Machine Translation

June 5th, 2012

Sabine Hunsicker
DFKI GmbH
sabine.hunsicker@dfki.de

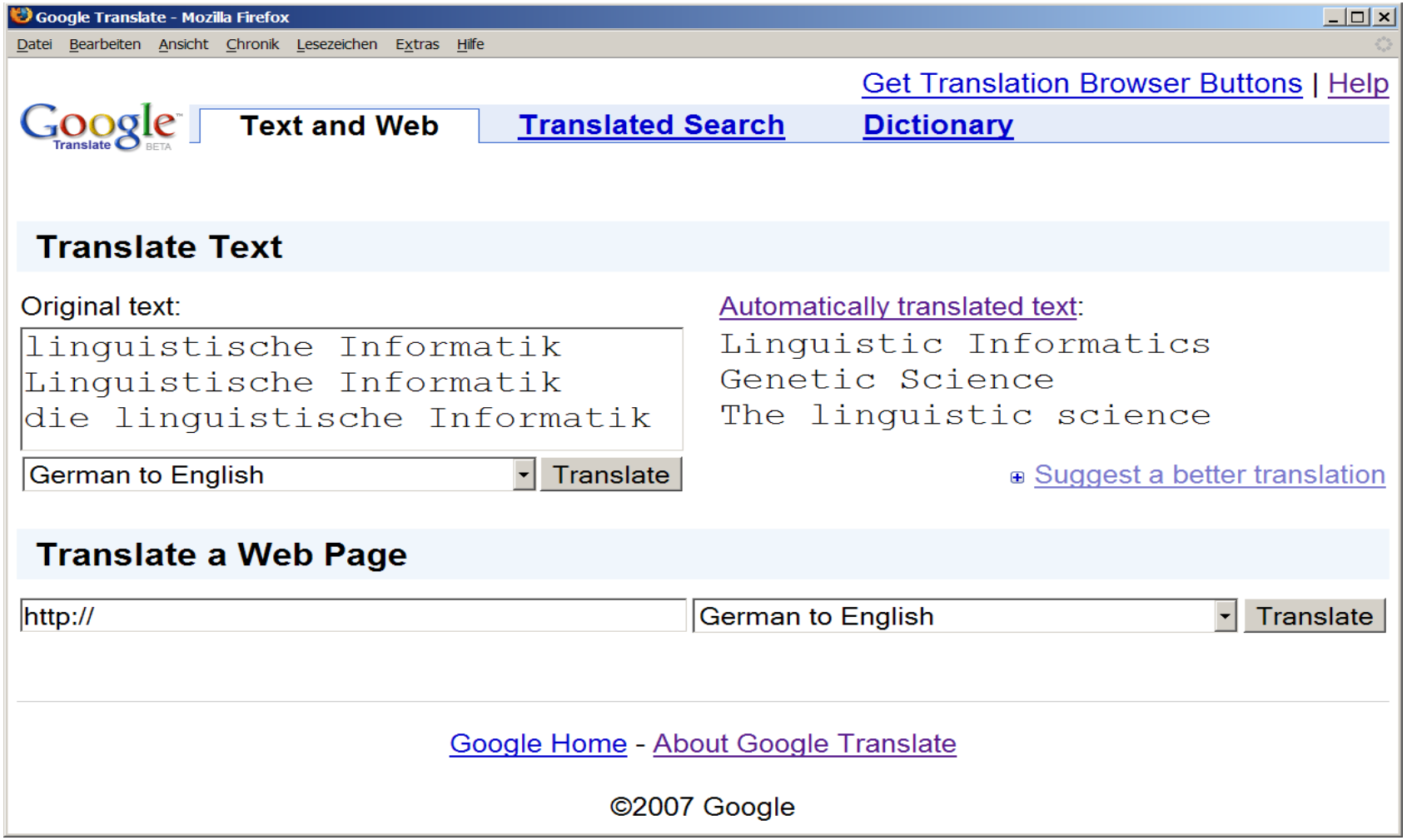
Language Technology II
SS 2012

Problems of SMT

UNIVERSITÄT DES SAARLANDES

- Factored and tree-based models can fix some of the problems of phrase-based SMT.
- But they can't fix them *reliably*:
  - We cannot ensure that a certain linguistic phenomenon is always translated in the same way.
- SMT translations cannot be predicted.
- We want to prevent errors, but how to enforce this?
  - Rules?

