Computational Linguistics

Machine Translation

Suhaila Saee & Bali Ranaivo-Malançon

Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak

August 2014

What is 'translation'?

Noun: translation

|trānz'ley-shun| |tran(t)s'ley-shun|

1. A written communication in a second language having the same meaning as the written communication in a first language
   - = interlingual rendition, rendering, version
   - ~ written account, written record
   - ⇒ caption, crib, mistranslation, retroversion, subtitle, supertitle, surtitle, trot

(SOURCE:WORDWEB PRO 1.4)

But do we still need to translate automatically?

- Domination of few languages in the Internet
- Users should be allowed to query in their native language
- Information should be available in all living languages
- Impact of globalisation, translation is in demand: "British translation software company SDL said it was seeing a boost from areas like web-content management, as more people around the world use the internet to communicate with multinationals." (*Reuters*, August 2, 2011)

What is Machine Translation (MT)?

The oldest CL application!

The use of a machine (computer) to translate a **source language** (**SL**) text into a **target language** (**TL**) text

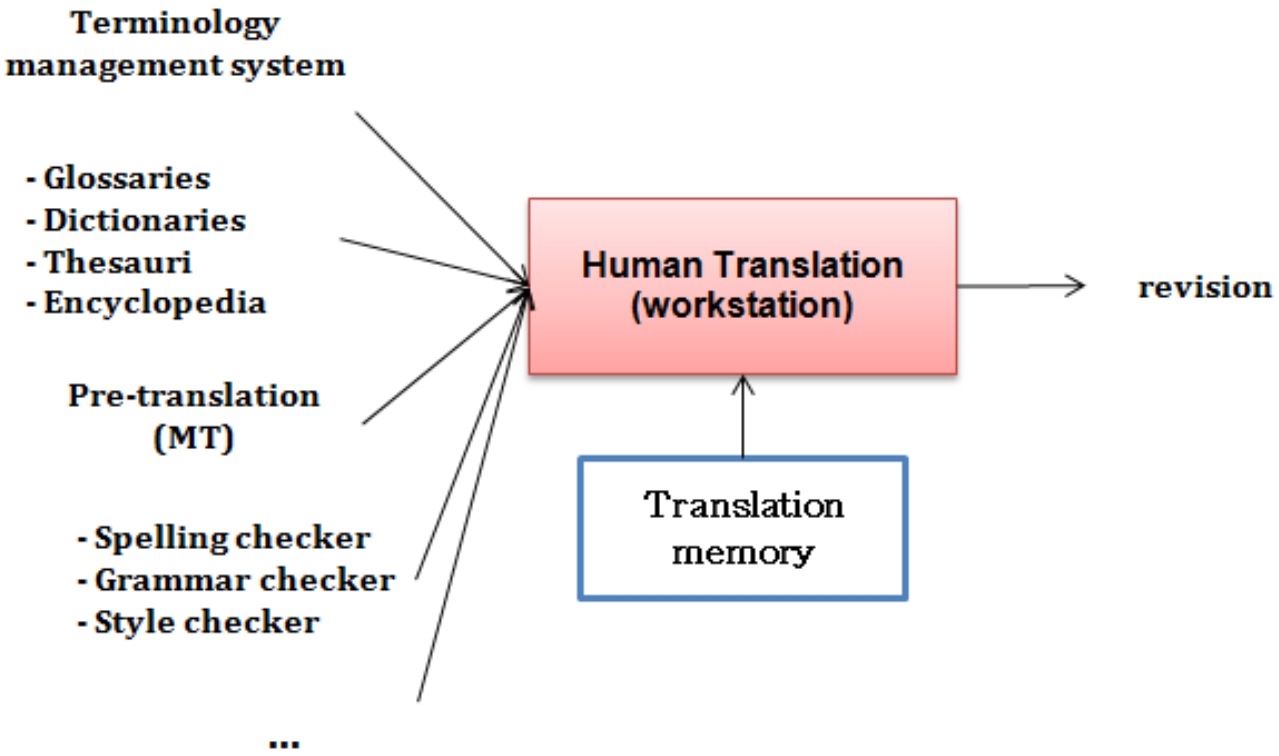
- the process of translating human languages by machine, with or without human assistance
- Written vs. Spoken translation
- Bilingual (two languages) vs. Multilingual MT (more than two languages)
- Unidirectional (e.g., English $\rightarrow$ Malay) vs. Bidirectional MT (e.g., English $\leftrightarrow$ Malay)

Classification of MT Systems

- Fully automatic (MT)
- Human aided machine translation (HAMT)
- Machine aided human translation (MAHT)
- The term Computer Aided human Translation (CAT) is sometimes used to cover HAMT and MAHT

Machine Aided Human Translation (MAHT)

