Machine Translation II: Hybrid Methods and Evaluation

January 19, 2009

Andreas Eisele
UdS Computerlinguistik & DFKI
eisele@dfki.de

Language Technology I
WS 2008/9

Hybrid Methods and Evaluation

UNIVERSITÄT DES SAARLANDES

Overview

- Hybrid MT architectures
  - Motivation
  - Architectures
- Evaluation techniques
  - Automatic Evaluation
  - Human Evaluation
  - Improvements via Meta-Evaluation

Motivation of Hybrid MT

UNIVERSITÄT DES SAARLANDES

Observation (early 90s): Approaches to MT have complementary PROs and CONs:

Table 1. Summary of Different Approaches to Machine Translation System

	Advantages	Disadvantages
Rule-Based	1. easy to build an initial system 2. based on linguistic theories 3. effective for core phenomena	1. rules are formulated by experts 2. difficult to maintain and extend 3. ineffective for marginal phenomena
Knowledge-Based	1. based on taxonomy of knowledge 2. contains an inference engine 3. interlingual representation	1. hard to build knowledge hierarchy 2. hard to define granularity of knowledge 3. hard to represent knowledge
Example-Based	1. extracts knowledge from corpus 2. based on translation patterns in corpus 3. reduces the human cost	1. similarity measure is sensitive to system 2. search cost is expensive 3. knowledge acquisition is still problematic
Statistics-Based	1. numerical knowledge 2. extracts knowledge from corpus 3. reduces the human cost 4. model is mathematically grounded	1. no linguistic background 2. search cost is expensive 3. hard to capture long distance phenomena

Source: Chen & Chen: A Hybrid Approach to Machine Translation System Design, Computational Linguistics and Chinese Language Processing, 1996

Strengths and Weaknesses of SMT vs. RMBT

UNIVERSITÄT DES SAARLANDES

(RBMT:translate pro ⟷ SMT:Koehn 2005, examples from EuroParl)

EN: ***I wish the negotiators continued success with their work in this important area.***

RBMT: ***Ich wünsche, dass die Unterhändler Erfolg mit ihrer Arbeit in diesem wichtigen Bereich fortsetzten.***

continued: Verb instead of adjective

SMT: ***Ich wünsche der Verhandlungsführer fortgesetzte Erfolg bei ihrer Arbeit in diesem wichtigen Bereich.***

three wrong inflectional endings

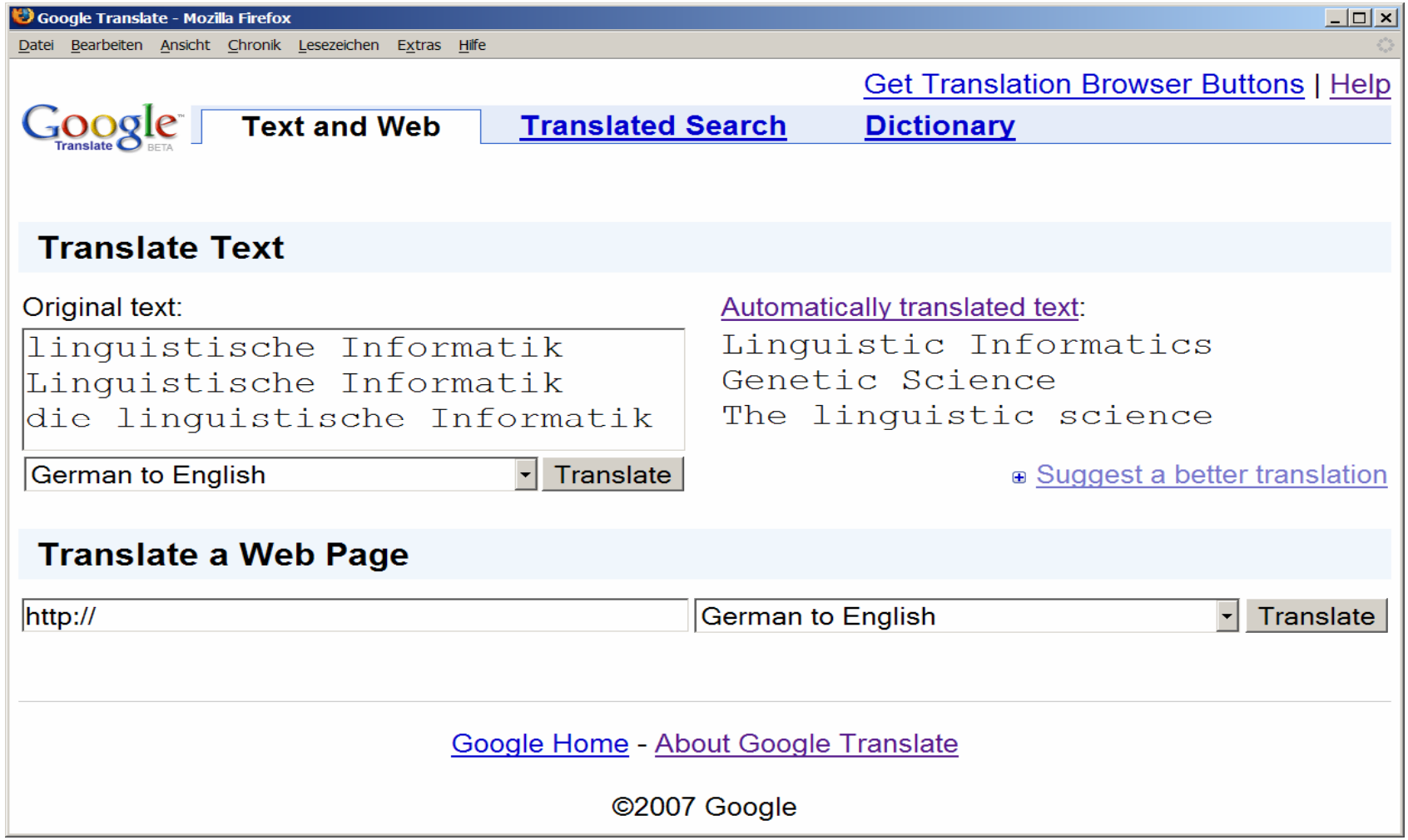
Strengths and Weaknesses of SMT vs. RMBT

