Dublin City University, Dublin, 15-Jun-2007

Text and Speech Translation at RWTH

Hermann Ney
P. Popovic, E. Matusov

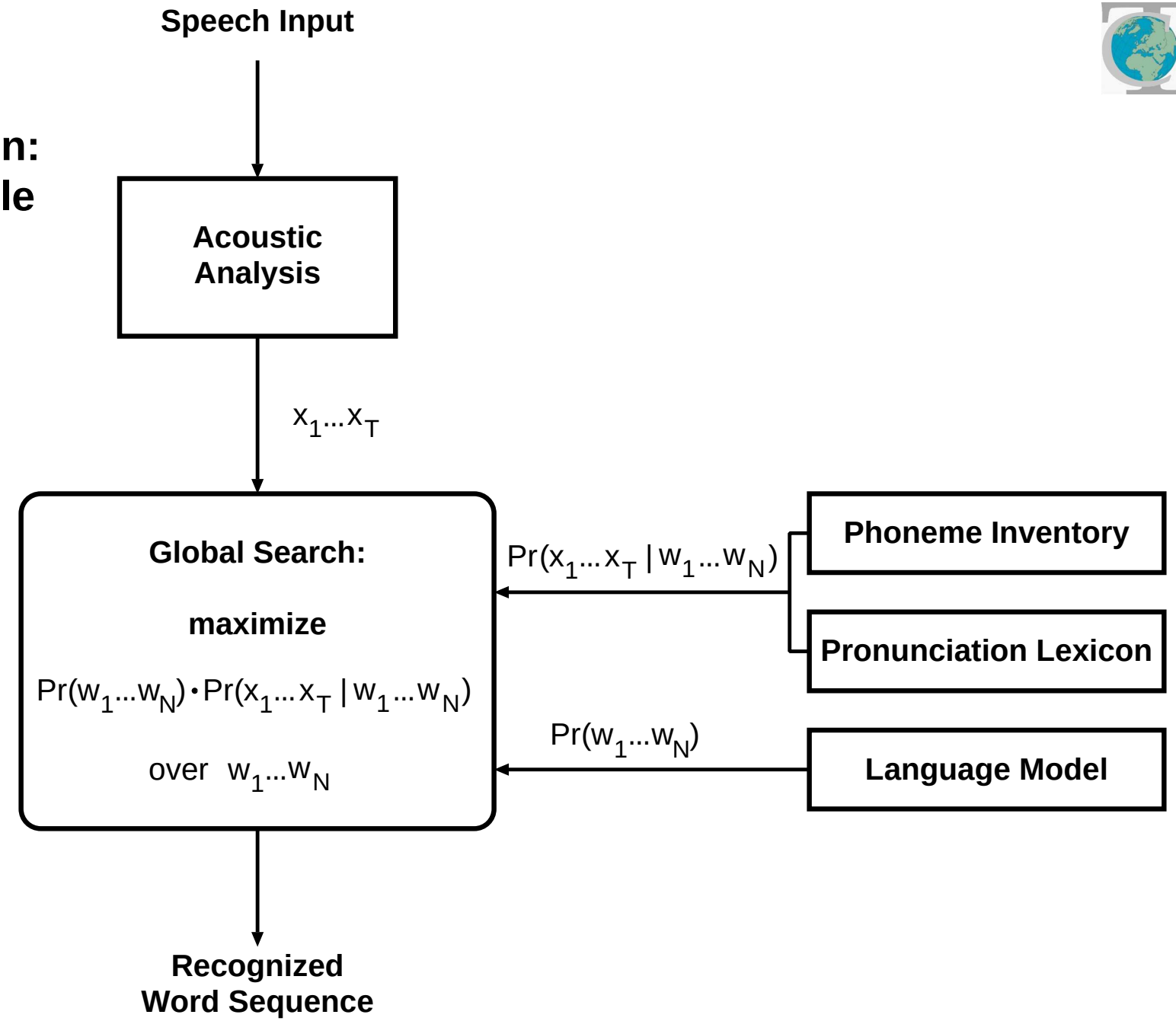
Lehrstuhl für Informatik VI
Human Language Technology and Pattern Recognition
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany

Contents

RWTH Topics: Human Language Technology and Pattern Recognition

- speech recognition:
  - acoustic modelling (VTLN, adaptation, ...)
  - language modelling
  - search
- language processing:
  - text translation
  - speech translation
  - language understanding
  - dialogue systems
  - information retrieval
- image processing:
  - optical character recognition
  - object recognition
  - content-based image retrieval
- sign language recognition and translation

