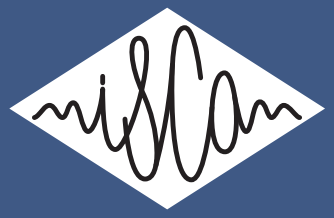
MAVIR Workshop, Madrid, November 18/19, 2013, UNED, National Distance Education University, Madrid

The Statistical Approach to Speech Recognition and Natural Language Processing: Achievements and Open Problems

Hermann Ney
Human Language Technology and Pattern Recognition

RWTH Aachen University, Aachen
DIGITEO Chair, LIMSI-CNRS, Paris

Outline

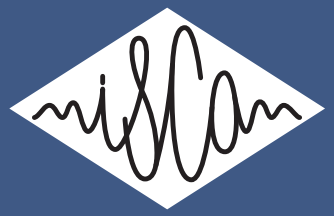


international speech communication association

promoting international speech communication, science and technology

ISCA: International Speech Communication Association

- **ISCA started as ESCA (European Speech Communication Association): March 27, 1988 by Rene Carree.**
- **purpose:**
  to promote Speech Communication Science and Technology,
  both in the industrial and academic areas,
  covering all the aspects of Speech Communication
  (acoustics, phonetics, phonology, linguistics, natural language processing, artificial intelligence, cognitive science, signal processing, pattern recognition, etc.
- **ISCA offers a wide range of services;**
  in particular Interspeech, ISCA workshops, SIGs (special interest groups)



international speech communication association

promoting international speech communication, science and technology

ISCA Objectives:

- **to stimulate scientific research and education,**
- **to organize conferences, courses and workshops,**
- **to publish, and to promote publication of scientific works,**
- **to promote the exchange of scientific views in the field of speech communication,**
- **to encourage the study of different languages,**
- **to collaborate with all related associations,**
- **to investigate industrial applications of research results,**
- **and, more generally, to promote relations between public and private, and between science and technology.**

1 History and Projects

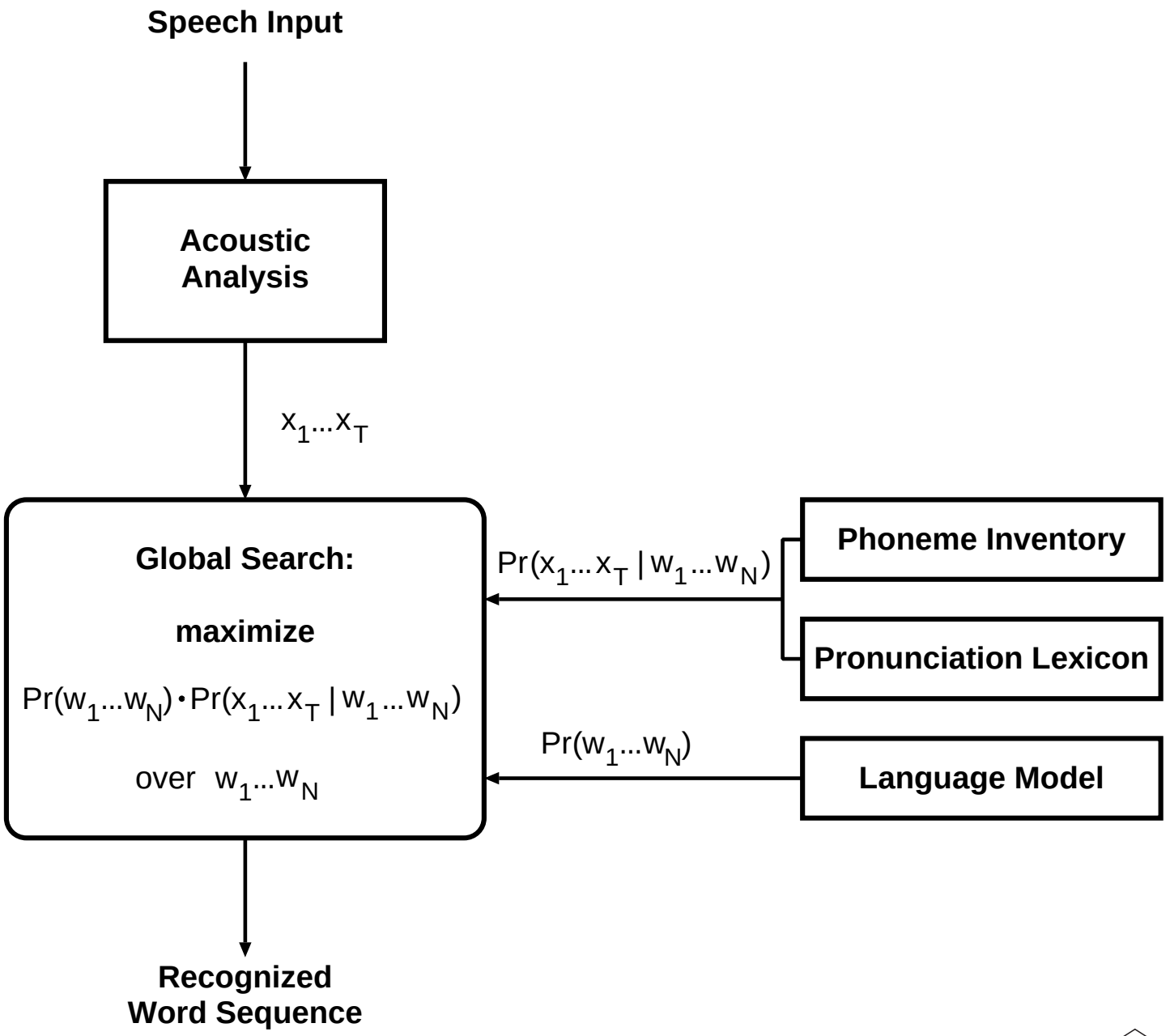
terminology: tasks in speech and natural language processing (NLP)

- **automatic speech recognition (ASR)**
- **optical character recognition (OCR: printed and handwritten text)**
- **machine translation (MT)**
- **document classification**
- **understanding of speech or language**

characteristic properties of these tasks (ASR, OCR, MT):

- **well-defined 'classification' tasks:**
  - **due to 5000-year history of (written!) language**
  - **well-defined classes: letters or words of the language**
- **easy task for humans**
  (ASR, OCR: at least in their native language!)
- **hard task for computers**
  (as the last 40 years have shown!)