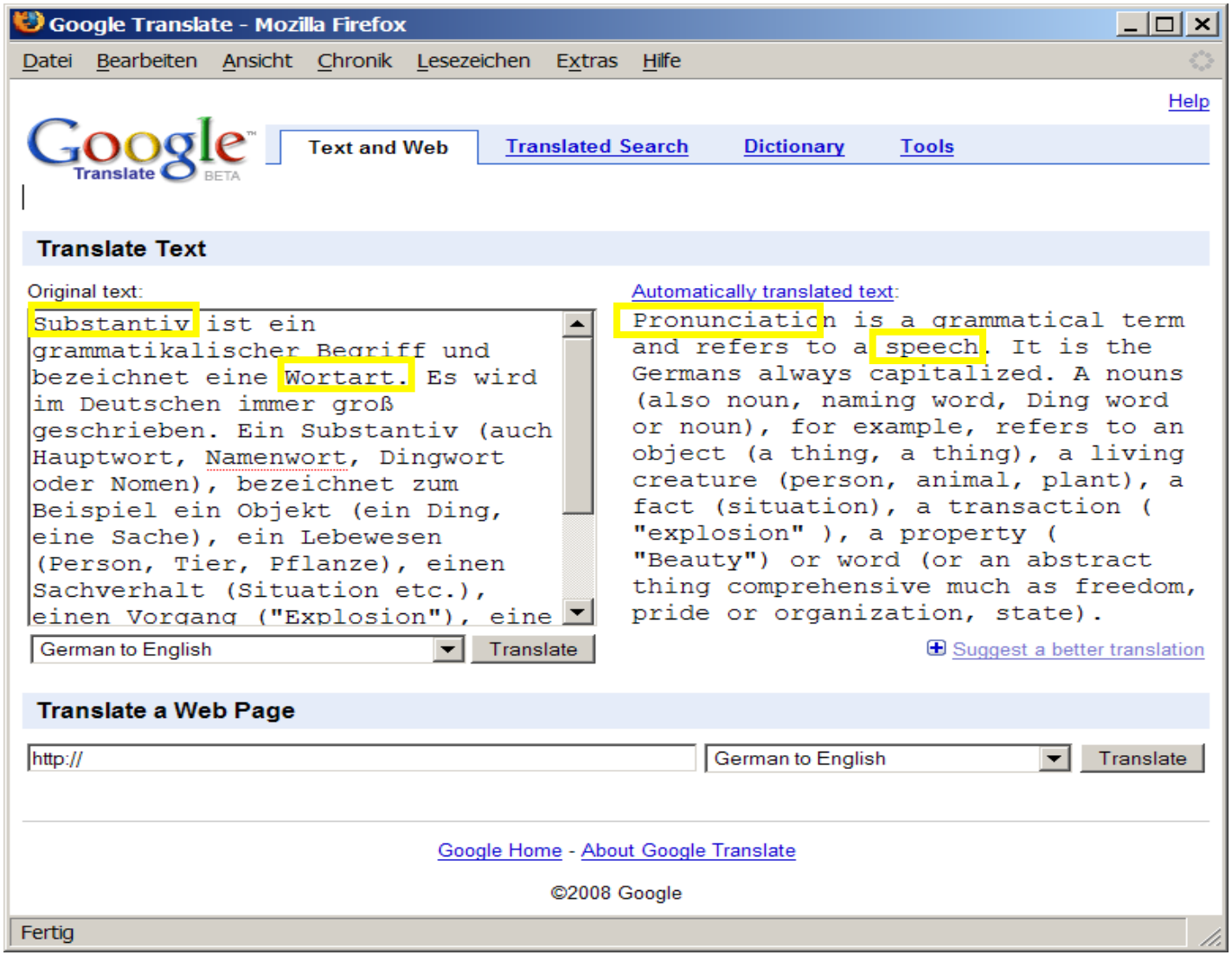
Problems with Lexical Reliability



[November 2007, corrected in the meantime]

More Examples of Reliability Problems

UNIVERSITÄT DES SAARLANDES



[January 2008, partly corrected in the meantime]

Problems of RBMT

UNIVERSITÄT DES SAARLANDES

- RBMT translations are predictable and reliable.
- Also the errors are: if a rule covering a linguistic phenomenon is missing, the system will always translate it incorrectly.
  - But rule base is difficult to adapt or extend.
- RBMT also gets many of the things SMT gets wrong, right.
- Do they make different mistakes?

Let's Compare …

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 &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.*

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	+	-

Knowledge Required for Translation

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.
- We need to combine both approaches.

Toward Hybrid Systems

UNIVERSITÄT DES SAARLANDES

- Both paradigms have different requirements:
  - RBMT requires a rule base and a lexicon to exist
  - SMT needs data
- We would prefer a deep integration, e.g. an analysis phase that uses both a rule-based grammar and a statistical parser.
- Research on deep integration of statistical and linguistic approaches is on-going.
- Let's focus on shallow approaches first.

Methods of Combining - Coupling

UNIVERSITÄT DES SAARLANDES

- Serial Coupling:
  - SMT + RBMT: Syntactic Selection
  - RBMT + SMT: Statistical Post-Editing
- Parallel Coupling:
  - MT_1, …, MT_n → select best output
  - Works on full sentences or smaller segments

