MT Summit IX, New Orleans, Sep. 23-27, 2003
Panel Discussion

HAVE WE FOUND THE HOLY GRAIL ?

Hermann Ney

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

Contents

1 Specific Questions

1. **Have we found the holy grail?**
2. **Will progress in data-driven MT continue unabated?**
3. **Has the data-driven paradigm been able to model information that was not present in rule-based systems?**
4. **Was the metric used to rank participating systems in the NIST competition fair ?**
5. **Is it correct that SMT has indeed surpassed traditional rule-based systems?**
6. **Are there niche applications for which SMT is well suited?**
7. **Is there a danger that SMT's recent success gives the impression that MT is a solved problem?**
8. **Would the NIST evaluation have been different for the language pair English-French?**
9. **What about rule-based component's in today's and future data-driven systems?**

2 Recent Projects: Speech and Language Translation

spoken language translation: joint projects (national, European, international: ATR, C-Star, Verbmobil, Eutrans, Nespole!, Fame, LC-Star, PF-Star, ...):

- **restricted domains:**
  appointment scheduling, conference registration, travelling, tourism information, ...
- **vocabulary size: 3 000 – 10 000 words**
- **best performing systems and approaches: data-driven**
  – example-based methods
  – finite-state transducers
  – statistical approaches
  e.g.: Verbmobil evaluation [June 2000]: better by a factor of 2

written language translation: US Tides project 2001-2004

- **unrestricted domain: press news, vocab.size $\cong$ 50 000 words**
- **language pairs: Chinese $\rightarrow$ English, Arabic $\rightarrow$ English**
- **performance [July 2003]:**
  best statistical systems are better than conventional/commercial systems

3 The Statistical Approach to NLP and MT

principles:

- **MT and other NLP tasks are complex tasks,**
  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 probabilistic dependencies between input x and decision c:

$$\begin{aligned} x \rightarrow \hat{c} &= \arg\max_{c} \left\{ pr(c|x) \right\} \\ &= \arg\max_{c} \left\{ pr(c) \cdot pr(x|c) \right\} \end{aligned}$$

- **resulting concept:**

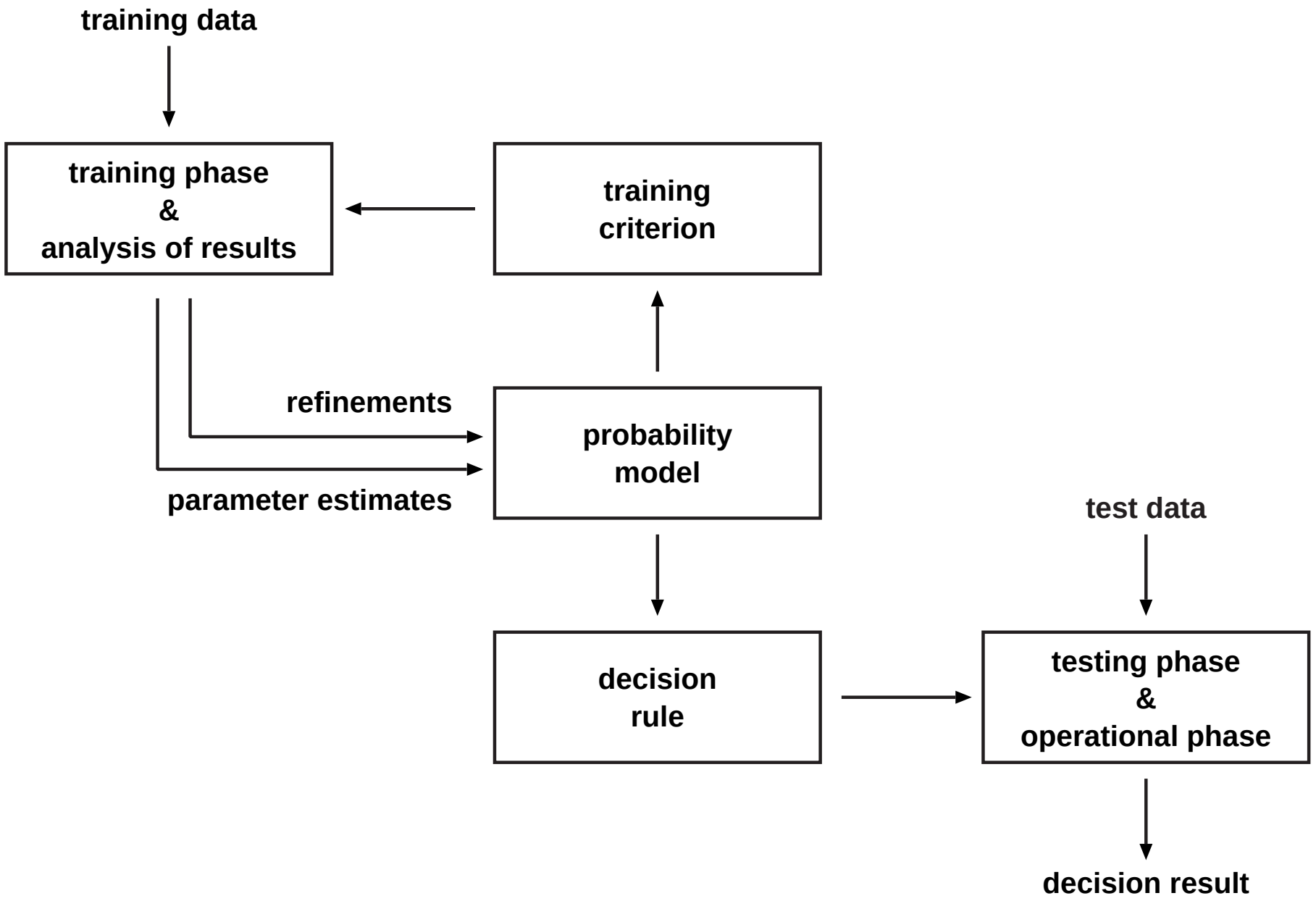
NLP = Statistics + (Linguistic ?) Modelling

