Introduction to Computational Linguistics

Frank Richter

fr@sfs.uni-tuebingen.de.

Seminar für Sprachwissenschaft

Eberhard Karls Universität Tübingen

Germany

What Makes Machine Translation Hard

- Lexical Ambiguity

What Makes Machine Translation Hard

- Lexical Ambiguity
- Lexical Gaps

What Makes Machine Translation Hard

- Lexical Ambiguity
- Lexical Gaps
- Syntactic Divergences between Source and Target Language

Problems: Word-to-Word Translations

English – German

The ticket office in the train station
Der Fahrkartenschalter im Bahnhof

öffnet wieder um ein Uhr.
re-opens at one o'clock.

Lexical Ambiguity: Open (1)

English	German
in store door	Offen
on new building	Neu eröffnet
open door	Tür öffnen
open golf tourney	Golfspiel eröffnen
open question	offene Frage
open job	freie Stelle

Lexical Ambiguity: Open (2)

English	German
loose ice	offenes Eis
blank endorsement	offene Befürwortung
private firm	offene Handelsgesellschaft
unfortified town	offene Stadt
blank cheque	offener Wechsel
to unbutton a coat	einen Mantel öffnen

Structural Divergence (1)

English – German

Max likes to swim.
NP VFIN INF

Max schwimmt gerne.
NP VFIN ADV

Structural Divergence (2)

Russian – English

Jego zovut Julian.
Him they call Julian.
They call him Julian.

Japanese – English

Kino ame ga futa.
Yesterday rain fell.
It was raining yesterday.

Differences in Word Order

English – German

Does it make sense to translate
Macht es Sinn

documents automatically ?
Dokumente automatisch zu übersetzen ?

MT: The Weaver Memo (1)

- Translation and Context

  If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words.

MT: The Weaver Memo (1)

- Translation and Context

  If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words.

  But if one lengthens the slit in the opaque mask, until one sees not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word.

MT: The Weaver Memo (2)

- Translation and Context

  The practical question is: “What minimum value of N will, at least, in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?”

MT: The Weaver Memo (2)

- Translation and Context

  The practical question is: “What minimum value of N will, at least, in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?”

- Translation and Cryptography

  ... it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the “Chinese code”.

MT: The Weaver Memo (3)

- Translation and Language Universals (Invariants)

  ... there are certain invariant properties which are, again not precisely, but to some statistically useful degree, common to all languages.

MT: The Weaver Memo (3)

- Translation and Language Universals (Invariants)

  ... there are certain invariant properties which are, again not precisely, but to some statistically useful degree, common to all languages.
  Thus may it be true that the way to translate Chinese to Arabic or from Russian to Portuguese, is not to attempt the direct route ... but down to the common base of human communication – the real but yet undiscovered universal language – and then to re-emerge by whatever particular route is convenient.

Strategies for Machine Translation

- Word-to-Word (Direct) Translation

Strategies for Machine Translation

- Word-to-Word (Direct) Translation
- Syntactic Transfer

Strategies for Machine Translation

- Word-to-Word (Direct) Translation
- Syntactic Transfer
- Semantic Transfer

Strategies for Machine Translation

- Word-to-Word (Direct) Translation
- Syntactic Transfer
- Semantic Transfer
- Interlingua Approach

Strategies for Machine Translation (2)

- Word-to-Word (Direct) Translation
  - simplest approach:

Strategies for Machine Translation (2)

- Word-to-Word (Direct) Translation
  - simplest approach:
  - may require only an electronic, bi-lingual dictionary

Strategies for Machine Translation (2)

- Word-to-Word (Direct) Translation
  - simplest approach:
  - may require only an electronic, bi-lingual dictionary
  - depending on the source and target languages and the dictionary, minimal morphological analysis and generation may be required.