UNIVERSITÄT DES SAARLANDES

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.*

Problems with Reliability of Lexicon Acquisition

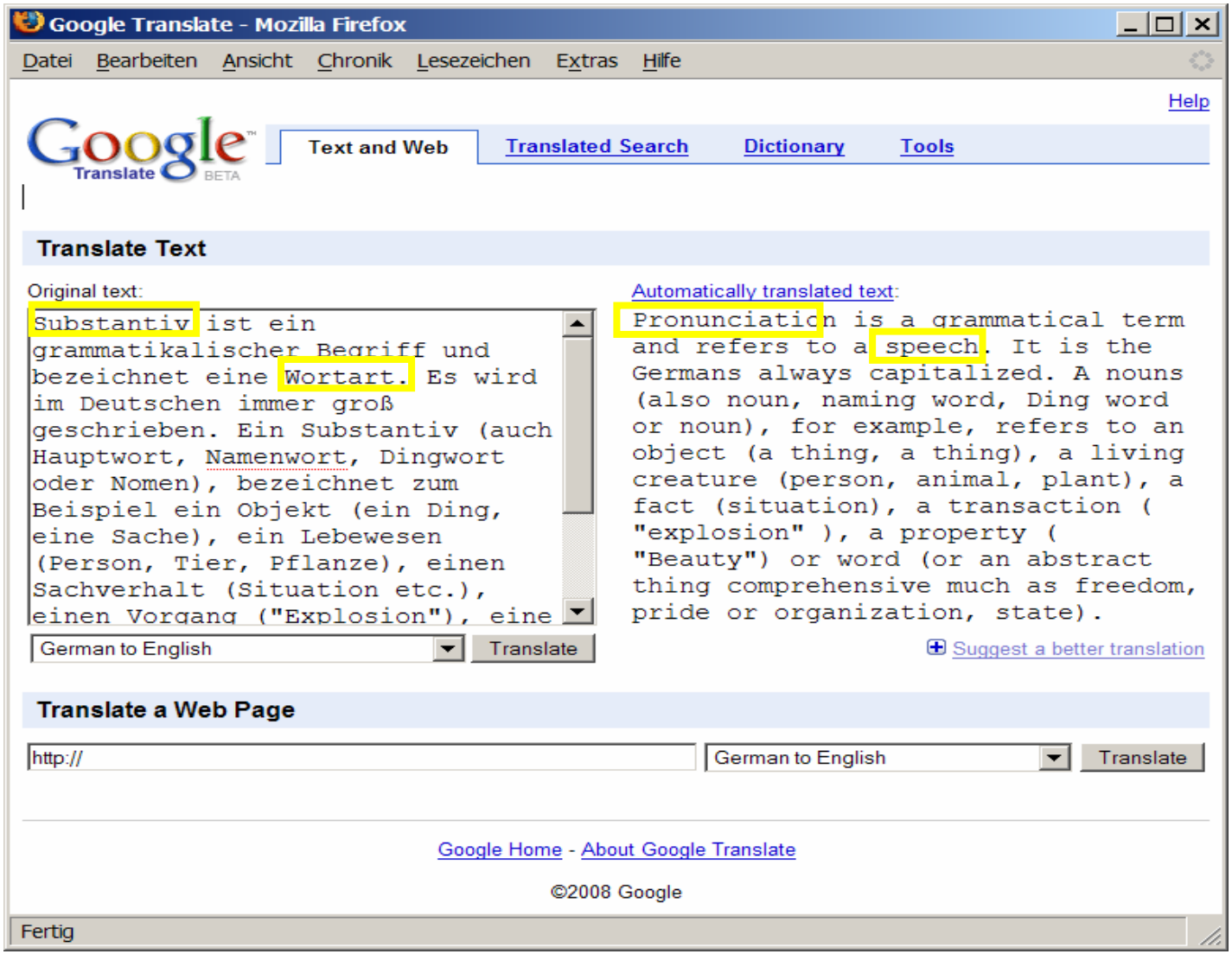
UNIVERSITÄT DES SAARLANDES



[November 2007, corrected in the meantime]

More Examples of Reliability Problems

UNIVERSITÄT DES SAARLANDES



[January 2008, partly corrected in the meantime]

Motivation for Hybrid Approaches to MT

UNIVERSITÄT DES SAARLANDES

In the early 90s, SMT and RBMT were seen in sharp contrast.

But advantages and disadvantages are complementary.

➔ Search for integrated methods is now seen as natural extension for both approaches

	RBMT	SMT
Syntax, Morphology	++	--
Structural Semantics	+	--
Lexical Semantics	-	+
Lexical Adaptivity	--	+
Lexical Reliability	+	-

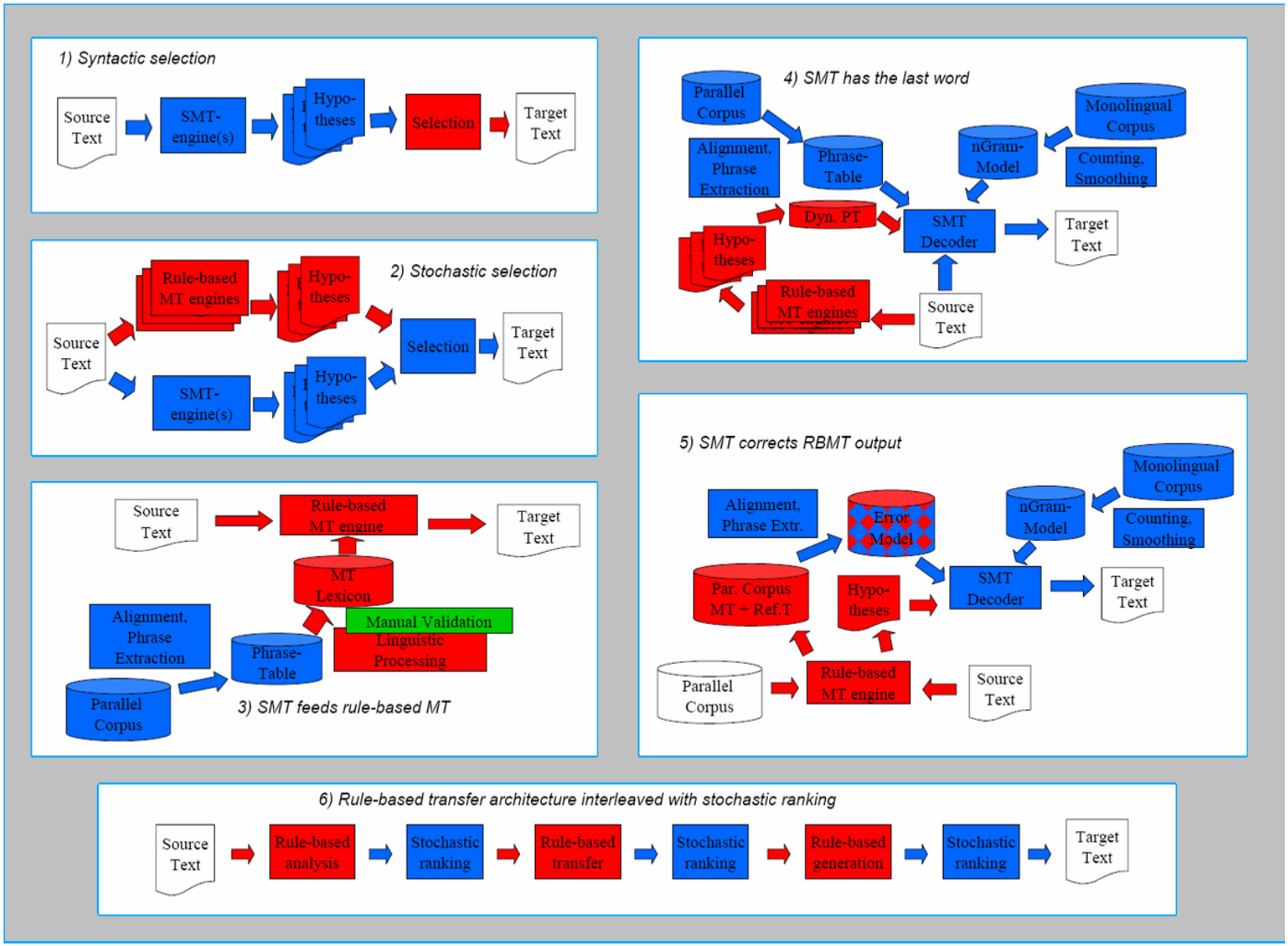
Motivation for Hybrid Approaches to MT

UNIVERSITÄT DES SAARLANDES

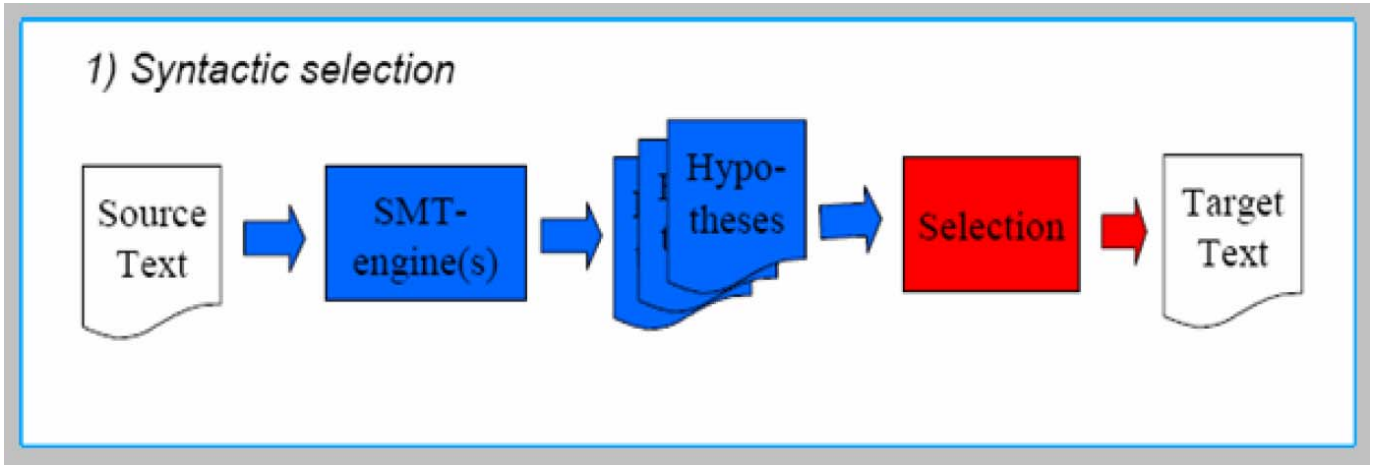
- Statistical and rule-based approaches address different types of knowledge:
  - Rule-based approaches focus on linguistic knowledge
  - Statistical approaches provide a holistic, integrated model that also incorporates (some) implicit knowledge of the world
- All available types of knowledge are urgently required, as the task is too difficult to ignore important aspects
- Research on a deep integration of statistical and linguistic approaches is required but this will take some time
- In the meantime, we can try to tinker with existing MT engines