The Statistical Approach: Key Components

- **decision rule:**
  requires maximization (sometimes hard!)
  and probability distribution $pr(c|x)$, which is unknown

- **probability model $p_\theta(c|x)$ or $p_\theta(c) \cdot p_\theta(x|c)$**
  is used to replace $pr(c|x)$ or $pr(c) \cdot pr(x|c)$

- **training criterion**
  to learn the unknown parameters θ from training data
  ideal goal: optimum performance


training data
training phase
&
analysis of results
training
criterion
refinements
parameter estimates
probability
model
test data
decision
rule
testing phase
&
operational phase
decision result

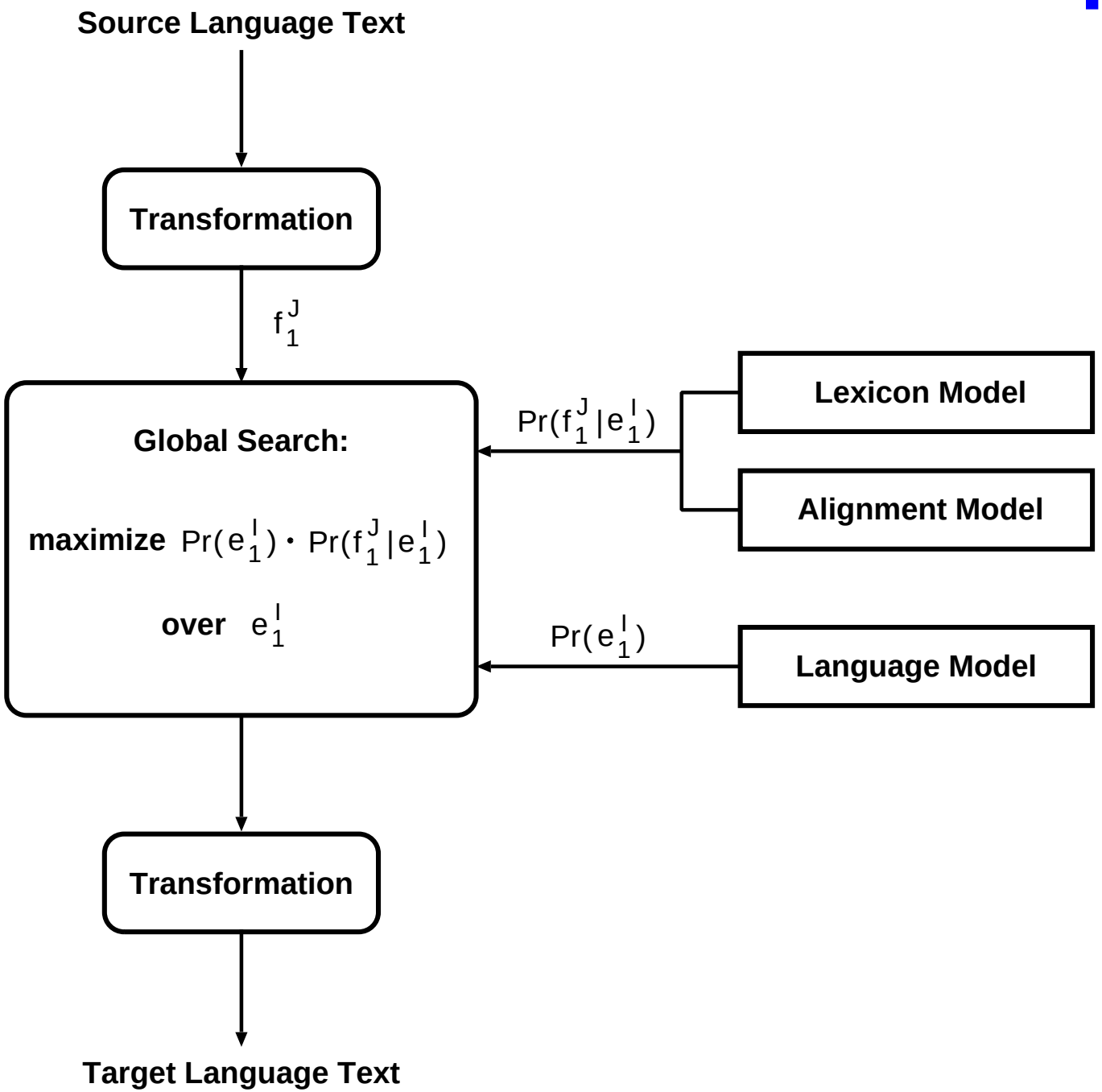
Advantages of Statistical Approach

- **holistic decision criterion:**
  - **exploits ALL (available) knowledge sources**
  - **is able to combine thousands of weak dependencies**
  - **handles interpendencies, 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:
virtually none of these statements applies to rule-based systems!

Machine Translation: Bayes Decision Rule



4 State of the Art in SMT

lot of progress in SMT:
best statistical systems are competitive with conventional, hand tailored systems

system components:

- **alignment and lexicon model:**
  - **training: IBM-1 to -5 and/or HMM: based on single words**
  - **symmetrization of roles of source and target languages**
- **extraction of phrases (alignment templates):**
  try to memorize all source/target phrases
- **language model:**
  word tri- and higher n-grams
- **generation (search):**
  beam search, with limited degree of non-monotinicity

performance:

- **use of phrases:**
  - **lion's share of the improvement**
  - **unclear: performance on unseen test data**
- **lack of syntactic structure**

Room for Improvements and Challenges

- **Bayes decision rule**