Statistical Approach to Automatic Speech Recognition (ASR)



four ingredients of the statistical approach to ASR:

- decision procedure (Bayes decision rule):
  - minimizes the decision errors
  - consistent and holistic criterion
  - no explicit segmentation
- models of probabilistic dependencies:
  - problem-specific (in lieu of 'big tables')
  - textbook statistics and much beyond ...
- model parameters are learned from examples:
  - statistical estimation and (any type of) learning
  - suitable training criteria
- search or decoding:
  find the most 'plausible' hypothesis

statistical approach to ASR:

$$\text{ASR} = \text{Modelling} + \text{Statistics} + \text{Efficient Algorithms}$$

25 years ago (1987):

- **SPICOS dialogue system (Siemens/Philips/IPO research):**
  1k-word vocabulary, continuous speech, network grammar,
  speaker dependent
- **IBM research prototype:**
  5k-word vocabulary, isolated (!!) words, trigram language model,
  speaker dependent

today (2012):

- **systems for broadcast news, lectures, conversations, ...:**
  - **100k-word vocabulary**
  - **natural speech**
  - **speaker independent, multi-speaker**
  - **many languages (E, F, S, G, ...; CH, AR, ...)**

Short History of ASR

- start of statistical approach around 1972 at IBM research
- steady improvement of statistical methods over 40 years
- controversial issues: about usefulness of
  – 'existing' theories/models from phonetics and linguistics
  – rule-based approaches from classical artificial intelligence

40 years of progress by improving the statistical methods (along with training criteria):

- Hidden Markov models (HMM) along with EM algorithm
- smoothing/regularization
- CART and phonetic decision trees
- discriminative training:
  MMI, Poveys's MPE, MCE, ...
- adaptation (unsupervised and supervision light training)
- neural networks and log-linear modelling
- machine learning?

From Speech to Language

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**

IBM's Jelinek did not care about Chomsky's ban:

- **1988: IBM starts building a statistical system for MT**
  (in opposition to linguistics and artificial intelligence)
- **task: Canadian Hansards: English/French parliamentary debates (text!)**
- **1994 DARPA evaluation:**
  – comparable to 'conventional' approaches (Systran)
  – results only for French $\rightarrow$ English
- **team went off to Renaissance Technologies (Hedge Fund)**

After IBM: 1992 – 2000

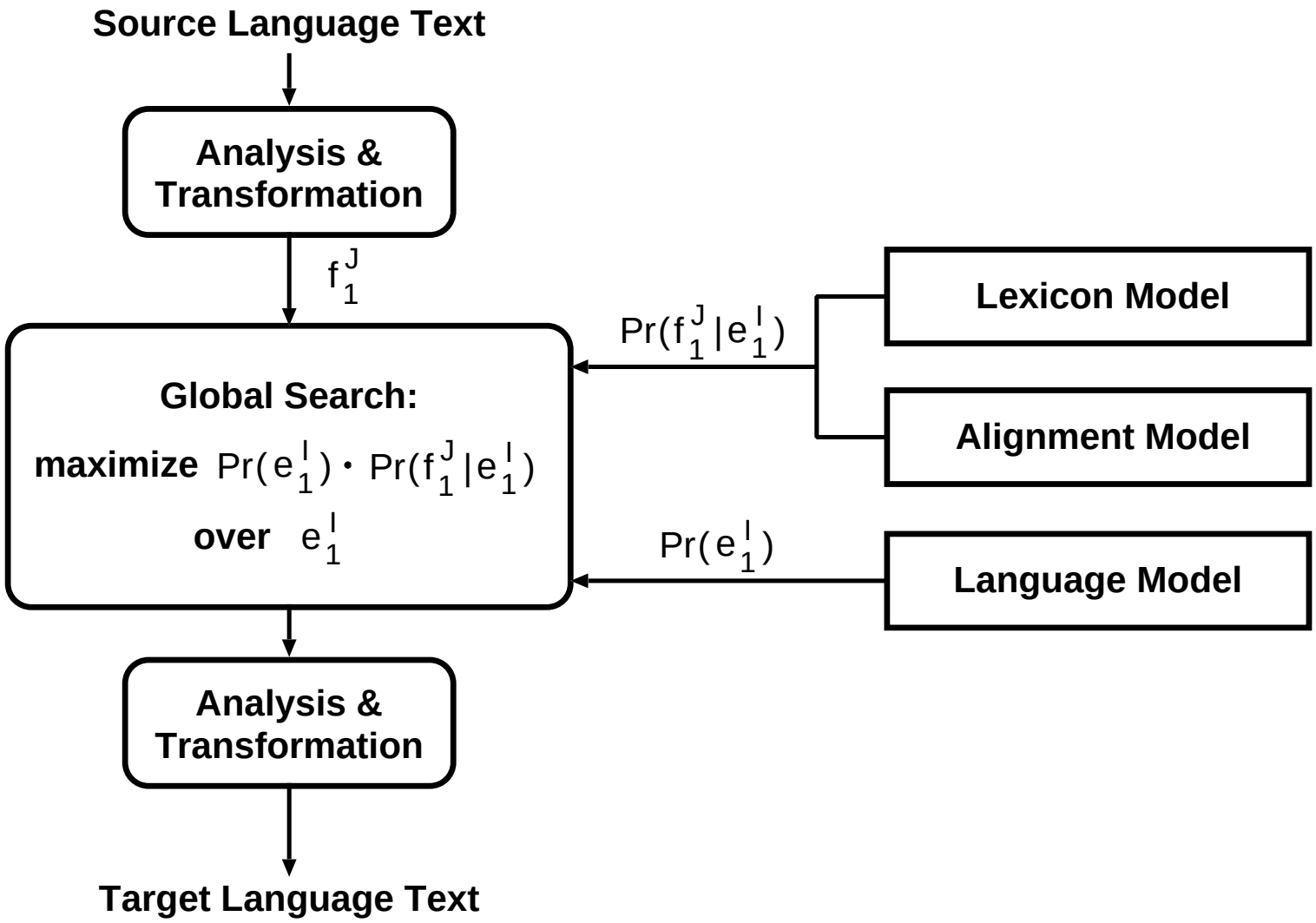
translation of SPEECH (vs. text):

- **specific task:**
  - **speech translation: ambitious task in general**
  - **therefore limited domains (data collected in lab): travelling, appointment scheduling,...**
- **justification for statistical approach: to cope with non-perfect input (as opposed to text input)**
- **projects:**
  - **CSTAR consortium**
  - **Verbmobil (German)**
  - **EU projects: Eutrans, PF-Star, LC-Star, ...**

side result:

statistical approach looks promising for text, too!