Speech Recognition: Bayes Decision Rule



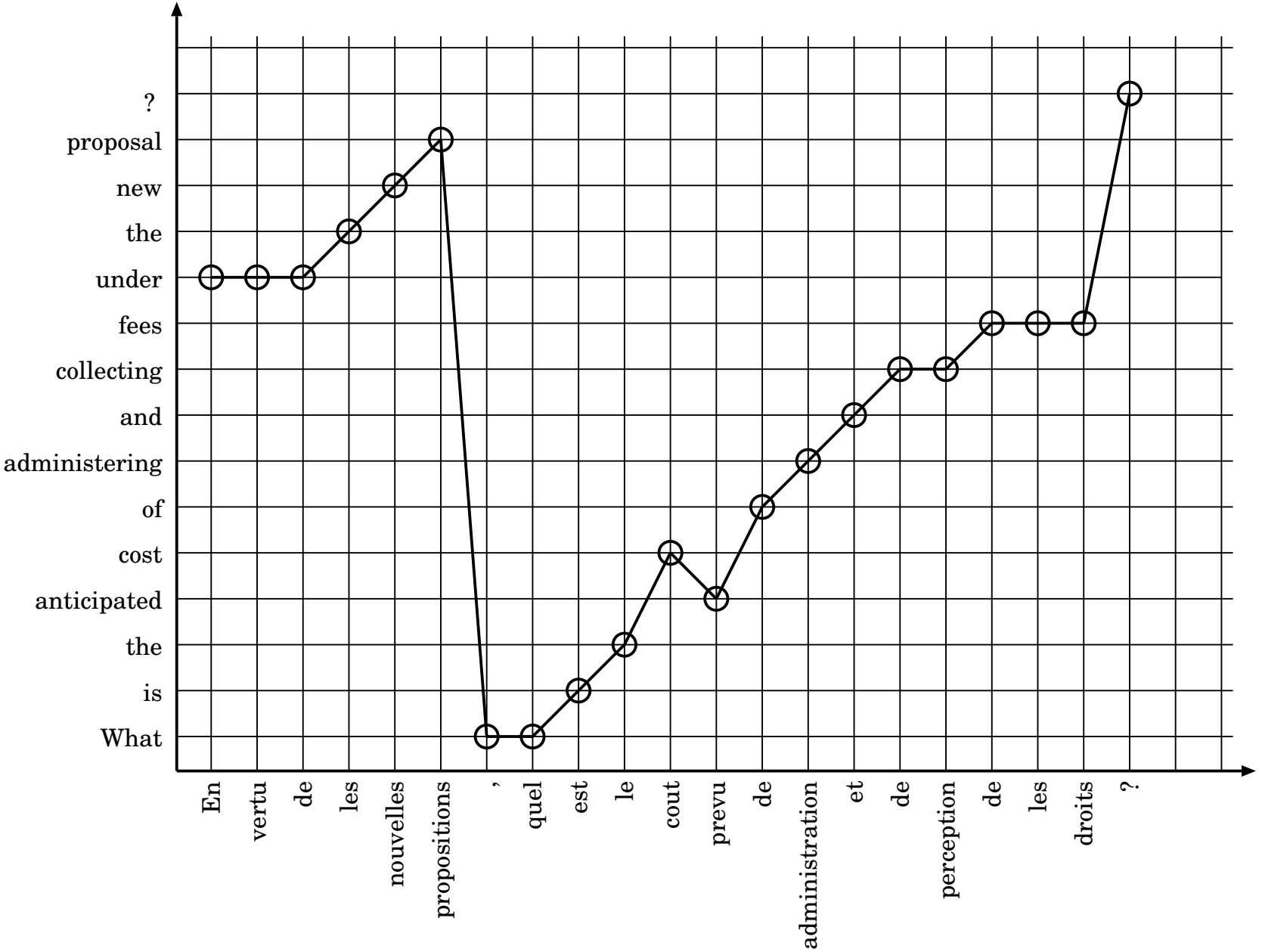
1 Statistical MT and TC-Star

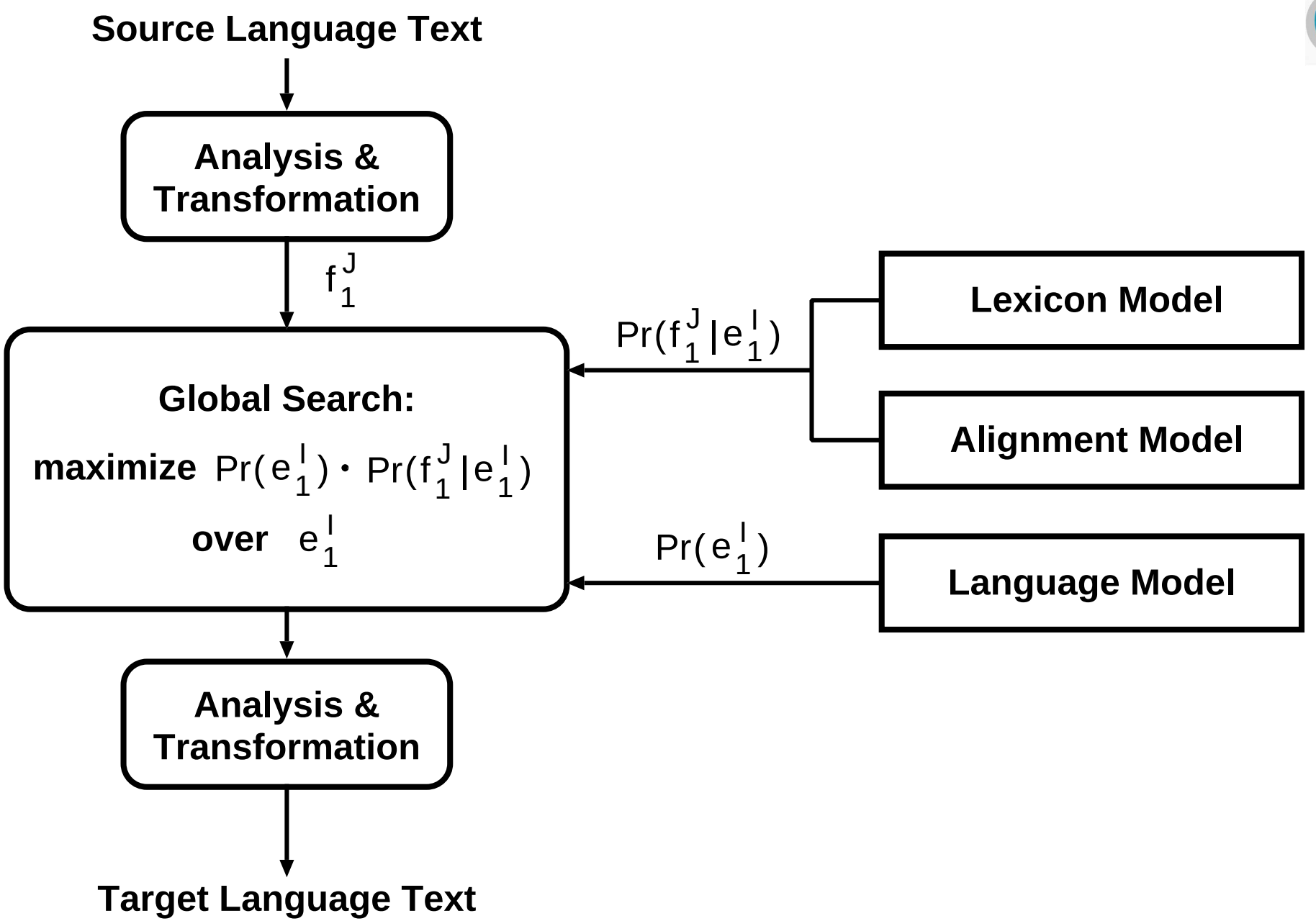
Bayes decision rule:

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

- **distributions $p(E)$ and $p(F|E)$:**
  - **– are unknown and must be learned**
  - **– complex: distribution over strings of symbols**
  - **– using them directly 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)**

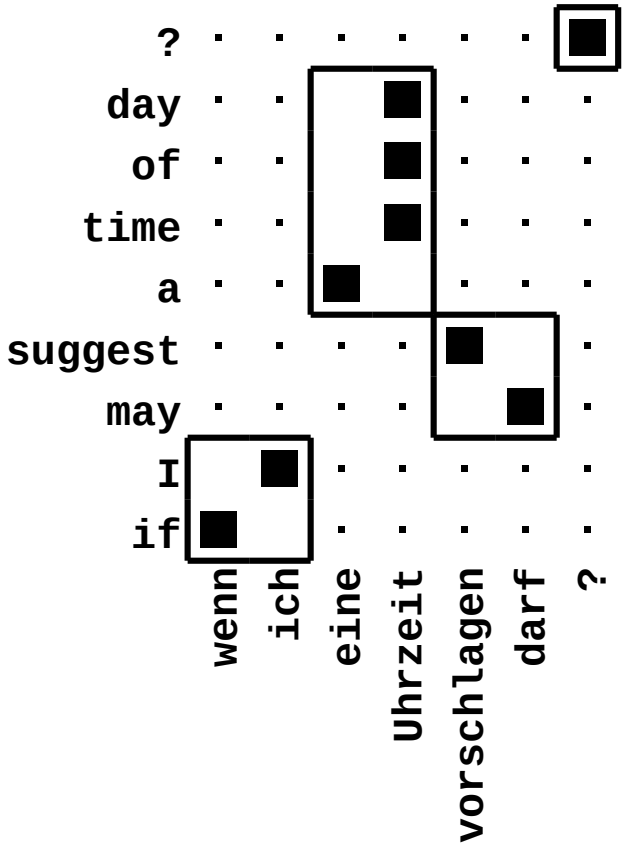
Example of Alignment (Canadian Hansards)



