Machine Translation

February 13, 2008

Andreas Eisele
UdS Computerlinguistik & DFKI
eisele@dfki.de

Foundations of Language Science and Technology
WS 2007/8

Machine Translation: Overview

UNIVERSITÄT DES SAARLANDES

LT1: Motivation and overview of MT paradigms, including rule-based, statistical, and hybrid techniques

- 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

FLST: Focus on translation task (linguistic issues), including some algorithmic aspects

- Differences between languages
- Typical difficulties in translation
- Treatment of ambiguity

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:

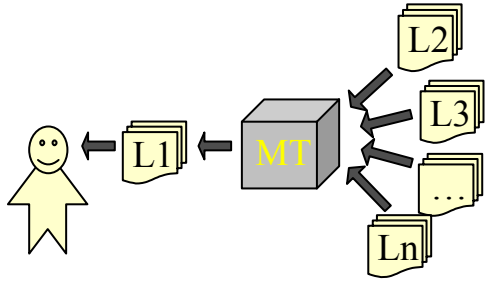
  See www.statmt.org

Use cases and requirements for MT

UNIVERSITÄT DES SAARLANDES

a) MT for assimilation

More words are spent in indicative translation

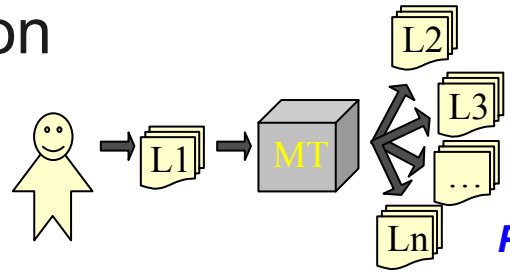


Robustness
Coverage

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

b) MT for dissemination

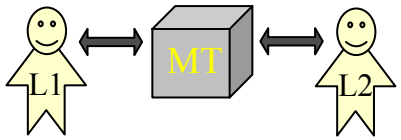
More money is spent in informative translation



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

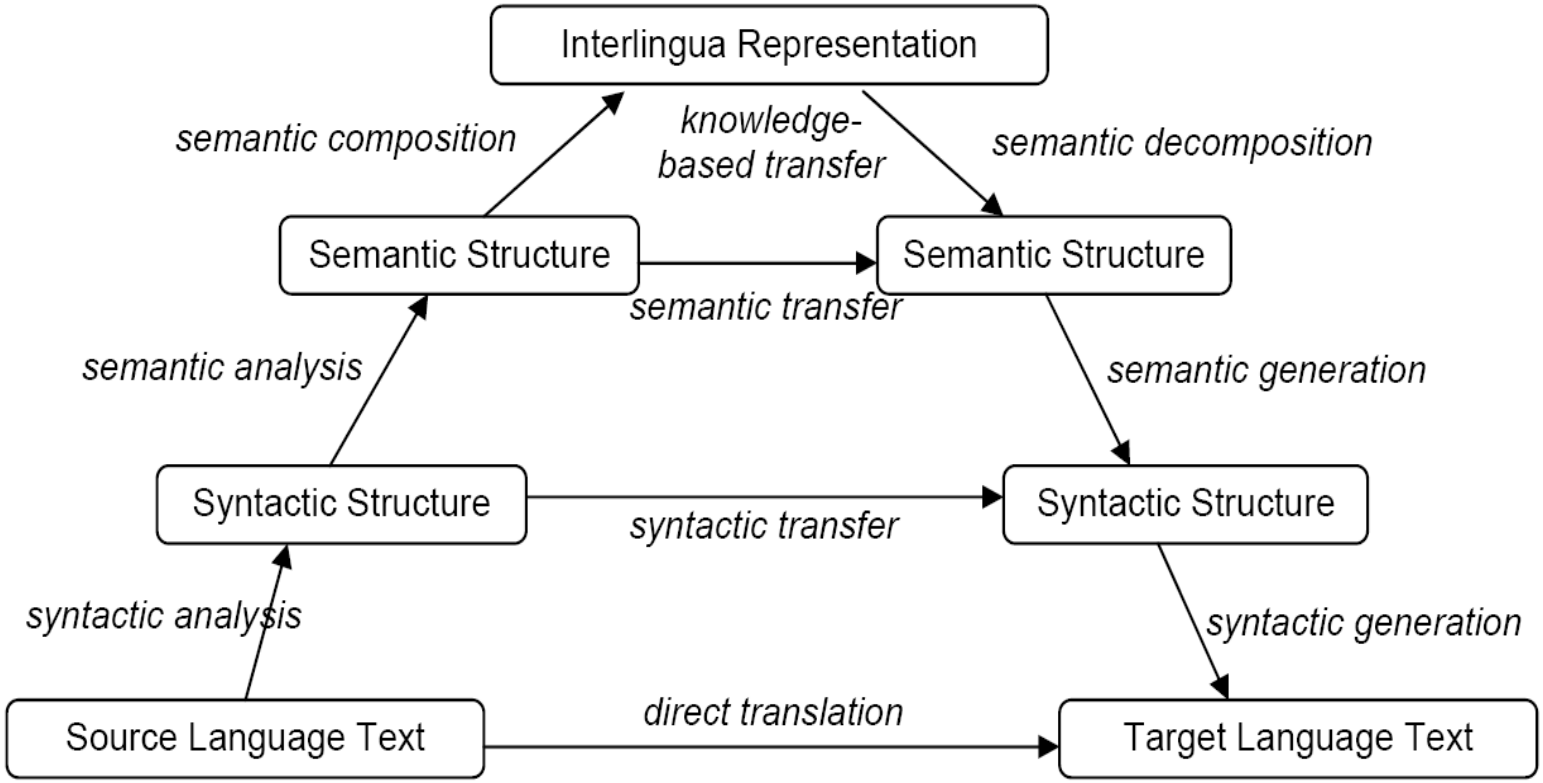
- Slides by John Hutchins:

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

Possible (rule-base) MT architectures

UNIVERSITÄT DES SAARLANDES

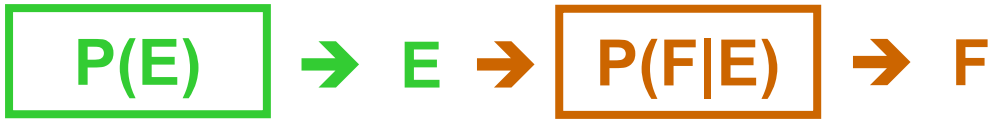
The „Vauquois Triangle“



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***

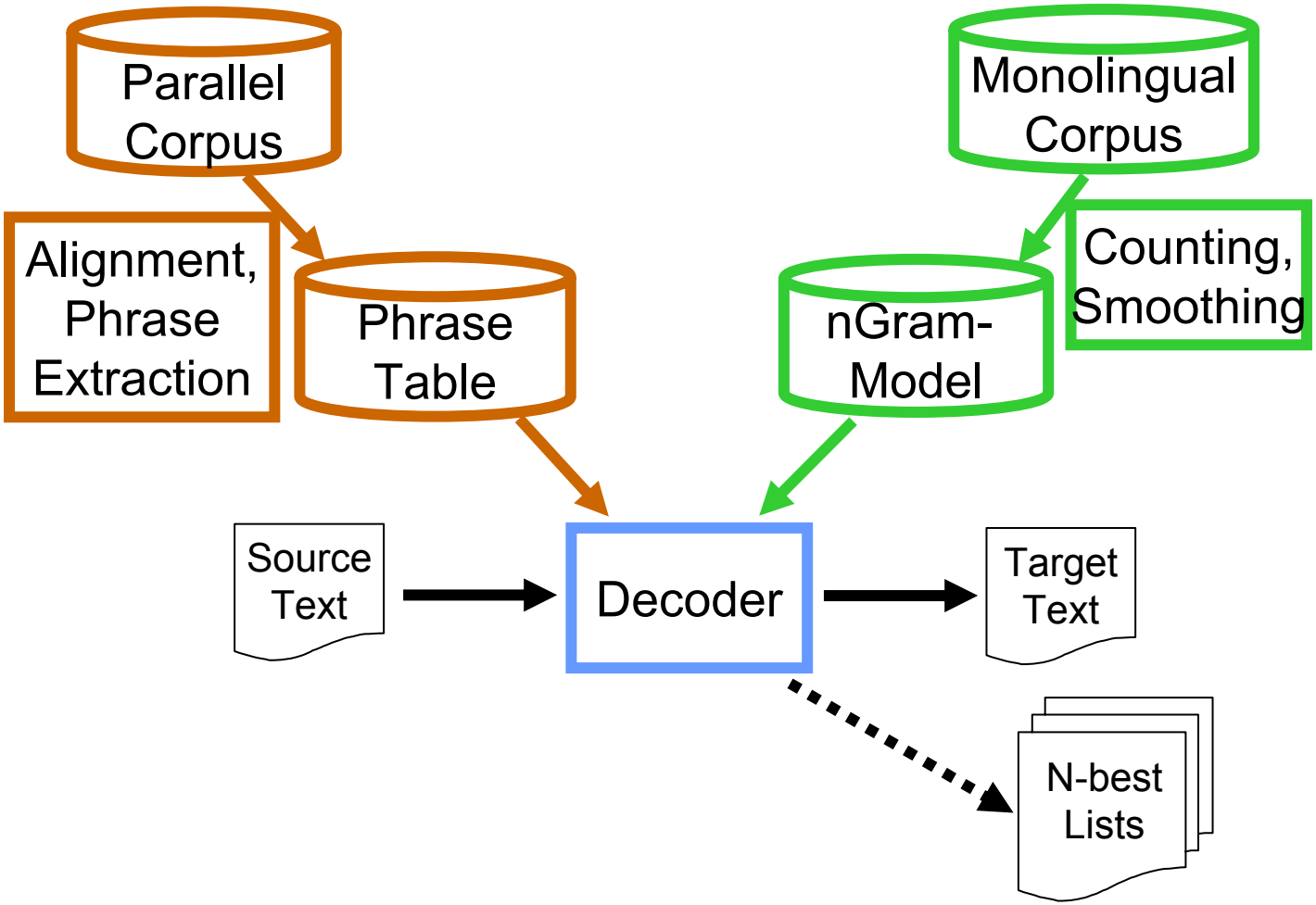
each has approximative solutions
nGram-Models $P(e_1 \ldots e_n) = \Pi P(e_i|e_{i-2}\ e_{i-1})$
Transfer of „phrases“ $P(F|E) = \Pi P(\mathbf{f}_i|\mathbf{e}_i)*P(d_i)$
Heuristic *(beam)* search

- 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



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.*

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	--	+
Lexical Reliability	+	-