Architecture of Statistical MT System (similar to speech recognition)



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?
- overall result: statistical MT highly competitive

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

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

EU Project TC-Star (2004-2007)

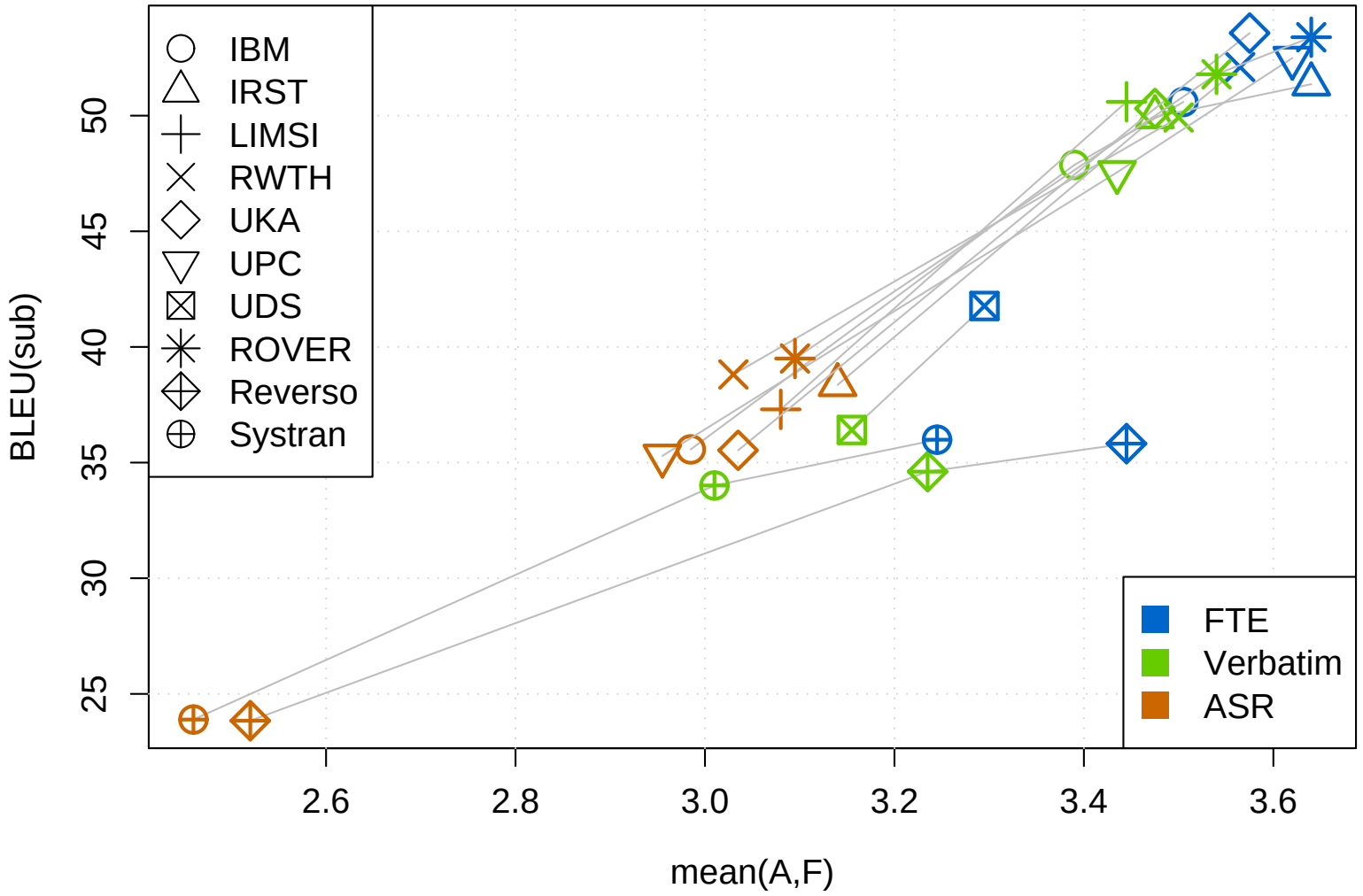
domain: speeches given in the European Parliament

- **work on a real-life task:**
  - **'unlimited' domain**
  - **large vocabulary**
- **speech input:**
  - **cope with non-grammatical input and disfluencies**
  - **handle recognition errors**
  - **sentence segmentation**
- **FIRST research prototype ever on speech translation for unlimited domain and real-life data**

experimental results:

good performance

E → S 2007: Human vs. Automatic Evaluation



More MT Projects 2001 – 2012

'unlimited' domain (real-life data) with associated evaluations:

- **TIDES 2001-04 funded by DARPA: written text (newswire): Arabic/Chinese to English**
- **GALE 2005-2011 (and BOLT 2012-2017) funded by DARPA (funding: 40 Mio US$ per year):**
  - **text and speech**
  - **Arabic/Chinese to English**
  - **ASR, MT and information extraction ('question answering')**
  - **performance measure: HTER (= human translation error rate)**
- **QUAERO 2008-2013 funded by OSEO France:**
  - **text and speech (news, lectures, discussions, ...)**
  - **more colloquial text and speech**
- **more EU projects and on text (after GOOGLE Translate!):**
  - **EuroMatrix and -Plus (...-2012): text MT for all 22 official EU languages**
  - **EU-Bridge (2012-2015): speech and language**
  - **...**

IWSLT 2011

- IWSLT: Int. Workshop on Spoken Language Translation
- TED lectures: from English to French
- automatic performance measures:
  - TER: error rate: the lower, the better.
  - BLEU: accuracy measure: the higher, the better.

