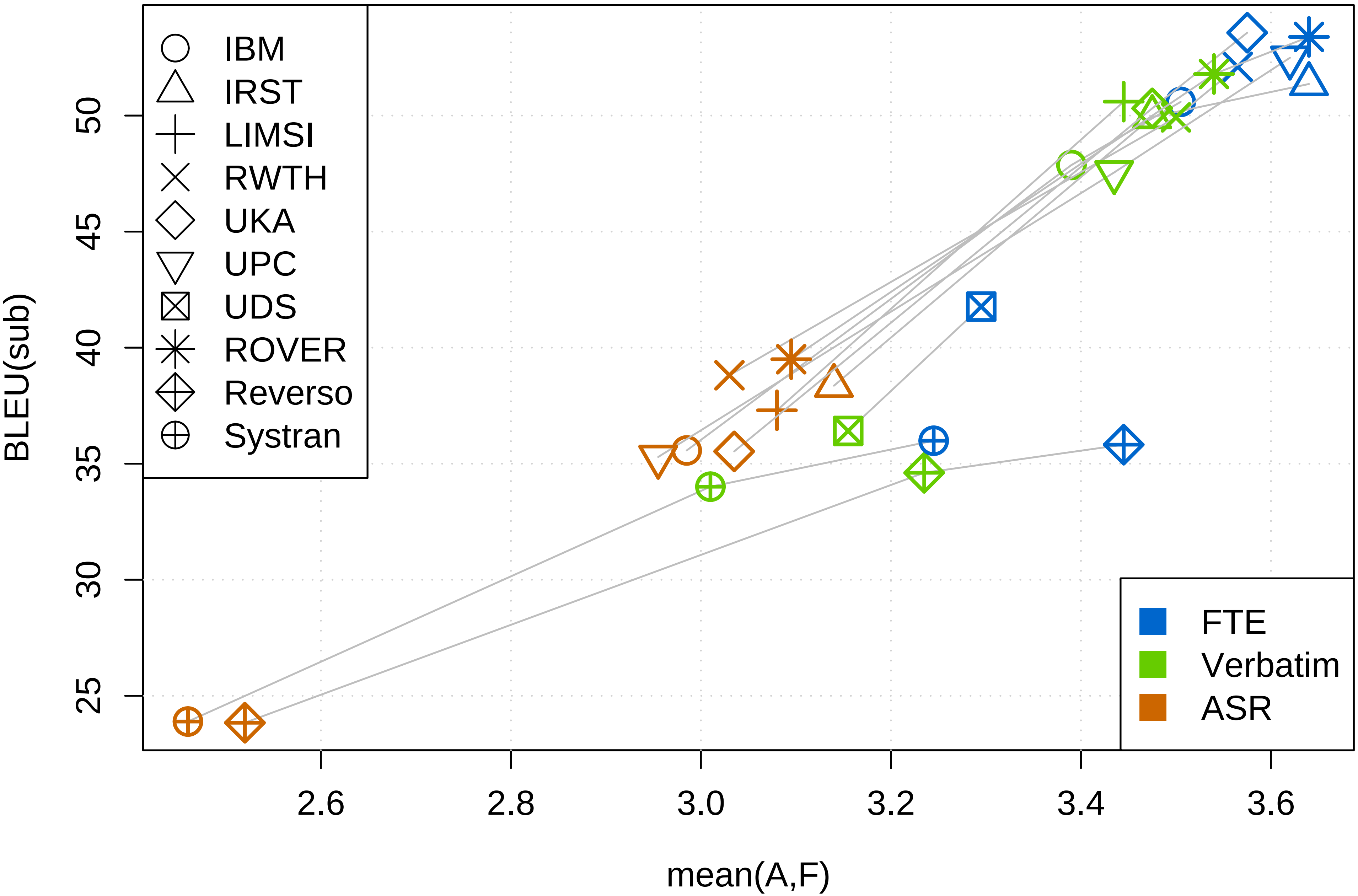
three types of input to translation:

- **ASR: (erroneous) recognizer output**
- **verbatim: correct transcription**
- **text: final text edition**
  (after removing effects of spoken language: false starts, hesitations, ...)

best results (system combination) of evaluation 2007:

Input	BLEU [%]	PER [%]	WER [%]
ASR (WER$= 5.9\%$)	44.8	30.4	43.1
Verbatim	53.5	25.8	35.5
Text	53.6	26.7	37.2

E → S 2007: Human vs. Automatic Evaluation



English $\longrightarrow$ Spanish: Human vs. Automatic Evaluation

observations:

- **good performance:**
  - **BLEU: close to 50%**
  - **PER: close to 30%**
- **fairly good correlation**
  between adequacy/fluency (human) and BLEU (automatic)
- **degradation:**
  from text to verbatim: none or small
  from verbatim to ASR: ΔPER corresponds to ASR errors

Today's Statistical MT

four key components in building today's MT systems:

- **training:**
  word alignment and probabilistic lexicon of (source,target) word pairs
- **phrase extraction:**
  find (source,target) fragments (='phrases') in bilingual training corpus
- **log-linear model:**
  combine various types of dependencies between F and E
- **generation (search, decoding):**
  generate most likely (='plausible') target sentence

ASR: some similar components (not all!)

3 Statistical MT

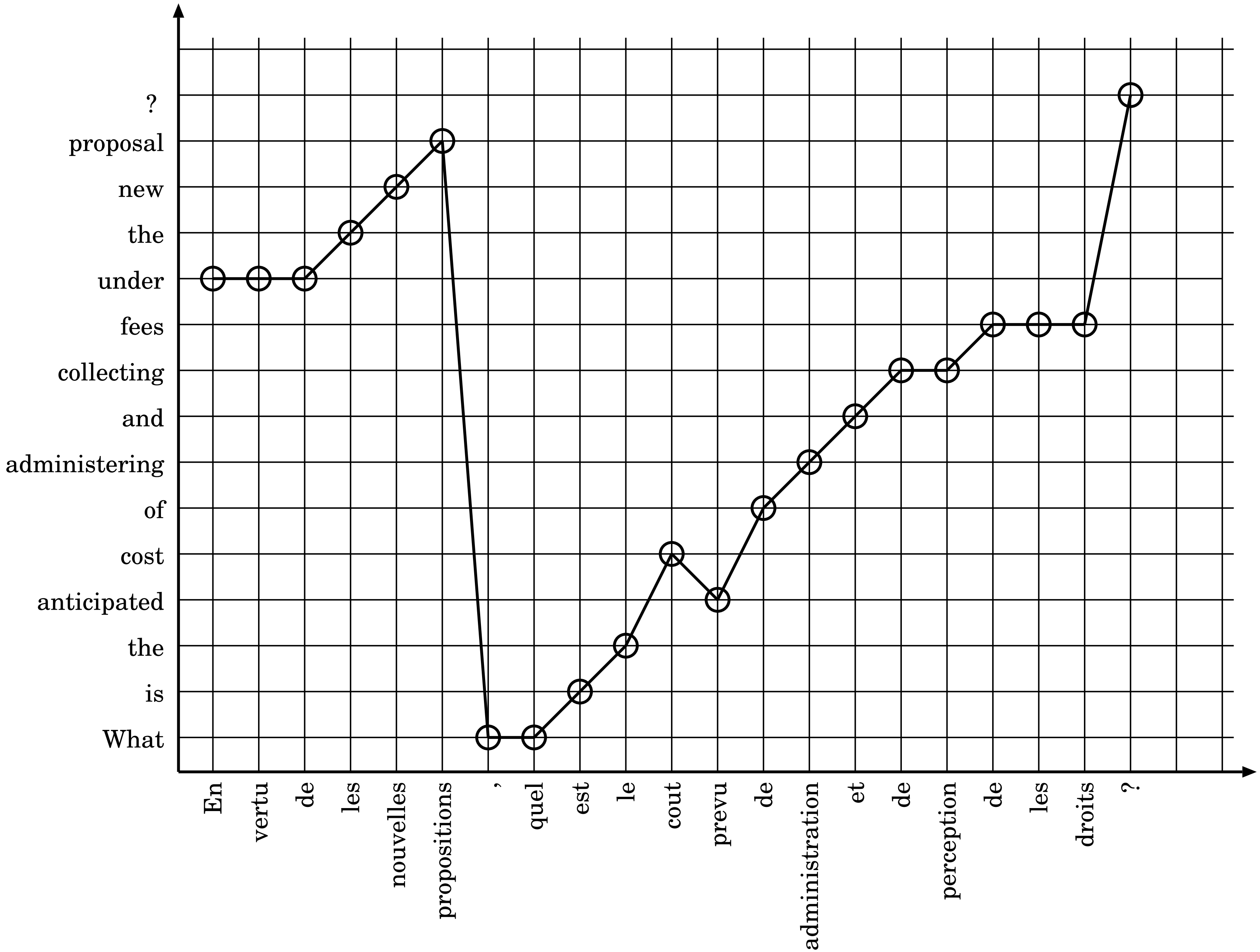
starting point: probabilistic models in Bayes decision rule:

$$F \rightarrow \hat{E}(F) = \arg\max_{E} \left\{ p(E|F) \right\} = \arg\max_{E} \left\{ p(E) \cdot p(F|E) \right\}$$