Strategies for Machine Translation (2)

- Word-to-Word (Direct) Translation
  - simplest approach:
  - may require only an electronic, bi-lingual dictionary
  - depending on the source and target languages and the dictionary, minimal morphological analysis and generation may be required.
  - no use of syntactic or semantic knowledge

Strategies for Machine Translation (3)

- Syntactic Transfer

Strategies for Machine Translation (3)

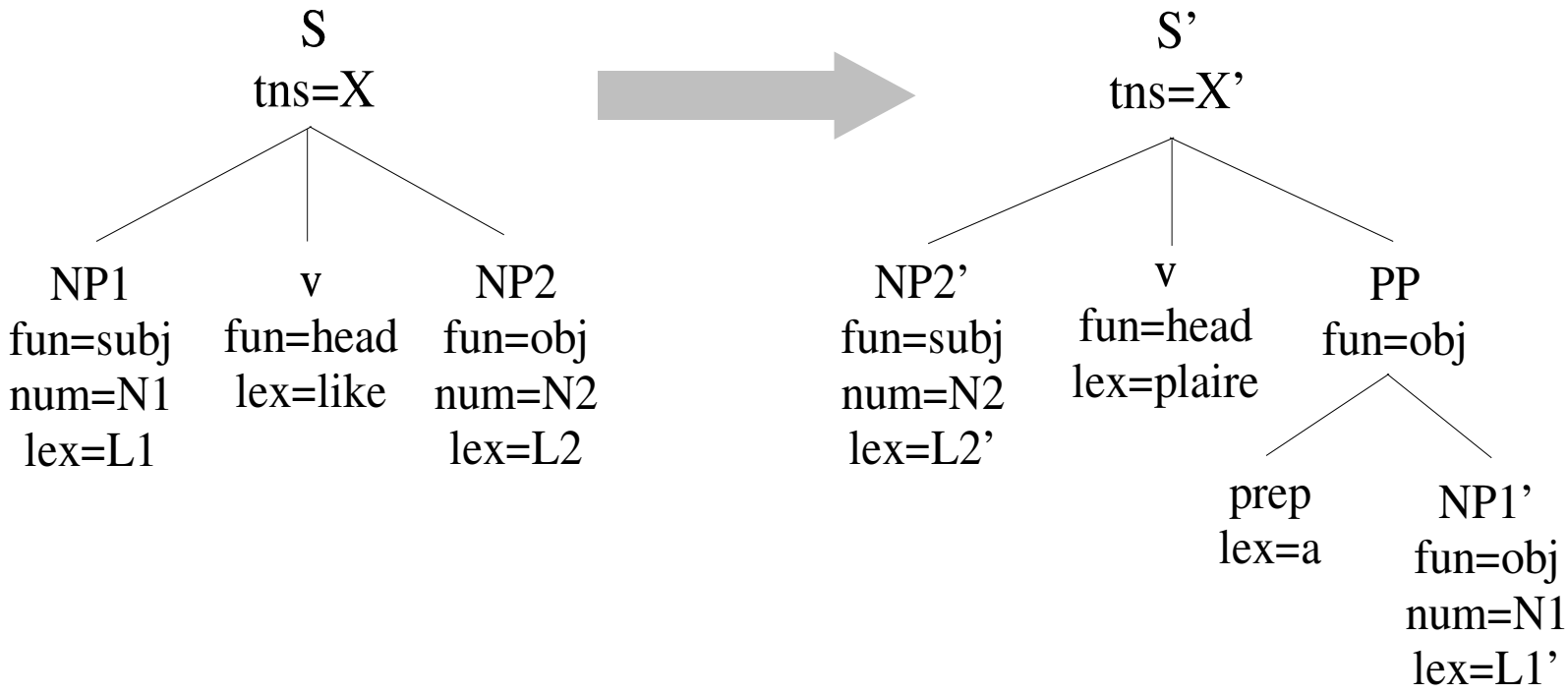
- Syntactic Transfer
  - requires syntactic analysis of the source language

Strategies for Machine Translation (3)

- Syntactic Transfer
  - requires syntactic analysis of the source language
  - requires a syntactic parser