System	Results 2011	
	BLEU [%]	TER [%]
Karlsruhe IT	37.6	41.7
LIMSI Paris	36.5	43.7
RWTH Aachen	36.1	43.7
MIT Cambridge	35.3	44.0
FBK Trento	34.9	44.7
U Grenoble	34.6	44.1
DFKI Saarbrücken	34.4	45.7

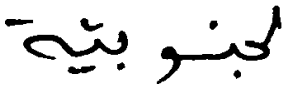
WMT 2012

- WMT: ACL Workshop on Machine Translation
- text input: German to English
- domain: news
- QUAERO systems: marked by *

System	Results 2012	
	BLEU [%]	TER [%]
* QUAERO SysCom	24.4	65.4
* Karlsruhe IT	23.4	66.3
* RWTH Aachen	23.3	65.9
U Edinburgh	22.9	67.0
* LIMSI Paris	22.8	67.7
Qatar CRI	22.6	66.8
DFKI Saarbrücken	20.7	70.5
JHU Baltimore	19.7	69.4
U Prague	20.0	71.3
U Toronto	14.0	76.1

Automatic Recognition: From Speech to Characters

image text recognition:
– define vertical slots over horizontal axis
– result: image signal = (quasi) one-dim. structure like speech signal

Language	Database	Example
French	RIMES	
Arabic	IfN/ENIT	
English	IAM	

2 Inside Statistical MT

from subsymbolic to symbolic processing:

- **so far: recognition of signals: speech and image**
- **consider the problem of translation:**
  - **– convert the text from a source language to a target language**
  - **– problem of symbolic processing**

machine translation: why a statistical approach?
answer: we need decisions along various dimensions:

- **select the right target word**
- **select the position for the target word**
- **make sure the resulting target sentence is well formed**

interaction: Bayes decision rule handles the interdependencies of decisions

conclusion: MT (like other NLP tasks) amounts to making decisions

scientific framework for making good decisions:
probability theory, statistical classification, statistical learning

Statistical MT

key ideas of statistical approach:

- **MT (like most NLP tasks) is a complex task, for which perfect solutions are difficult (compare: all models in physics are approximations!)**
- **consequence: use imperfect and vague knowledge and try to minimize the number of decision errors**
- **statistical decision theory and Bayes decision rule using prob. dependencies between source sentence $F = f_1^J = f_1 ... f_j ... f_J$ and target sentence $E = e_1^I = e_1 ... e_i ... e_I$:**

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

- **resulting concept:**

MT = (Linguistic?) Modelling + Statistics + Efficient Algorithms

Statistical MT: Methodology

Bayes decision rule:

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

important aspects in the decision rule:

- **two INDEPENDENT prob. distributions (or stat. knowledge sources):**
  $p(F|E)$**: translation model:**
  link to source sentence ('adequacy')
  $p(E)$**: language model:**
  well-formedness of target sentences ('fluency')
  i.e. its syntactic–semantic structure

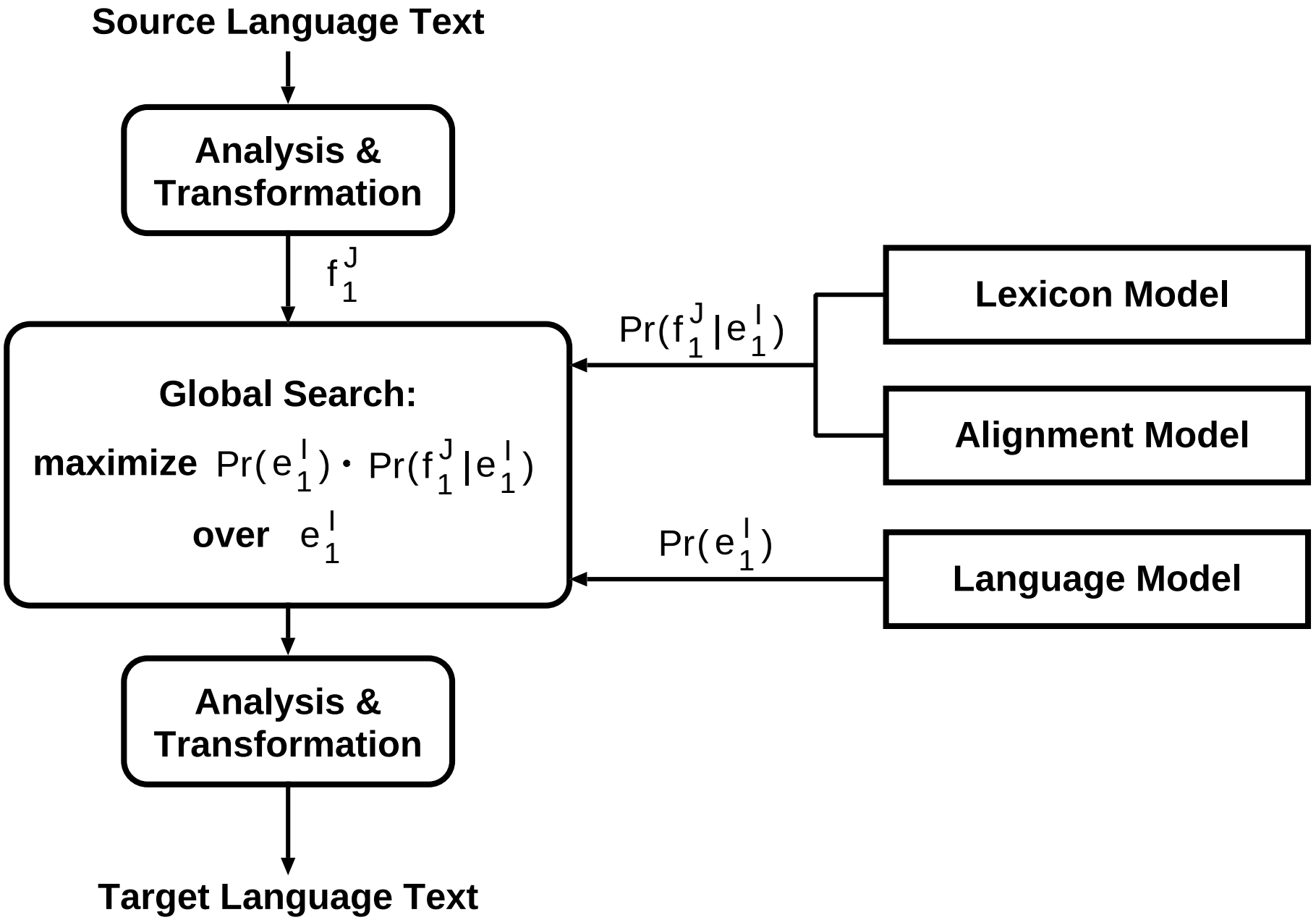
  Why this decomposition?
  each of these can be modelled separately