3.1 Training

- **distributions $p(E)$ and $p(F|E)$:**
  – are unknown and must be learned
  – complex: distribution over strings of symbols
  – using them directly is not possible (sparse data problem)!
- **therefore: introduce (simple) structures by decomposition into smaller 'units'**
  – that are easier to learn
  – and hopefully capture some true dependencies in the data
- **example: ALIGNMENTS of words and positions:**
  bilingual correspondences between words (rather than sentences)
  (counteracts sparse data and supports generalization capabilities)

Example of Alignment (Canadian Hansards)



standard procedure:

- **sequence of IBM-1,...,IBM-5 and HMM models:**
  (conferences before 2000; Comp.Ling.2003+2004)
- **EM algorithm (and its approximations)**
- **implementation in GIZA++**

remarks on training:

- **based on single word lexica** $p(f|e)$ **and** $p(e|f)$**;**
  no context dependency
- **simplifications:**
  only IBM-1 and HMM

alternative concept for alignment (and generation):

ITG approach [Wu ACL 1995/6]

HMM: Recognition vs. Translation

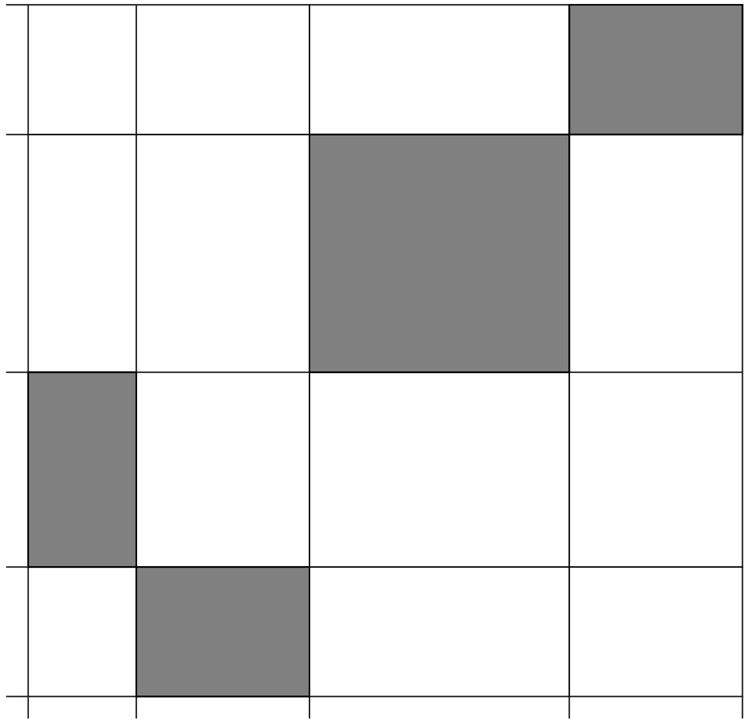
speech recognition	text translation
$Pr(x_1^T \| T, w) =$ $\sum_{s_1^T} \prod_t [p(s_t \| s_{t-1}, S_w, w)\ p(x_t \| s_t, w)]$	$Pr(f_1^J \| J, e_1^I) =$ $\sum_{a_1^J} \prod_j [p(a_j \| a_{j-1}, I)\ p(f_j \| e_{a_j})]$
time $t = 1, ..., T$ observations x_1^T with acoustic vectors x_t states $s = 1, ..., S_w$ of word w path: $t \rightarrow s = s_t$ always: monotonous	source positions $j = 1, ..., J$ observations f_1^J with source words f_j target positions $i = 1, ..., I$ with target words e_1^I alignment: $j \rightarrow i = a_j$ partially monotonous
transition prob. $p(s_t \| s_{t-1}, S_w, w)$ emission prob. $p(x_t \| s_t, w)$	alignment prob. $p(a_j \| a_{j-1}, I)$ lexicon prob. $p(f_j \| e_{a_j})$

3.2 Phrase Extraction

segmentation into two-dim. ’blocks’

blocks have to be “consistent” with the word alignment:

- words within the phrase cannot be aligned to words outside the phrase
- unaligned words are attached to adjacent phrases



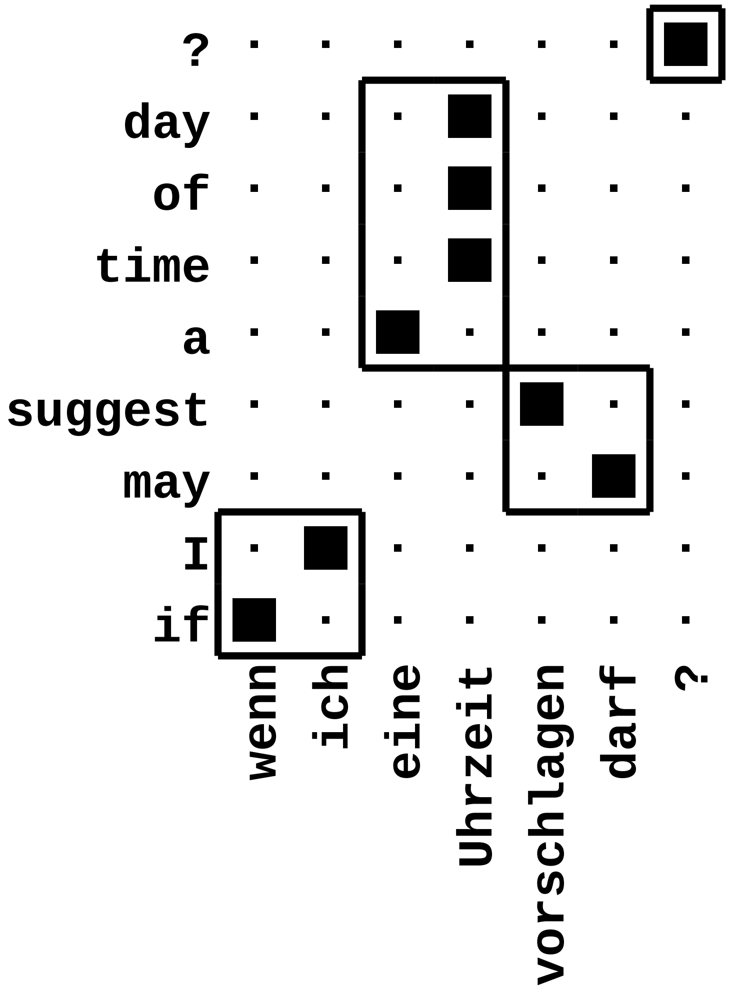
purpose: decomposition of a sentence pair (F, E) into phrase pairs $(\tilde{f}_k, \tilde{e}_k), k = 1, ..., K$:

$$p(E|F) \;=\; p(\tilde{e}_1^K|\tilde{f}_1^K) \;=\; \prod_k p(\tilde{e}_k|\tilde{f}_k)$$

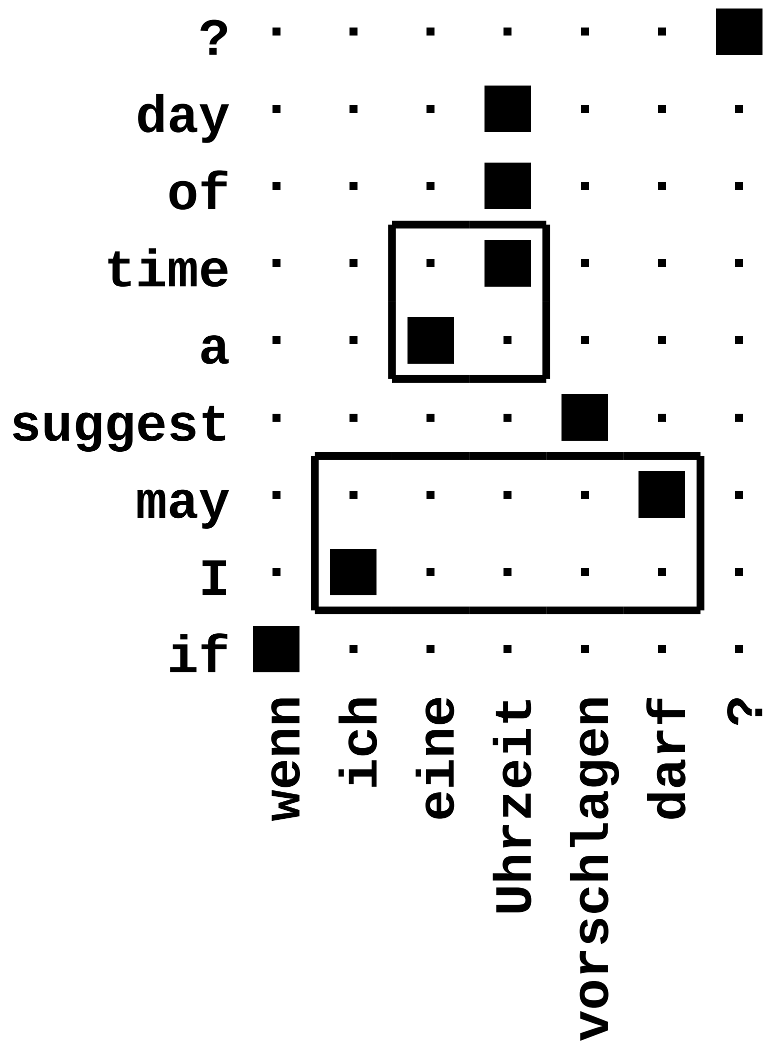
(after suitable re-ordering at phrase level)