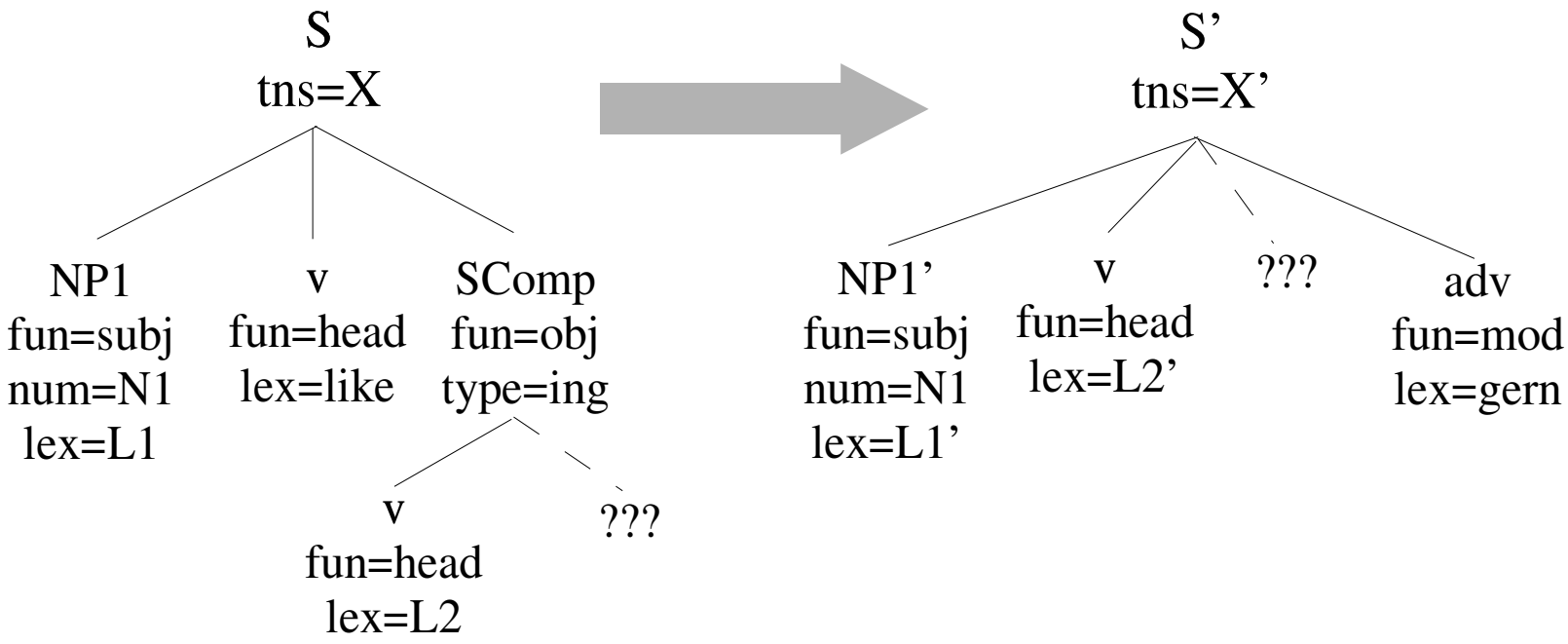
Syntactic Transfer Trees

An Example of a Transfer Tree for English *like* and French *plaire*



Syntactic Transfer Trees (2)

An Example of a Transfer Tree for English *like to* ⟨ *V* ⟩ and German ⟨ *V* ⟩ *gern*



Strategies for Machine Translation (4)

- Semantic Transfer
  - requires syntactic and semantic analysis of the source language

Strategies for Machine Translation (4)

- Semantic Transfer
  - requires syntactic and semantic analysis of the source language
  - requires language-dependent meaning representation language

Strategies for Machine Translation (4)

- Semantic Transfer
  - requires syntactic and semantic analysis of the source language
  - requires language-dependent meaning representation language
  - language-dependent rules that relate source language meaning representations to target language meaning representations

Strategies for Machine Translation (4)

- Semantic Transfer
  - requires syntactic and semantic analysis of the source language
  - requires language-dependent meaning representation language
  - language-dependent rules that relate source language meaning representations to target language meaning representations
  - requires language generation component which maps target language meaning representations to output sentences

Strategies for Machine Translation (5)

- Semantic Transfer
  - synthesis typically performed in two stages: semantic synthesis (resulting in syntactic trees) and morphological synthesis (resulting in strings of inflected word forms).

