Machine Translation

January 14, 2008

Andreas Eisele
UdS Computerlinguistik & DFKI
eisele@dfki.de

Language Technology I
WS 2007/8

Machine Translation: Overview

UNIVERSITÄT DES SAARLANDES

- Relevance of MT, typical applications and requirements
- History of MT
- Basic approaches to MT
  - Rule/grammar based
  - Statistical
  - Example-based
  - Hybrid, multi-engine
- Evaluation techniques

Sources for Information

UNIVERSITÄT DES SAARLANDES

- MT in general, history:
  - http://www.MT-Archive.info: Electronic repository and bibliography of articles, books and papers on topics in machine translation and computer-based translation tools, regularly updated, contains over 3300 items
  - Hutchins, Somers: An introduction to machine translation. Academic Press, 1992, available under http://www.hutchinsweb.me.uk/IntroMT-TOC.htm
- MT systems:

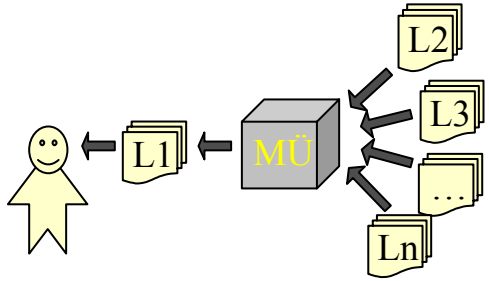
  Compendium of Translation Software, see http://www.hutchinsweb.me.uk/Compendium.htm
- Statistical Machine Translation:

  See www.statmt.org

Use cases and requirements for MT

UNIVERSITÄT DES SAARLANDES

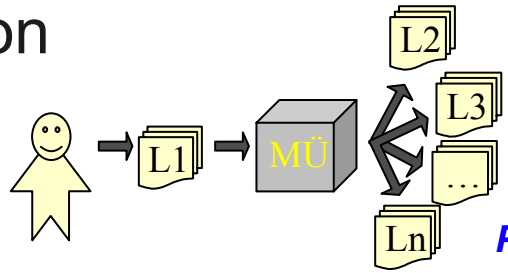
a) MT for assimilation



Robustness
Coverage

Daily throughput of online-MT-Systems > 500 M Words

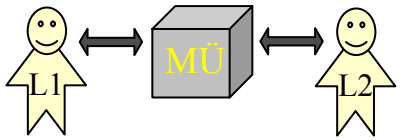
b) MT for dissemination



Textual quality

Publishable quality can only be authored by humans; Translation Memories & CAT-Tools mandatory for professional translators

c) MT for direct communication



Speech recognition, context dependence

Topic of many running and completed research projects (VerbMobil, TC Star, TransTac, …)
US-Military prepares deployment of systems for spoken MT

History of Machine Translation

UNIVERSITÄT DES SAARLANDES

- Slides by John Hutchins:

  http://www.hutchinsweb.me.uk/SUSU-2007-1-ppt.pdf

Existing MT systems for EU languages

UNIVERSITÄT DES SAARLANDES

Situation in early 2005, almost all systems are rule-based

From Hutchins: Compendium of Translation Software, 2005

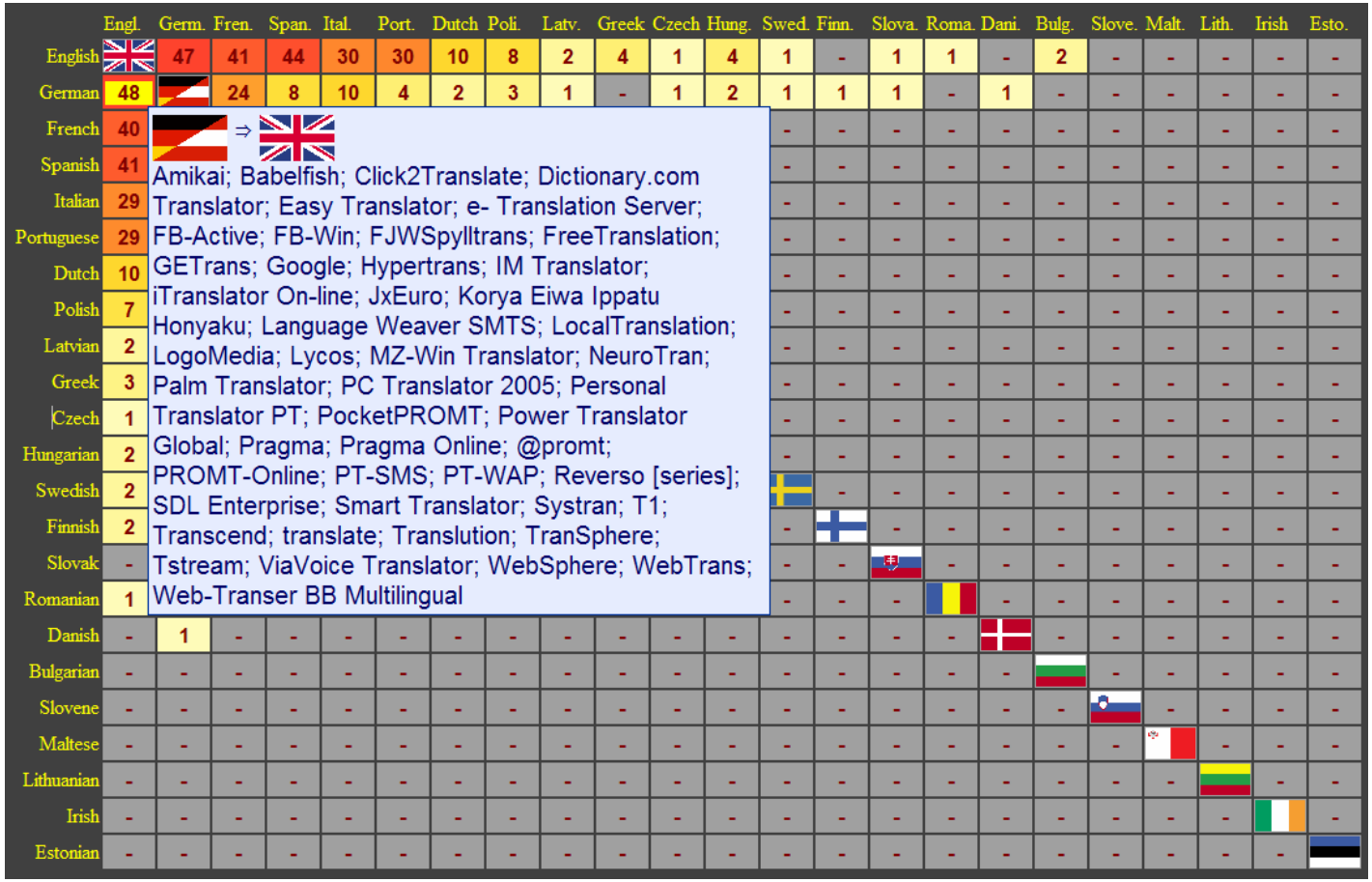
	Engl.	Germ.	Fren.	Span.	Ital.	Port.	Dutch	Poli.	Latv.	Greek	Czech	Hung.	Swed.	Finn.	Slova.	Roma.	Dani.	Bulg.	Slove.	Malt.	Lith.	Irish	Esto.
English		47	41	44	30	30	10	8	2	4	1	4	1	-	1	1	-	2	-	-	-	-	-
German	48		24	8	10	4	2	3	1	-	1	2	1	1	1	-	1	-	-	-	-	-	-
French	40	23		11	13	8	4	1	1	3	1	-	-	-	-	-	-	-	-	-	-	-	-
Spanish	41	7	11		9	8	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Italian	29	10	13	9		4	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Portuguese	29	5	7	8	4		1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dutch	10	2	4	1	1	1		-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Polish	7	2	1	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Latvian	2	1	1	1	1	1	1	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-
Greek	3	-	3	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-
Czech	1	1	1	-	1	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-
Hungarian	2	2	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-
Swedish	2	1	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-
Finnish	2	1	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-
Slovak	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
Romanian	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-
Danish	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Existing MT systems for EU languages