- **generation: = search = maximization over** E
  generate target sentence with the largest posterior probability

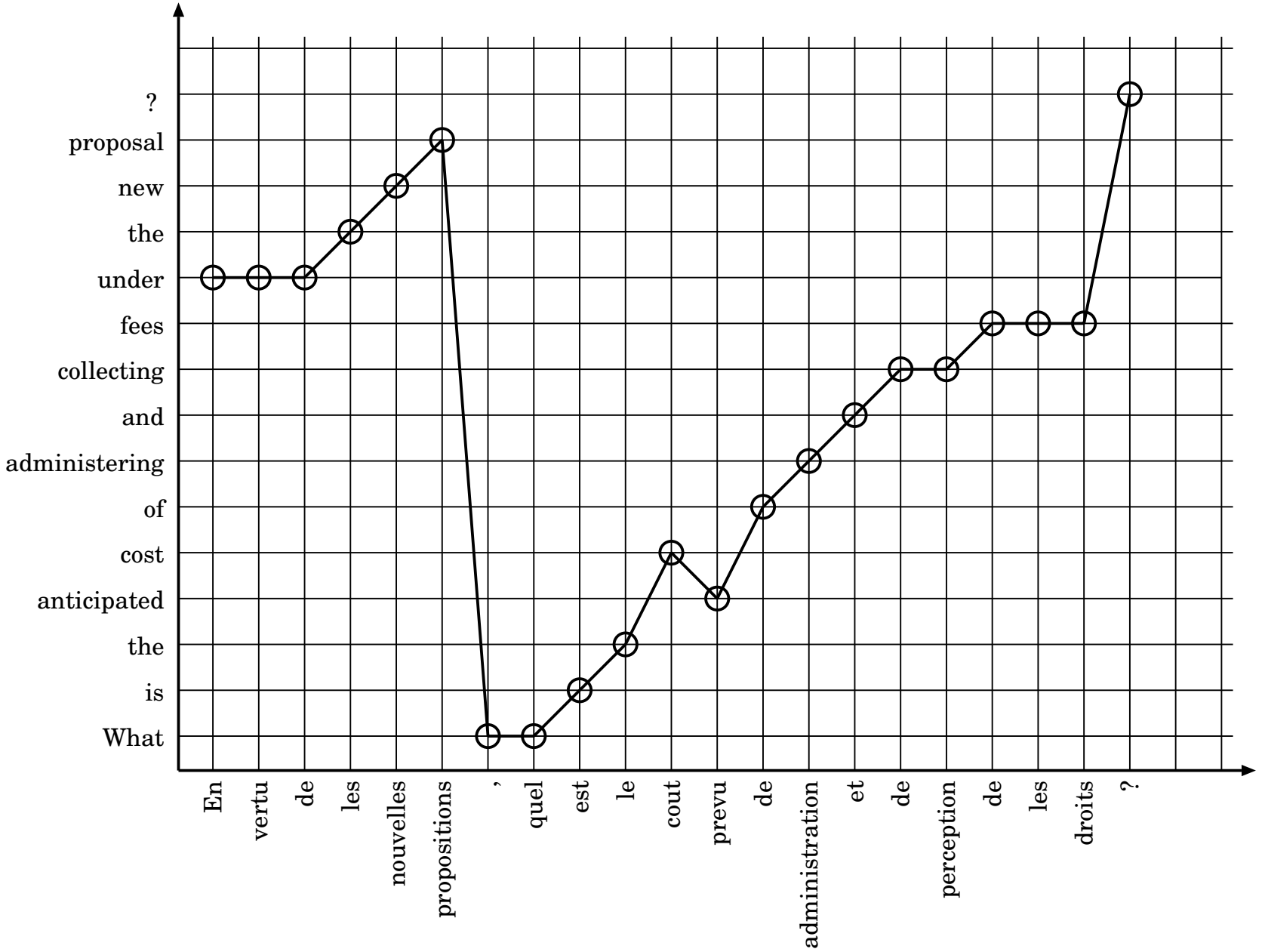
Statistical MT: Methodology

- 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)

$$\begin{aligned} p(F|E) &= \sum_A p(F, A|E) \\ &= \sum_A p(A|E) \cdot p(F|E, A) \end{aligned}$$


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

Example of Alignment (Canadian Hansards)



HMM: Recognition vs. Translation

speech recognition	text translation
$Pr(x_1^T \mid T, w) =$ $\sum_{s_1^T} \prod_t [p(s_t \mid s_{t-1}, S_w, w)\ p(x_t \mid s_t, w)]$	$Pr(f_1^J \mid J, e_1^I) =$ $\sum_{a_1^J} \prod_j [p(a_j \mid a_{j-1}, I)\ p(f_j \mid e_{a_j})]$
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: monotone	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: $j \rightarrow i = a_j$ sometimes: montone
transition prob. $p(s_t \mid s_{t-1}, S_w, w)$ emission prob. $p(x_t \mid s_t, w)$	alignment prob. $p(a_j \mid a_{j-1}, I)$ lexicon prob. $p(f_j \mid e_{a_j})$