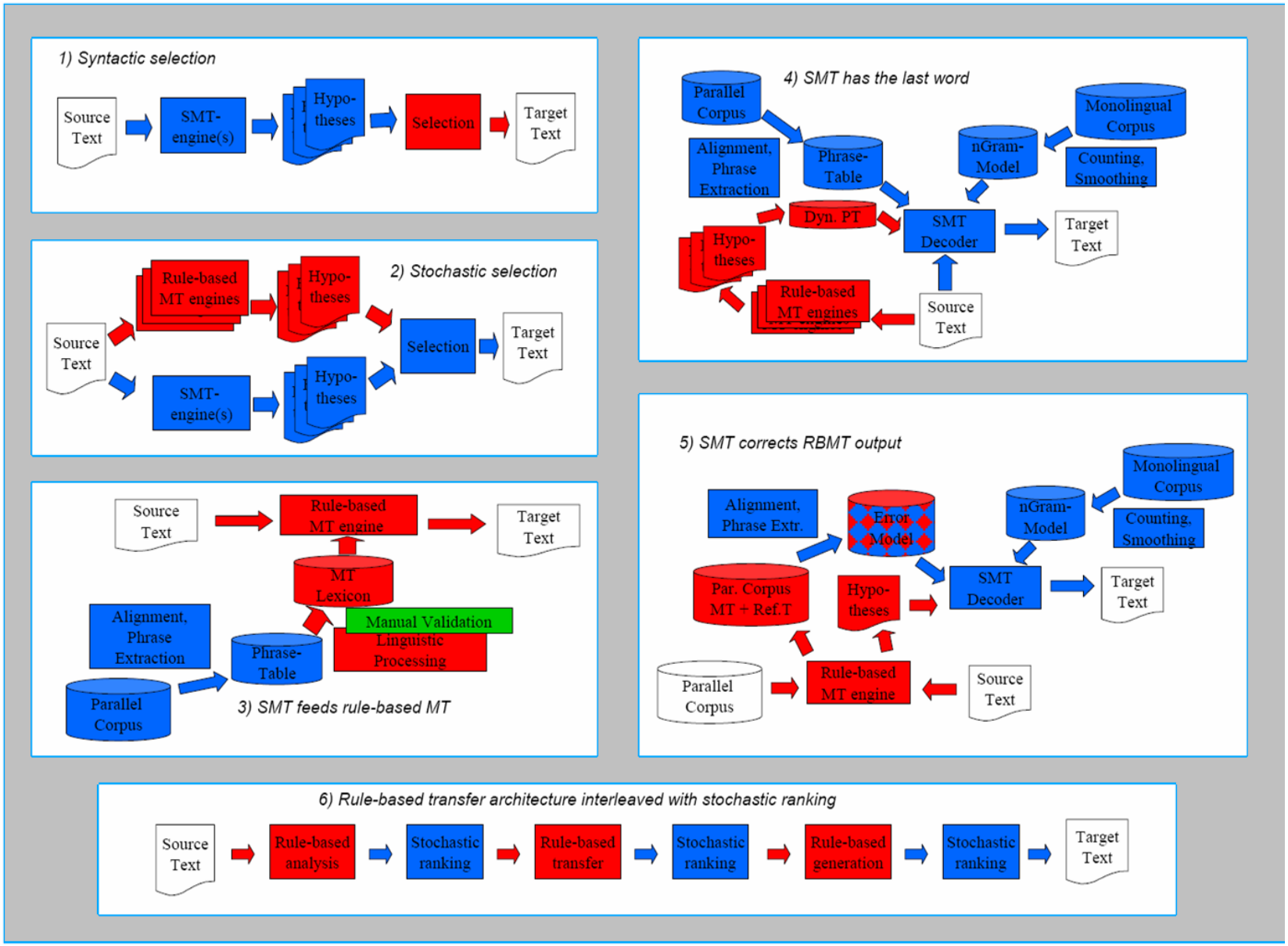
Motivation for hybrid MT (2)