- Intended for professional translators
- MAHT tools include

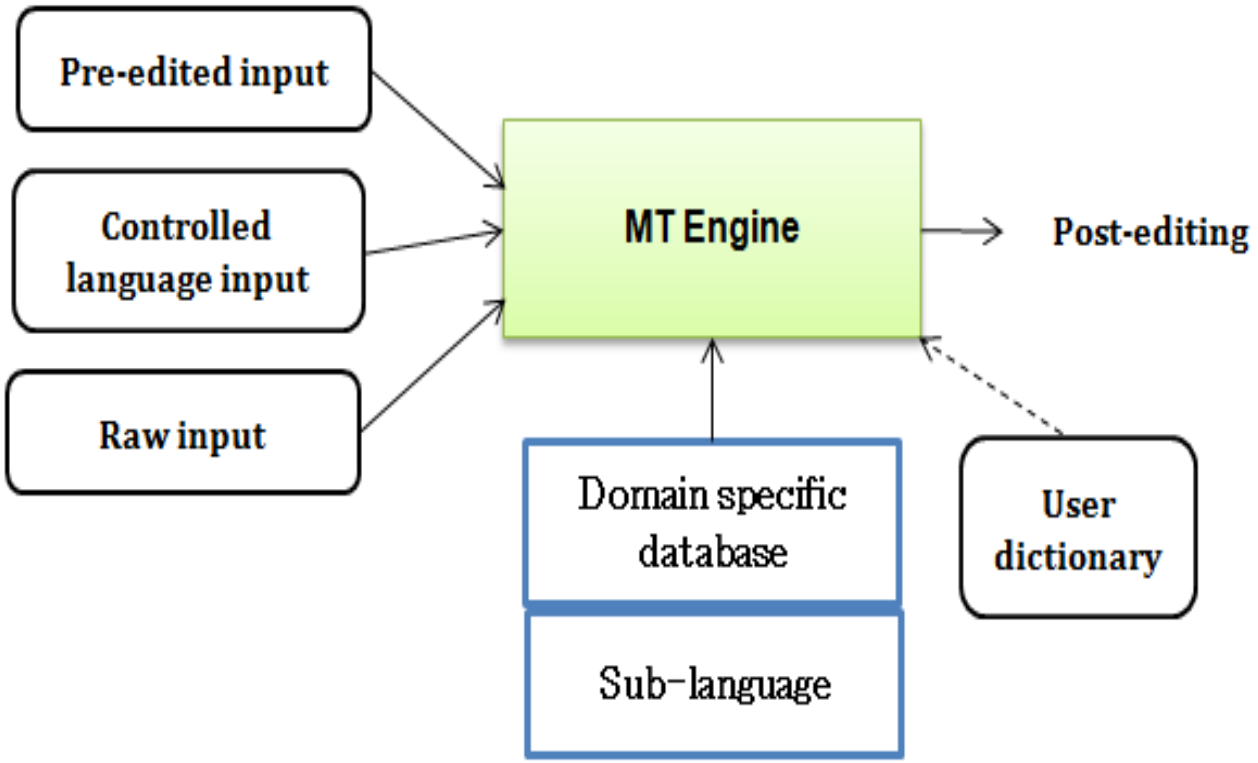


Translation Memory

- It is an aid for human translators
- It stores and indexes existing translations
- Before translating new text
  - Check to see if you have translated it before
  - If so, reuse the original translation

Human Aided Machine Translation (HAMT)

- The translation process is controlled by the machine
- The human assistance takes place during the translation process



HAMT (cont'd)

- **Pre–editing** - adjusting the source text
  - For example, removing typos, segmenting long sentences into short sentences, fixing up punctuation, tagging non-translatable items, marking grammatical categories of homographs, flagging or substituting foreign words, etc.
- **Post–editing** - correcting the output of an MT system to an agreed standard
- Interactive - an MT system, which will pause and ask the user to resolve the problem of ambiguity

Lexical Ambiguity

book the flight $\Rightarrow$ reservar
read the **book** $\Rightarrow$ libro

Example:
the box was in the **pen**
the **pen** was on the table

Different Word Orders

- English word order is SUBJECT $\Rightarrow$ VERB $\Rightarrow$ OBJECT
- Japanese word order is SUBJECT $\Rightarrow$ OBJECT $\Rightarrow$ VERB

Example:
English: *IBM bought Lotus*.
Japanese: *IBM Lotus bought*.

Pronoun Resolution

The computer outputs the data; it is fast.

⇓

La computadora imprime los datos; es rapida.

The computer outputs the data; it is stored in ascii.