Methods of Combining - Extensions

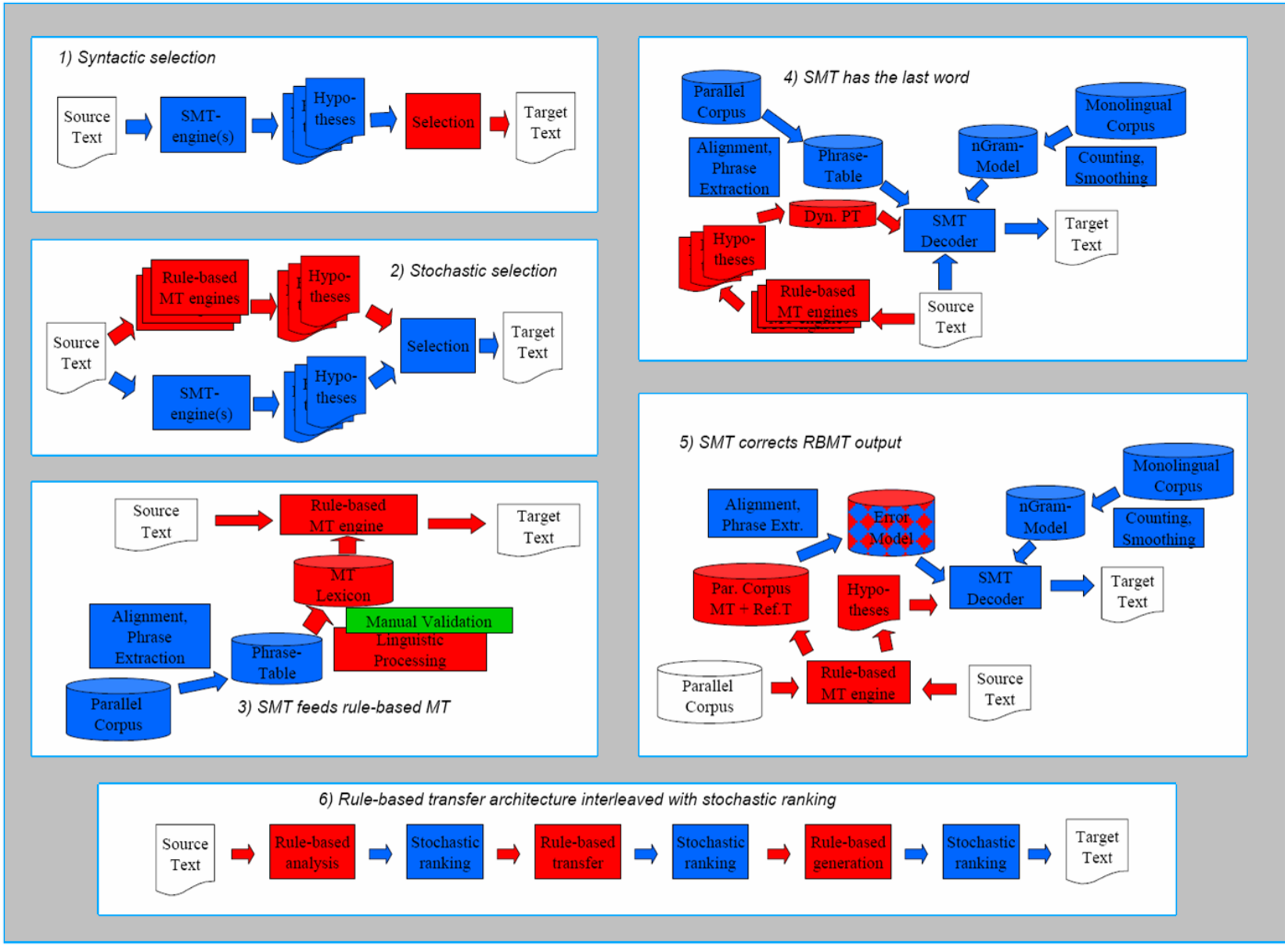
UNIVERSITÄT DES SAARLANDES

- Extensions to RBMT
  - Pre-Editing: learning new lexicon entries or new rules
  - Core Extensions: adapt rule-based components such as transfer to be able to process probability information learned from a corpus

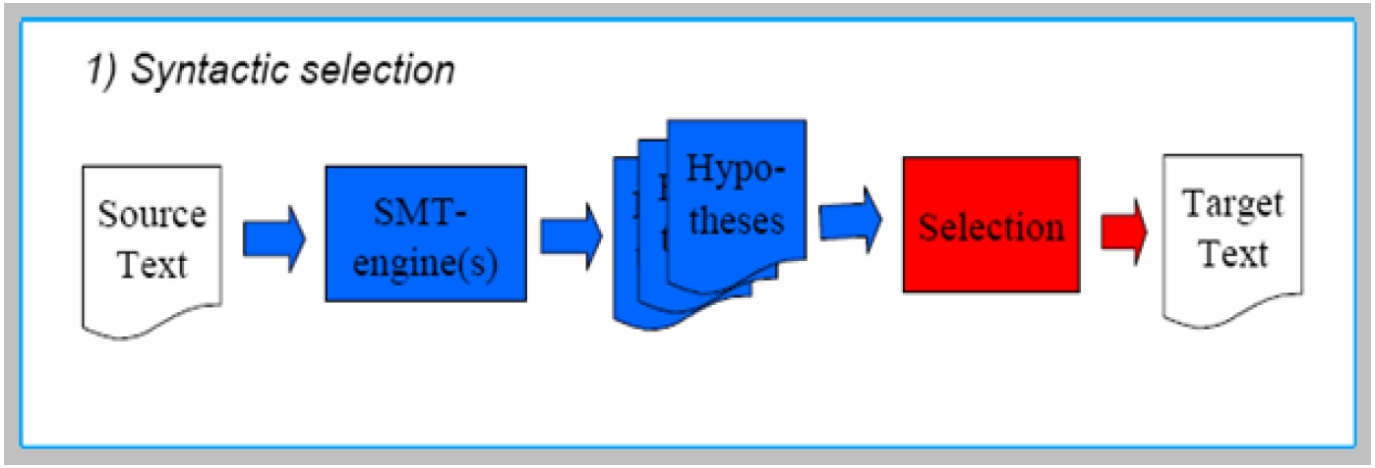
- Extensions to SMT
  - Pre-Editing: lemmatise corpus (cf. factored models); compound splitting; reordering
  - Core Extensions: import RBMT resources into the phrasetable; improving decoding using target grammars