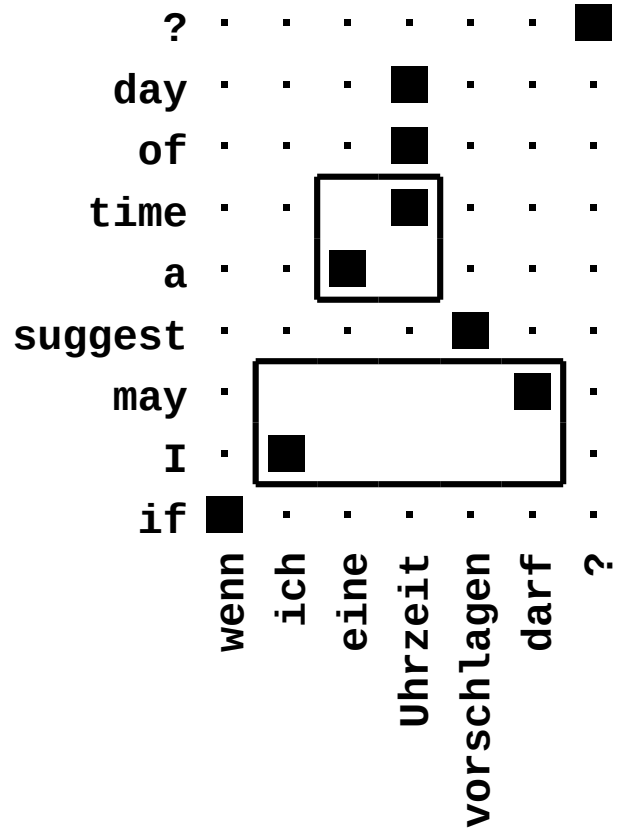

Source Language Text
Analysis & Transformation
f_1^J
Global Search:
maximize $Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)$
over e_1^I
$Pr(f_1^J|e_1^I)$
Lexicon Model
Alignment Model
$Pr(e_1^I)$
Language Model
Analysis & Transformation
Target Language Text

Phrase Extraction: Example

possible phrase pairs:

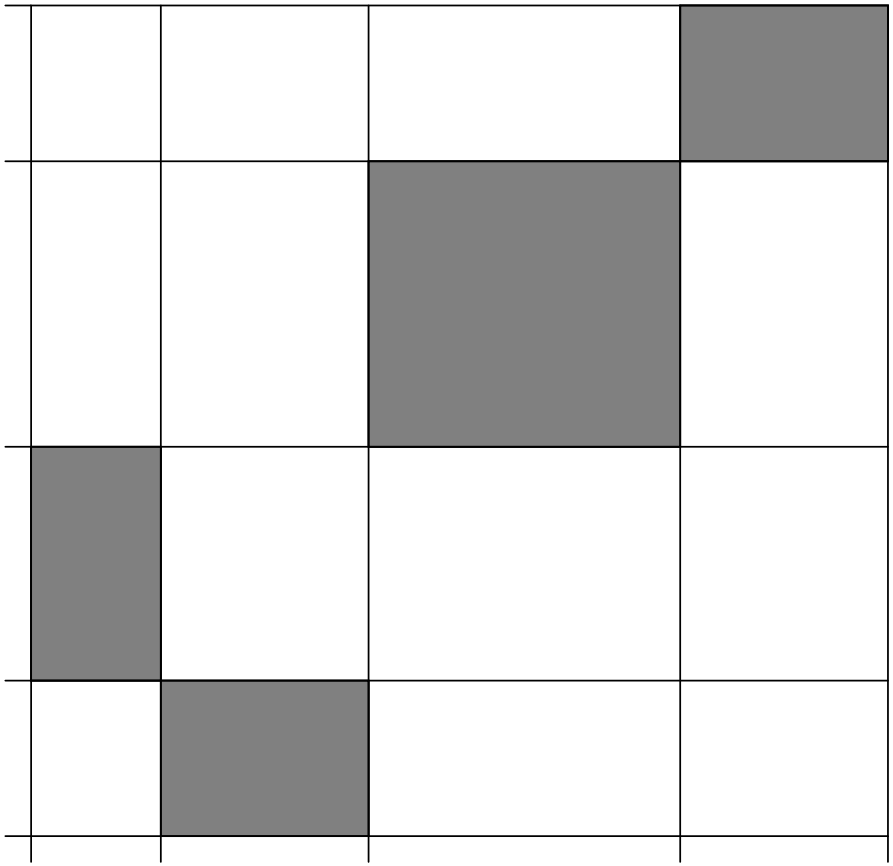


impossible phrase pair:



Translation Using Bilingual Phrases

segmentation into two-dim. ’blocks’
with constraints:
no empty phrases, no gaps
and no overlaps

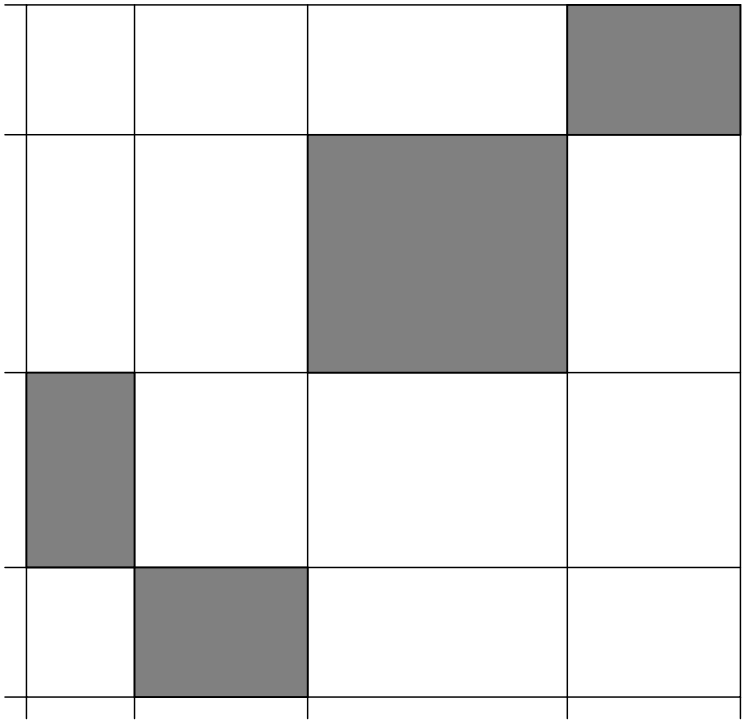


operations with interdependencies:
– find segment boundaries
– allow re-ordering in target language
– find most ’plausible’ sentence

similar to: memory-based and
example-based translation

Phrase-Based Approach: RWTH

segmentation into two-dim. 'blocks' $k := 1, ..., K$



translation based on segmentation and 5 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 (trigram) LM**