Strategies for Machine Translation (5)

- Interlingua Approach
  - source language input is mapped to a language-neutral (quasi-universal) meaning representation language

Strategies for Machine Translation (5)

- Interlingua Approach
  - source language input is mapped to a language-neutral (quasi-universal) meaning representation language
  - requires syntactic and semantic analysis of the source language into interlingua

Strategies for Machine Translation (5)

- Interlingua Approach
  - source language input is mapped to a language-neutral (quasi-universal) meaning representation language
  - requires syntactic and semantic analysis of the source language into interlingua
  - requires language generation component which maps interlingua to output sentences

Strategies for Machine Translation (5)

- Interlingua Approach
  - source language input is mapped to a language-neutral (quasi-universal) meaning representation language
  - requires syntactic and semantic analysis of the source language into interlingua
  - requires language generation component which maps interlingua to output sentences
  - synthesis typically performed in two stages: semantic synthesis from the interlingua (resulting in syntactic trees) and morphological synthesis (resulting in strings of inflected word forms).

Interlingua Representation for Motion Verbs

He walked across the road.
Il traversa la rue a pied.

$$\begin{bmatrix} \text{PRED} = \langle \text{MOTION} \rangle \\ \text{TENSE} = \text{PAST} \\ \text{AGENT} = \begin{bmatrix} \text{PRED} = \text{PRON} \\ \text{NUM} = \text{SING} \\ \text{PERS} = 3 \\ \text{SEX} = \text{MALE} \end{bmatrix} \\ \text{INSTR} = \begin{bmatrix} \text{PRED} = \langle \text{FOOT} \rangle \end{bmatrix} \\ \text{LOC} = \begin{bmatrix} \text{PRED} = \langle \text{CROSS} \rangle \\ \text{OBJ} = \begin{bmatrix} \text{PRED} = \langle \text{ROAD} \rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Interlingua Representation for Motion Verbs (2)

They flew from Gatwick.
Ils partirent par avion de Gatwick.

$$\begin{bmatrix} \text{PRED} = \langle \text{ MOTION} \rangle \\ \text{TENSE} = \text{PAST} \\ \text{AGENT} = \begin{bmatrix} \text{PRED} = \text{PRON} \\ \text{NUM} = \text{PLUR} \\ \text{PERS} = 3 \end{bmatrix} \\ \text{INSTR} = \begin{bmatrix} \text{PRED} = \langle \text{ PLANE} \rangle \end{bmatrix} \\ \text{LOC} = \begin{bmatrix} \text{PRED} = \langle \text{LEAVE} \rangle \\ \text{OBJ} = \begin{bmatrix} \text{PRED} = \text{GATWICK} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Interlingua Representation for Verbs (1)

English *wall* German *Wand* (inside a building)
Mauer (outside)

English *river* French *rivière* (general term)
fleuve (major river, flowing into sea)

Interlingua Representation for Verbs (2)

English	*leg*	Spanish	*pierna* (human)
			pata (animal,table)
			pie (chair)
			etapa (of a journey)
		French	*jambe* (human)
			patte (animal,insect)
			pied (chair,table)
			étape (journey)

Interlingua Representation for Verbs (3)

English	*blue*	Russian	*goluboi* (pale blue) *sinii* (dark blue)
French	*louer*	English	*hire* or *rent*
French German	*colombe* *Taube*	English	*pigeon* or *dove*
German	*leihen*	English	*borrow* or *lend*