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

Some hybrid MT architectures

= SMT Module
= RBMT Module

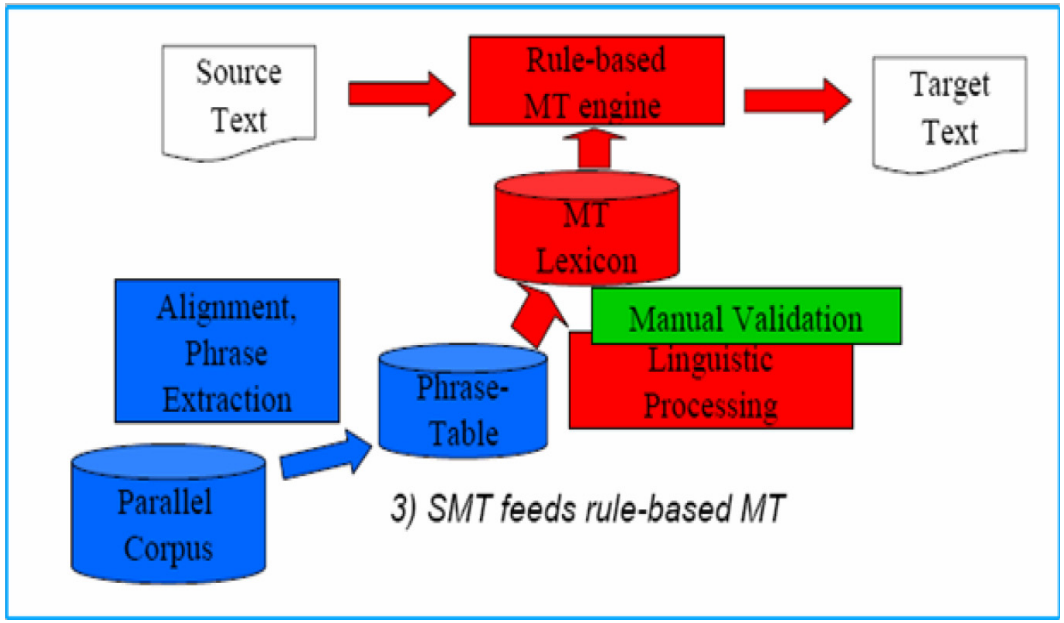


SMT feeds rule-based MT

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

Differences between Languages

Languages can differ in many ways (studied in language typology)

Morphology:

Morpheme-to-word ratio:
Isolating ← → Synthetic ←→ Polysynthetic

Segmentability:
Agglutinative ←→ Fusion Language

Syntax:

Word order: SVO vs. SOV vs. VSO vs. V2 vs. Unconstrained (+ case marked)

Whether to use determiners or not

Head-marking vs. dependent-marking:

English	the	man-^{A}'s	^{H}house
Hungarian	az	ember	^{H}ház-^{A}a
	the	man	house-his