7 free parameters: 5 exponents + phrase/word penalty

Challenges for TC-Star

domain: speeches given in the European Parliament

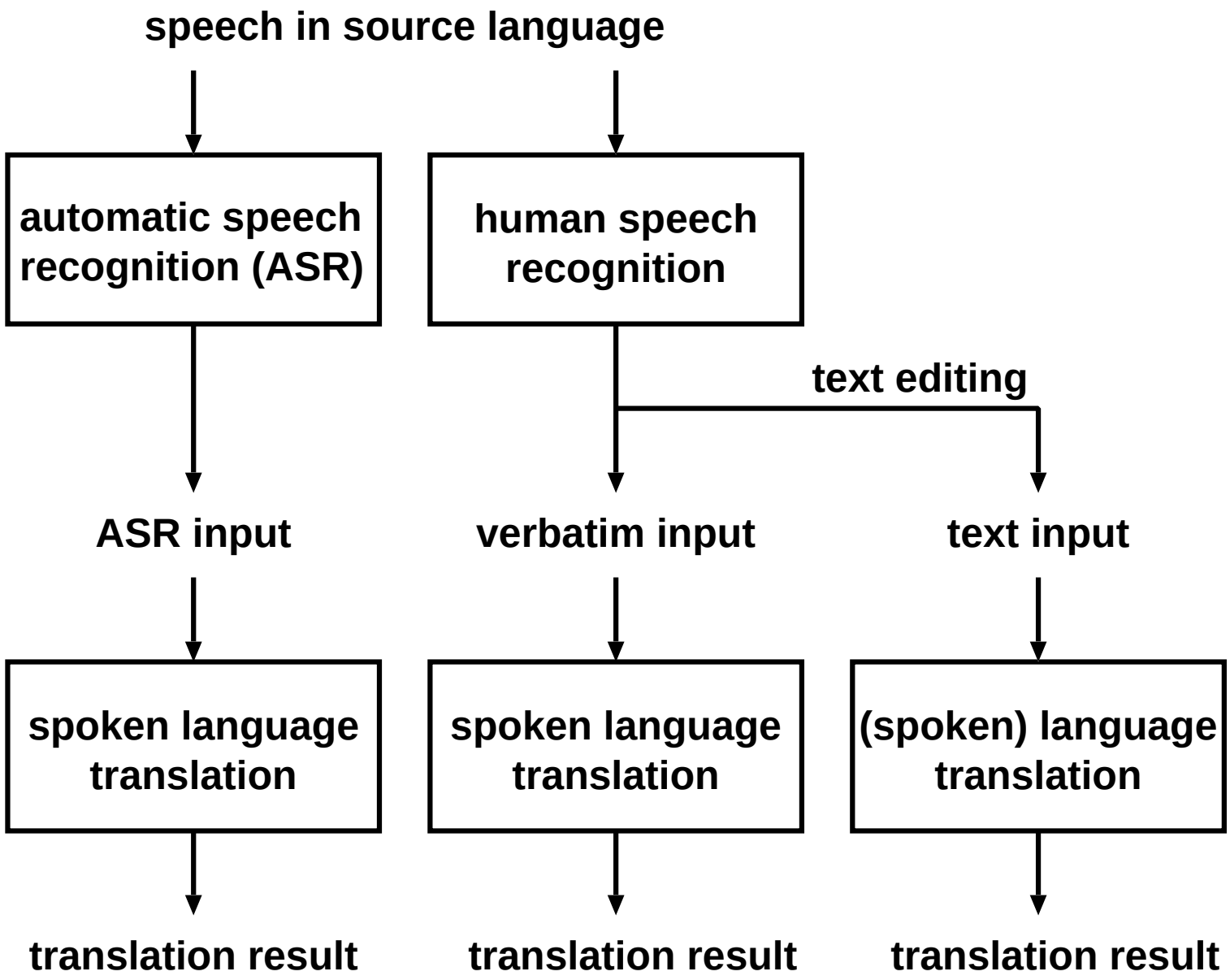
- **work on a real-life task:**
  - **– unlimited domain**
  - **– large vocabulary**
- **speech input:**
  - **– cope with disfluencies**
  - **– handle recognition errors**
- **sentence segmentation**
- **reasonable performance**

TC-Star: Approaches to MT

- 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 SLT**
  - **generate improved output from several MT engines**
  - **problem: word re-ordering**

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


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

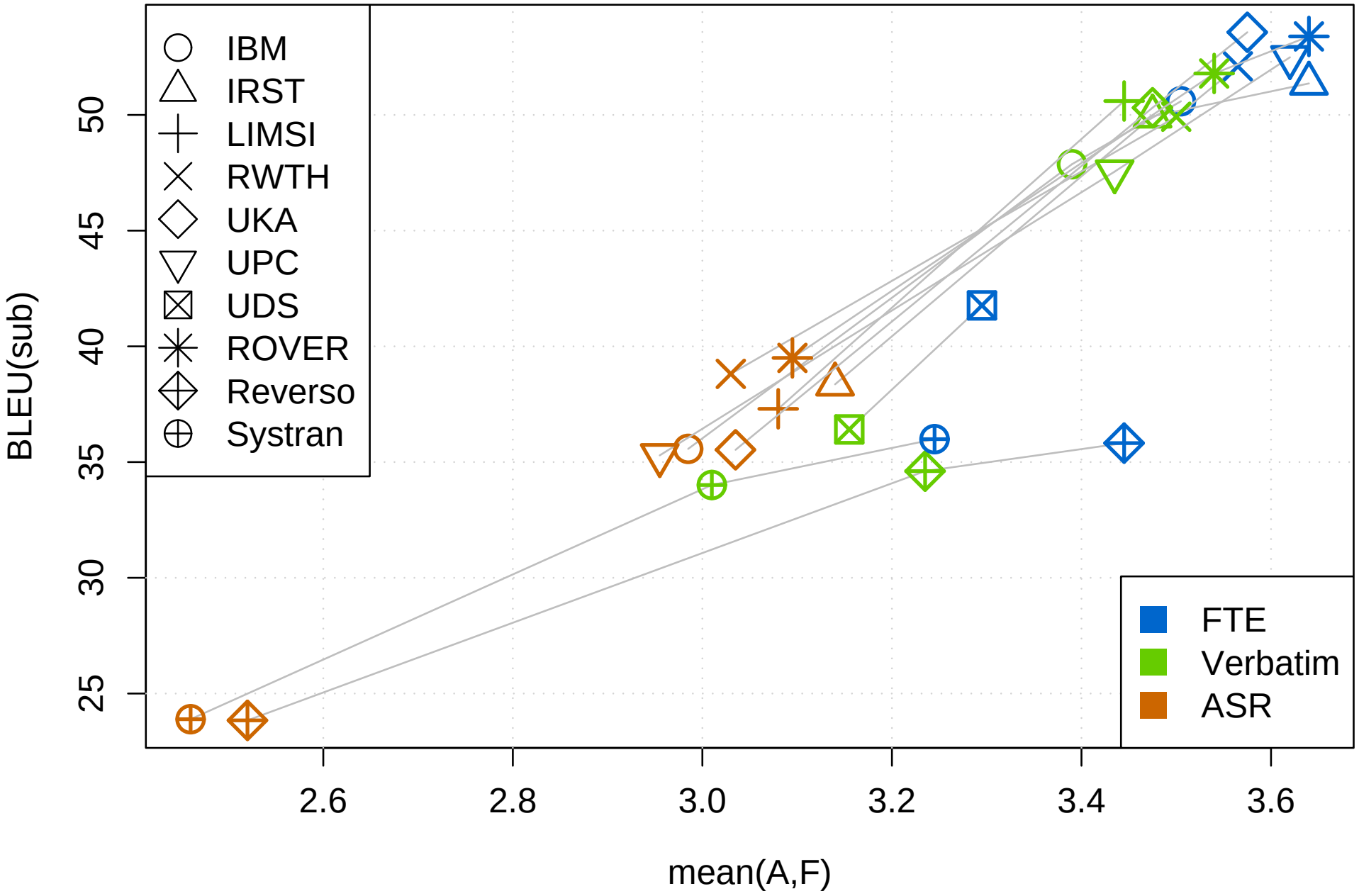
three types of input to translation:

- 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 (Text) 2007: Human vs. Automatic Evaluation



English → 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: no or small
  from verbatim to ASR: ΔPER corresponds to ASR errors

under TC-Star conditions:

critical questions and effects:

- **memory effect:**
  effect of phrase length ?
- **translation with no phrases**

Effect of Phrase Length

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

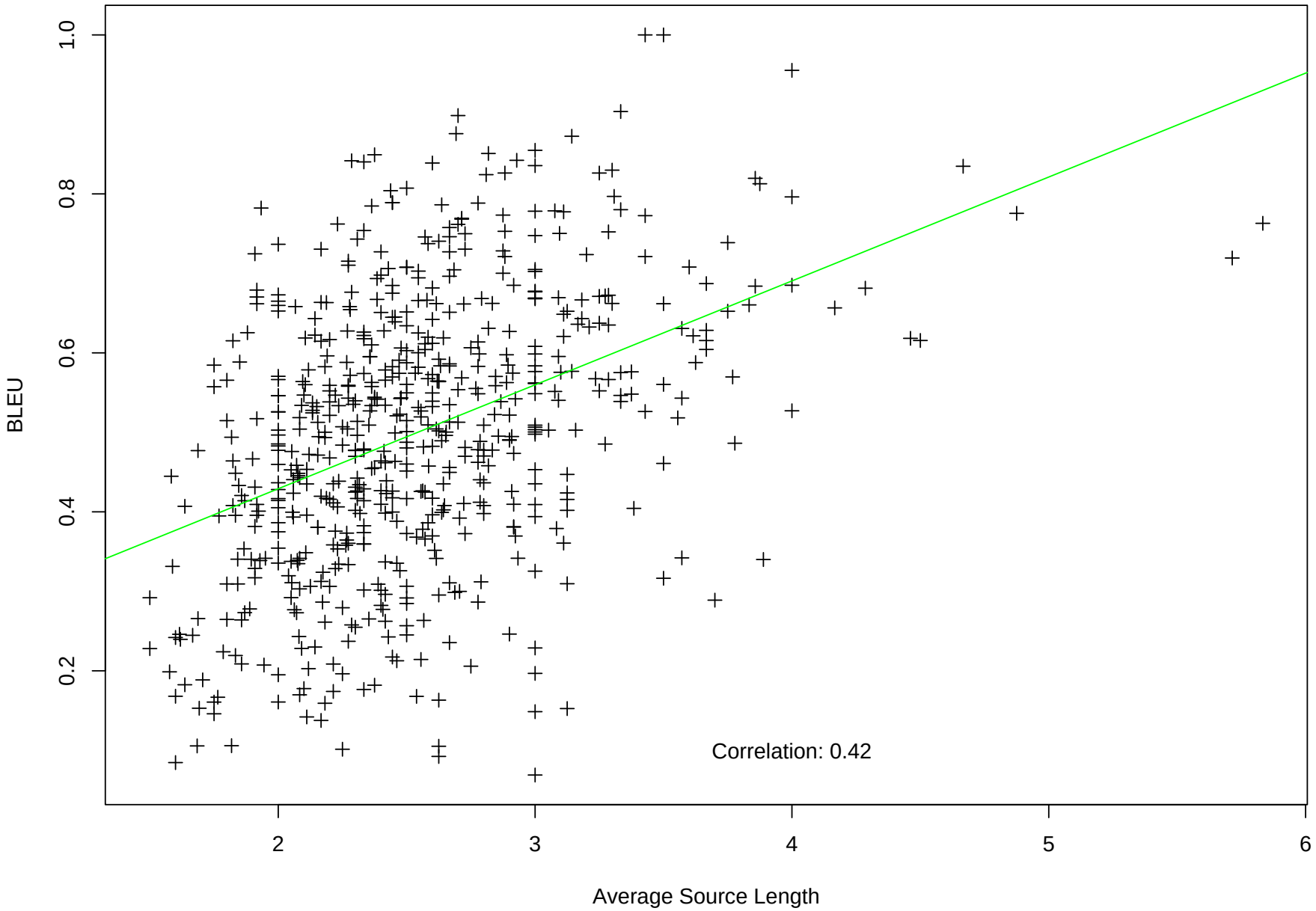
experimental analysis:

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

experimental results:

clear effect: 40% to 80%

Effect of Phrase Length (J>20)



Translation with No Phrases

How important are the phrases?

translation experiment without any phrases:

$$\hat{e}_1^J = \underset{e_1^J}{\operatorname{argmax}} \left\{ \log p(e_1^J) + \lambda \sum_j p(e_j|f_j) \right\}$$

using

- **single-word lexicon extracted from the final GIZA++ alignment**
- **word-by-word translation: fixed alignment!**

Spanish-English

Spanish→English	WER	PER	BLEU
full SMT	34.5	25.5	54.7
word lexicon	47.6	31.7	34.0
+ re-order adjective	43.4	31.2	40.8
English→Spanish	**WER**	**PER**	**BLEU**
full SMT	39.7	30.6	47.8
word lexicon	57.7	43.0	26.1
+ re-order adjective	55.5	41.6	29.1

results:

- **PER is increased: by 6% for S→E and 11 for E→S**
- **ΔBLEU $\cong$ 2 ΔPER**

German-English

German→English	WER	PER	BLEU
full SMT	66.9	45.9	24.4
word lexicon	71.1	49.0	14.5
+ re-order verbs	70.0	48.8	15.2
English→German	**WER**	**PER**	**BLEU**
full SMT	72.6	54.2	18.2
word lexicon	82.3	61.9	10.3
+ re-order verbs	81.4	61.8	10.5

results for G-E:

- **baseline method: worse performance than S-E**
- **PER is increased: by 3% for G→E and 8 for E→G**
- **ΔBLEU $\cong$ 2 ΔPER for G→E**

2 From Text to Speech: What is Different? (Matusov)

- **vocabulary normalization**
- **sentence segmentation**
- **handling of punctuation marks**
  (5 PMs: . ? , ; :)
- **handling of ASR errors/ambiguities**
  (N-best list, word lattice, confusion network)

Vocabulary Normalization

(trivial) requirement:
ASR and MT must use the same vocabulary and the same representation

- **normalization and adaptation steps:**
  - **remove hesitations, false starts, filled pauses, ...**
  - **normalize abbreviations (U. S. A. $\longrightarrow$ USA)**
  - **normalize the genetive ′s**
  - **normalize numbers (’nine’ vs. 9), times, dates, ...**
  - **normalize prefixes/suffixes for Arabic or word segmentation for Chinese**
  - **...**
- **additional steps in training:**
  add speech transcriptions and translations to the MT training data

Sentence Segmentation

(trivial) requirement for speech input:
sentence-like units are needed in MT!

- **automatic segmentation algorithm:**
  - **– adapt a word trigram language model**
  - **– acoustic features: pause duration, prosody, ...**
- **use length constraints (minimum/maximum) or length distribution (dynamic programming algorithm)**

Methods for PM Prediction

- method A: target sentence postprocessing (= MT without PM)
  - remove PMs from bilingual phrase pairs in training
  - perform MT without any PMs
  - insert PMs in target sentence by extended language model
- method B: source sentence processing (= MT with PMs)
  - preserve PMs in training of bilingual phrase pairs
  - insert PMs in the source sentence by language model and acoustic/prosodic features
- method C: MT itself generates PMs:
  - remove PMs from source side of bilingual phrase pairs in training
  - target sentence is generated with PMs (search!)

Evaluation Methodology

- evaluation and (most) automatic measures need sentence segmentation and alignment with reference translation
- potential problem:
  sentence segmentation errors will result in a mismatch
  between reference translations and test translations
- solution: synchronize/align reference and test sequences (maybe VERY long) using extended edit distance [tool by Matusov et al., 2005]

re-segmentation for evaluation

Experiments for SS and PM

TC-STAR English-to-Spanish Task (monotone translation; eval 2006)

transcription	segmentation	PM prediction	BLEU [%]	WER [%]	PER [%]
verbatim	manual	–	45.2	43.3	32.2
ASR	manual/aligned	B: source sent.	37.8	50.6	37.6
	automatic	B: source sent.	36.7	51.2	38.1
		C: by MT	36.1	51.5	38.6
		A: target sent.	36.3	51.3	38.4

results:

- **automatic segmentation (vs. manual):**
  only a small degradation in MT quality
- **PM prediction:**
  all three methods produce comparable results

Word lattice translation results (BTEC task)

Italian-to-English (ASR-WER: 21.0%, [Matusov et al., ICASSP 2006]):

System:	Input (test corpus):	WER	PER	BLEU
PBT	single best	32.4	27.2	55.4
	word lattice (30-40)	31.9	28.0	54.7
	+ add acoustic + LM scores	30.6	26.6	56.2
	+ re-opt all exponents	29.8	25.8	57.7

- monotone phrase-based translation:
  always best results on the BTEC Italian-English task
- improvements by using lattice: BLEU 55.4% $\rightarrow$ 57.7%

Experiments: Lattice and Reordering

Chinese-to-English BTEC task (ASR WER: 42.0%, [Matusov et al., ICASSP 2006]):

Reordering in MT	Input	BLEU [%]	WER [%]	PER [%]
monotone	1-Best	31.1	62.1	52.7
	Lattice (30-40)	34.1	58.3	48.1
phrase skip 1	1-Best	33.1	61.3	51.7
	Lattice (30-40)	35.1	57.7	47.2

- **'phrase skip 1': limited reordering: significantly improves translation quality for language pair with different word order (C-E)**

Translation results on the Spanish-to-English TC-STAR task (eval 2006)

Word lattices: ASR+MT system (ASR WER: 9.0%):

Input	WER	PER	BLEU
ROVER-ASR + single best (POS-based reordering)	51.3	36.7	37.5
RWTH-ASR single best (monotone)	53.0	36.7	36.1
RWTH-ASR word lattice (monotone)	52.6	36.5	36.4

Confusion networks: IRST ASR+MT system (ASR WER: 22.4%):

Input	WER	PER	BLEU
single best	50.0	39.2	37.6
confusion network (5-10)	49.5	38.6	39.2

result:

improvements in MT quality seem to be higher for high ASR-WER

Bayes Rule for Speech Translation

three levels in spoken language translation:

$$\begin{array}{ccccc} \text{speech} & \rightarrow & \text{source text} & \rightarrow & \text{target text} \\ X & \rightarrow & F & \rightarrow & E \end{array}$$

question: how to handle recognition errors? what is the right interface?

re-consider Bayes decision rule:

$$\begin{aligned} X \rightarrow \hat{E} &= \arg\max_E p(E|X) = \arg\max_E \{p(E) \cdot p(X|E)\} \\ &= \arg\max_E \left\{p(E) \cdot \sum_F p(F|E) \cdot p(X|F,E)\right\} \\ &\cong \arg\max_E \left\{p(E) \cdot \max_F \{p(F|E) \cdot p(X|F)\}\right\} \end{aligned}$$

mild assumptions:
– knowing E in addition to F does not help: $p(X|F) = p(X|F,E)$
– maximum approximation

Challenges

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

remarks:
– source language model $p(F)$ does not show up
– translation model $p(F|E)$: can we use the same models as for text?
– coupling of recognition and translation: efficient search
– improvement over serial coupling: how much?

so far: output of recognizer beyond single best:
– N-best list
– word graph or lattice
– confusion network

3 Translation With Scarce Resources (Popovic)

two aspects of statistical MT:

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

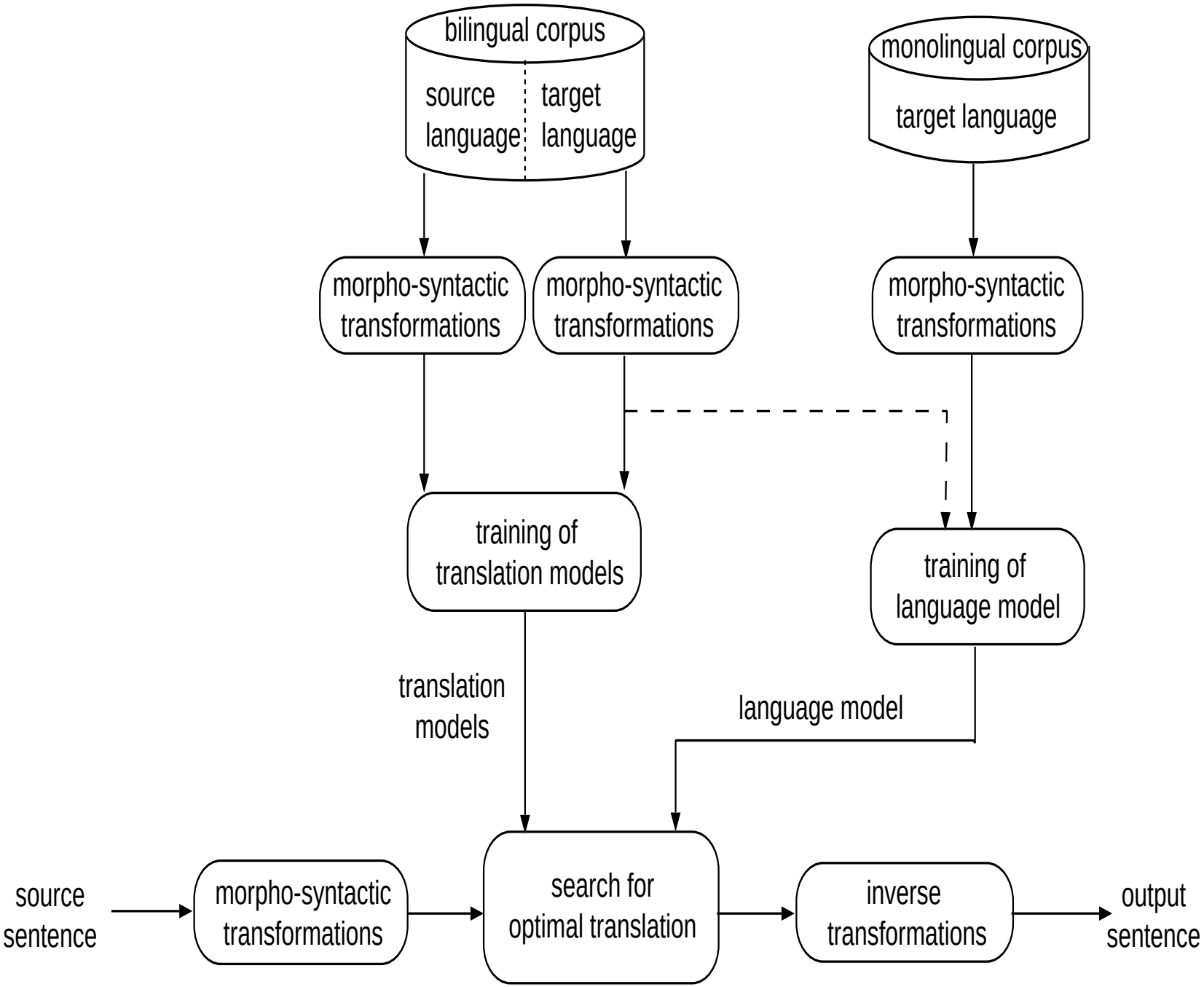
$$\hat{E} = \arg\max_{E}\{p(E) \cdot p(F|E)\}$$

- **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!)


bilingual corpus
source language
target language
monolingual corpus
target language
morpho-syntactic transformations
morpho-syntactic transformations
morpho-syntactic transformations
training of translation models
training of language model
translation models
language model
source sentence
morpho-syntactic transformations
search for optimal translation
inverse transformations
output sentence

Spanish↔English Corpora

- EPPS training corpus:
  subsets: 1.3M (full), 13k, 1k
- general purpose dictionary downloaded from the web (full forms):

	Spanish	English
entries	52637	
distinct phrases	31126	30761
distinct words	33762	35715
one-word entries	47980	45236

- test corpus: TC-star evaluation 2005

Spanish↔English
Morpho-syntactic transformations

Adjective treatment:

- **POS-based local reorderings of nouns and adjectives**
- **replacing Spanish adjectives with their base forms**

Spanish	**original:**	**motivos económicos y políticos**
	reordered:	**económicos y políticos motivos**
	+ adjective base form:	**económico_ y político_ motivos**
English	**original:**	**economic and political reasons**
	reordered:	**reasons economic and political**

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

English→Spanish	WER	PER	BLEU	OOVs
dictionary	67.6	55.9	14.1	16.2
+adjective treatment	65.7	54.5	16.5	16.2
1k	60.1	47.4	23.9	9.4
+dictionary	56.0	43.2	28.3	4.8
+adjective treatment	53.9	42.0	30.6	4.8
13k	49.6	37.4	36.2	2.6
+dictionary	48.6	36.3	37.2	1.8
+adjective treatment	47.3	35.7	39.1	1.8
1.3M	39.7	30.6	47.8	0.25
+adjective treatment	39.6	30.5	48.3	0.25

observations for S→E:

- **similar effects as for direction SE**
- **improvements through dictionary and morpho-syntactic information are slightly smaller**
  - **translation into a 'highly' inflected language is more difficult**
  - **Spanish has a rather free word order**

German↔English Corpora

- Europarl corpus of ACL-SMT workshop 2005: 700k and 1k sentence pairs
- general purpose dictionary (Chemnitz U + Verbmobil; full forms)

	German	English
entries	292497	
distinct phrases	165775	161596
distinct words	138253	82457
one-word entries	247103	177446

- test corpus: ACL-SMT workshop 2005

German→English
Morpho-syntactic transformations

POS-based long range reorderings of verbs

	German	English
original sentence	ich habe diese Frage bereits beantwortet.	I have already answered this question.
reordered sentence	ich habe beantwortet diese Frage bereits.	I have already this question answered.