Phrase Extraction: Example

possible phrase pairs:



impossible phrase pairs:



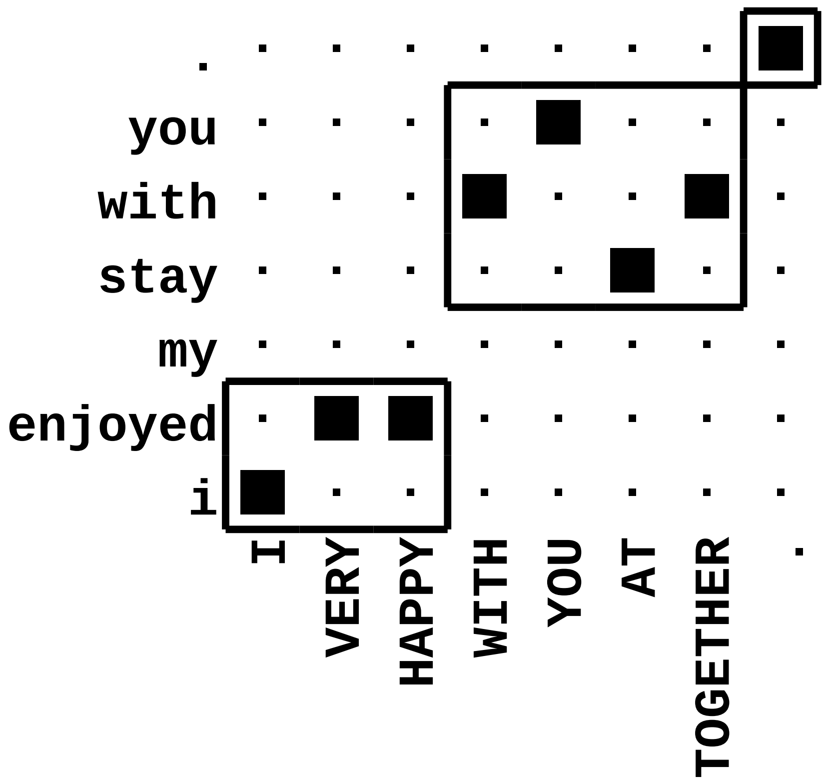
Example: Alignments for Phrase Extraction

source sentence 我 很 高兴 和 你 在 一起 .

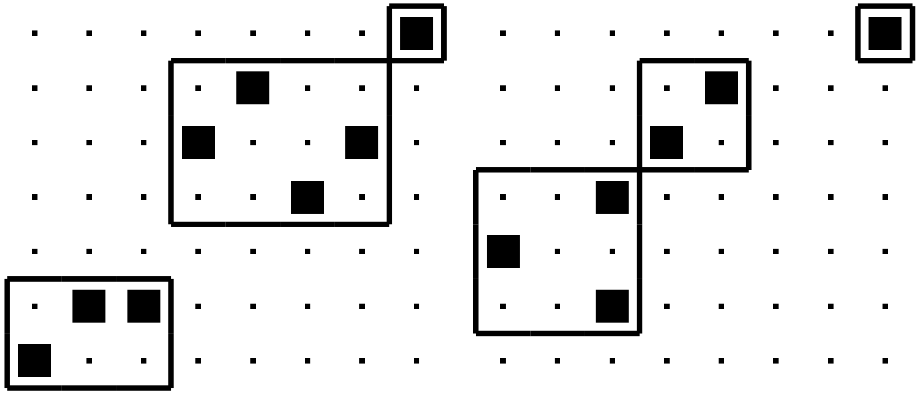
gloss notation **I VERY HAPPY WITH YOU AT TOGETHER .**

target sentence **I enjoyed my stay with you .**

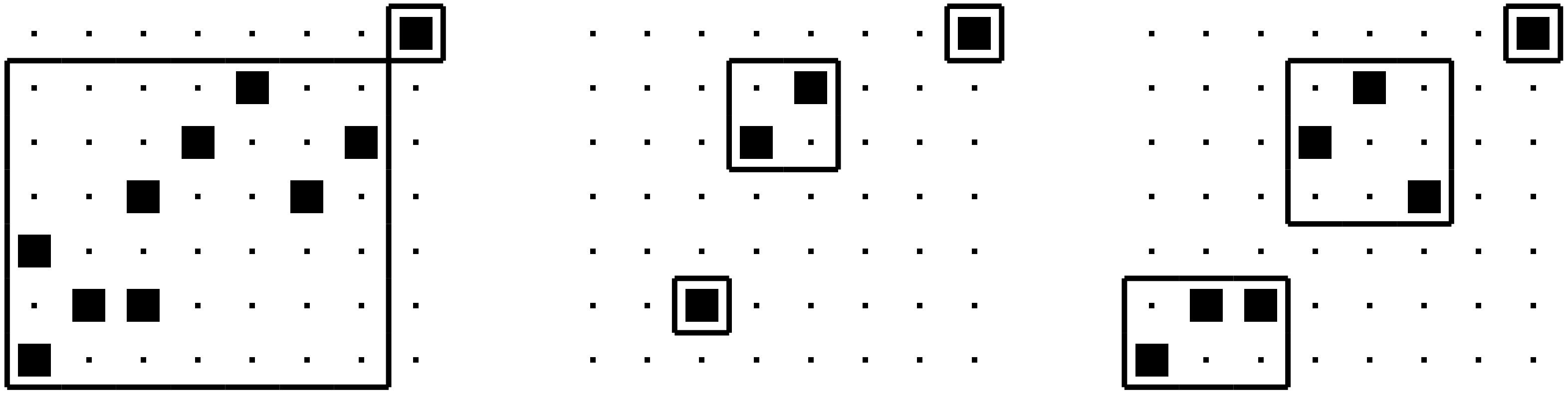
Viterbi alignment for $F \rightarrow E$:



Example: Alignments for Phrase Extraction



Viterbi: $F \longrightarrow E$ Viterbi: $E \longrightarrow F$



union intersection refined

Alignments for Phrase Extraction

most alignment models are asymmetric:
$F \longrightarrow E$ and $E \longrightarrow F$ will give different results

in practice: combine both directions using a simple heuristic

- ***intersection*: only use alignments where both directions agree**
- ***union*: use all alignments from both directions**
- ***refined*: start from *intersection* and include adjacent alignments from each direction**

effect on number of extracted phrases and on translation quality (IWSLT 2005)

heuristic	# phrases	BLEU[%]	TER[%]	WER[%]	PER[%]
union	489 035	49.5	36.4	38.9	29.2
refined	1 055 455	54.1	34.9	36.8	28.9
intersection	3 582 891	56.0	34.3	35.7	29.2

3.3 Phrase Models and Log-Linear Scoring

combination of various types of dependencies using log-linear framework (maximum entropy):

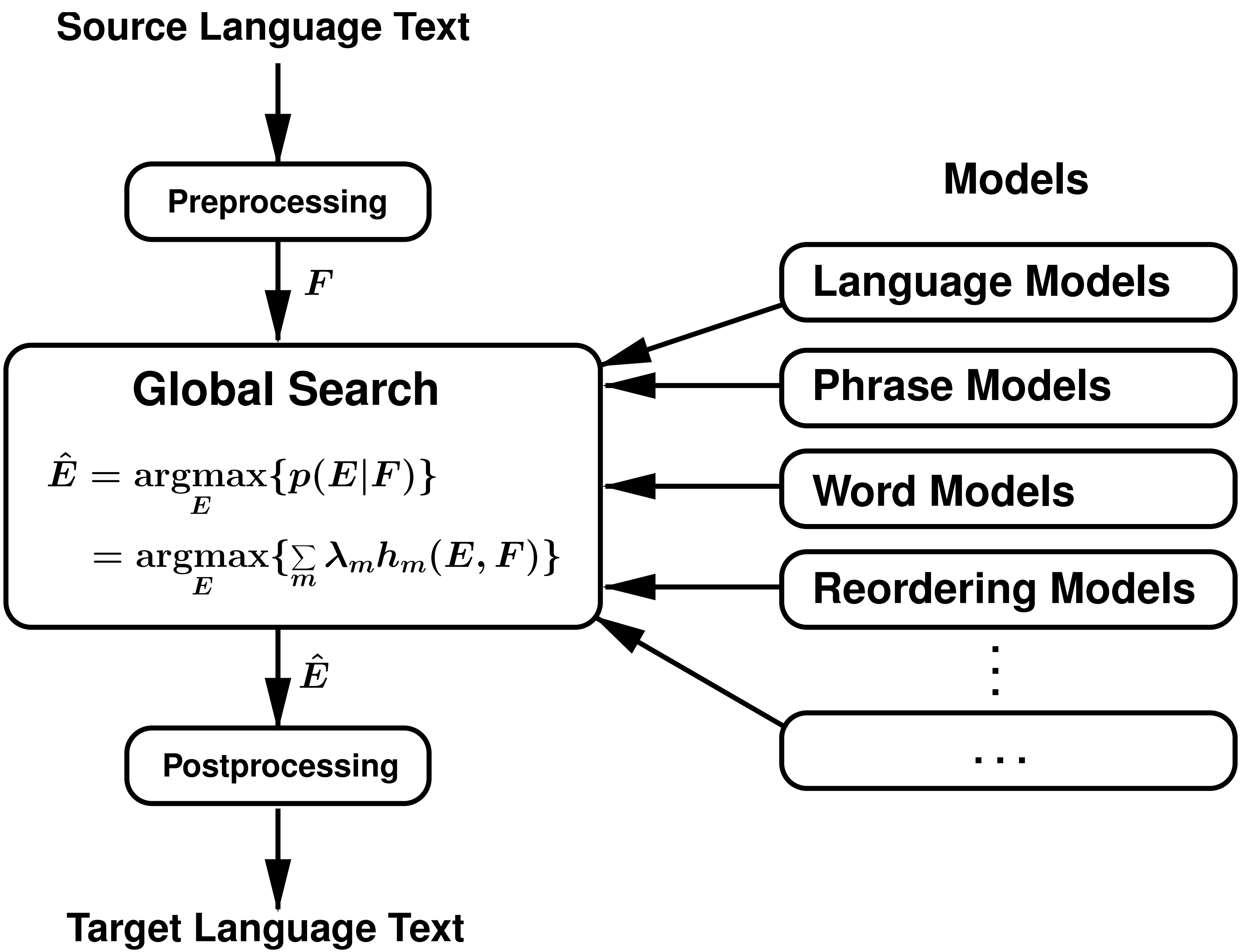
$$p(E|F) = \frac{\exp\left[\sum_m \lambda_m h_m(E,F)\right]}{\sum_{\tilde{E}} \exp\left[\sum_m \lambda_m h_m(\tilde{E},F)\right]}$$