Hybrid MT Architectures

= SMT Module
= RBMT Module



Syntactic Selection



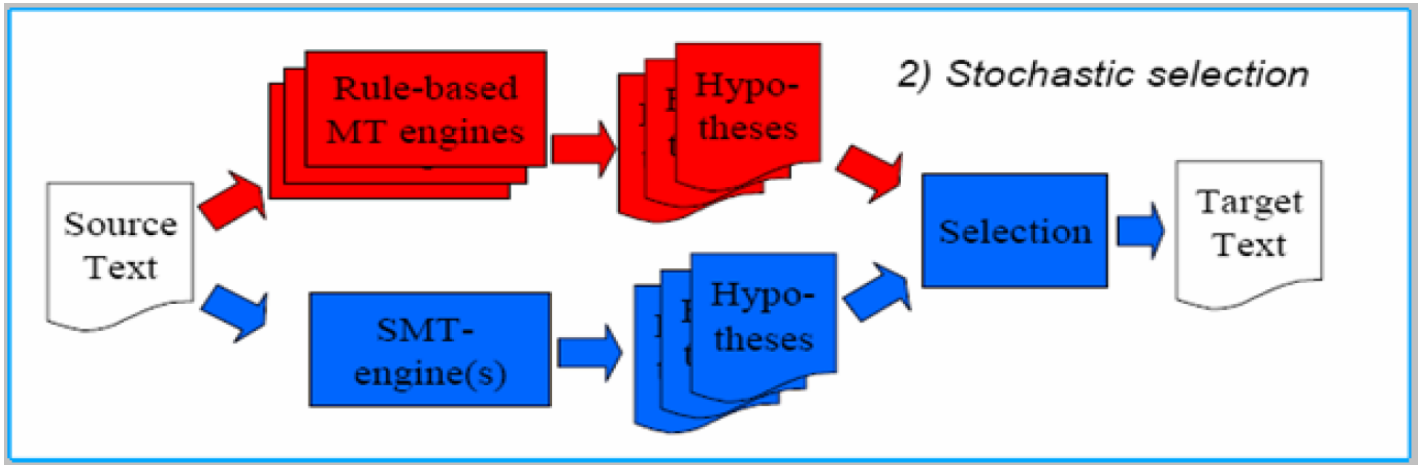
Motivation: SMT output is often syntactically ill-formed

➔ 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:

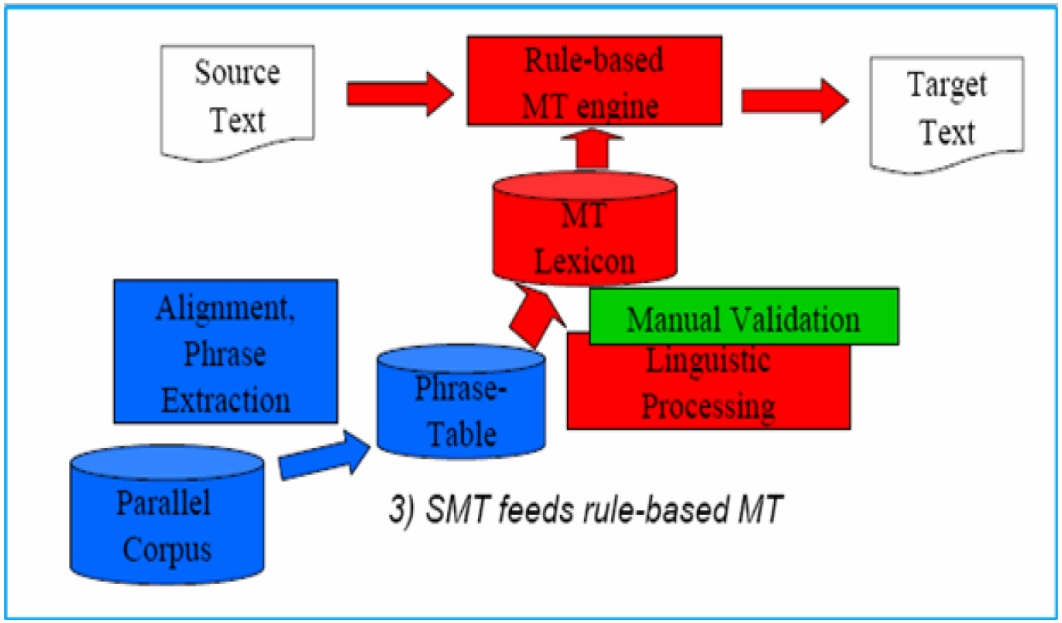
- Works mainly for short utterances, where one of the candidates may be good enough (VerbMobil)
- Different candidates may have problems in different parts of the sentence, granularity of decisions too coarse

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):