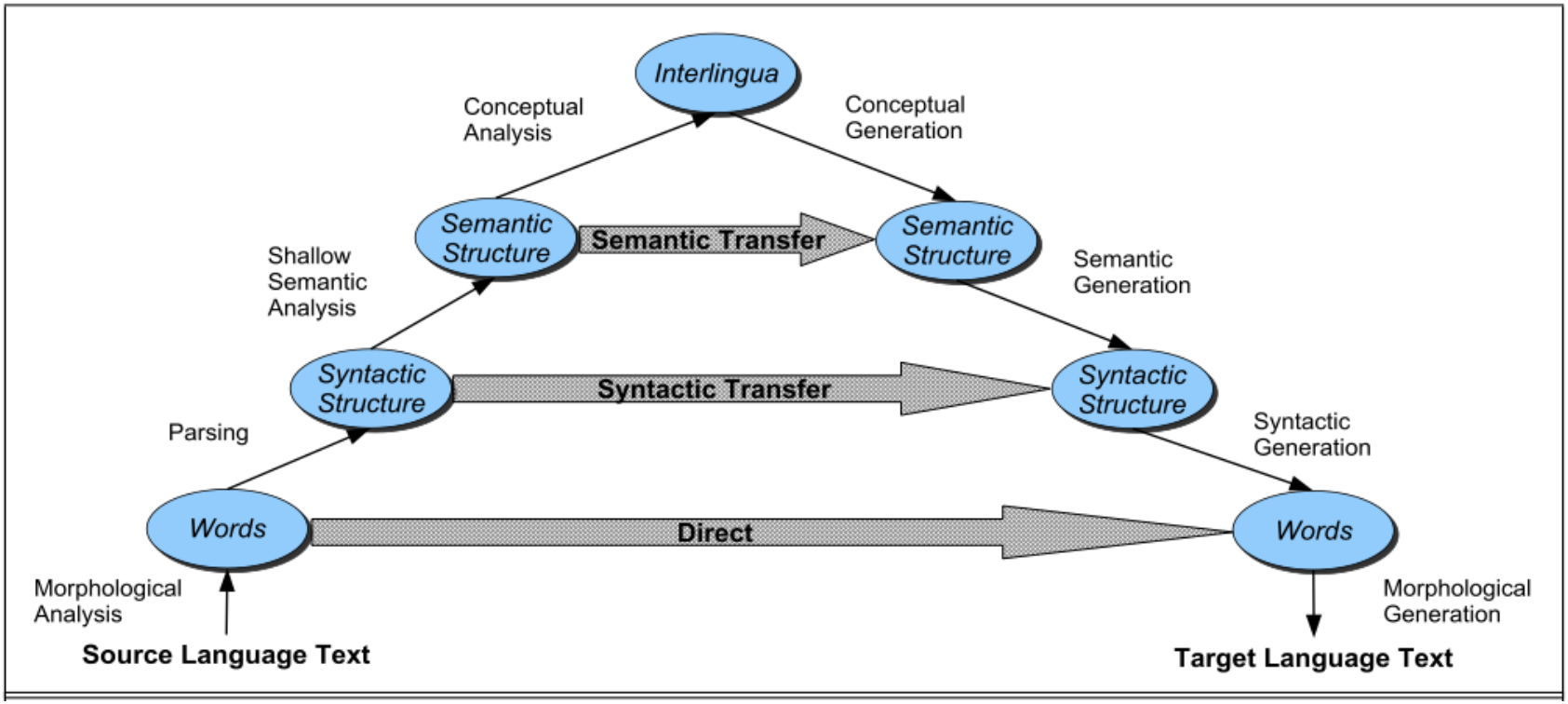
⇓

La computadora imprime los datos; estan almacendos en ascii

Rule–based MT (RBMT)

- Dominant approach in 1970s and 1980s
- Requires human expertise to write rules based on grammar, manual dictionaries, etc
- Three translation techniques in RBMT:
  1. direct translation
  2. transfer-based translation
  3. interlingua-based translation

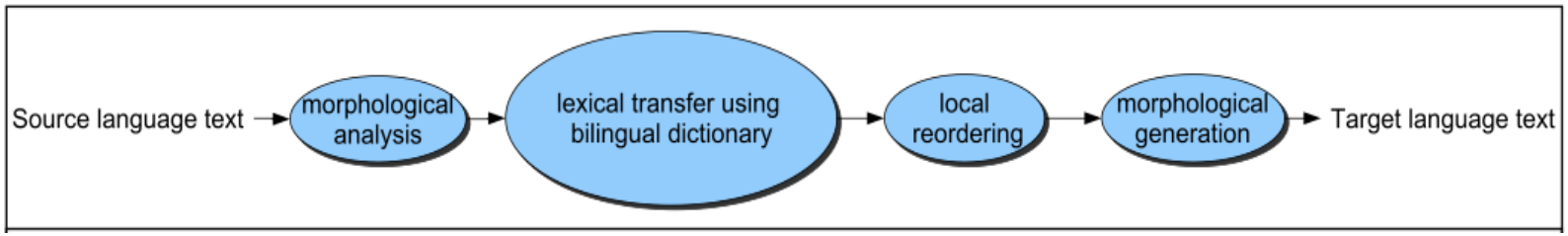
Vauquois Triangle



(SOURCE: JURAFSKY & MARTIN, 2008: SPEECH AND LANGUAGE PROCESSING)