UNIVERSITÄT DES SAARLANDES

Situation in early 2005, almost all systems are rule-based

From Hutchins: Compendium of Translation Software, 2005



	Engl.	Germ.	Fren.	Span.	Ital.	Port.	Dutch	Poli.	Latv.	Greek	Czech	Hung.	Swed.	Finn.	Slova.	Roma.	Dani.	Bulg.	Slove.	Malt.	Lith.	Irish	Esto.
English		47	41	44	30	30	10	8	2	4	1	4	1	-	1	1	-	2	-	-	-	-	-
German	48		24	8	10	4	2	3	1	-	1	2	1	1	1	-	1	-	-	-	-	-	-
French	40												-	-	-	-	-	-	-	-	-	-	-
Spanish	41												-	-	-	-	-	-	-	-	-	-	-
Italian	29												-	-	-	-	-	-	-	-	-	-	-
Portuguese	29												-	-	-	-	-	-	-	-	-	-	-
Dutch	10												-	-	-	-	-	-	-	-	-	-	-
Polish	7												-	-	-	-	-	-	-	-	-	-	-
Latvian	2												-	-	-	-	-	-	-	-	-	-	-
Greek	3												-	-	-	-	-	-	-	-	-	-	-
Czech	1												-	-	-	-	-	-	-	-	-	-	-
Hungarian	2												-	-	-	-	-	-	-	-	-	-	-
Swedish	2													-	-	-	-	-	-	-	-	-	-
Finnish	2												-		-	-	-	-	-	-	-	-	-
Slovak	-												-	-		-	-	-	-	-	-	-	-
Romanian	1												-	-	-		-	-	-	-	-	-	-
Danish	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Motivation for rule-based MT