with 'models' (feature functions) $h_m(E,F), m = 1, ..., M$

Bayes decision rule:

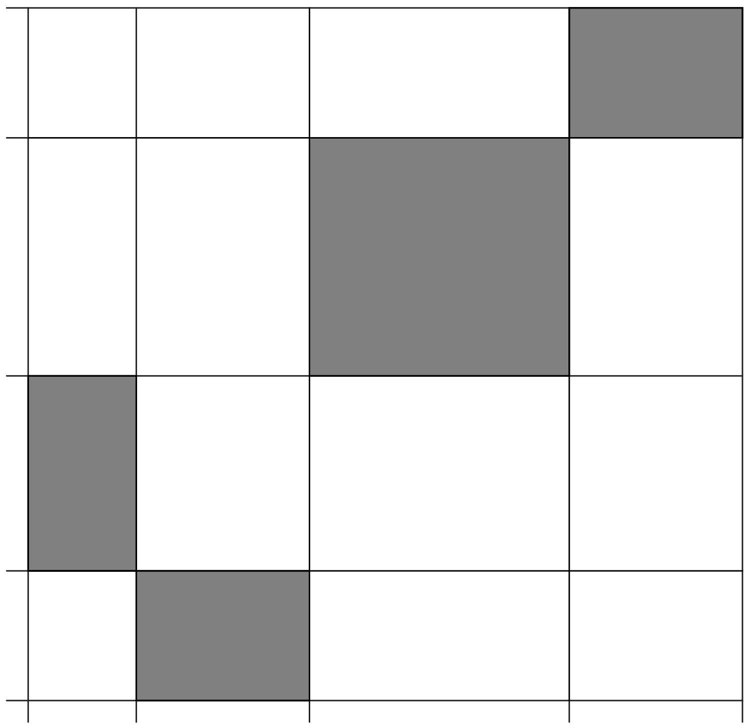
$$\begin{aligned} F \rightarrow \hat{E}(F) &= \operatorname*{argmax}_E \left\{ p(E|F) \right\} = \operatorname*{argmax}_E \left\{ \exp\left[\sum_m \lambda_m h_m(E,F)\right] \right\} \\ &= \operatorname*{argmax}_E \left\{ \sum_m \lambda_m h_m(E,F) \right\} \end{aligned}$$

consequence:
– do not worry about normalization
– include additional 'feature functions' by checking BLEU ('trial and error')


Source Language Text
Preprocessing
F
Global Search
$\hat{E} = \underset{E}{\operatorname{argmax}}\{p(E|F)\}$
$= \underset{E}{\operatorname{argmax}}\{\sum_{m} \lambda_m h_m(E,F)\}$
$\hat{E}$
Postprocessing
Target Language Text
Models
Language Models
Phrase Models
Word Models
Reordering Models
⋮
. . .

Phrase Model Scoring

most models $h_m(E, F)$ are based on segmentation into two-dim. 'blocks' $k := 1, ..., K$



five baseline models:

- phrase lexicon in both directions:
  - $p(\tilde{f}_k|\tilde{e}_k)$ and $p(\tilde{e}_k|\tilde{f}_k)$
  - estimation: relative frequencies
- single-word lexicon in both directions:
  - $p(f_j|\tilde{e}_k)$ and $p(e_i|\tilde{f}_k)$
  - model: IBM-1 across phrase
  - estimation: relative frequencies
- monolingual (fourgram) LM

7 free parameters: 5 exponents + phrase/word penalty

history:

- **Och et al.; EMNLP 1999:**
  – alignment templates ('with alignment information')
  – and comparison with single-word based approach
- **Zens et al., 2002: German Conference on AI, Springer 2002;**
  phrase models used by many groups
  (Och $\longrightarrow$ ISI/Koehn/...)

later extensions,
mainly for rescoring N-best lists:

- **phrase count model**
- **IBM-1 $p(f_j|e_1^I)$**
- **deletion model**
- **word n-gram posteriors**
- **sentence length posterior**

Experimental Results: Chin-Engl. NIST

		BLEU[%]	
Search	**Model**	**Dev**	**Test**
monotone	4-gram LM + phrase model $p(\tilde{f}\|\tilde{e})$	31.9	29.5
	+ word penalty	32.0	30.7
	+ inverse phrase model $p(\tilde{e}\|\tilde{f})$	33.4	31.4
	+ phrase penalty	34.0	31.6
	+ inverse word model $p(e\|\tilde{f})$ (noisy-or)	35.4	33.8
non-monotone	+ distance-based reordering	37.6	35.6
	+ phrase orientation model	38.8	37.3
	+ 6-gram LM (instead of 4-gram)	39.2	37.8

Dev: NIST'02 eval set; Test: combined NIST'03-NIST'05 eval sets

Re-ordering Models

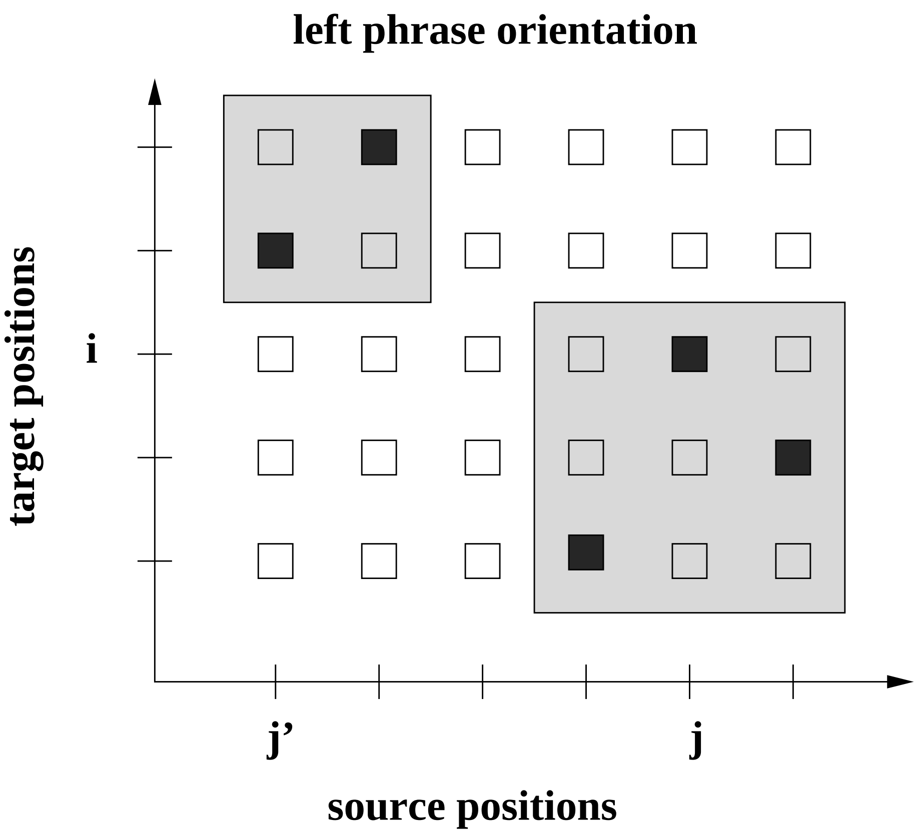
soft constraints ('scores'):