Verb-framed vs. satellite-framed:

EN: The bottle floated out.
ES: *La botella salió flotando.*
The bottle exited floating.

Differences in Specificity of Expressions

UNIVERSITÄT DES SAARLANDES

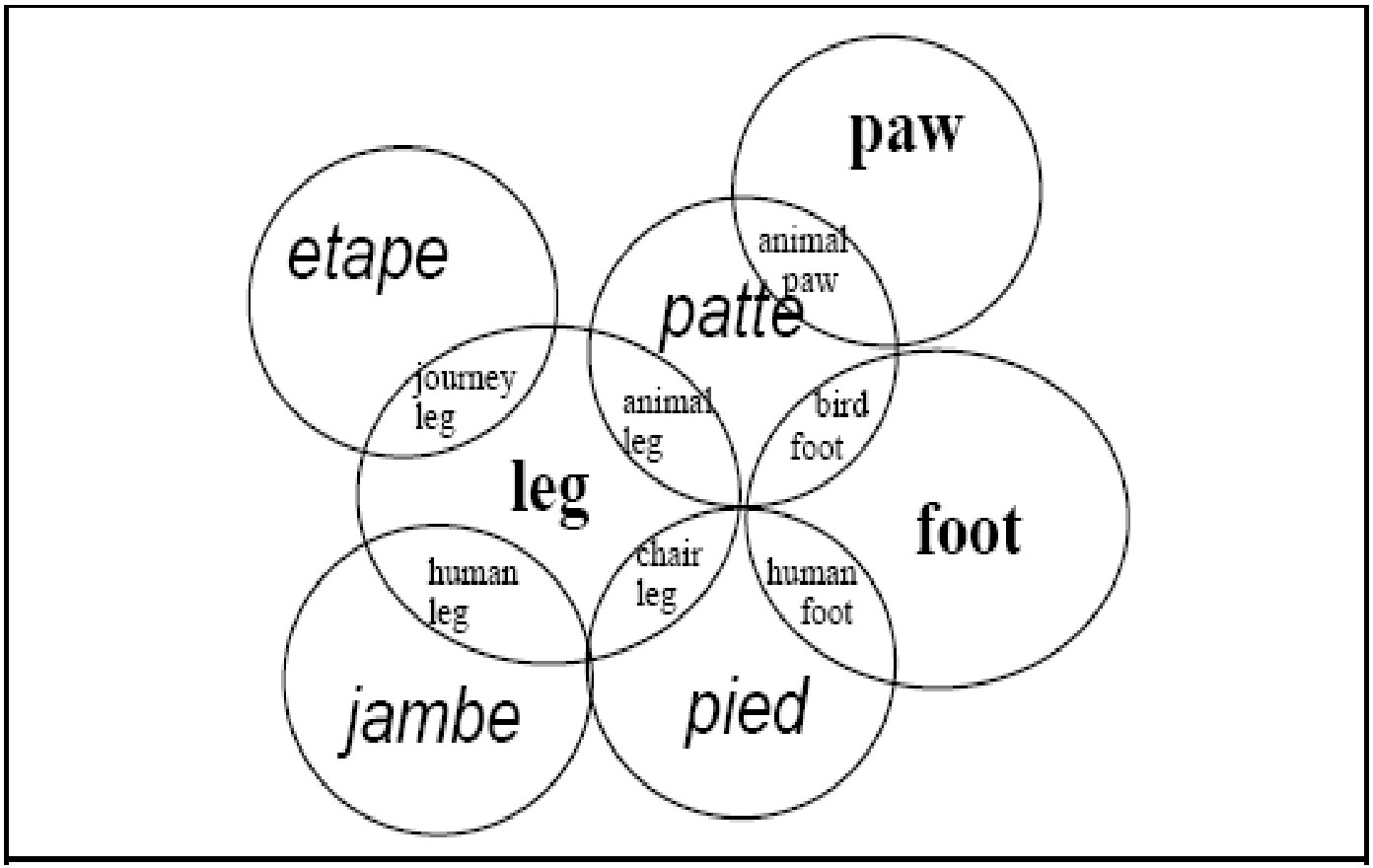
English	*brother*	Japanese	*otooto* (younger)
		Japanese	*oniisan* (older)
		Mandarin	*gege* (older)
		Mandarin	*didi* (older)
English	*wall*	German	*Wand* (inside)
		German	*Mauer* (outside)
English	*know*	French	*connaître* (be acquainted with)
		French	*savoir* (know a proposition)
English	*they*	French	*ils* (masc)
		French	*elles* (fem)
German	*berg*	English	*hill*
		English	*mountain*
Mandarin	*tā*	English	*he*
		English	*she*
		English	*it*