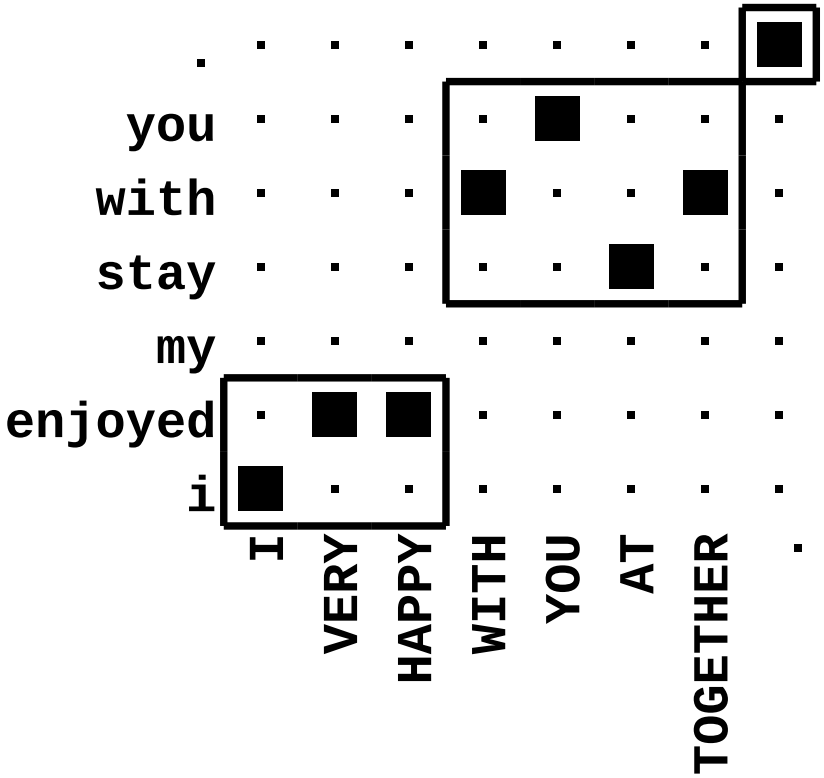
From Words to Phrases

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 $F \rightarrow E$:

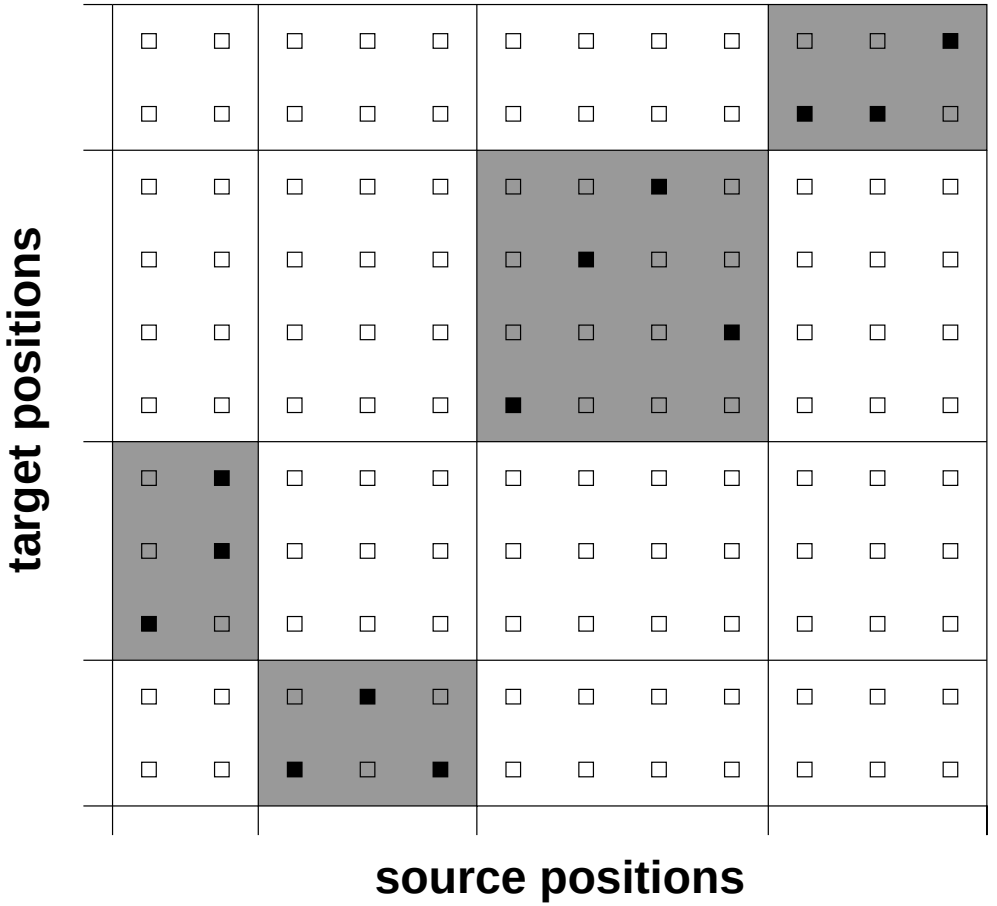


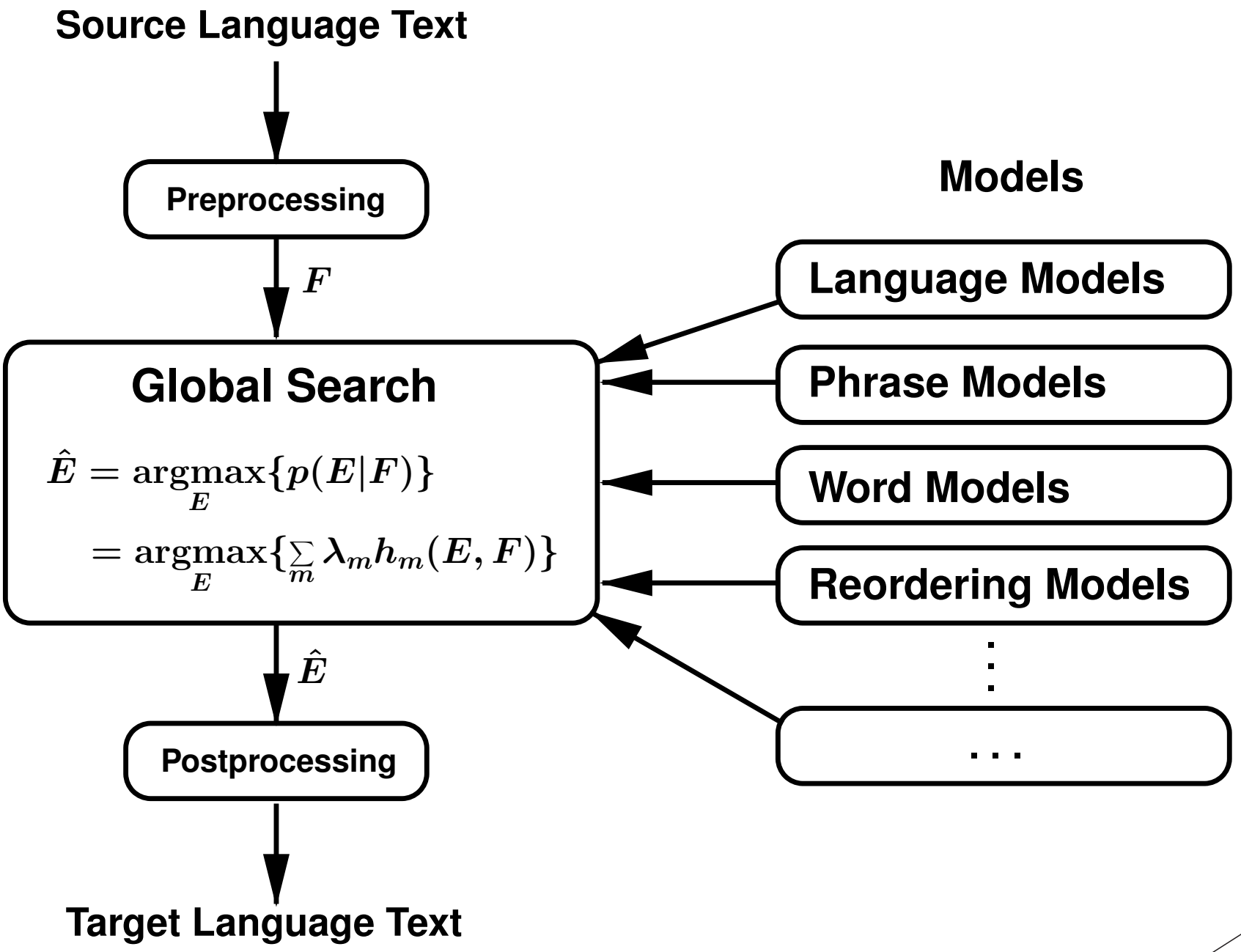
From Words to Phrases (Segments)