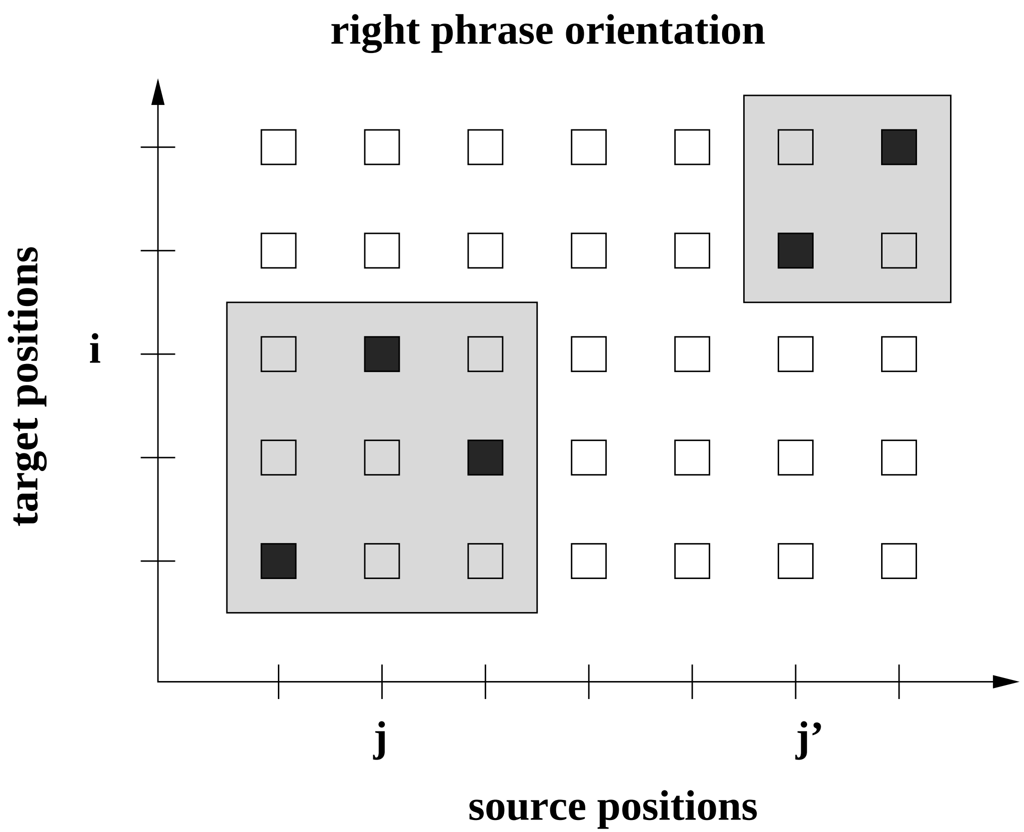
- **distance-based reordering model**
- **phrase orientation model**

hard constraints (to reduce search complexity):

- **level of source words:**
  - local re-ordering
  - IBM (forward) constraints
  - IBM backward constraints
- **level of source phrases:**
  - IBM constraints (e.g. #skip=2)
  - side track: ITG constraints

Phrase Orientation Model





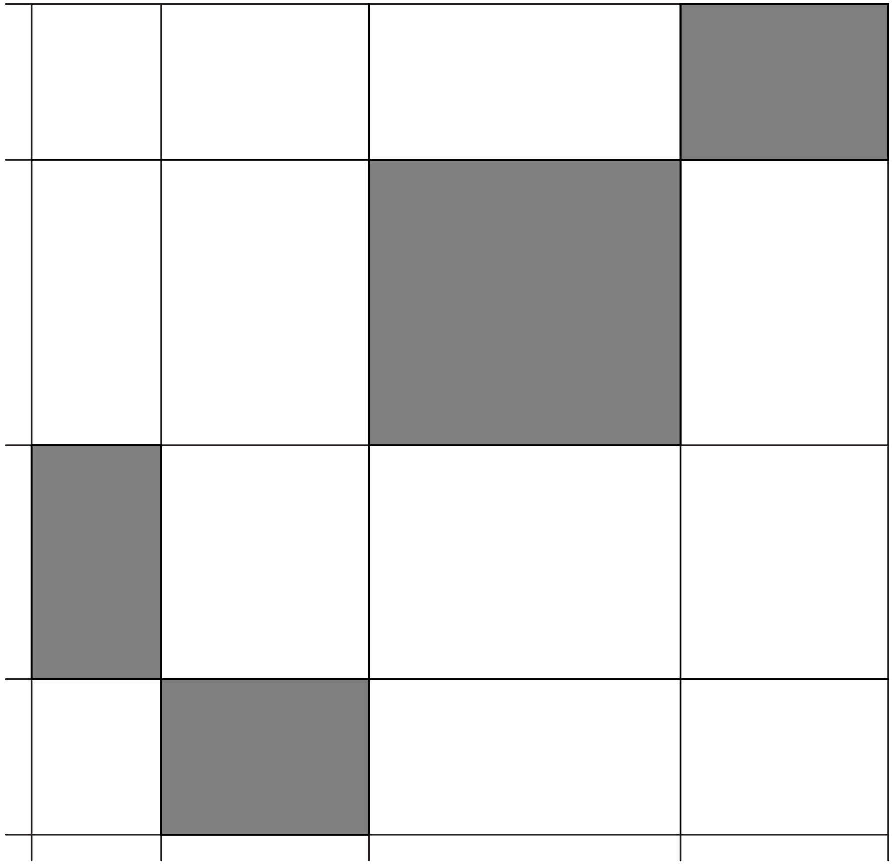
Re-ordering Constraints

dependence on specific language pairs:

- **German - English**
- **Spanish - English**
- **French - English**
- **Japanese - English (BTEC)**
- **Chinese - English**
- **Arabic - English**

3.4 Generation

constraints:
no empty phrases, no gaps
and no overlaps

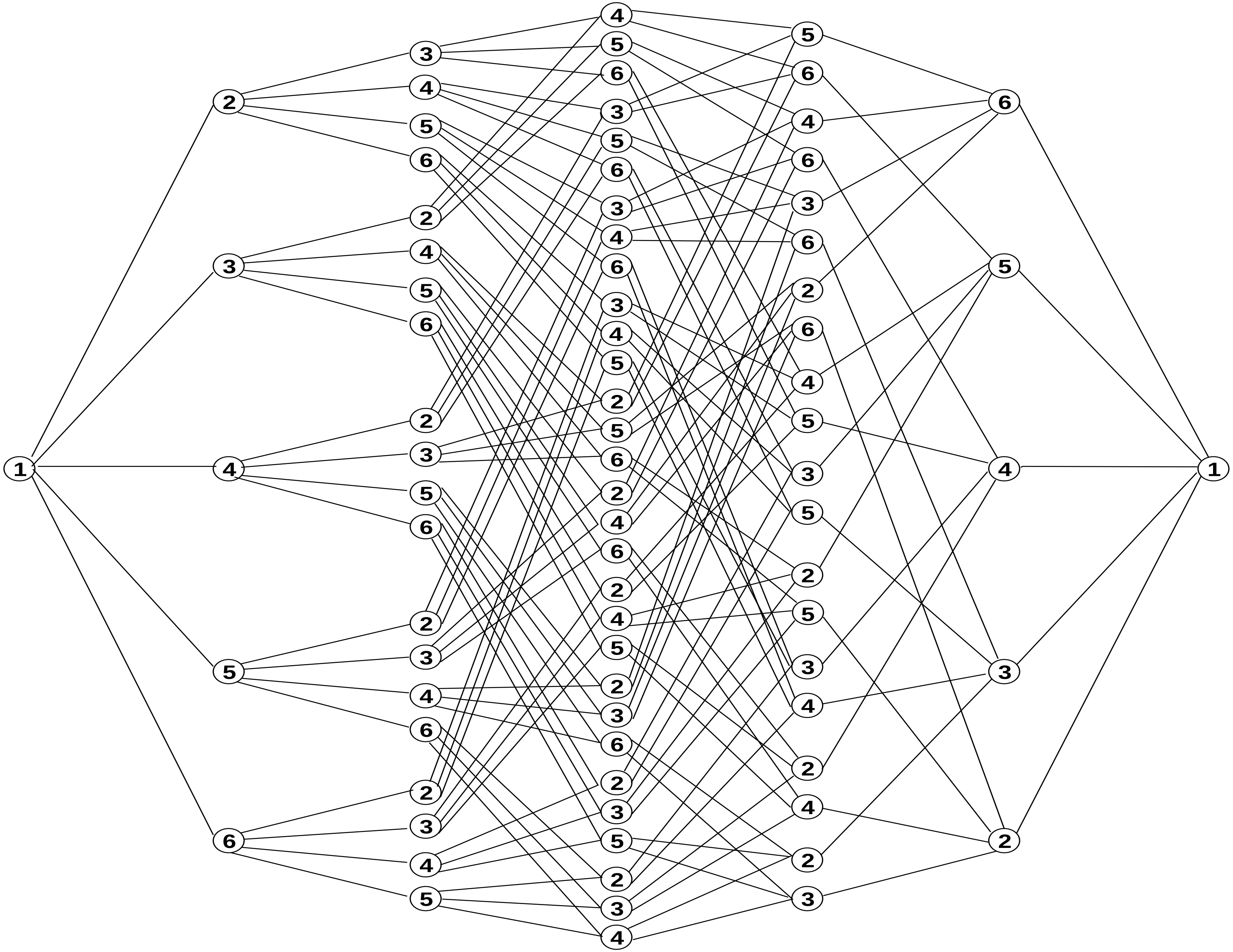


operations with interdependencies:
– find segment boundaries
– allow re-ordering in target language
– find most 'plausible' sentence

similar to: memory-based and
example-based translation

search strategies:
(Tillmann et al.: Coling 2000, Comp.Ling. 2003; Ueffing et al. EMNLP 2002)