Direct Translation

- A **word-to-word** translation using a huge **bilingual dictionary**



(SOURCE: JURAFSKY & MARTIN, 2008: SPEECH AND LANGUAGE PROCESSING)

- PROBLEMS
  - There is **no one-to-one correspondence** between words in different languages
  - Languages have **different word orders**

Direct Translation (cont'd)

Example from: (JURAFSKY & MARTIN, 2008)

English → Spanish

Input	Mary didn't slap the green witch
Morphological analysis	Mary DO-PAST not slap the green witch
Lexical Transfer	Maria PAST no dar una bofetada a la verde bruja
Local reordering	Maria no dar PAST una bofetada a la bruja verde
Morphological generation	Maria no daba una bofetada a la bruja verde

Transfer Based Translation

Handles translation into a sequence of three steps:

1. **SL analysis**: SL is syntactically parsed
2. **Transfer**
   - **Syntactic transfer rules** - modify the source parse tree to resemble the target parse tree
   - **Lexical transfer rules**: based on a bilingual dictionary
3. **TL synthesis** (or generation)

Transfer Based Translation Problems

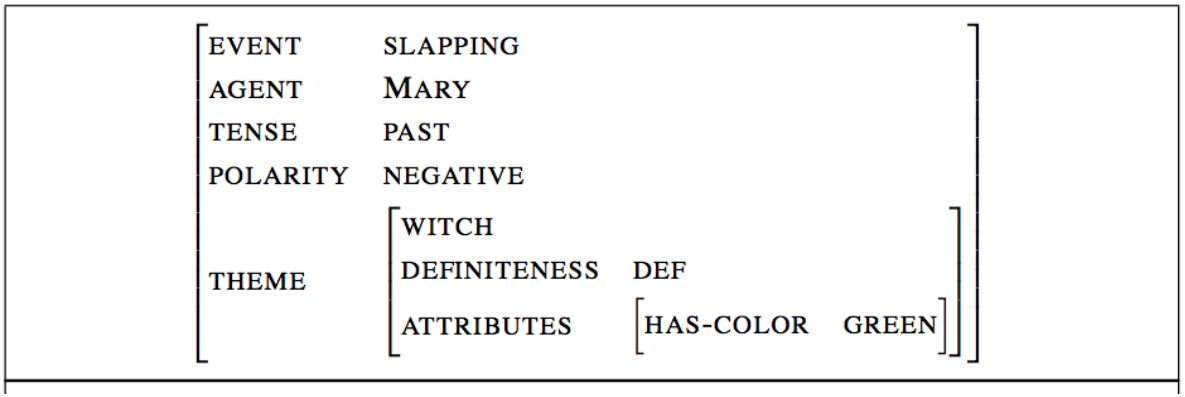
- Lots of grammar engineering (writing rules for each pair of languages)
- Requires a distinct set of transfer rules for each pair of languages
- Failure at one analysis stage may mean no output
- Complexity of tree transduction rules

Interlingua Translation

- aka **Knowledge–based translation**
- It does not rely on literal translations
- Translation is done via an **interlingua**
- To treat translation as a process of **extracting** the meaning of the input and then **expressing** that meaning in the target language

An interlingua is a knowledge representation formalism that is independent of the way particular languages express meaning

Interlingua Translation (cont'd)



(SOURCE: JURAFSKY & MARTIN, 2008: SPEECH AND LANGUAGE PROCESSING)

- PROBLEMS
  - Nature of interlingua: natural, artificial, logical? language-neutral or language-universal? complex
  - Requires a large amount of hand-coded lexical knowledge which may not be available

Shortcomings of RBMT

- Expensive: development and maintenance of appropriate large scale grammatical and lexical resources
- Slow: takes years to be developed
- Knowledge acquisition bottleneck
- Quality and level of linguistic detail required for various cases of disambiguation
- All RBMT systems are very hard to extend or to adapt to new languages

Example Based MT (EBMT)

- First proposed by Makoto Nagao in 1981 to translate abstracts (Japanese $\rightarrow$ English)
- BASIC IDEA:
  - Collect a bilingual corpus of translation pairs (= examples)
  - Use a best match algorithm to find the closest example to the SL text

