HYBRID MACHINE TRANSLATION

Hans Uszkoreit

German Research Center for Artificial Intelligence (DFKI)

joint work with

Christian Federmann, Yu Chen, Andreas Eisele,
Silke Theison and Sabine Hunsicker

Observed Progress

☆ Considerable progress in Statistical Machine Translation:

Although the SMT systems are not yet better than the best rule-based systems, they overcome central shortcomings and they can be produced much faster and cheaper.

☆ But also progress in linguistic processing:

Progress in parsing, morphology and generation has been rather remarkable. Robust wide coverage analysis becomes feasible.

Available language resources and tools for producing them have considerably improved.

☆ Less progress but nevertheless increased use of Rule-Based MT:

Growing number of institutional users. Adaptation to special tasks.

Current Trends

- ☆ increase of linguistic structure and knowledge in SMT
- ☆ increase of statistical methods for disambiguation and lexical selection in RBMT
- ☆ increase of number and power of systems combinations
- ☆ increase of research and number of approaches in Hybrid MT
- ☆ exploitation of social computing through data feedback by humans, less by active learning

Major Bottlenecks in Processing Methods

for SMT

- ☆ no adequate solutions for non-local grammatical phenomena such as free word order, long-distance dependencies, ellipsis, complex coordination, etc
- ☆ no adequate solutions for (lexical and syntactic) gaps in training data

for RBMT

- ☆ no adequate solutions for disambiguation, semantic selection, style, usage preferences
- ☆ no adequate solutions for gaps in lexicon and grammar

Respective Advantages

RBMT	SMT
☆ large development effort	☆ small development effort
☆ systems for few languages	☆ systems for many languages
☆ gaps depend on developers	☆ gaps depend on training data
☆ problems with lexical choice	☆ better lexical choice
☆ better grammatical structure	☆ frequent grammatical errors

Examples

Englisch	RMBT: translate pro	SMT: Koehn 2005
We seem sometimes to have lost sight of this fact.	*Wir scheinen manchmal Anblick dieser Tatsache verloren zu haben.*	*Manchmal scheinen wir aus den Augen verloren haben, diese Tatsache.*
The leaders of Europe have not formulated a clear vision.	*Die Leiter von Europa haben keine klare Vision formuliert.*	*Die Führung Europas nicht formuliert eine klare Vision.*
I would like to close with a procedural motion.	*Ich möchte mit einer verfahrenstechnischen Bewegung schließen.*	*Ich möchte abschließend eine Frage zur Geschäftsordnung ε.*