  for translating a source sentence f_1^J into a target sentence e_1^I:

$$\underset{e_1^I}{\operatorname{argmax}}\, Pr(e_1^I|f_1^J) \;=\; \underset{e_1^I}{\operatorname{argmax}}\{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\}$$

  – optimizes sentence errors, not word errors or BLEU/NIST score

  challenge:
  – decision rule closer to word errors or BLEU/NIST score ?
  – training criterion ?

- **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, ...**
- **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 ?

comparison with speech recognition (1973-2003):

- **most of the progress: by pure statistical modelling**
- **some progress: by weak acoustic-phonetic knowledge**
- **no progress: by classical rule-based and AI methods**

prediction (?) for machine translation:
improvements by progress in pure statistical modelling:

- **more training data (counteracts estimation problems)**
- **improved training criteria and training algorithms**
- **by better modelling the data inherent dependenccies (more structured models) (program for 20-200 years?)**

5 Answers

- **SMT is the right direction, there is no inherent ceiling, but it is still a long way to go (20–200 years?)**
- **advantages of statistical MT: better decisions, processing lots of data, performance feedback**
- **If done correctly, SMT must result in the best performance due to the coupling of training and performance criterion**
- **fair comparison:**
  – many aspects: time, effort, ...
  – evaluation metric: not perfect, but of secondary importance
- **specific applications for SMT: rapid system development (if parallel corpus exists)**
- **hybrid systems: in theory yes, in practice ??? (see speech recognition)**
- **funding: Being too successful is not good for funding.**

THE END