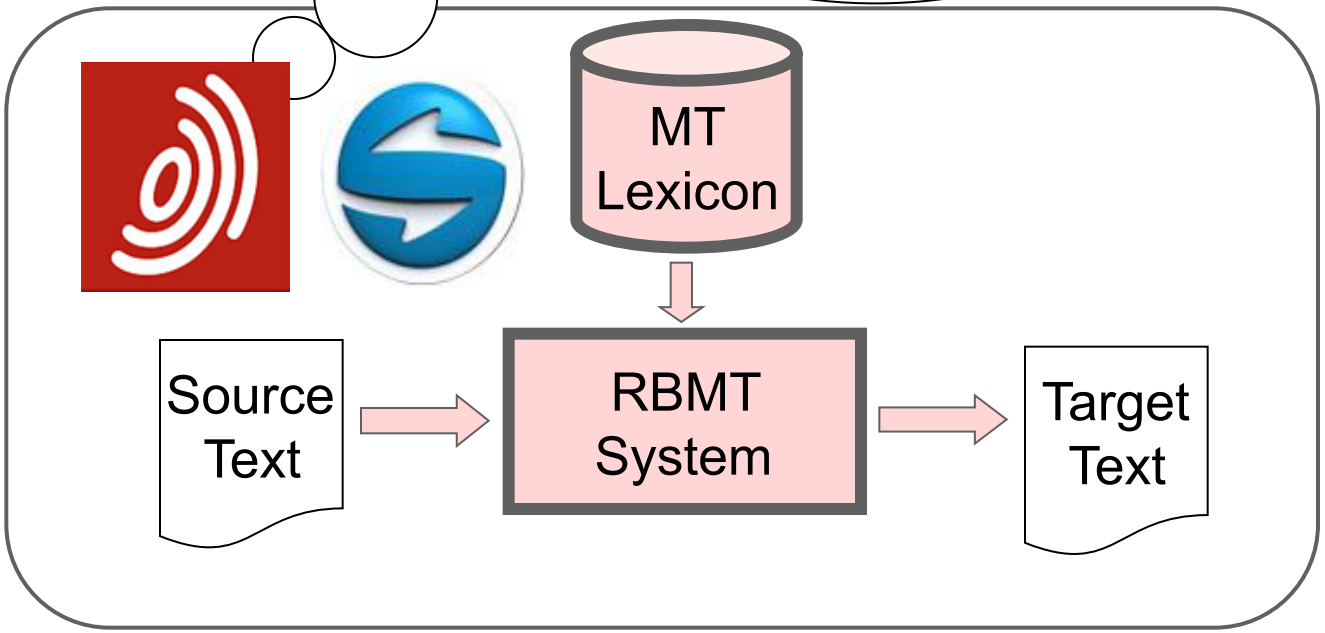
6000 employees from > 30 countries in Munich, The Hague, Berlin, Vienna, Brussels