Some hybrid MT architectures

= SMT Module
= RBMT Module



Syntactic Selection



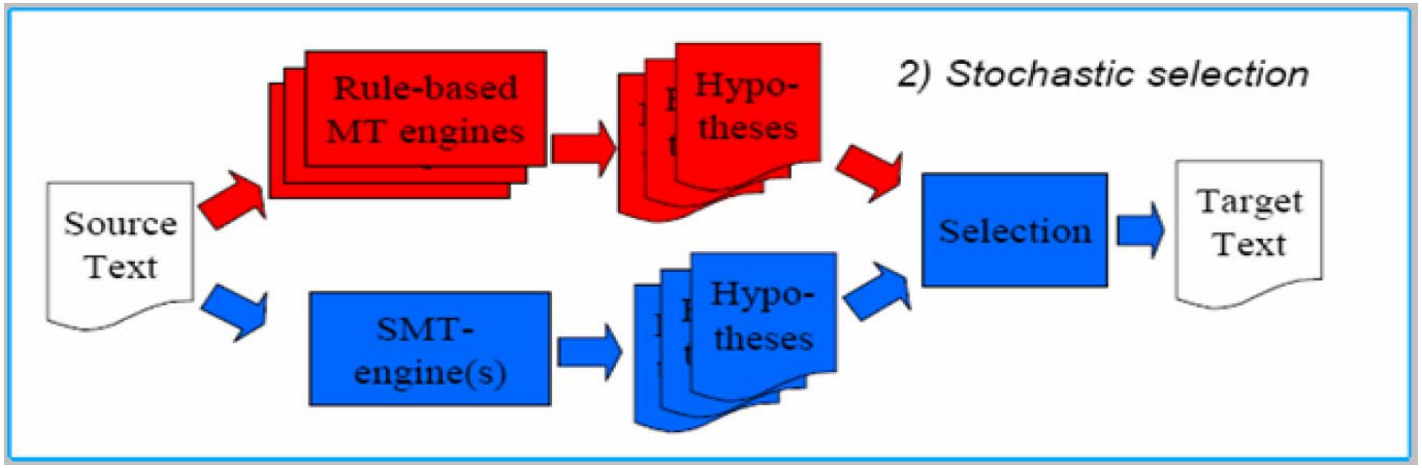
Motivation: SMT output often syntactically ill-formed

➔ Selection mechanism in SMT „generate and test“ should be enriched with syntactic knowledge

BUT:

- syntactic parsers not (yet) robust enough
- High computational cost of processing many ill-formed candidates

Stochastic Selection



Motivation: Selection from an increased number of candidates can improve overall quality

BUT:

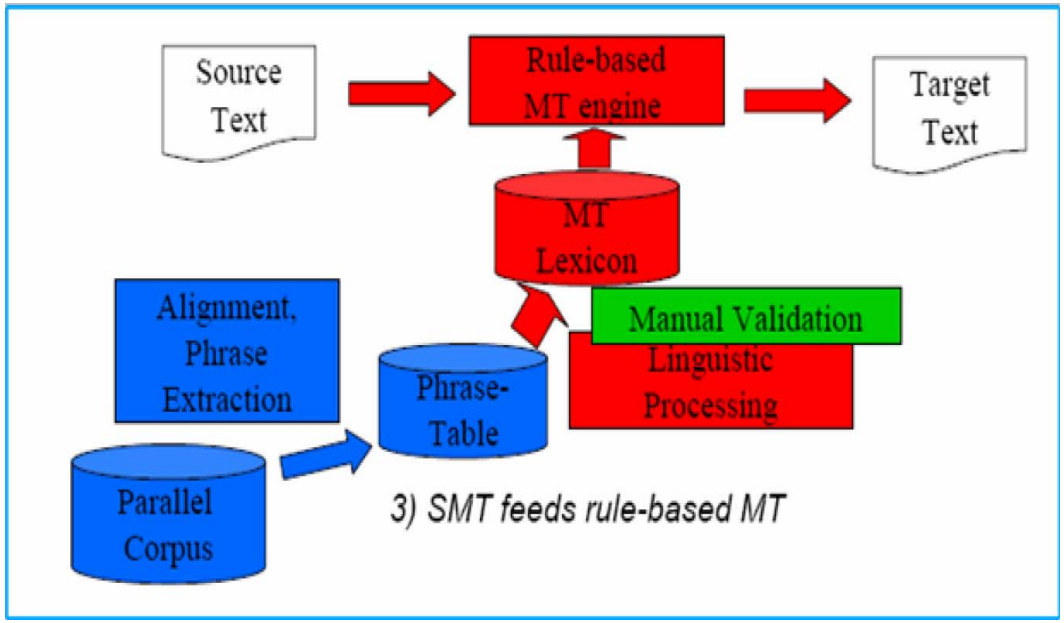
- Works mainly for short utterances, where one of the candidates may be good enough (VerbMobil)
- Different candidates may have problems in different parts of the sentence, granularity of decisions too coarse

SMT feeds rule-based MT

UNIVERSITÄT DES SAARLANDES

Motivation:

- Adapting RBMT to new domains requires lots of new lexical entries that are difficult to write manually
- SMT techniques can help to partially automate this process



BUT:

- Not all required information can be learned from data
- Errors in examples/SMT alignment may creep in, but RBMT has no mechanism to discard implausible outcomes
- Some manual effort is required

Corpu

European Patent Office (EPO):

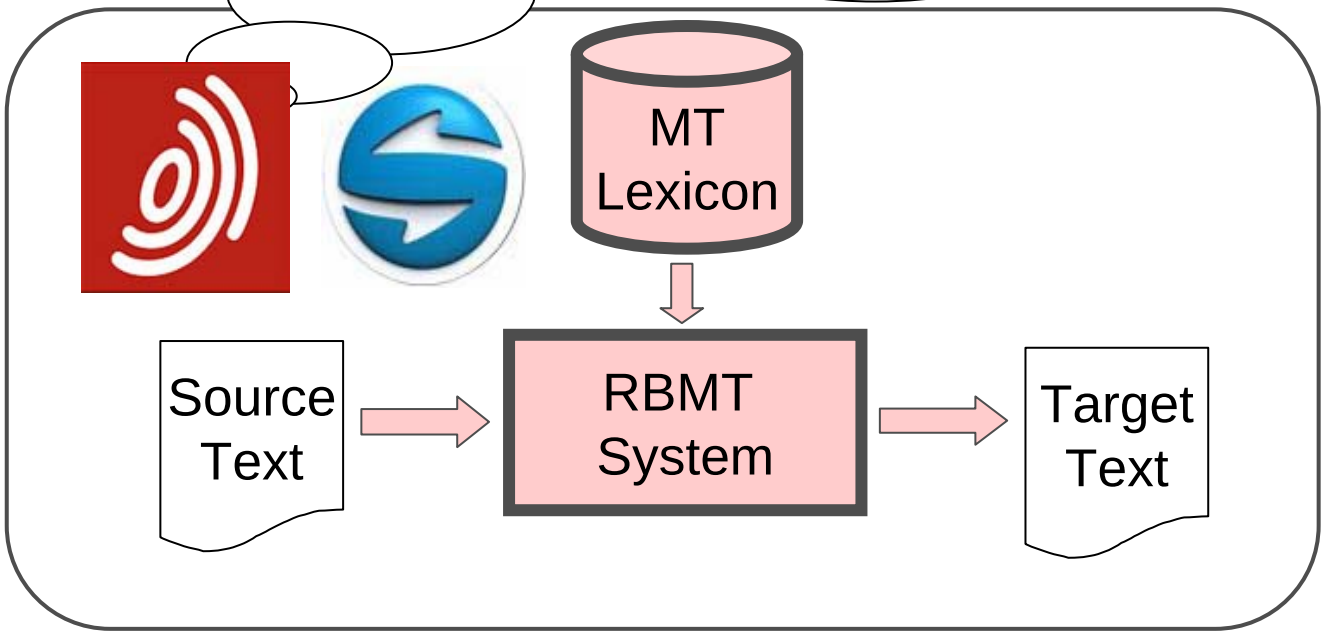
6000 employees from > 30 countries in Munich, The Hague, Berlin, Vienna, Brussels