- source language German:
  rules for moving infinitives, past participles, finite verbs, negation particles
- source language English:
  rules for moving infinitives and past participles

German→English
Morpho-syntactic transformations

corpus-based splitting German compound words:
frequency-based method as described in [Koehn & Knight 03]

- **consider decomposition of potential compound words into single words**
- **decision based on frequencies of compound word and single words**

compound word	split words
Treibhauseffekt	Treibhaus_ _effekt
Treibhauseffektgase	Treibhauseffekt_ _gase

German→English	WER	PER	BLEU	OOVs
dictionary	78.5	59.2	11.7	10.2
+reorder verbs	77.1	58.6	12.5	10.2
+split comp	76.8	57.9	12.8	9.6
1k	78.5	60.3	11.6	16.4
+dictionary	75.8	55.9	14.6	6.6
+reorder verbs	74.9	55.4	15.0	6.6
+split comp	74.2	54.5	15.7	5.6
700k	66.9	45.9	24.4	0.8
+reorder verbs	65.4	45.7	25.5	0.8
+split comp	65.1	45.2	25.5	0.7

- for comparison: 24.8% BLEU
  for best system of ACL-SMT workshop 2005 (U of Washington)
- similars effect as for the language pair S-E,
  in particular performance gap caused by OOV effect!
- improvements by
  dictionary and morpho-syntactic transformations

English→German	WER	PER	BLEU	OOVs
dictionary	83.1	65.7	8.9	4.2
+reorder verbs	82.8	65.6	9.3	4.2
1k	85.5	68.1	8.4	10.2
+dictionary	81.9	64.0	10.8	2.2
+reorder verbs	81.8	64.0	11.0	2.2
700k	72.6	54.2	18.2	0.2
+reorder verbs	71.6	53.8	18.4	0.2

- comparison: no other system available
- translation into German is more difficult:
  - Germain is a 'highly' inflected language
  - word compounds
  - word order (especially for verbs)

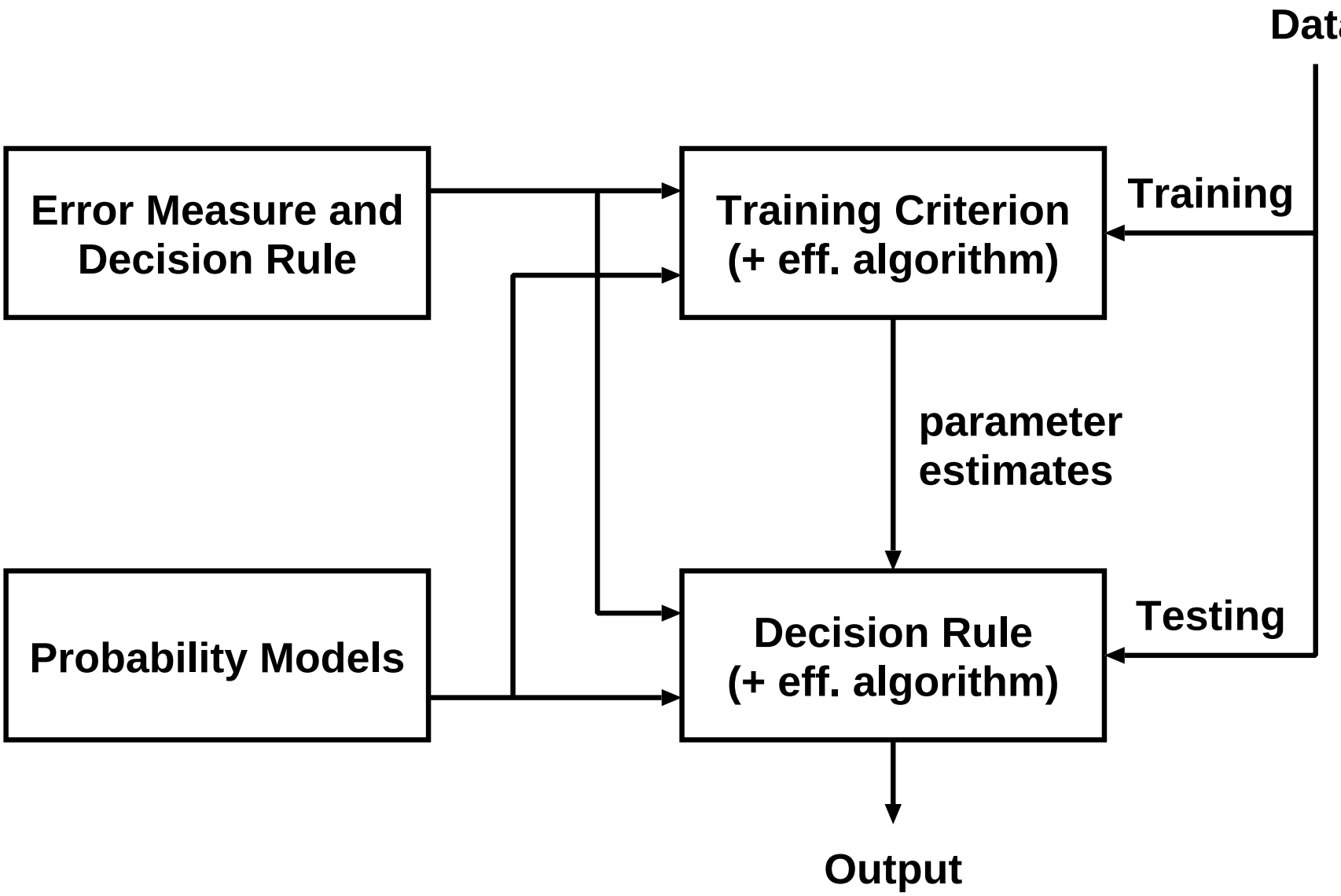
4 Statistical MT: Limitations and Future Directions

Advantages of statistical NLP: two attractive properties:

- **holistic decision criterion:**
  - **exploits all (available) dependencies (=knowledge sources)**
  - **is able to combine thousands/millions of weak dependencies**
  - **handles interdependences, ambiguities and conflicts**
- **powerful training methods:**
  - **training criterion is linked to PERFORMANCE**
  - **fully AUTOMATIC procedures (no human involved !)**
  - **HUGE amounts of data can be exploited**

note:
these statements do not apply to rule-based systems!

Statistical NLP: Four Key Ingredients



Limitations of Today's Statistical MT

- poor probab. models:
  model = dependency = condit. probab. distribution
  = structure + free parameters
  – translation model:
  alignment models: no morphology, no syntax
  lexicon model: no morphology, no context dependency
  – language model: no grammar
- poor generation method:
  – suboptimal decision rule:
  * maximum approximation
  * loss function different from BLEU/NIST or WER/PER
  – suboptimal search/decoding: GLOBAL maximum is not found
- poor training method:
  – suboptimal criterion (e.g. max.lik. rather than BLEU/NIST or WER/PER)
  – suboptimal algorithm (iterative procedure!)
  – limited or even insufficient amount of training data
  (statistical estimation problem)

Statistical MT: Room for Improvements

- **alignment and lexicon models (in training):**
  challenges:
  - **introduction of context dependency:**
    intra- and inter-sentence level
  - **integration of morphology and -syntax**
  - **reordering based on syntactic structure**
- **phrases (alignment templates):**
  good for seen test data $\Rightarrow$ memory-based translation

  - **challenge: design models with good generalization capabilities,**
    i.e. which work well on UNSEEN test data
  - **challenge: consistent framework for implicit segmentation,**
    words-phrases balance, ...

Statistical MT: Room for Improvements

- language model:
  - monolingual grammar to improve the syntactic structure
  - explicit link with word alignment and reordering
  - bilingual grammar
- generation (or search):
  not a problem for present models,
  but what about more complex models in the future ?

THE END