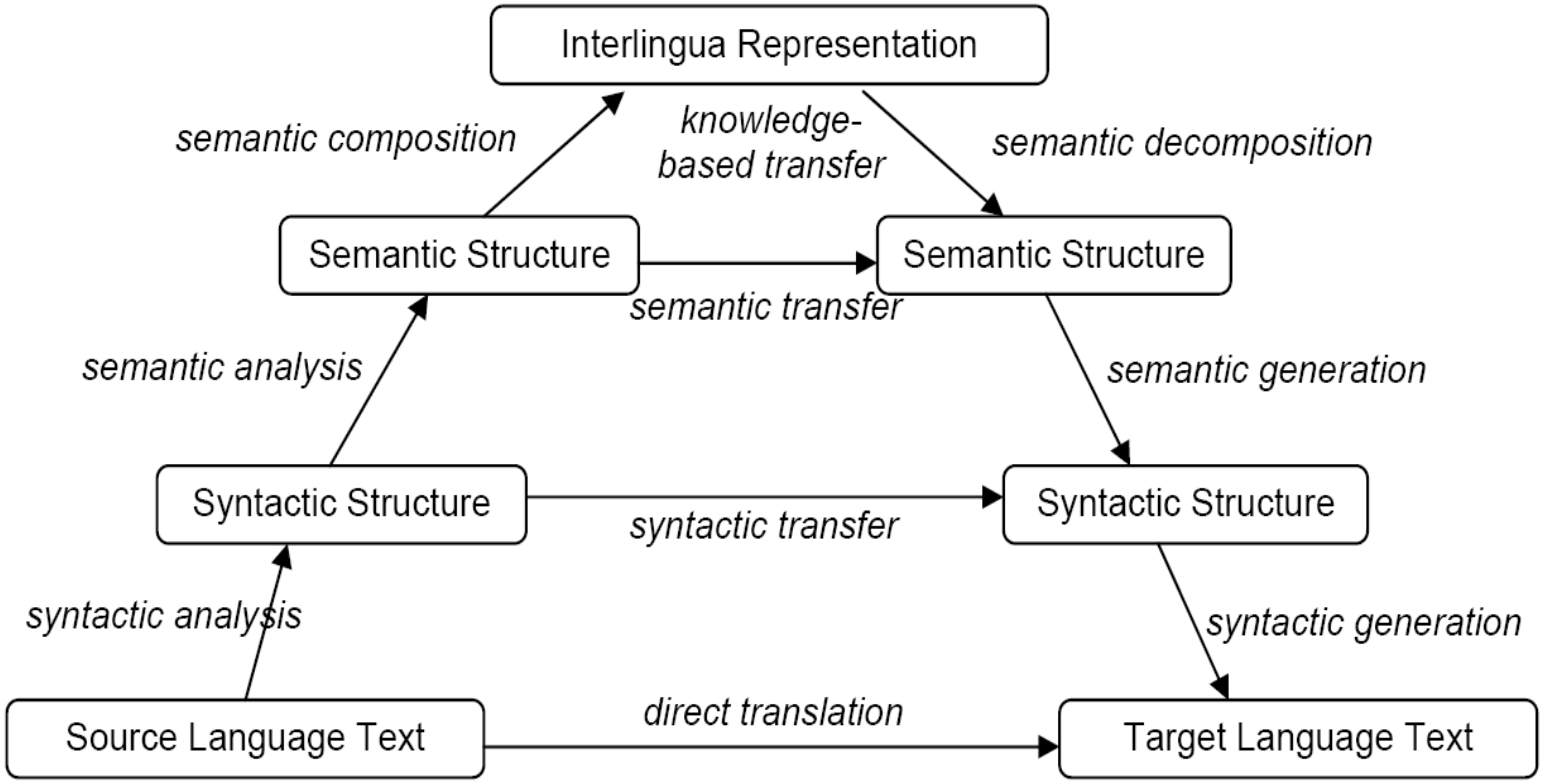
UNIVERSITÄT DES SAARLANDES

- Good translation requires knowledge of linguistic rules
  - …for understandig the source text
  - …for generating well-formed target text
- Rule-based accounts for certain linguistic levels exist and should be used, especially for
  - Morphology
  - Syntax
- Writing one rule is better than finding hundreds of examples, as the rule will apply for new, unseen cases

Possible (rule-base) MT architectures

UNIVERSITÄT DES SAARLANDES

The „Vauquois Triangle“



Motivation for statistical MT

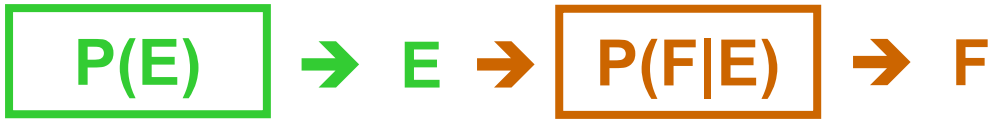
UNIVERSITÄT DES SAARLANDES

- Good translation requires knowledge and decisions on many levels
  - syntactic disambiguation (POS, attachments)
  - semantic disambiguation (collocations, scope, word sense)
  - reference resolution
  - lexical choice in target language
  - application-specific terminology, register, connotations, good style …
- Rule-based models of all these levels are very expensive to build, maintain, and adapt to new domains
- Statistical approaches have been quite successful in many areas of NLP, once data has been annotated
- Learning from existing translation will focus on distinctions that matter (not on the linguist's favorite subject)
- Translation corpora are available in rapidly growing amounts
- SMT *can* integrate rule-based modules (morphologies, lexicons)
- SMT *can* use feed-back for on-line adaptation to domain and user preferences

Statistical Machine Translation

UNIVERSITÄT DES SAARLANDES

- Based on *„distorted channel“* Paradigm (successful for pattern- and speech recognition)



- Decoding: Given observation **F**, find most likely cause **E***

$$E^* = \text{argmax}_E\ P(E|F) = \text{argmax}_E\ P(E,F) = \text{argmax}_E\ P(E) * P(F|E)$$