use of into two-dim. 'blocks':
beyond original IBM approach

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**




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
. . .

Log-Linear Combination of Models

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:

$$F \rightarrow \hat{E}(F) = \operatorname*{argmax}_E \left\{p(E|F)\right\} = \operatorname*{argmax}_E \left\{\sum_m \lambda_m h_m(E,F)\right\}$$

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

System Combination

concept for combining translations from several MT engines:

- **align the system outputs:**
  non-monotonic 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

RWTH Research Topics

- training:
  - replace IBM-2 by homogeneous HMM
  - swap source and target languages
- phrase training by
  forced alignments
- search:
  – reordering constraints
  – word graph and N-best list
  – DP beam search vs. (pure) A^*
- log-linear combination of models
- various types of phrase models:
  extraction, with/without alignment, ...
- morpho-syntactic analysis:
  German language
- system combination

Details:

- authors:
  Och, Niessen, Tillmann, Vogel,
  Ueffing, Zens, Leusch,
  Stein, Vilar, Matusov,
  Mauser, ...
- Comp. Ling. conferences:
  COLING; ACL, EACL,
  EMNLP, AMTA, EAMT, ...
- journals:
  IEEE Trans. on SAP;
  Machine Translation;
  Comp. Linguistics

Ongoing Work

approaches (RWTH and other teams):

- **improved phrase training by forced alignments**
- **consistent modeling of lexical dependencies (to replace phrase table)**
- **neural network: for language model and for translation model**

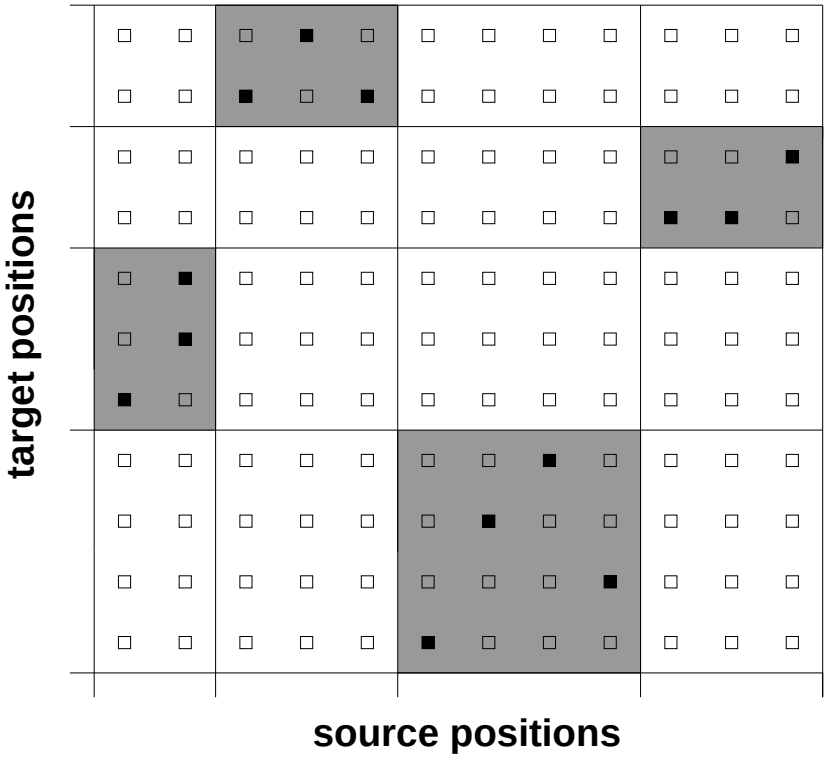
- **hierarchical phrases (=phrases with 'gaps'):**
  – long-distance dependencies
  – syntactic dependencies
- **integration of morphosyntax**
- **learning from mono-lingual data ('deciphering' approach)**
- **...**

open question: amount/type of 'classical' linguistic models ?

Phrase Training

today's SMT systems:

- word alignments as obtained by GIZA++
- heuristic methods for phrase extraction for both 'conventional' and 'gappy' phrase-based systems
- main effect: sophisticated 'translation memory' by automatic extraction
- unsatisfactory: inconsistency between training and translation



more details: Wuebker et al. ACL 2010

3 Conclusions

What have we learned?

- **steady improvements of models and methods (ASR: 40 years)**
- **lion's share of the improvements:**
  - **– better understanding of the modelling and the learning problems**
  - **– more efficient algorithms for learning and search ('generation')**
- **room for future improvements:**
  - **– better understanding of interaction of levels: frames, phones, words**
  - **– from log-linear models to neural networks**
  - **– better training criteria, linked to performance**

Methodology has been sucessfully applied to a large variety of tasks:

- **speech recognition**
- **character recognition**
- **machine translation**
- **gesture recognition (sign language)**
- **...**

Example: Can we do MT without statistics?