To Illustrate the Idea

- Input: He buys a book on language technology
- Existing examples
  - He buys a notebook - Il achète un ordinateur portable
  - I read a book on language technology - Je lis un livre sur la technologie de la langue
- Output: Il achète un livre sur la technologie de la langue

EBMT vs Translation Memory (TM)

- TM: similar to EBMT
- Both EBMT and TM involve matching the input against a database of real examples and identifying the closest matches
- BUT
  - In TM: translator decides what to do with the proposed matches
  - In EBMT: the automatic process continues by identifying corresponding translation fragments, and then recombining these to give the target text

Main processes of EBMT

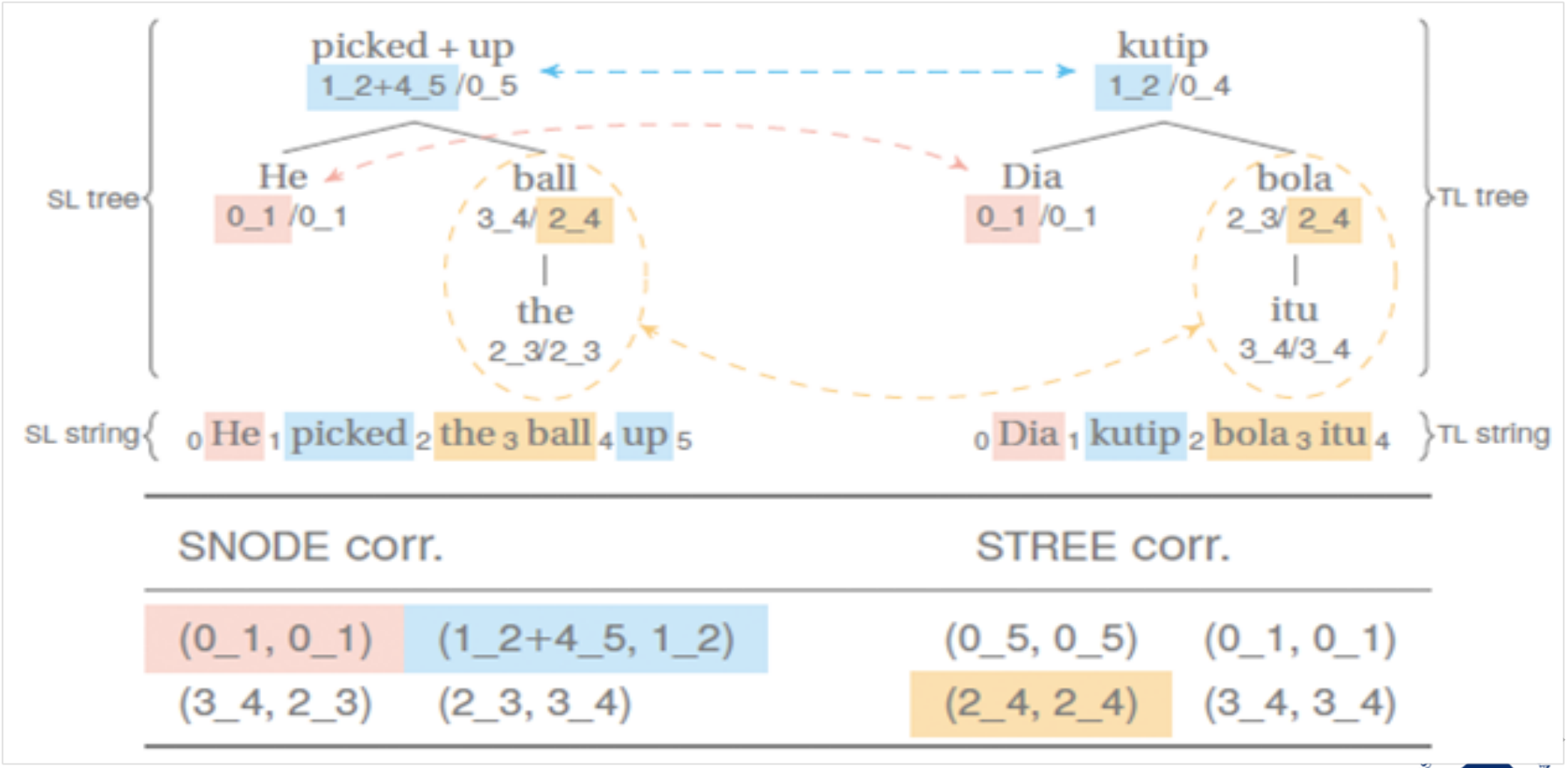
1. **Matching** - searching for fragments of the source text in the reference corpus
2. **Alignment** - identifying the corresponding translation fragments
3. **Recombination** - composing these translation fragments into the appropriate target text

How Examples in EBMT are stored?