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***

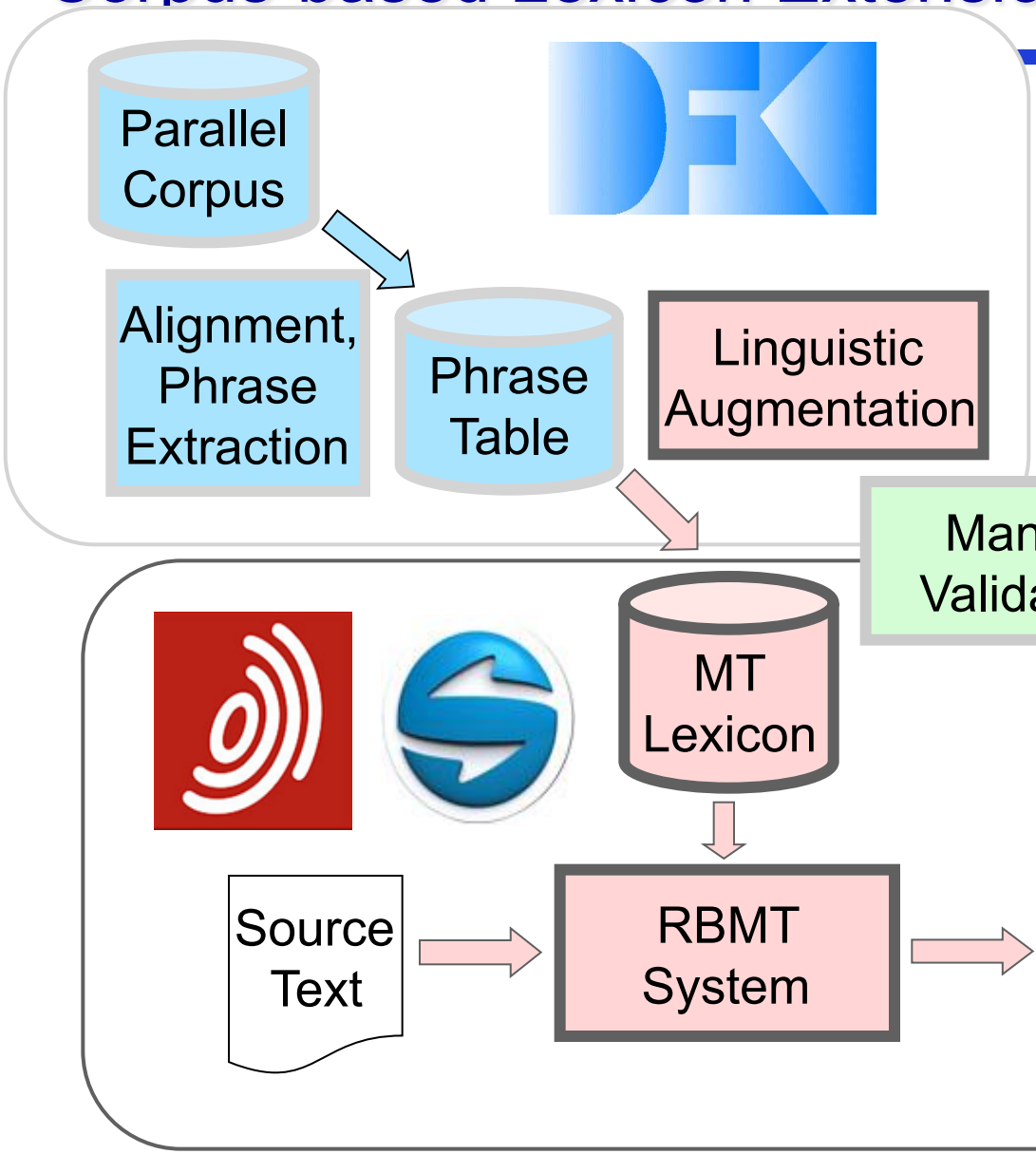
NL ↔ ***EN***

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

Problems with Using SMT

UNIVERSITÄT DES SAARLANDES

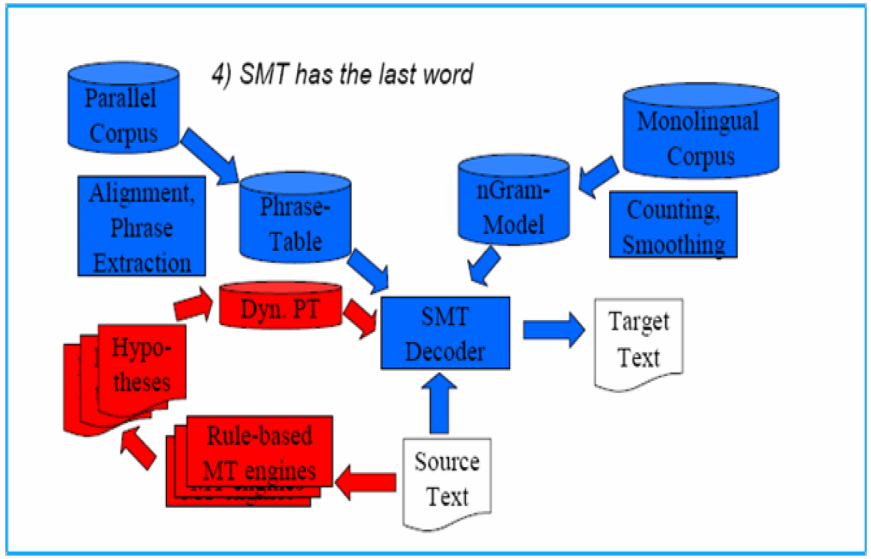
- The phrasetable does not contain only phrases in the linguistic sense.
- But adding malformed lexicon entries will hurt the translation quality of the rule-based sentence.
- We need to invest effort into making sure that the SMT data is well-formed.
- But manual validation is expensive.
- What other resources could we use?

Introducing TermEx/LiSTEX

UNIVERSITÄT DES SAARLANDES

- In EuroMatrixPlus we developed a term extraction tool which can be used to extend the coverage of an RBMT system.
- This tool creates term lists in a format that can be used by the Lucy RBMT system for importing terms.
- But: TermEx doesn't use the phrasetable, instead it uses the analysis trees from the RBMT system.
  - We extract proper linguistic phrases from the trees on both sides.

RBMT feeds SMT

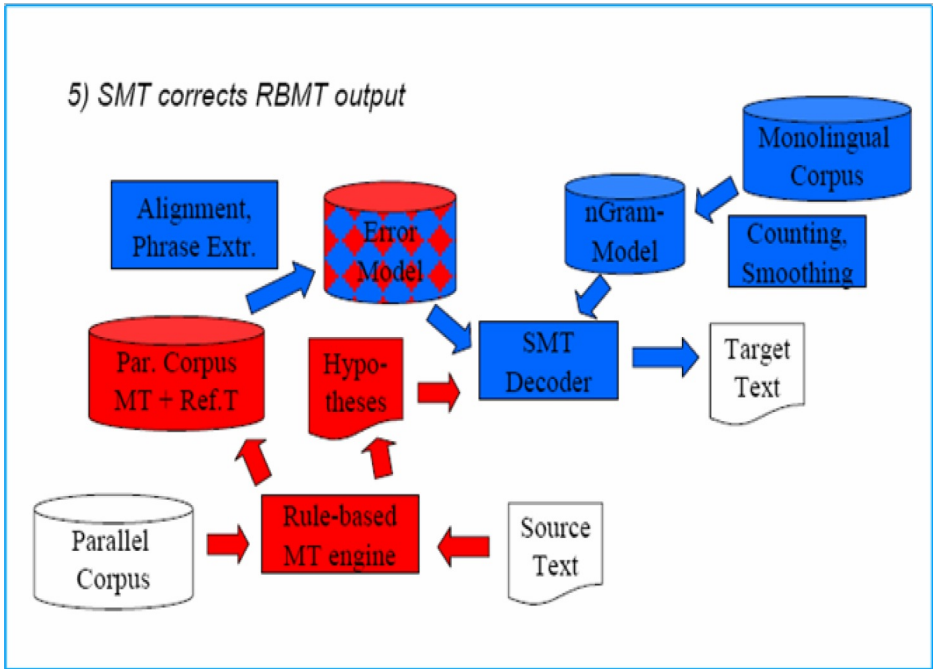


Motivation: SMT can only know what is in the training data, RBMT systems often contain extensive lexical knowledge