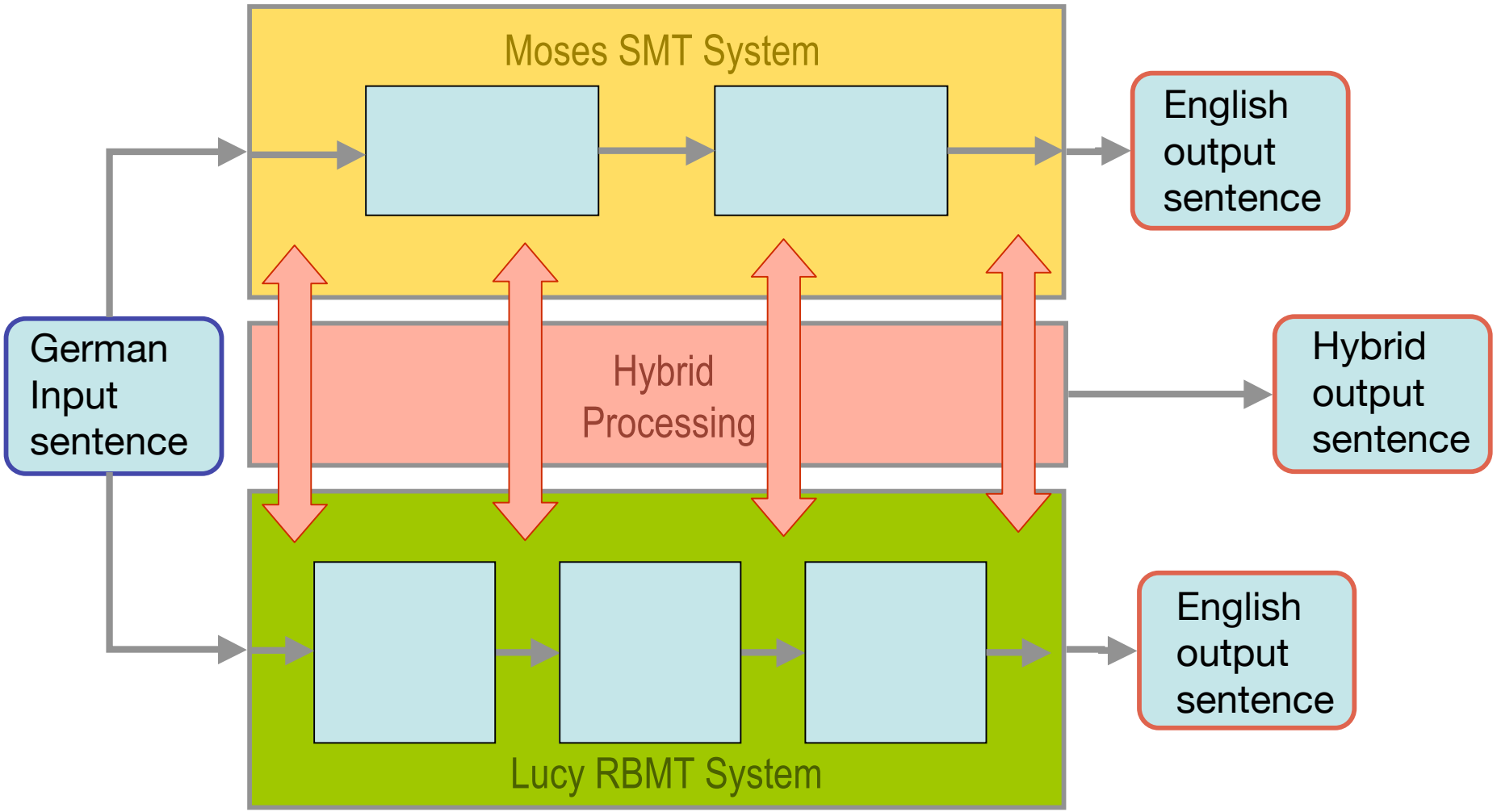
Major Bottlenecks in Evaluation

- ☆ lack of reliable and diagnostic automatic evaluation methods
- ☆ lack of evaluation metrics that reflect the usefulness/economic value with respect to actual application settings

Hybrid Approach

- ☆ Open Source SMT plattform Moses
- ☆ Proprietory RBMT system Lucy (vormals METAL, Comprendium, ...)