- consider any MT system, e.g. rule-based or statistical:
  we want to change/add a component in the system
- question: how to optimize the interaction of this component
  with the whole system?
- answer: by EXPERIMENTALLY tuning the WHOLE system
  (no matter: rule-based or statiscal) for optimum performance
  $\rightarrow$ statistics

- advantage of statistical approach:
  – tuning can be done automatically!
  – WHOLE set of parameters/components
- in practice: huge mathematical problem
  due to interaction of various components and performance criterion

Statistical Approach Revisited

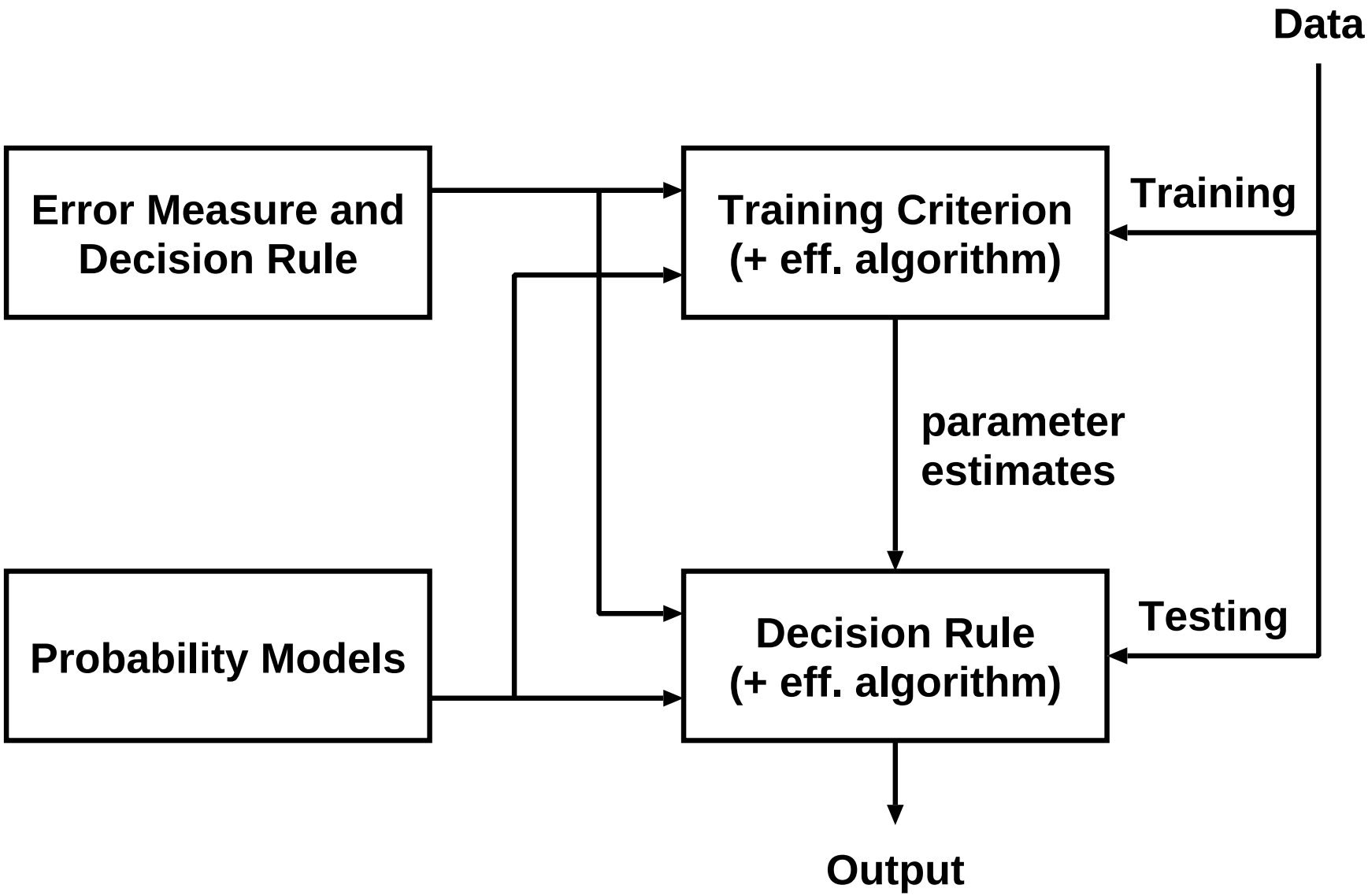
four key ingredients:

- form of Bayes decision rule:
  cost function = performance measure
- probability models:
  (mutual) dependencies between data and within data
  $\rightarrow$ problem-specific knowledge (e.g. from phonetics and linguistics)
- training criterion
  along with optimization strategy
- generation ('search', 'decoding')
  along with efficient strategy

Why does a system make errors?
none of the four components is perfect!

Statistical Approach to NLP Revisited

four key ingredients:



Misconceptions about Statistical MT (SMT)

typical objections and criticisms:

- **SMT has reached a plateau, there is no more improvement.**
  $\rightarrow$ The problem is not easy! MANY details have to be worked out!
- **SMT just remembers the examples.**
  $\rightarrow$ No, there is generalization due to the model!
- **SMT output depends only on the size and quality of the data;**
  $\rightarrow$ No, the output depends on the models, the training criterion, ...
- **To improve SMT, we need 'deep structural' models.**
  $\rightarrow$ Yes and No: we need better models, but what type of models?

Beyond 'Orthodox' Statistics

- huge number of free parameters:
  - statisticians prefer a few parameters
  - not enough training data
  - interaction between these parameters
- performance (= error rate) of the whole system matters and not quality of parameter estimates
- task: more 'predictive' than 'descriptive'
- problem-specific knowledge required: how much?
- computational efficiency matters:
  - training procedure
  - search (or generation) process

Towards Better Models

promising directions:

- **Yes, we need better models that extract more information/dependencies from the data.**
- **These models can be related to existing phonetic/linguistic theories, but they might also be very much different.**
- **These models have to be extracted from data and verified on data!**
- **Theses models might require a DEEP integration and require research on STATISTICAL decision theory along with efficient algorithms and implementations.**

- **examples of such approaches for MT:**
  - better integration of morphosyntax
  - long-distance dependencies
  - consistent lexicon models (’phrase table’)
  - ...

THE END