- Pairs of strings with no additional information
- Annotated constituency tree
- Dependency tree pairs
- LFG f-structure pairs
- Tree-to-string systems
- Generalised examples or Translation templates/patterns - Similar examples are combined and stored as a single "generalised" example

Example of Annotated Tree Structures

Example from (AL-ADHAILEH AND TANG, 2002): Synchronous Structured String-Tree Correspondence (S-SSTC)



Example of Generalised Examples

Example:
John Miller flew to Frankfurt on December 3rd.
$< firstname >< lastname > flewto < city > on < month >< ordinal >$

Strengths of EBMT

- Not domain specific
- (In principle) Very efficient
- Potentially multilingual
- Correspondences can be found from raw data
- Examples give well structured output

Weaknesses of EBMT

- Depends on good bilingual corpus, which might not be highly available
- Suitability of examples - how to choose appropriate examples for the database?
- The calculation of the best match might be a complicated and lengthy process
- If the information needed for a translation is not included in the database, no adequate translation can be produced

Statistical Machine Translation (SMT)

Statistical Machine Translation (SMT) uses examples of translations and statistical models to learn how to translate texts

Why Choose SMT?

- Economic reasons: - Low cost; Rapid prototyping
- Practical reasons: - Many language pairs don't have NLP resources, but do have parallel corpora
- Quality reasons:
  - Uses chunks of human translated as its building blocks
  - When very large data sets are available produces state of the art results

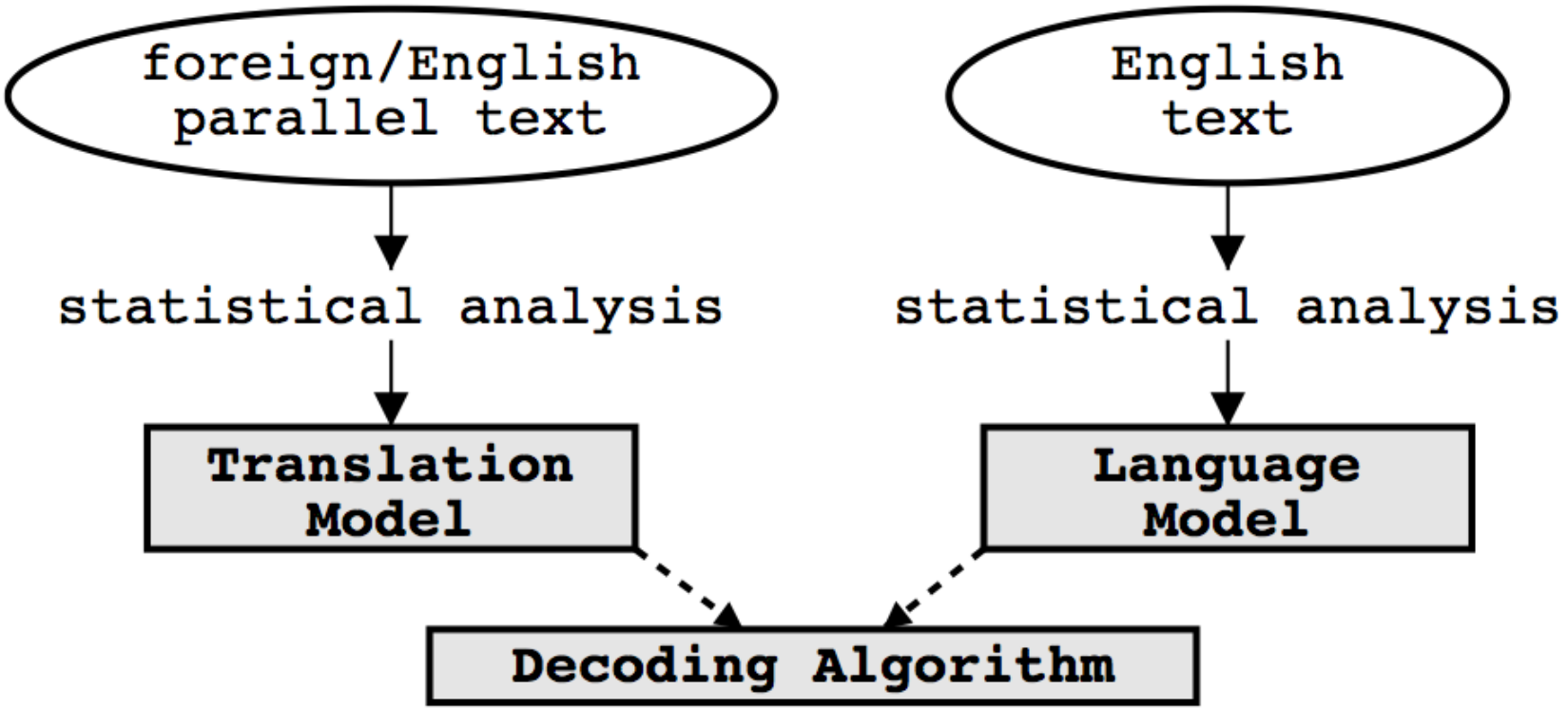
Data: Parallel Corpus

- We assume a coprus of example translations. A parallel corpus:
  - Source-target sentence pairs: for every source sentence its target translation.
- Human translation is the norm.