Plans for Hybrid Processing



First Experiments

- ☆ SMT Postediting of output of Lucy RBMT

- ☆ controlled substitution of phrases in the RBMT output by SMT phrase table

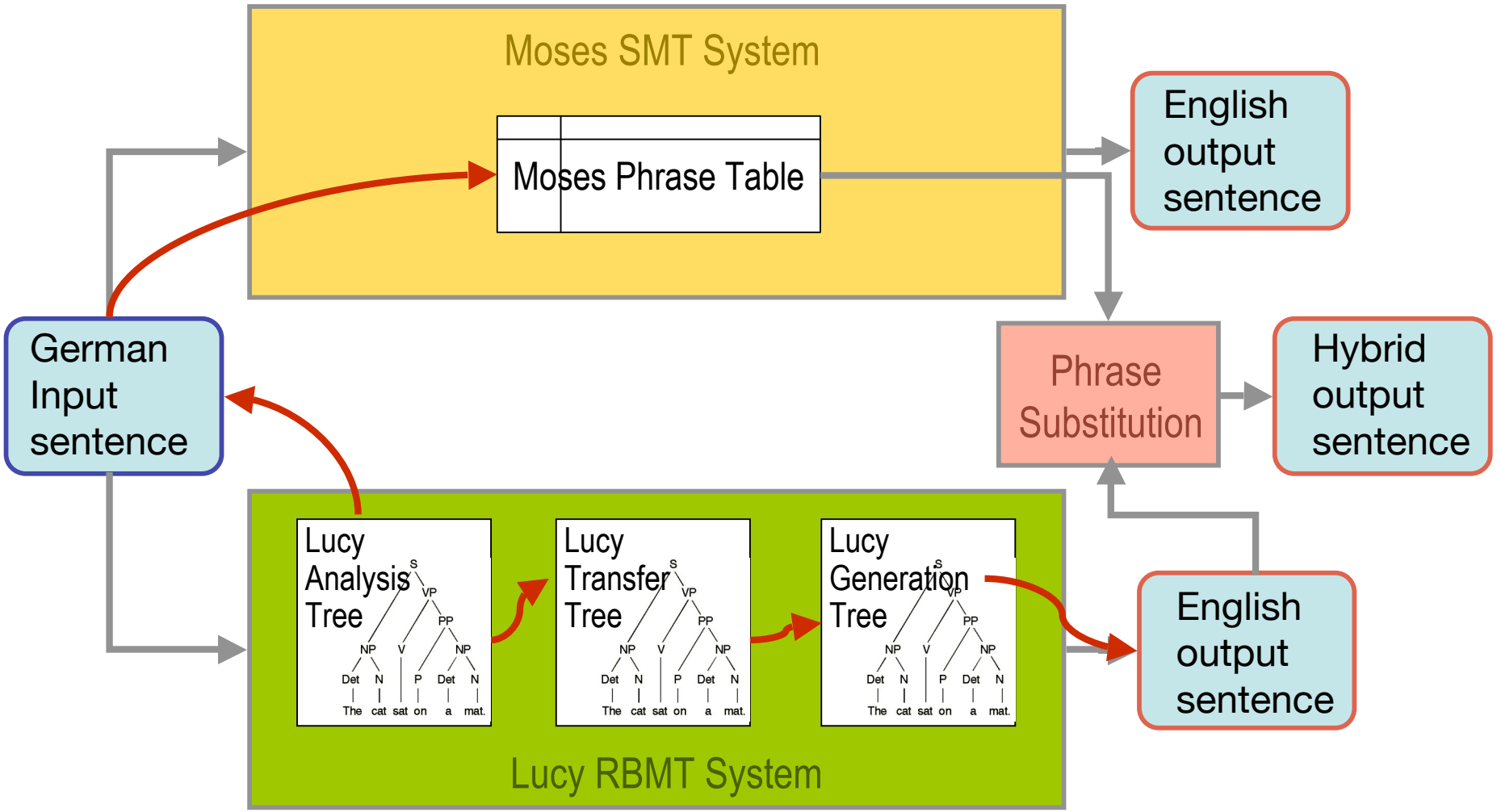
Approach

- ☆ we start by substituting noun phrases
- ☆ criteria for substitution:
  - category
  - alignment
  - morphological fit
  - length
  - complexity
  - probability in the phrase table
  - probability in the language model

Experiment

- ☆ German -> English
- ☆ 2525 sentences
- ☆ taken from the test set of the EuroMatrix WMT Shared Task 2009

Architecture



Evaluation

Automatic Evaluation

System	BLEU
Lucy	11.78
Lucy SPE	11.92
Hybrid	12.17

Ranking by Human Evaluators

System	Ranked 1/2/3 (in %)			Average
Lucy	47.74	38.75	13.5	1.64
Lucy SPE	27.00	40.50	32.5	2.10
Hybrid	61.50	30.00	8.5	1.49

Error Analysis

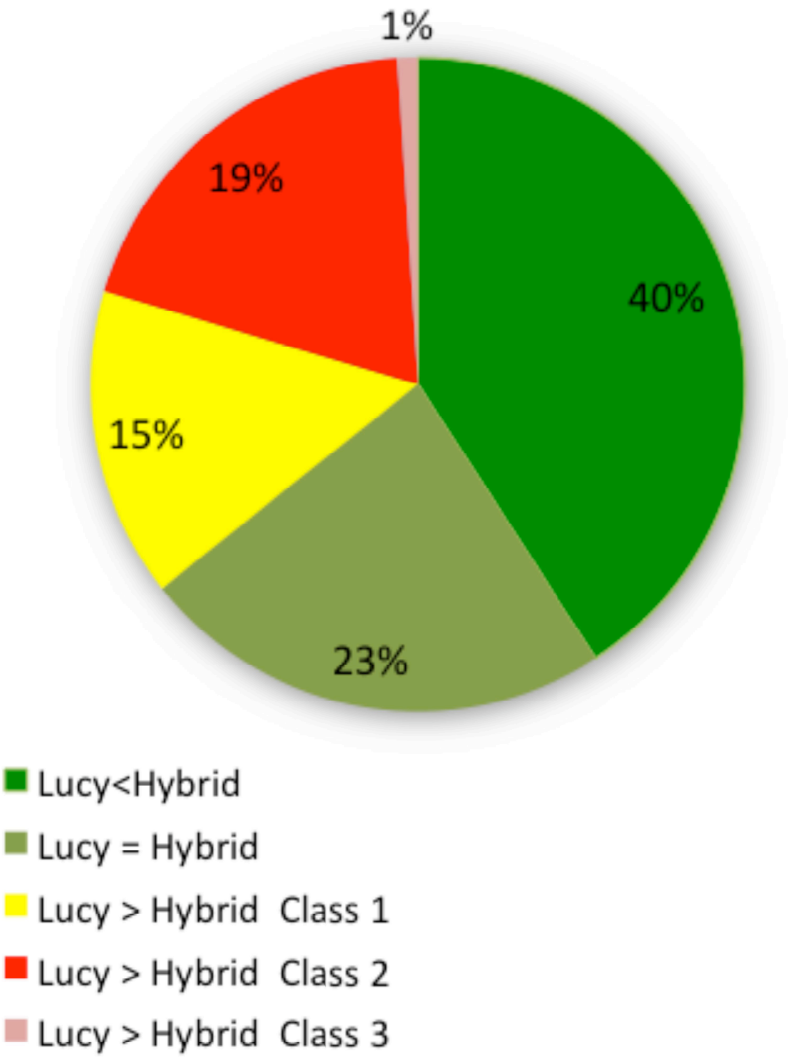
Improvement, i.e. substitution better than Lucy.

Preservation, i.e. substitution equal to Lucy.

Class 1 Error The result is correct "content-wise", but the syntactic structure degrades. Destroyed agreement, double prepositions, etc. We consider these errors not very harmful as they can easily be fixed.

Class 2 Error Due to bad input from the SMT system. Because of the nature of the algorithm, these errors cannot be avoided! Some may be prevented by employing several SMT systems.

Class 3 Error Substitution process goes, astray, because of, e.g., tokenization, problems. It will take more time to fix, errors of this class.



Conclusion and Outlook

- ☆ Conclusions
  - results are encouraging
  - search space for optimization is large
  - informative evaluation is hard
- ☆ Next Steps
  - we are now trying machine learning methods
  - we are improving diagnostic evaluation
  - we will include other phrase types
  - we will include additional criteria and additional knowledge sources such as terminologies and TMs