Collection of > 60 Mio. patent documents

130000 patent applications/year (2006)

Prepares translation service for patent documents

Call for tenders & ***selection test***, fall 2005



Language pairs

DE ↔ ***EN***

ES ↔ ***EN***

FR ↔ ***EN***

IT ↔ ***EN***

planned:

EL ↔ ***EN***

PT ↔ ***EN***

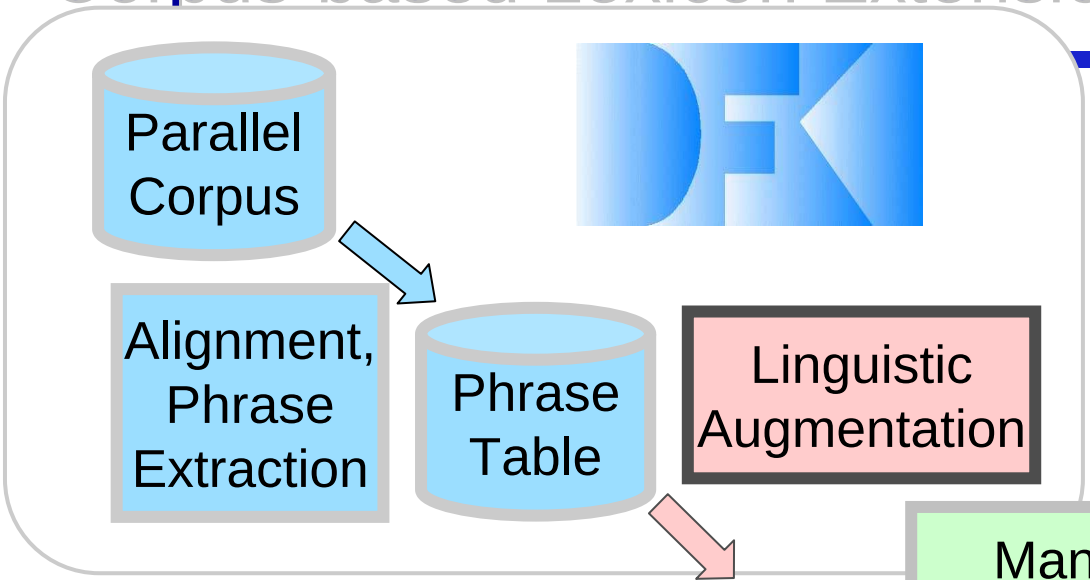
NL ↔ ***EN***

RO ↔ ***EN***

FR ↔ ***DE***

FR ↔ ***ES***

Corpus-based Lexicon Extension for RBMT



SMT-Technology with linguistic knowledge helps rule-based MT-System

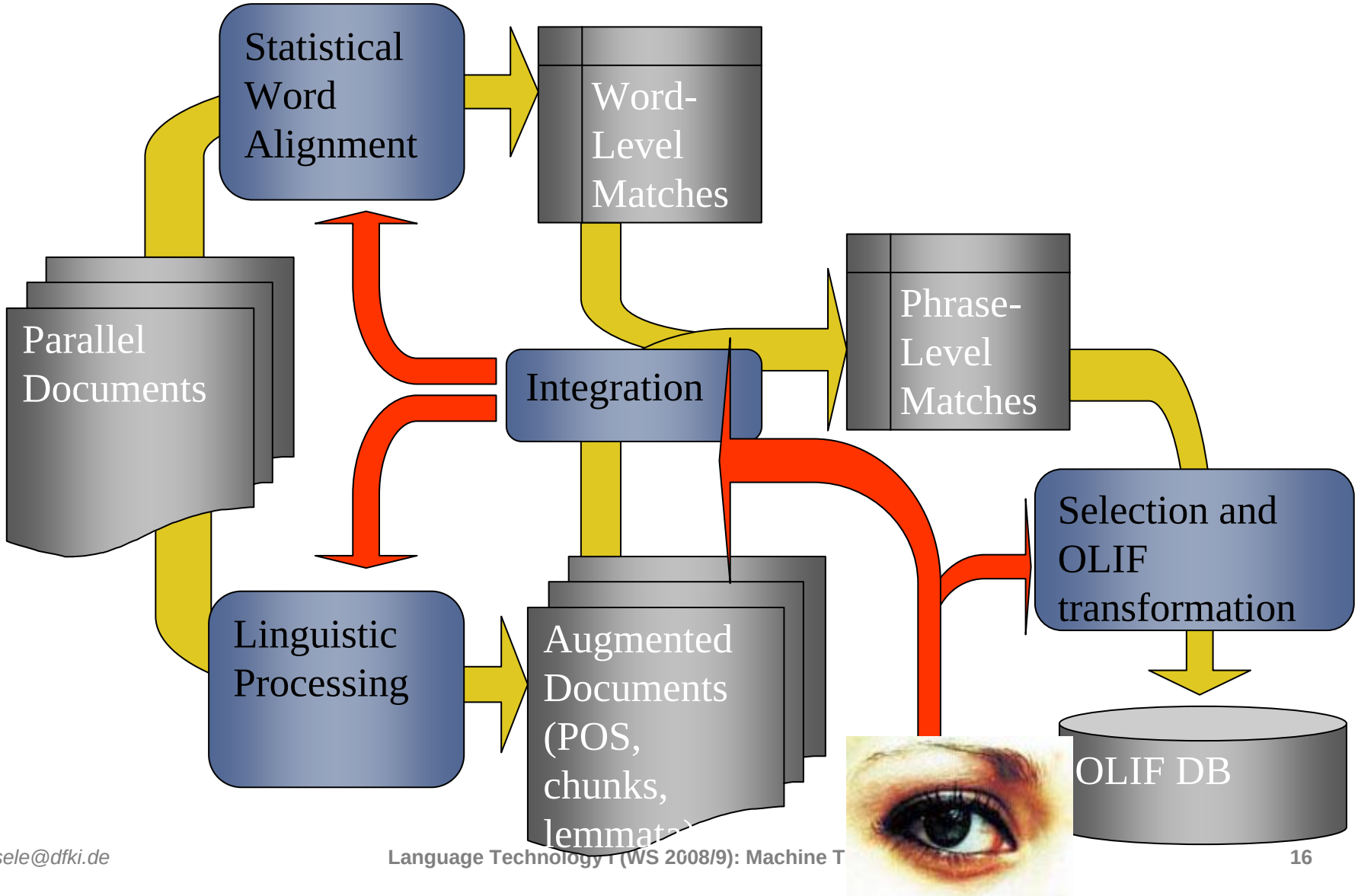
Language pairs

DE ↔ EN
ES ↔ EN
FR ↔ EN
IT ↔ EN

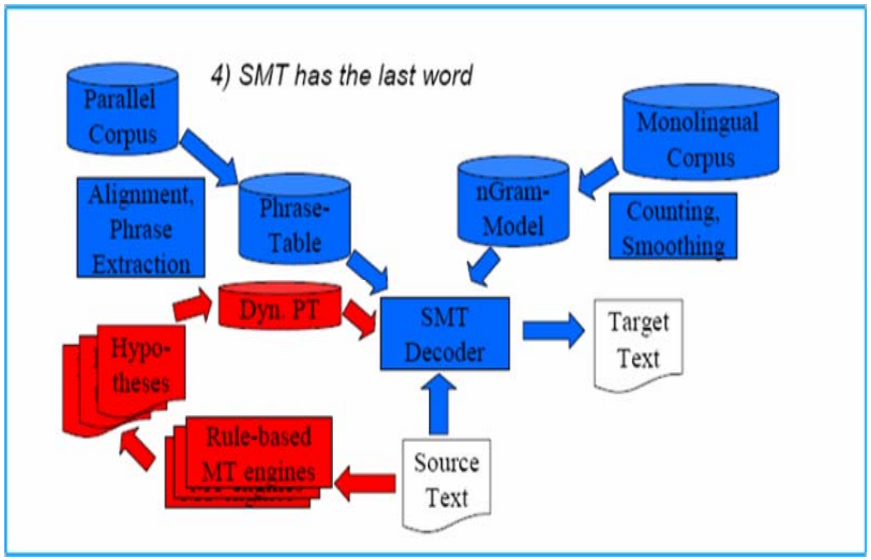
planned:

EL ↔ EN
PT ↔ EN
NL ↔ EN
RO ↔ EN
FR ↔ DE
FR ↔ ES

Terminology Extraction for MT: Architecture



RBMT feeds SMT

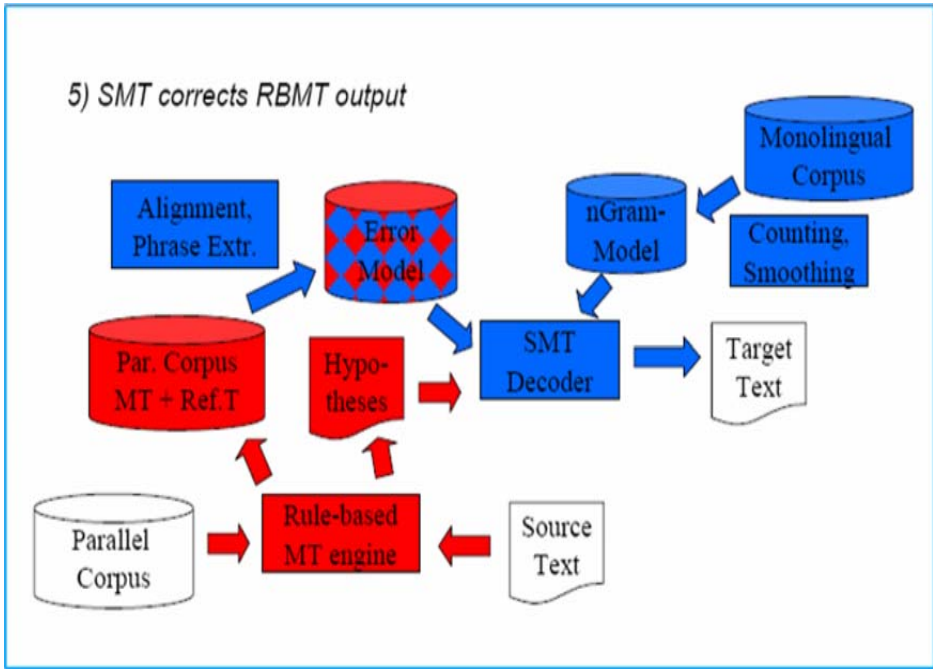


Motivation: SMT can only know what is in the training data,
RBMT systems often contain extensive lexical knowledge

BUT:

Architecture can fix lexical gaps, but will not covercome problems with syntactically ill-formed candidates

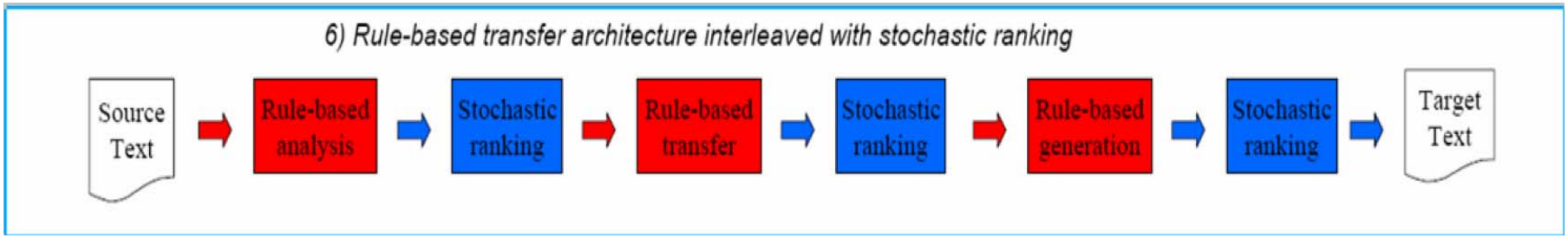
Statistical post-correction



Motivation: Errors in RBMT can be systematic/regular, may be fixed automatically. Target language model helps to find most natural wording in context

BUT: Sometimes RBMT messes a sentence completely up, no hope to repair these cases via SMT

Transfer architecture with stochastic ranking



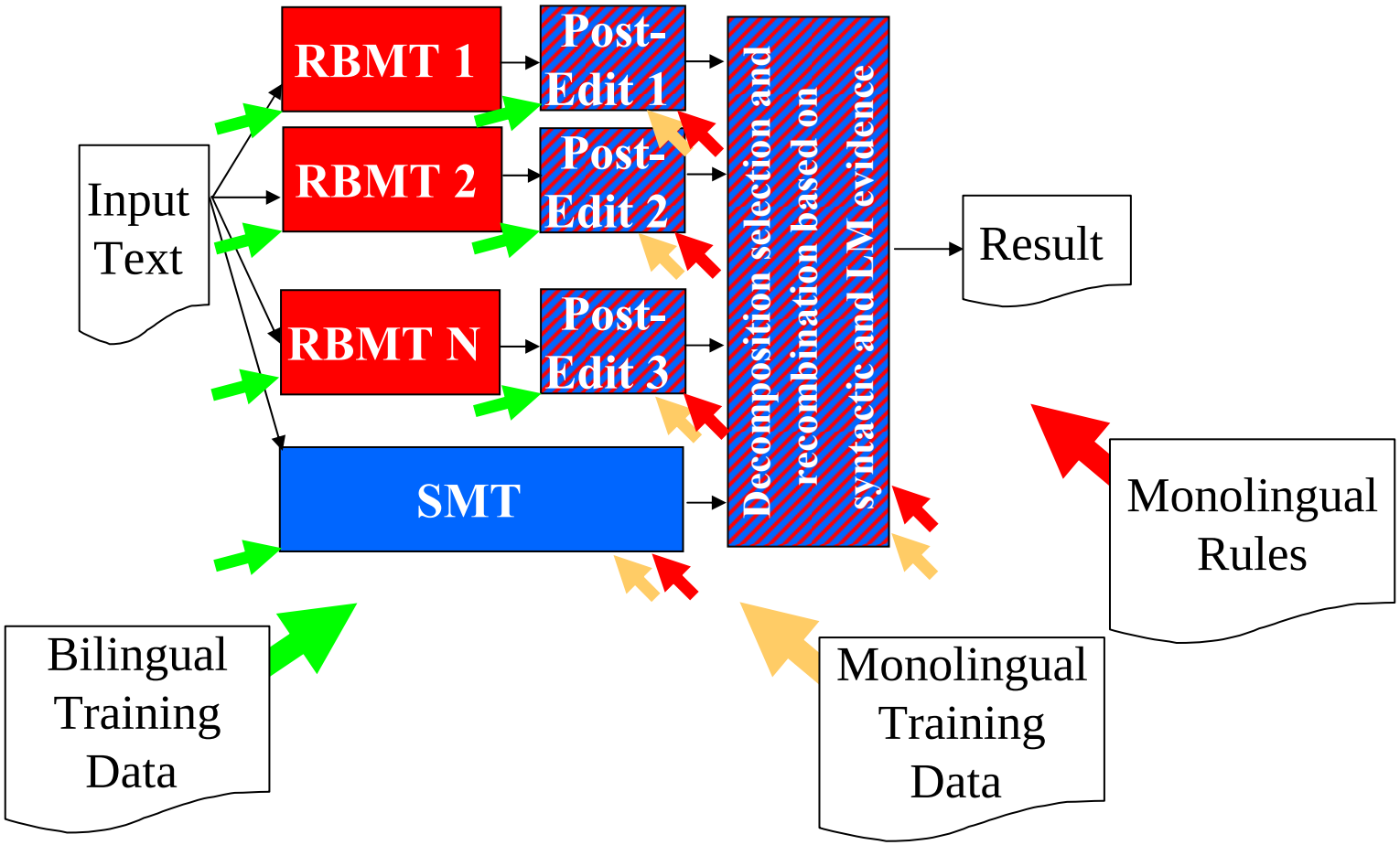
Motivation: Fine-grained combination of statistical and linguistic evidence on all levels requires a closely coupled implementation