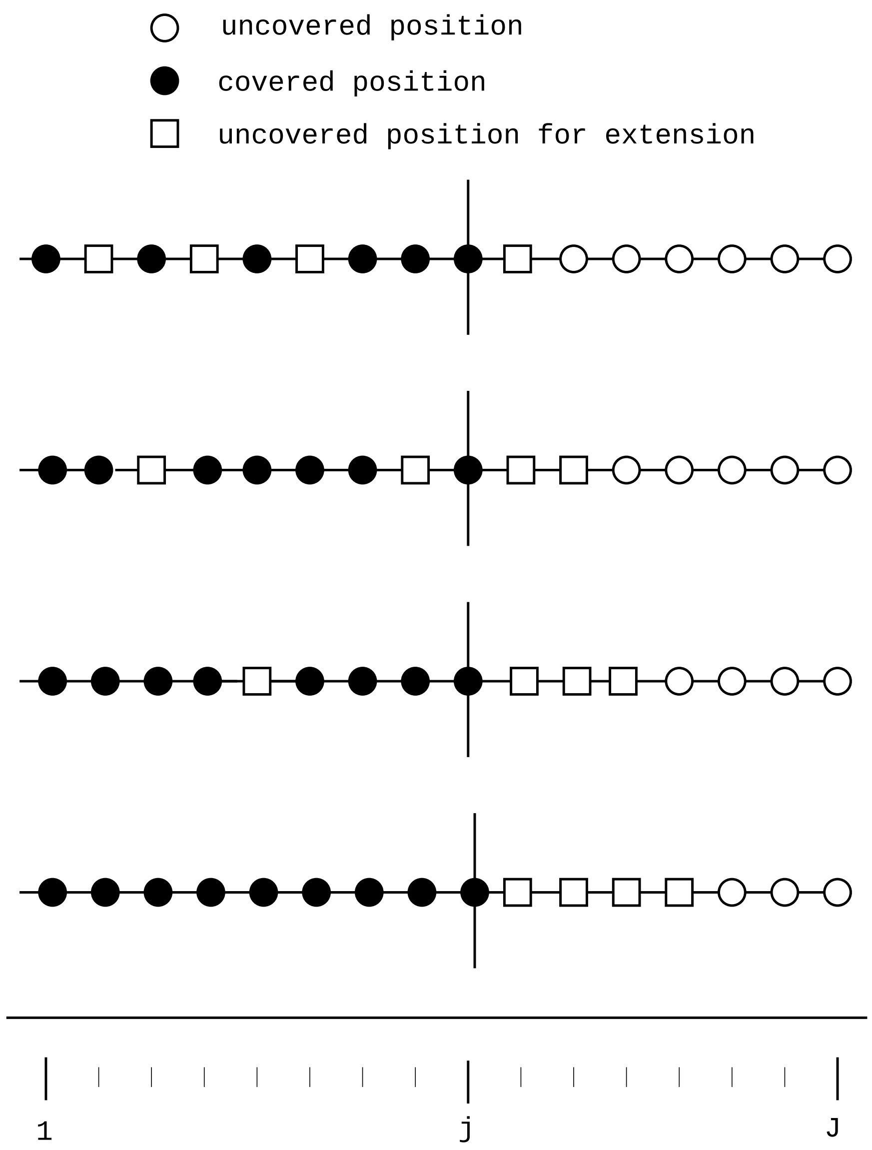
Travelling Salesman Problem: Redraw Network (J=6)



Reordering: IBM Constraints

IBM constraints:
'#skip=3'

result: limited reordering lattice



DP-based Algorithm for Statistical MT

extensions:
– phrases rather than words
– rest cost estimate for uncovered positions

input: source language string $f_1 ... f_j ... f_J$

for each cardinality $c = 1, 2, ..., J$ **do**

    for each set $C \subset \{1, ..., J\}$ **of covered positions with** $|C| = c$ **do**

        for each target suffix string $\tilde{e}$ **do**

            – evaluate score $Q(C, \tilde{e}) := ...$

            – apply beam pruning

traceback:

    – recover optimal word sequence

dynamic programming beam search:

- **build up hypotheses of increasing cardinality:**
  each hypothesis $(C, \tilde{e})$ has two parts:
  coverage hyp. (C) + lexical hyp. $(\tilde{e})$
- **consider and prune competing hypotheses:**
  – with the same coverage vector
  – with the same cardinality
  – additional: observation pruning

Effect of Phrase Length

How does the translation accuracy depend on the length of the 'matching' phrases?

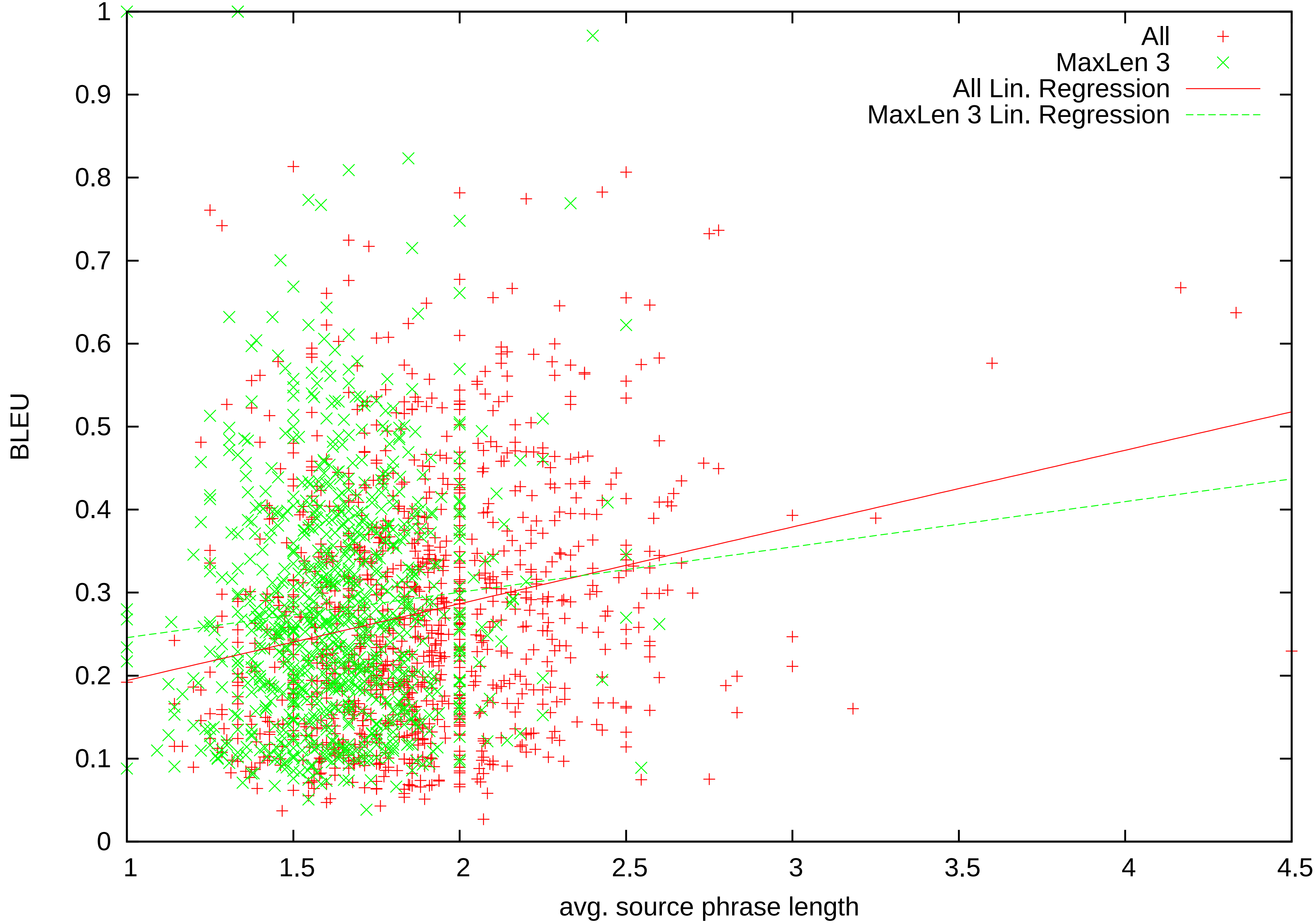
experimental analysis:

- **measure BLEU separately for each sentence**
- **curve:**
  plot BLEU vs. average length of matching phrases

experimental results:

phrase length $1 \rightarrow 3$: BLEU from 20% to 40%

Effect of Phrase Length (Chin.-Engl. NIST)



Conclusions about Statistical MT

memory effect:

- **more and longer matching phrases:**
  help improve translation accuracy
- **today's SMT is closer to example/memory-based MT**
  than 10 years ago

most important difference to example/memory-based MT:

- **consistent scoring**
  (handles weak interdependencies and conflicting requirements)
- **fully automatic training**
  (starting from a sentence-aligned bilingual corpus)

4 Recent Extensions

- **system combination**
- **gappy phrases**
- **statistical MT without data?**

4.1 System Combination

concept for combining translations from several MT engines:

- **align the system outputs:**
  non-monotone alignment (as in training)
- **construct a confusion network from the aligned hypotheses**
- **use weights and language model**
  to select the best translation

- **use of 'adapted' language model:**
  adaptation to translated test sentences
- **10-best lists of each individual system as input**

first work presented at EACL 2006;
(similar approaches in GALE)

Build Confusion Network

Example:

(1+3) system hypotheses with weights	0.25 would your like coffee or tea 0.35 have you tea or coffee 0.10 would like your coffee or 0.30 I have some coffee tea would you like
alignment and re-ordering	have|would you|your $|like coffee|coffee or|or tea|tea would|would your|your like|like coffee|coffee or|or $\|tea I\|$ would|would you|your like|like have|$ some|$ coffee|coffee $|or tea|tea

Extract Consensus Translation

- introduce confidence factors for each system and "vote"

confusion network	$	would	your	like	$	$	coffee	or	tea
	$	have	you	$	$	$	coffee	or	tea
	$	would	your	like	$	$	coffee	or	$
	I	would	you	like	have	some	coffee	$	tea
voting	$/0.7	would/0.65	you/0.65	$/0.35	$/0.7	$/0.7	coffee/1.0	or/0.7	tea/0.9
	I/0.3	have/0.35	your/0.35	like/0.65	have/0.3	some/0.3		$/0.3	$/0.1

- refinements:
  - use each system output as primary reference (combine several confusion networks)
  - include language model

Results

combination of 5 MT systems developed for the GALE 2007 evaluation (Arabic NIST05, case-insensitive):

	PER [%]	BLEU [%]	TER [%]
worst system	33.9	44.2	47.4
best system	28.4	55.3	38.9
combination	27.7	57.1	36.8

- **often: improvements,**
  in particular for ERROR measures (like PER)
- **word re-ordering and alignment:**
  sentence structure is not always preserved
- **“adapted” language model gives a bonus to n-grams present in the original phrases**
- **question: What is the human performance?**

Effect of individual system combination components:
(TC-STAR 2007 evaluation data, English-to-Spanish, verbatim condition)

	BLEU[%]	WER[%]	PER[%]	NIST
worst single system	49.3	39.8	30.0	9.95
best single system	52.4	36.7	27.9	10.45
system combination:				
single confusion net (uniform weights)	53.0	35.3	27.1	10.60
+ manual weight	53.4	35.5	27.0	10.62
+ union of all confusion nets	53.8	35.6	26.8	10.60
+ adapted LM	54.3	35.2	27.4	10.65
+ automatic weight optimization	54.5	35.5	27.5	10.62

Shortcomings of Present MT Rover

Task: TC-STAR 2006 Spanish-to-English evaluation data, 300 sentences

"Human MT Rover": human experts generate the output sentence.

System	BLEU[%]	WER[%]	PER[%]	NIST
worst single system	52.0	35.8	27.2	9.33
best single system	54.1	34.2	25.5	9.47
system combination	55.2	32.9	25.1	9.63
“human” system combination	58.2	31.5	24.3	9.85

result: room for improvement:
– BLEU: from 54.1% to 58.2% (human) vs. 55.2% (automatic)
– both for lexical choices (PER) and word order

4.2 Gappy Phrases

concept:

- **allow for gaps in the phrase pairs**
- **effect: long-distance dependencies**

history:

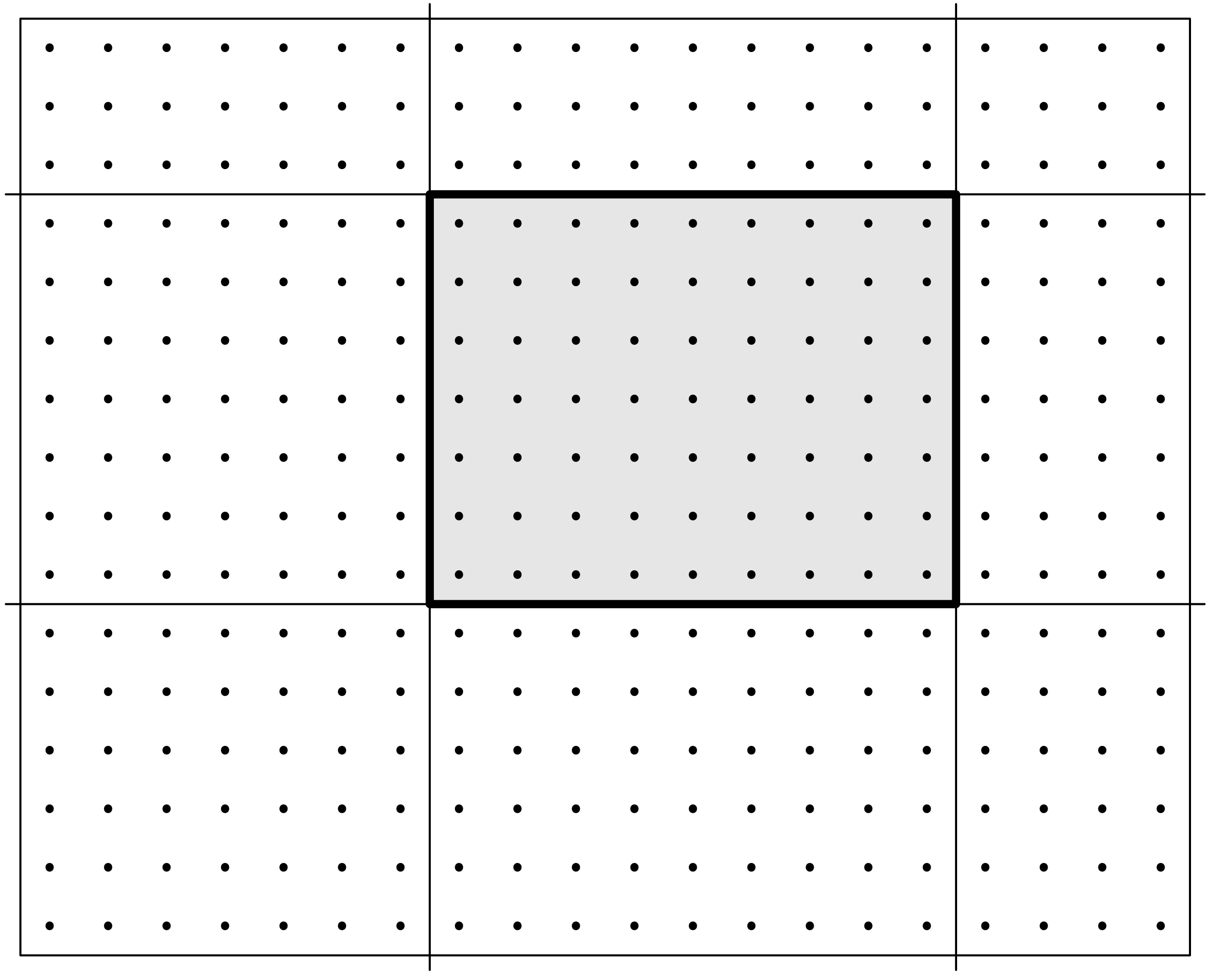
- **McTait & Trujillo 1999: discontiguous translation patterns**
- **U. Block 2000 (Verbmobil): (translation) pattern pairs**
- **R. Zens: diploma thesis 2002, RWTH Aachen (unpublished)**
- **D. Chiang 2005: hierarchical phrases**