English *rice*	Malay	*padi* (unharvested grain)
		beras (uncooked)
		nasi (cooked)
		emping (mashed)
		pulut (glutinous)
		bubor (cooked as a gruel)
English *wear*	Japanese	*kiru* (generic)
		haoru (coat or jacket)
		haku (shoes or trousers)
		kaburu (hat)
		hameru (ring or gloves)
		shimeru (belt or tie or scarf)
		tsukeru (brooch or clip)
		kakeru (glasses or necklace)

Translation into a language using more specific expression requires us to make decisions that may be rather difficult.

(Examples taken from Jurafsky & Martin and Hutchins & Somers)

Differences in Conceptual Space



Different expressions in French and English: Jurafsky & Martin's visualisation of data from Hutchins & Somers

More bilingual lexical differences

UNIVERSITÄT DES SAARLANDES

- bilingual lexical ambiguity (more than one equivalent, whether ambiguous in SL or not):
  - river: fleuve/rivière
  - Taube: dove/pigeon
  - Schraube: screw/bolt/propellor
  - corner: coin or angle; Ecke or Winkel
  - light: léger, clair, facile, allumer, lumière, lampe, feu
  - look: regarder, chercher, sembler
- lexical gaps
  - dacha, cottage, marmelade, vodka, etc.
  - snub: infliger un affront; verächtlich behandeln, or: derb zurückweisen
  - het Turks kennen: to know Turkish
  - kenner van het Turks: *knower of Turkish, someone who knows Turkish
- **Solved (?) by contextual rules (RBMT), or examples (EBMT), or frequencies and 'language models' (SMT)**