The idea is to learn from the parallel corpus a probabilistic model that enables translation.

Components in SMT

Three components : **translation model, language model & decoder**



(SOURCE: CALLISON-BURCH, INTRO TO STATISTICAL MT, MT MARATHON, 2008)

Statistical MT (SMT)

- We are translating from a foreign language sentence F to English
- In a probabilistic model, the best English sentence $\hat{E}$ is the one whose probability $P(E|F)$ is the highest
- We can rewrite this via Bayes rule:

$$\hat{E} = \arg\max_E P(E|F) \quad (1)$$

$$= \arg\max_E \frac{P(E)P(F|E)}{P(F)} \quad (2)$$

$$= \arg\max_E P(E)P(F|E) \quad (3)$$

- In Equation (2), we can ignore $P(F)$ since F is a constant

What the Probabilities Represent?

- $P(E)$ is the **language model**:
  - Assigns a higher probility to grammatical sentences
  - Estimated using monolingual corpora
- $P(F|E)$ is the **translation model**
  - Assigns a higher probility to sentences that have corresponding meaning
  - Estimated using bilingual corpora

Models in SMT

- Language models $P(E)$
  - (Smoothed) **N–gram language models**
- Translation models $P(F|E)$
  - Word-based models
  - Syntax-based models
  - **Phrase–based models**

Advantages of SMT

- Requires little human expertise
- Always give a possible translation: approximation of a possible translation

Disadvantages of SMT

- Depends totally on a good quality and quantity of bilingual or multilingual corpora
- Output are often ungrammatical
- Might tend to be domain-specific unless the corpus is very large and contains text from different domain (e.g. technical text, newspaper, novels, etc.)