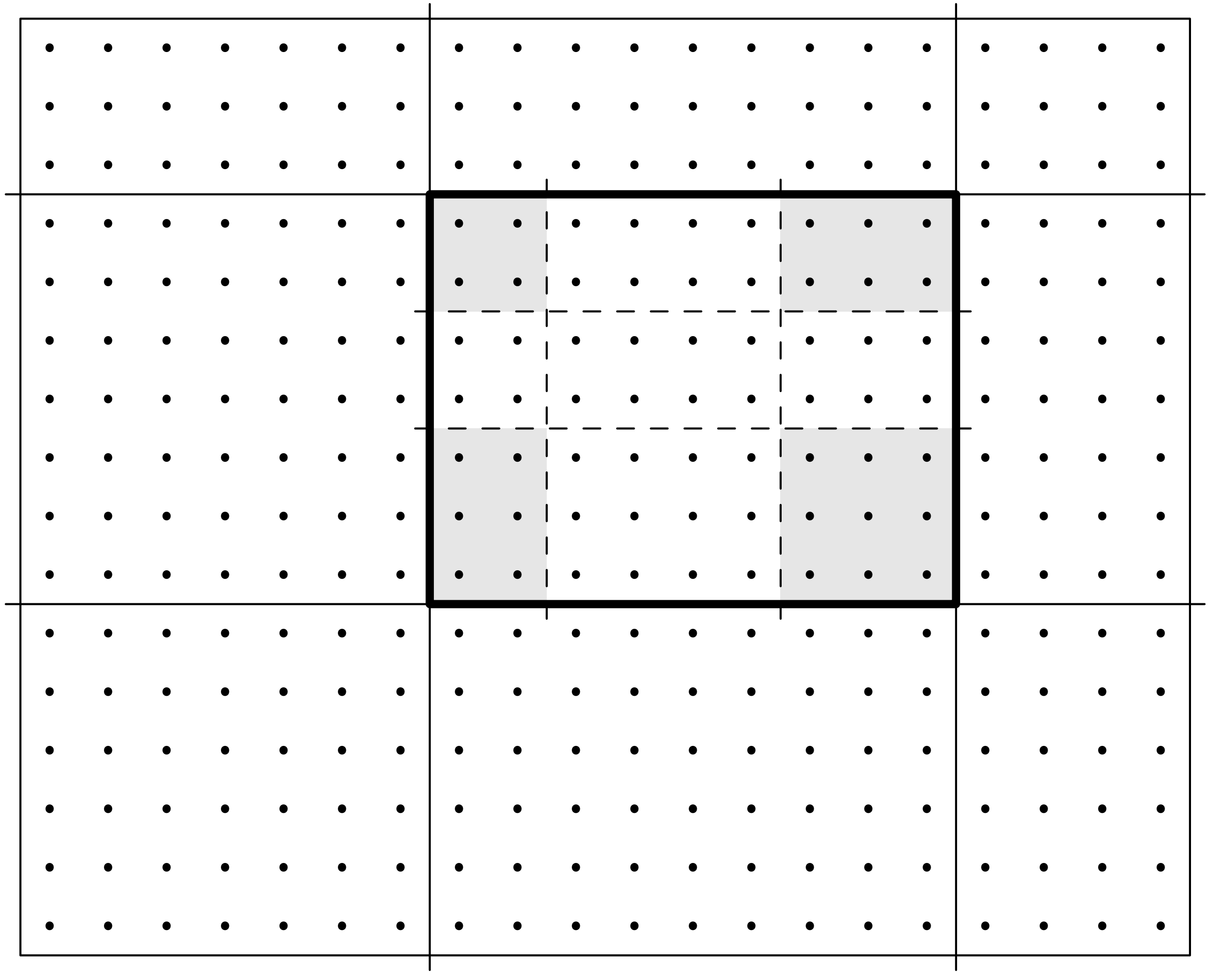
so far: (source,target) phrase pairs (α, β) without gaps:

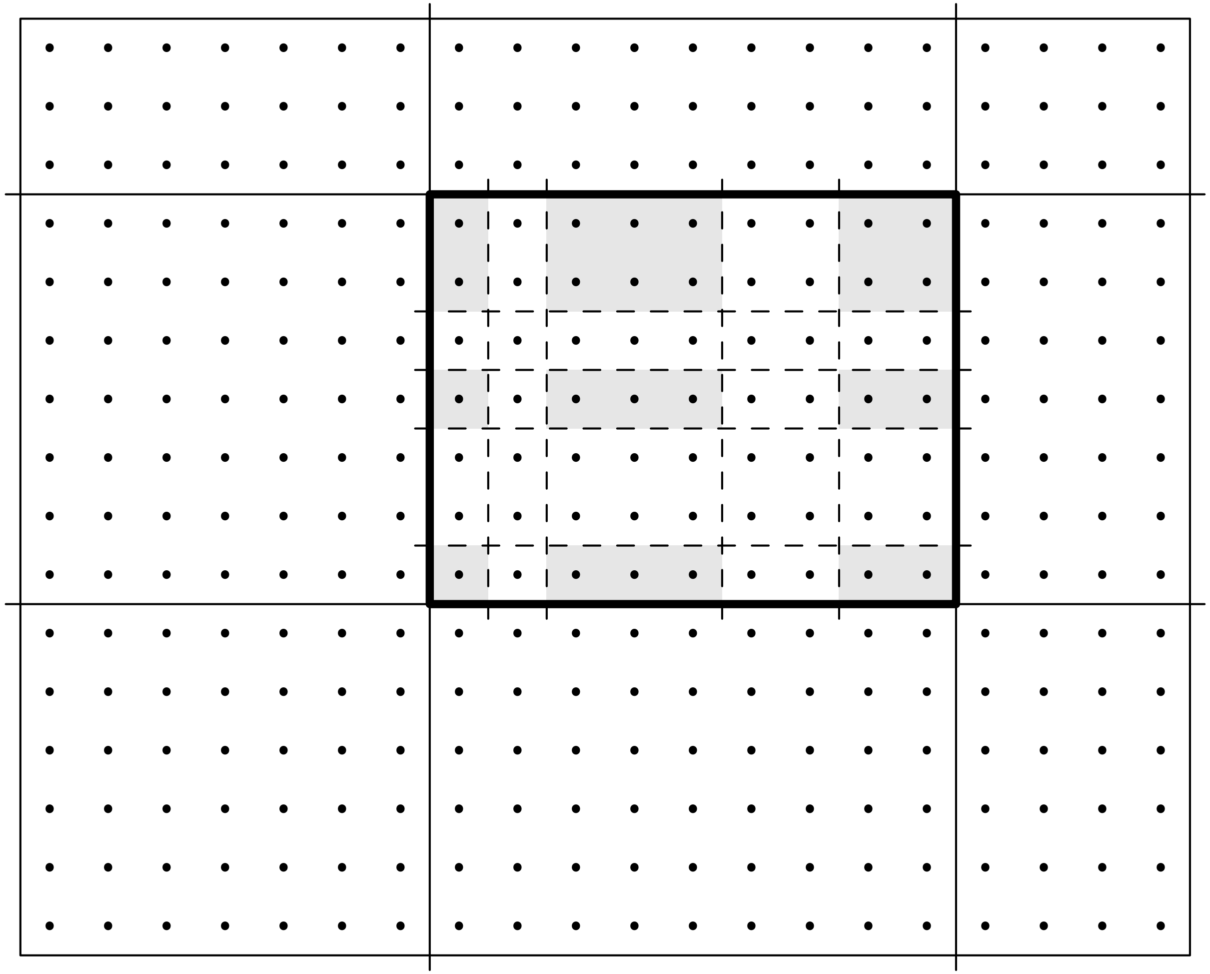
$$p(\beta|\alpha)$$

discontiguous phrase pairs $(\alpha_1 A \alpha_2, \beta_1 B \beta_2)$ WITH gaps (A, B):

$$p(\beta_1 B \beta_2 | \alpha_1 A \alpha_2) \;=\; p(A|B) \cdot p(\beta_1\_\beta_2 | \alpha_1\_\alpha_2)$$







ongoing work:

- **heuristics for gappy phrase extraction**
- **scoring of phrase models**
- **generation (search):**
  top-down vs. bottom-up, efficiency,...

Preliminary Experimental Results

IWSLT 2007, Chinese-to-English task

System	BLEU	TER	WER	PER
mono.PBT	29.6	56.0	58.3	48.9
best PBT	37.2	48.0	48.7	44.3
gappy PBT	35.0	50.5	51.3	46.4

Examples:

best PBT	Please tell me how to get there.
gappy PBT	Do you have any cancellation, please let me know.
Reference	If there is a cancellation, please let me know.

best PBT	Take me to a hospital?
gappy PBT	What should I take to go to the hospital?
Reference	What should I take with me to the hospital?

4.3 Statistical MT With No/Scarce Resources

two aspects of statistical MT:

- **decision process (from source F to target E):**

$$\hat{E} = \arg\max_{E}\{p(E) \cdot p(F|E)\}$$

- **learning the probability models:**
  – **language model $p(E)$: monolingual corpus**
  – **lexicon/translation model $p(F|E)$: bilingual corpus**

idea:

- **bilingual corpus: sometimes difficult to get**
- **substitute: conventional bilingual dictionary (and use uniform prob. distributions)**

consequence: morphology and morphosyntax helpful (all SMT systems use full-form words!)

Spanish→English	WER	PER	BLEU	OOVs
dictionary	60.4	49.3	19.4	20.7
+adjective treatment	56.4	46.8	23.8	18.9
1k	52.4	40.7	30.0	10.6
+dictionary	48.0	36.5	36.0	6.8
+adjective treatment	44.5	34.8	40.9	5.9
13k	41.8	30.7	43.2	2.8
+dictionary	40.6	29.6	46.3	2.4
+adjective treatment	38.3	29.0	49.6	2.2
1.3M	34.5	25.5	54.7	0.14
+adjective treatment	33.5	25.2	56.4	0.14

observations:

- **significant effect of OOV words:**
  difference in PER is largely caused by OOV effect!
- **reasonable translation quality using small corpora**
  dictionary and morpho-syntactic information are helpful

Summary

today's statistical MT:

- **IBM models for word alignment: learning from bilingual data**
- **from words to phrases:**
  phrase extraction, scoring models and generation (search) algorithms
- **experience with various tasks and 'distant' language pairs**
- **text + speech**

helpful conditions:

- **availability of bilingual corpora**
- **automatic evaluation measures**
- **public evaluation campaigns**
- **more powerful computers**
  and algorithms/implementations

THE END