BUT:

- Chain can only be as good as the weakest link
- Difficult to avoid mismatches between representations when hand-crafting grammars
- Many existing processing components are designed for deterministic processing; building up forests of alternative solutions may require redesign of algorithms

Competition vs. Integration

Ideas presented so far are independent, combinations are possible



Many combinations of techniques ➔ big effort for systematic tuning

Pre-emptive division of labor between modules

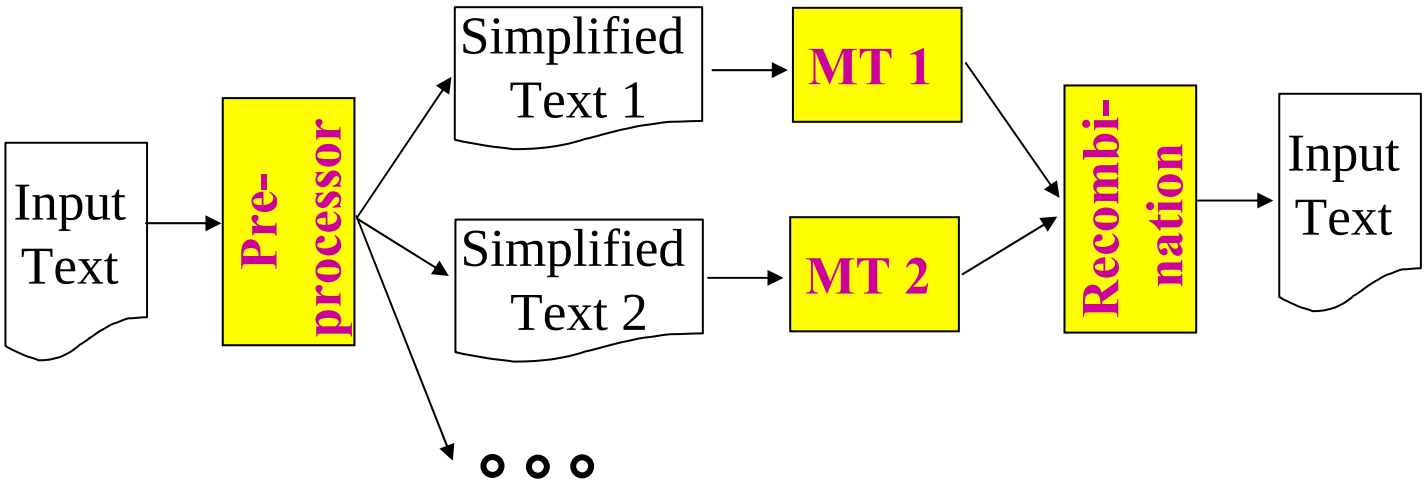
UNIVERSITÄT DES SAARLANDES

The idea:

- So far, we send the input text unmodified through many MT systems, try to make sense of (partially erroneous) output after errors have been made
- Sometimes, a slight modification of the input can prevent errors from happening, e.g. by
  - replacing named entities unknown to the engine by place-holders
  - simplifying technical noun-phrases
  - treating special cases (numbers, names) in special ways
- Statistics of error types can be used to find out specific weaknesses and best way to distribute work over engines

Pre-emptive division of labor

Schematic architecture



Actually already used in simplified form (e.g. for markup processing, numbers, proper names)

Open questions:

- Can we learn what to send through MT system from examples?
- What kind of pre-processing is adequate (should be robust ***and*** linguistically informed)

Evaluation of MT systems

UNIVERSITÄT DES SAARLANDES

Two types of MT evaluation

- Human („subjective“)
- Automatic („objective“)

The evaluation dilemma:

- Manual evaluation is meaningful, but expensive, tedious, and error-prone, not useful for regression testing
- Automatic evaluation is repeatable, objective, but not necessarily relevant; better systems may have worse scores

We need to

- lower the effort for manual evaluation,
- increase the quality of automatic evaluation,
- or do both

Reasons for MT Evaluation

UNIVERSITÄT DES SAARLANDES

"More has been written about MT evaluation over the past 50 years than about MT itself"

[Y. Wilks, according to Hovy e.a.]

MT evaluation may serve different purposes

It may help to decide

- whether to apply MT at all
- which of a set of systems to use for a given task
- which problems/error to focus on in further development of one system
- how to combine systems in a hybrid architecture

Types of MT Evaluation

UNIVERSITÄT DES SAARLANDES

- Relative vs. absolute evaluation
  - which system is better? vs.
  - rate system X on a scale from 0 (useless) to 100 (perfect)
- Adequacy evaluation
  - will system X fit a given purpose?
- Task-based evaluation
  - can users of system X achieve a given task?
- Diagnostic evaluation
  - which phenomena are/aren‘t handled correctly?
- Performance evaluation
  - measure performance in specific areas in more detail
- Black-Box vs. Glass-Box
  - does evaluation see only in-/output or also the internal representations?

Subjective Evaluation

UNIVERSITÄT DES SAARLANDES

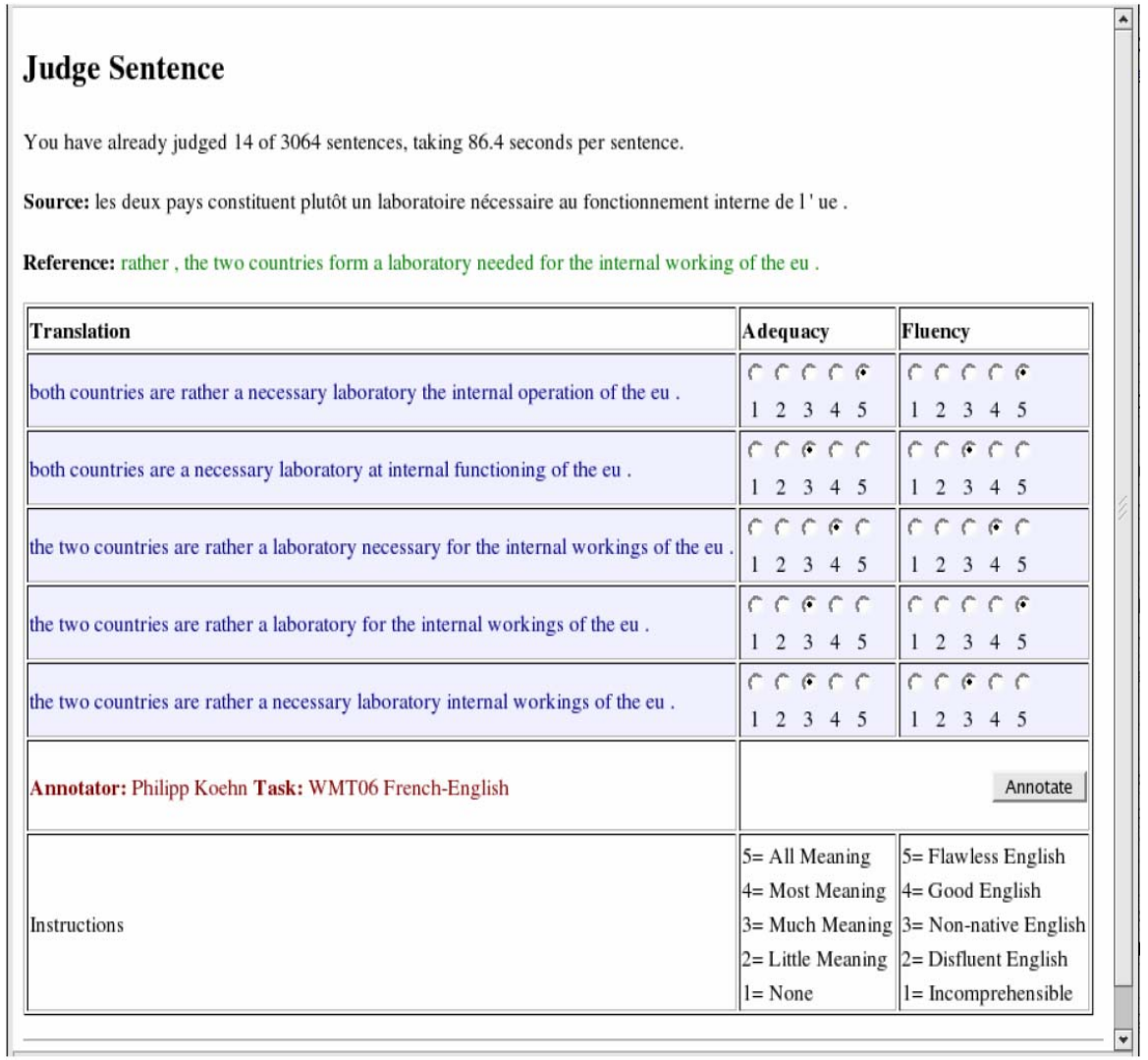
Main focus traditionally on two aspects:

- Adequacy
  - „Is the output equivalent to the input“ (in what sense?)
- Fluency
  - „Is the output well-formed in the target language?“

Subjective MT Evaluation in Practice

UNIVERSITÄT DES SAARLANDES

Koehn/Monz 2006 distributed the burden of manual evaluation over the participants in the shared MT task, using a web-based evaluation interface



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Problems of Subjective Evaluation

UNIVERSITÄT DES SAARLANDES

- Task is very tedious
- Inter-annotator agreement could be better
- Long sentences are particularly hard to judge
- Linguistic expertise of the evaluators not exploited

Manual Error Analysis

Human evaluators may give more specific diagnosis of problems [Vilar e.a. 2006]

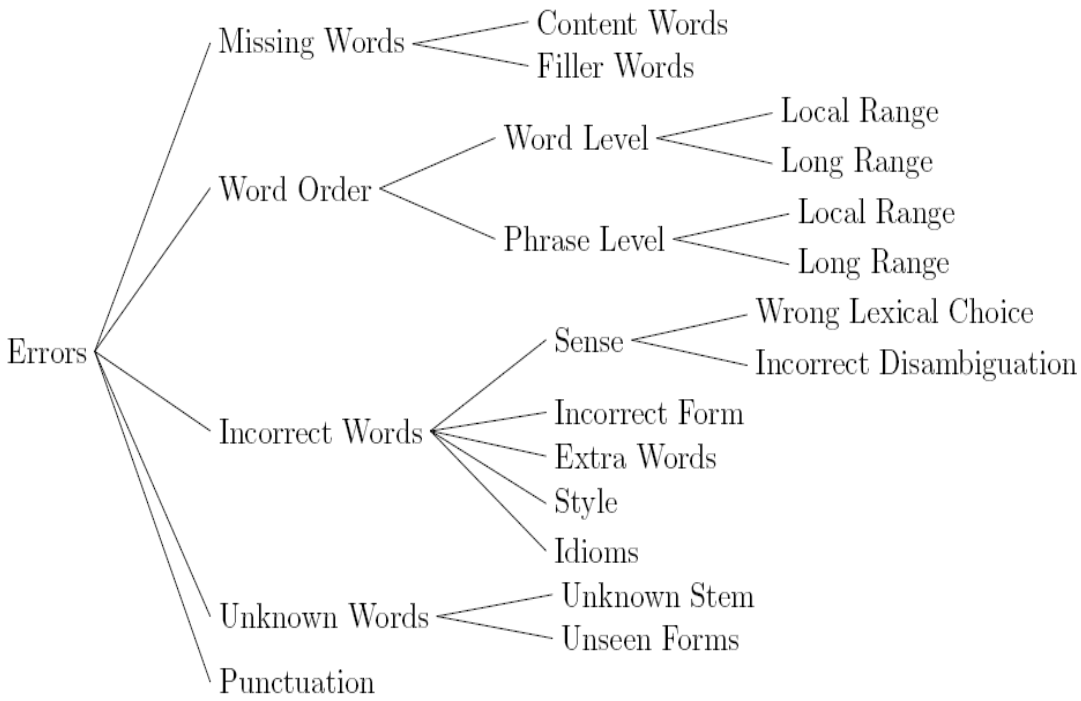


Figure 1: Classification of translation errors.

Automatic Scoring of MT Quality

UNIVERSITÄT DES SAARLANDES

Main Idea:

Given a “good” (reference) translation, quality of machine translation output boils down to the question of similarity

This is a monolingual problem, may be easier than the original question

Textual similarity may be measured automatically

Various simple error metrics have been successfully used in speech recognition (Word error rate, …)

Evaluation for SMT development

UNIVERSITÄT DES SAARLANDES

Development cycle of an SMT system [Och 2000]

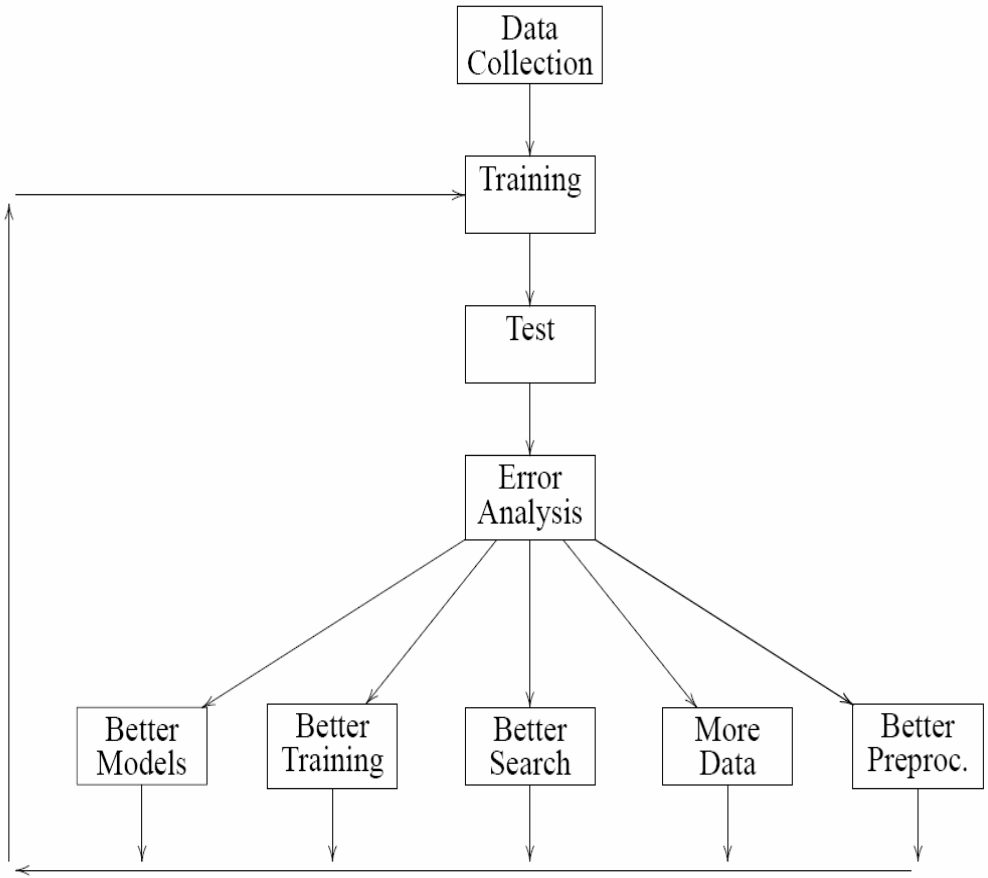


Figure 3.1: Development cycle of a statistical MT system.

The BLEU score

BLEU = Bilingual Evaluation Understudy

Goals:

- Measure the similarity of an MT result with reference translation(s)
- Can deal with multiple reference translations
- Take word order into account (more informed than position-independent word error rate)
- Allow for major reordering (less strict than word error rate/ Levenshtein distance)

Main ideas:

Combine n-gram **precision** for multiple n (typically 1..4)

Approximate **recall** via so-called **brevity penalty**

BLEU score

UNIVERSITÄT DES SAARLANDES

See http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf for details, the main formulas are as follows:

We first compute the geometric average of the modified n-gram precisions, p_n, using n-grams up to length N and positive weights w_n summing to one.