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

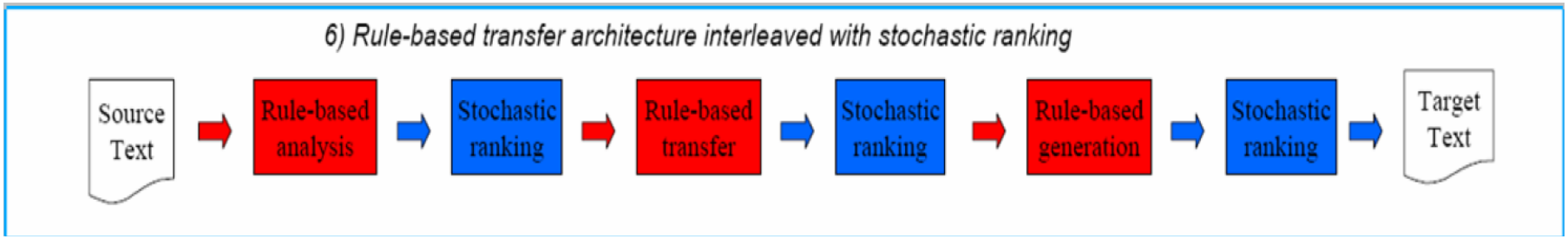
BUT: Sometimes RBMT messes a sentence completely up, no hope to repair these cases via SMT

Parse Errors

UNIVERSITÄT DES SAARLANDES

- Sometimes the grammar puts out an incorrect analysis:
  - I wish the negotiators continued success with their work in this important area
  - Ich wünsche, dass die Unterhändler Erfolg mit ihrer Arbeit in diesem wichtigen Bereich fortsetzten
- To fix these errors, we need to go back to the source and re-analyse (either using an SMT fallback or choosing a different RBMT analysis).
- But how to recognise parse errors, if they lead to grammatical output?

Transfer architecture with stochastic ranking



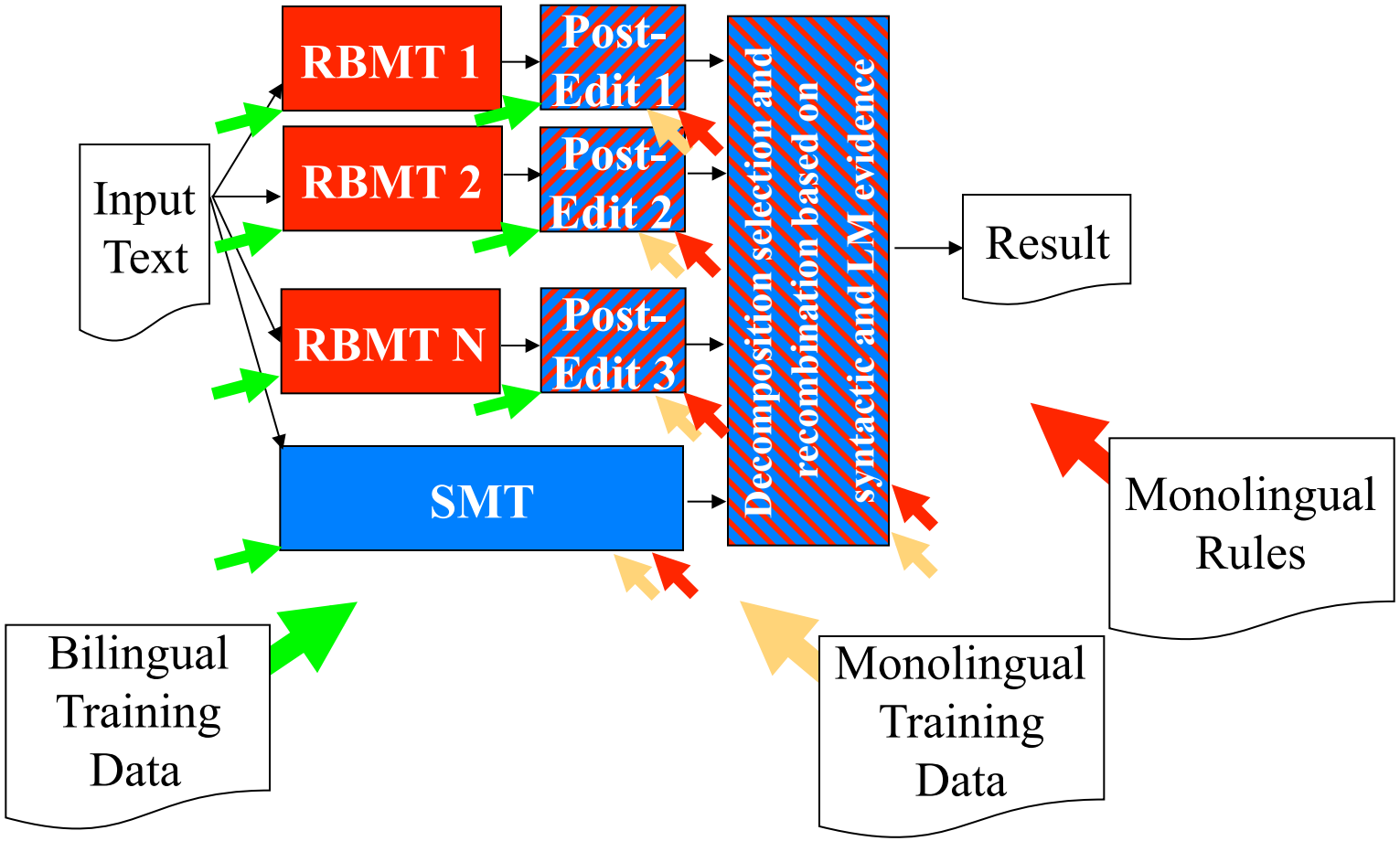
Motivation: Fine-grained combination of statistical and linguistic evidence on all levels requires a closely coupled implementation