➔ Three subproblems
Model of **P(E)**
Model of **P(F|E)**
Search for **E***

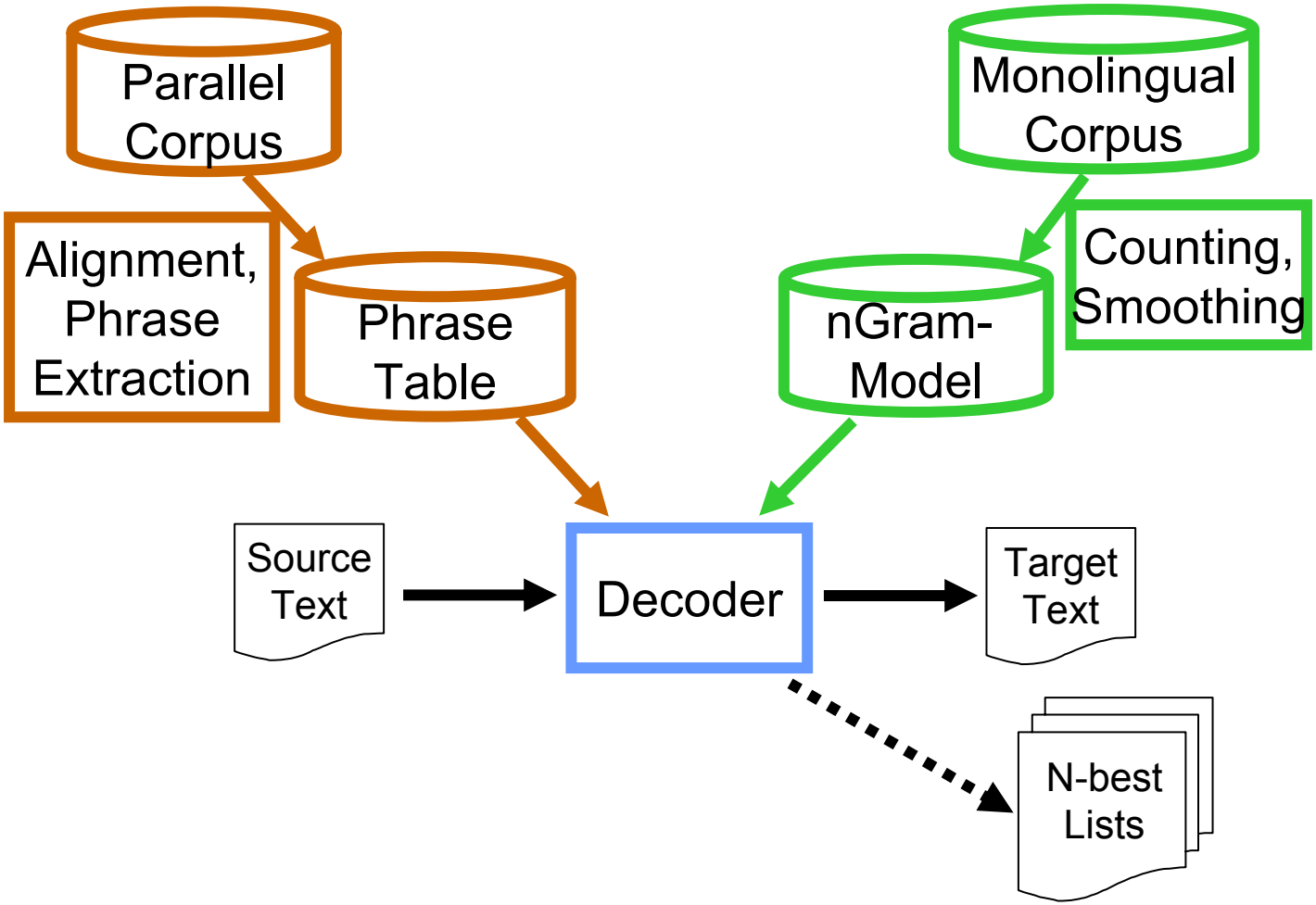
each has approximative solutions
nGram-Models $P(e_1 \ldots e_n) = \Pi P(e_i|e_{i-2}\ e_{i-1})$
Transfer of „phrases“ $P(F|E) = \Pi P(\mathbf{f}_i|\mathbf{e}_i)*P(d_i)$
Heuristic *(beam)* search

- Models are trained with (parallel) corpora, correspondences (alignments) between languages are estimated via EM-Algorithm (GIZA++, F.J.Och)

Statistical Machine Translation

UNIVERSITÄT DES SAARLANDES

schematic architecture



Rule-based vs. statistical translation (1)

UNIVERSITÄT DES SAARLANDES

(RBMÜ:translate pro ⟷ SMÜ:Koehn 2005, examples from EuroParl)

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

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

continued: Verb instead of adjective

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

three wrong inflectional endings

Rule-based vs. statistical translation (2)

UNIVERSITÄT DES SAARLANDES

Englisch	RMBÜ: translate pro	SMÜ: 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.*

Motivation for hybrid MT (1)

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	++	--
Structural Semantics	+	--
Lexical Semantics	-	+
Lexical Adaptivity	--	+

Motivation for hybrid MT (2)

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

Exercises

UNIVERSITÄT DES SAARLANDES

- Investigate two on-line MT engines by sending sentences through both of them and comparing the results

  Use http://www.google.com/language_tools as a typical SMT system and compare with http://babelfish.altavista.com/ (based on Systran)
- Try to find typical errors of both systems. Try variations of input sentences to find out whether the systems „understand“ what they translate