Next, let c be the length of the candidate translation and r be the effective reference corpus length. We compute the brevity penalty BP,

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

The ranking behavior is more immediately apparent in the log domain,

$$\log \textsc{Bleu} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

In our baseline, we use $N = 4$ and uniform weights $w_n = 1/N$.

See http://www.statmt.org/wmt06/shared-task/multi-bleu.perl for a practical implementation.

Why BLEU is popular

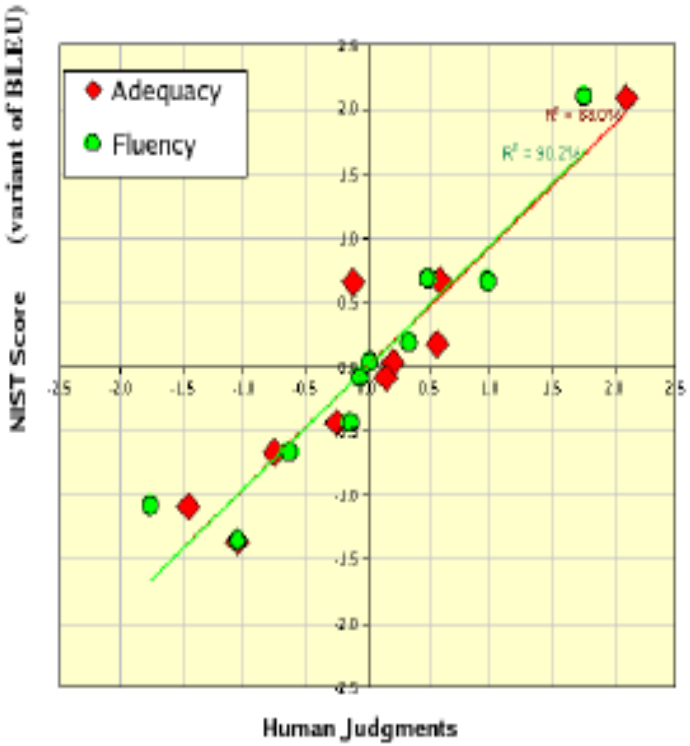


Figure 8.8: Correlation between an automatic metric (here: NIST) and human judgment (fluency, adequacy). Illustration by George Doddington.

From http://cio.nist.gov/esd/emaildir/lists/mt_list/msg00065.html

Why BLEU is controversial

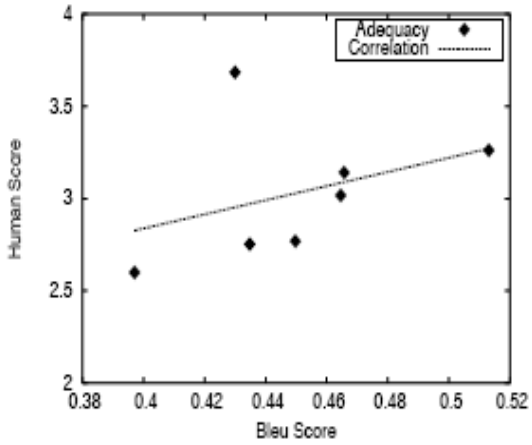


Figure 2: Bleu scores plotted against human judgments of adequacy, with $R^2 = 0.14$ when the outlier entry is included

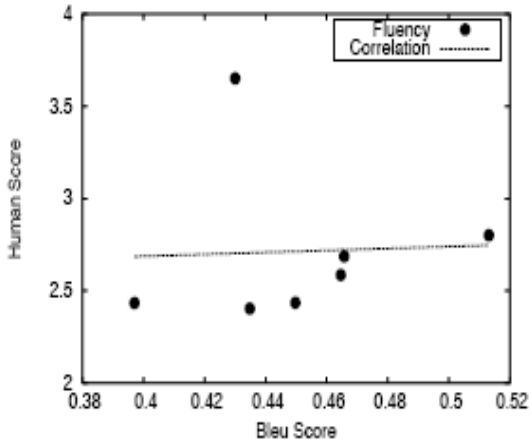


Figure 3: Bleu scores plotted against human judgments of fluency, with $R^2 = 0.002$ when the outlier entry is included

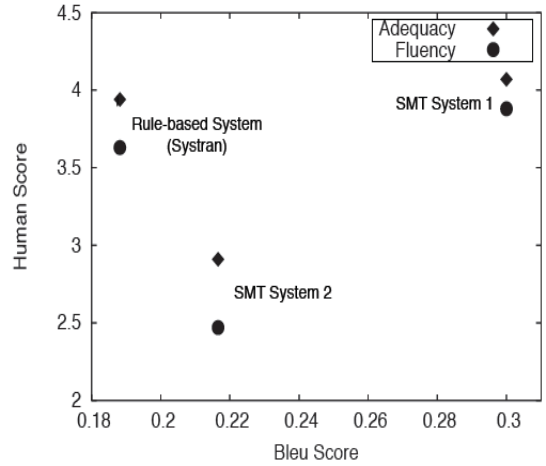


Figure 4: Bleu scores plotted against human judgments of fluency and adequacy, showing that Bleu vastly underestimates the quality of a non-statistical system

From: Re-evaluating the Role of BLEU in Machine Translation Research, Chris Callison-Burch, Miles Osborne, Philipp Koehn, EACL 2006
http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/bleu2006.pdf

BLEU Score Example

Three machine translation systems generate the following texts:

SYS1: She cannot be used as a basis for the installation of a European constitution .
SYS2: It cannot a basis for the establishment of a European constitution .
SYS3: It can form the basis for a European constitution .

Assume that for automatic evaluation we also have access to the following two reference translations:

REF1: It cannot serve as a basis for the establishment of a European constitution .
REF2: It can not serve as a basis for the introduction of a European constitution .

Sketch how the BLEU-4 score for the given translation candidates will be computed. What are the 1- ... 4-gram accuracies that will enter into the computation? Insert appropriate numbers into the slots in the following lines. You do not need to compute the brevity penalty for this exercise.

	1-grams	2-grams	3-grams	4-grams
SYS1:	__/15	__/14	__/13	__/12
SYS2:	__/12	__/11	__/10	__/9
SYS3:	__/10	__/9	__/8	__/7

MT Meta Evaluation

UNIVERSITÄT DES SAARLANDES

Problem: High-quality automatic MT evaluation would help tremendously, but is currently out of reach

Idea: Push development of better MT evaluation methods via competitive meta-evaluations

- Use MT + human evaluation results as training + test data
- Measure agreement (correlation) of automatic scores with human judgements

MT Meta Evaluation

UNIVERSITÄT DES SAARLANDES

Examples from recent evaluation campaign (WMT08)

	RANK	CONST	YES/NO	OVERALL
meteor-ranking	**.81**	.72	.77	**.76**
ULCh	.68	.79	.82	**.76**
meteor-baseline	.77	.75	.74	.75
posbleu	.77	**.8**	.66	.74
pos4gramFmeasure	.75	.62	.82	.73
ULC	.66	.67	**.84**	.72
DR	.79	.55	.76	.70
SR	.79	.53	.76	.69
DP	.57	.79	.65	.67
mbleu	.61	.77	.56	.65
mter	.47	.72	.68	.62
bleu	.61	.59	.44	.54
svm-rank	.21	.24	.35	.27

Table 8: Average system-level correlations for the automatic evaluation metrics on translations into English

	RANK	CONST	YES/NO
DP	.514	.527	.536
DR	.500	.511	.530
SR	.498	.489	.511
ULC	.559	**.554**	**.561**
ULCh	**.562**	.542	.542
alignment-prob	.517	.538	.535
mbleu	.505	.516	.544
meteor-baseline	.512	.520	.542
meteor-ranking	.512	.517	.539
mter	.436	.471	.480
pos4gramFmeasure	.495	.517	.52
posbleu	.435	.43	.454
svm-human-ref	.542	.541	.552
svm-pseudo-ref	.538	.538	.543
svm-rank	.493	.499	.497

Table 10: The percent of time that each automatic metric was consistent with human judgments for translations into English