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-Processing

UNIVERSITÄT DES SAARLANDES

- So far, we send the input text to the MT system without any modifications.
- Afterward we need to make sense of (partially erroneous) output after errors have been made.
- But, e.g. for the RBMT systems, we know what kind of errors they make.
- Can we simplify the input to reduce the risk of errors?

Pre-Processing II

UNIVERSITÄT DES SAARLANDES

- Statistics of error types can be used to find out specific weaknesses and best way to distribute work over engines.

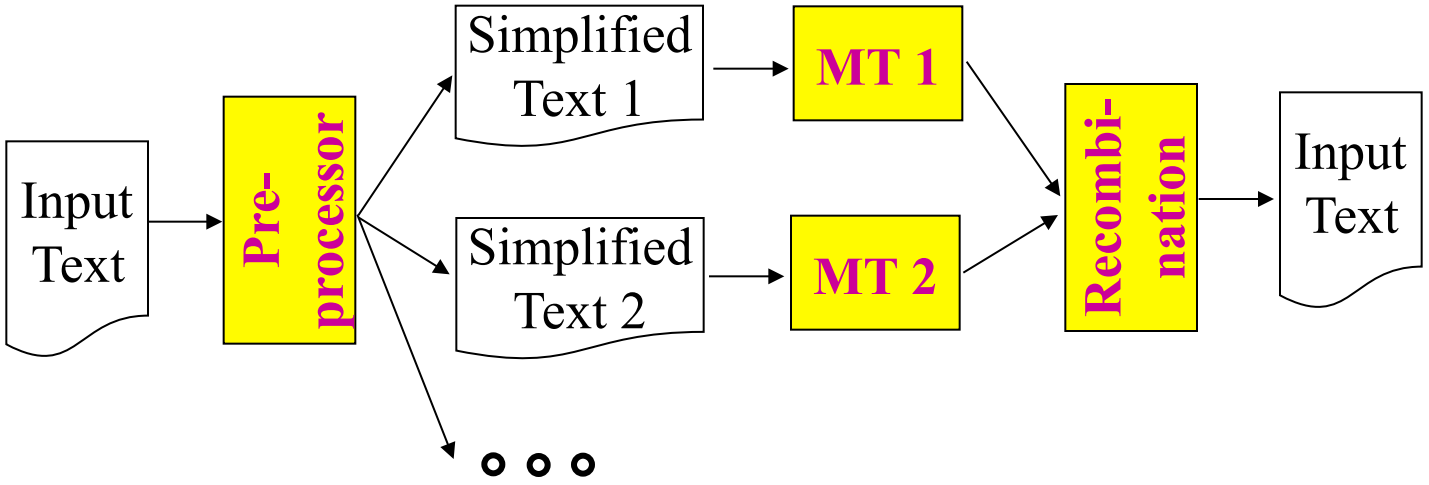
- Slight modifications of the input can prevent errors from happening, e.g. by
  - replacing named entities unknown to the engine by place-holders
  - simplifying technical noun-phrases
  - treating special cases (numbers, names) in special ways

Tools Used in Pre-Processing

UNIVERSITÄT DES SAARLANDES

- We can integrate external terminology databases to ensure lexical reliability & equivalence.
  - We can use XML mark-up to force a particular translation option to be used.
- We can use tools from both paradigms to annotate the input text with additional information.
- We can create different simplified texts and merge the translations.

Pre-emptive division of labor



- Simplified form: 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)

Hybrid Systems - Outlook

UNIVERSITÄT DES SAARLANDES

- To get qualitative good translations, we need both world knowledge (SMT) and linguistic expertise (RBMT).
- There are different ways to combine MT systems.
- Deep integration is most promising, but it's also very difficult to integrate both paradigms.
- We can pre-process texts to prevent (known) error types.
- Texts can be written in a way that they avoid linguistic phenomena which have proven to be difficult (*controlled language*).