Problems with structural ambiguity

UNIVERSITÄT DES SAARLANDES

- (1) Peter mentioned the book I sent to Mary [ambiguous for HT]
  - Peter mentioned the book which I sent to Mary
  - Peter mentioned to Mary the book which I sent [to Peter/David]
- (2a) We will meet the man you told us about yesterday [unambiguous for HT]
  - … the man you told us about yesterday
- (2b) We will meet the man you told us about tomorrow [unambiguous for HT]
  - we will meet tomorrow the man...
- (3) pregnant women and children [unambiguous for HT]
  - des femmes et des enfants enceintes [produced by MT system]
- (4a) Smog and pollution control are important factors
- (4b) Smog and pollution control is under consideration
- (4c) The authorities encouraged smog and pollution control

- **Often, problems such as (1), (2), and (3) are problematic for RBMT, but they may be 'solved' by SMT 'language model' and by EBMT databases. But problem (4c) requires 'knowledge' (i.e. rule-based KBMT)**

Sometimes translation is very hard

UNIVERSITÄT DES SAARLANDES

„Ist das Deine Cousine?“ „Nein, ich habe keine Cousine“

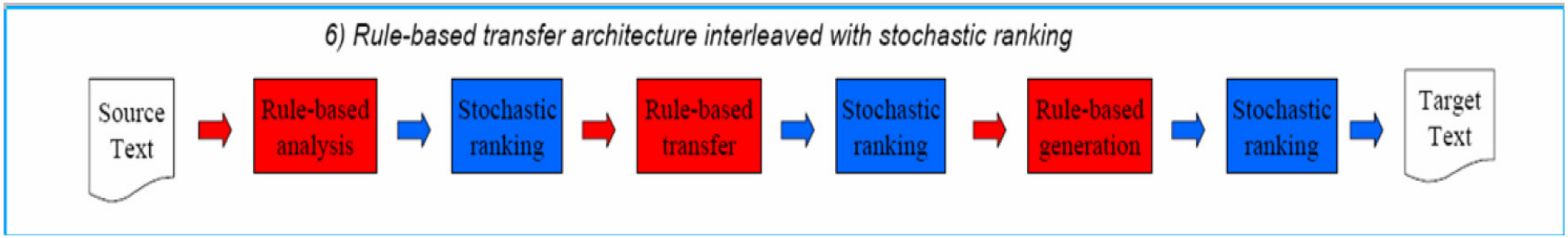
→ „Is this your cousin?“ „No, I don't have any cousin“

wrong meaning

→ „Is this your cousin?“ „No, I don't have any female cousin“

strange style, wrong connotation

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

Treatment of ambiguity

UNIVERSITÄT DES SAARLANDES

Current systems face a trade-off between efficiency and accuracy

Early binding:
Fast, but decisions are based on insufficient information

Late binding:
Decisions are (typically) better, but much computation is spent in useless paths

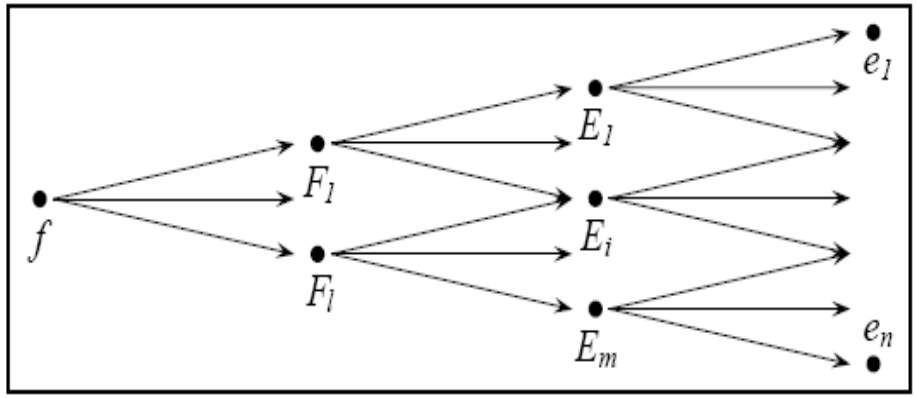


Figure 3: Abstract fan-out tree: each processing component operates non-deterministically, and distinct inputs can, in principle, give rise to equivalent outputs.

From: [Oepen e.a.: Towards Hybrid Quality-Oriented Machine Translation]