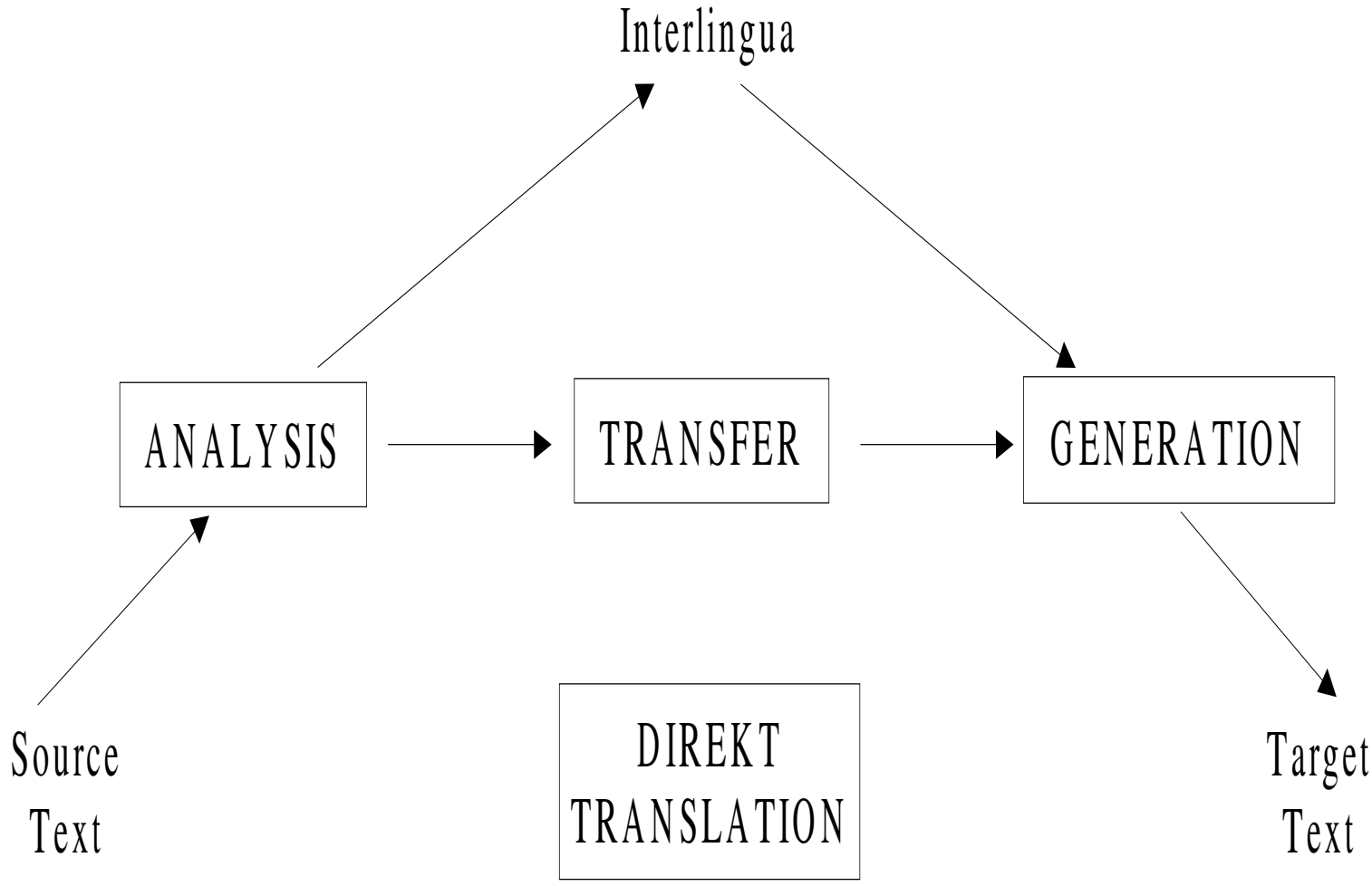
Interlingua Representation for Verbs (4)

English *rice* Malay *padi* (unharvested grain)
beras (uncooked)
nasi (cooked)
emping (mashed)
pulut (glutinous)
bubor (cooked as a gruel)

Interlingua Representation for Verbs (5)

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)

The Vauquois Triangle



Strategies for Machine Translation

Modules required in an all-pairs MTS

Number of languages	Analysis modules	Generation modules	Transfer modules	Total modules
2	2	2	2	6
3	3	3	6	12
4	4	4	12	20
5	5	5	20	30
...				
9	9	9	72	90
n	n	n	n(n-1)	n(n+1)