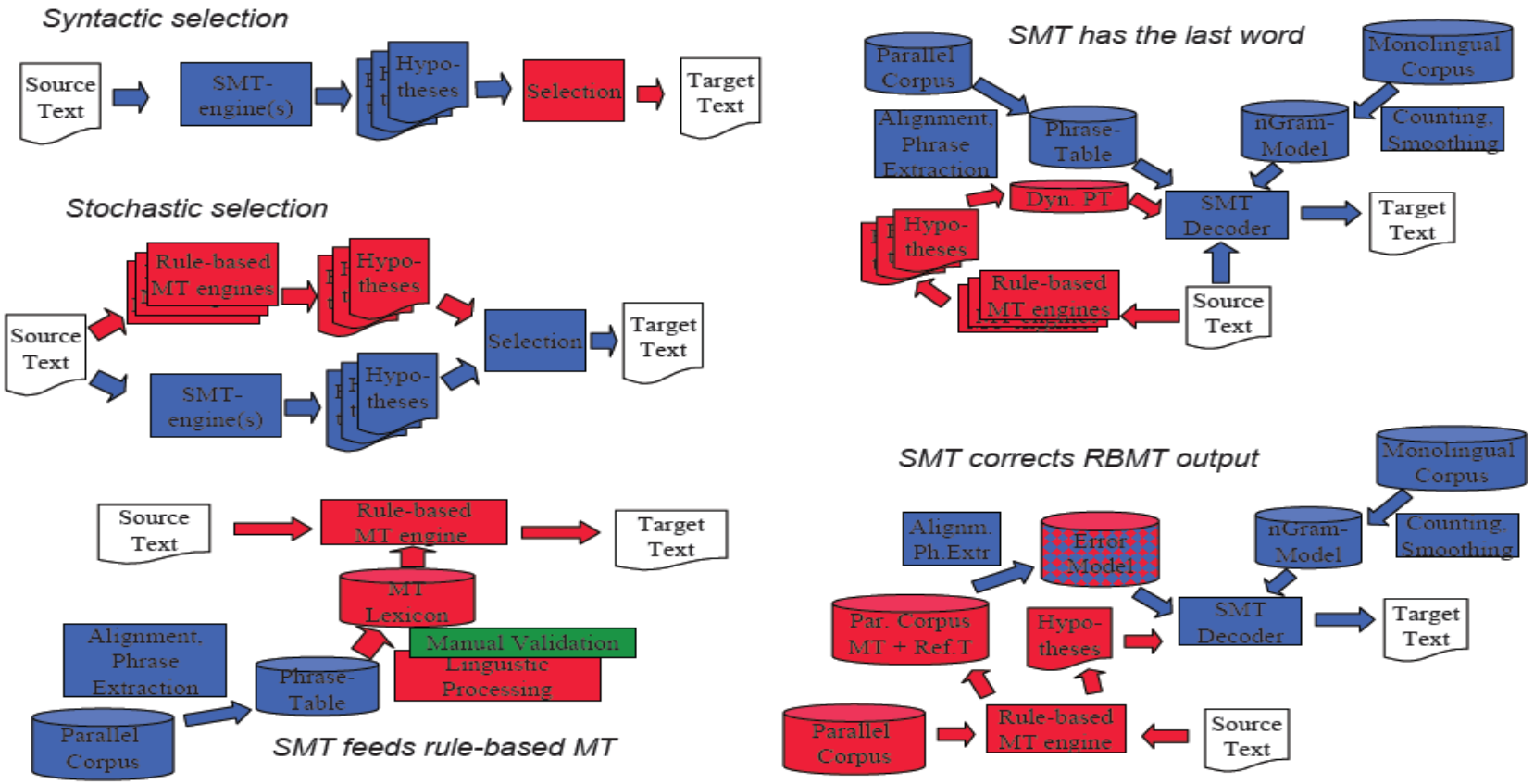
Hybrid MT

- Combining RBMT and SMT
- Motivations

	Syntax	Structural semantics	Lexical semantics	Lexical adaptivity
RBMT	++	+	-	- -
SMT	- -	- -	+	+
EBMT	-	- -	-	++

(SOURCE: A. EISELE, *HYBRID MACHINE TRANSLATION TRANSLATION: COMBINING RULE RULE-BASED AND STATISTICAL MT SYSTEMS*, 1ST MT MARATHON, EDINBURGH, 2007)

Overview of of Some Hybrid Architectures



(SOURCE: A. EISELE, *HYBRID MACHINE TRANSLATION TRANSLATION: COMBINING RULE-BASED AND STATISTICAL MT SYSTEMS*, 1ST MT MARATHON, EDINBURGH, 2007)

Alignment, Bitext, & Parallel Corpus

- A **bitext** is a pair of texts that correspond to each other in one way or another
- **Alignment** is the automatic mapping of corresponding segments in a bitext
- A **parallel corpus** consists of several bitexts

Types of Alignment

- **Monotonic alignment**: no crossings
  - The corresponding segments in a bitext must occur in the same order
  - i.e., the n^{th} segment on one side corresponds to the n^{th} segment on the other side of the bitext
- **Flat alignment**: all segments are disjoint
  - i.e., they do not overlap in any way
- **Hierarchical tree alignment**: smaller segments are nested within larger segments

Levels of Text Alignment

Text alignment can be done at many levels:

- Document alignment - mapping between corresponding documents
- Paragraph alignment - mapping between corresponding paragraphs
- Sentence alignment - mapping between corresponding sentences
  - A **bead** is a group of sentences in one language that corresponds in content to a group of sentences in another language
- Phrase alignment - mapping between corresponding phrases
- Word alignment mapping between corresponding words
- Character alignment - mapping between corresponding characters

Measuring the Quality of MT Systems

- Quality is difficult to measure in the context of MT
  - No absolute way to measure how "correct" a translation is
  - Many "correct" answers as there are translators
- Common way to measure quality: compare the output of automated translation to a human translation of the same document BUT
  - One human translator will translate the document significantly differently than another human translator

What do we Expect from MT?

- **Adequacy** and **informativeness**: preserve meaning
- **Fluency**: how close the generated translation resembles one that would be created by a fluent speaker of the TL
- **Grammaticality**
- Evaluation is difficult!
  - What is the best translation? (language variation!)
  - Subjective aspects (What is "fluent"? Clarity? Style?)
  - What is "grammatical"?
  - What is "adequate"? (Is it possible to be adequate?)

(SOURCE: TIEDEMANN, J. (2009), "MACHINE TRANSLATION - RULE-BASED MT AND MT EVALUATION")

Manual Evaluation

Compare MT engines:

- Ask actual users to rate translations
  - Human judgements: adequacy and fluency
- Statistics over user responses
- Separate evaluations of adequacy and fluency
- Requires guidelines
- Task specific evaluation

(SOURCE: TIEDEMANN, J. (2009), "MACHINE TRANSLATION - RULE-BASED MT AND MT EVALUATION")

Automatic Evaluation

- Comparison of MT output with reference translations
- Approximations by measuring overlaps
- Strong bias but useful for rapid development
- Many metric measures: BLEU, NIST, METEOR, WER, PER, TER, ROUGE, precision, recall, etc.

(Source: Tiedemann, J. (2009), "Machine Translation - Rule-based MT and MT evaluation")

Current Status

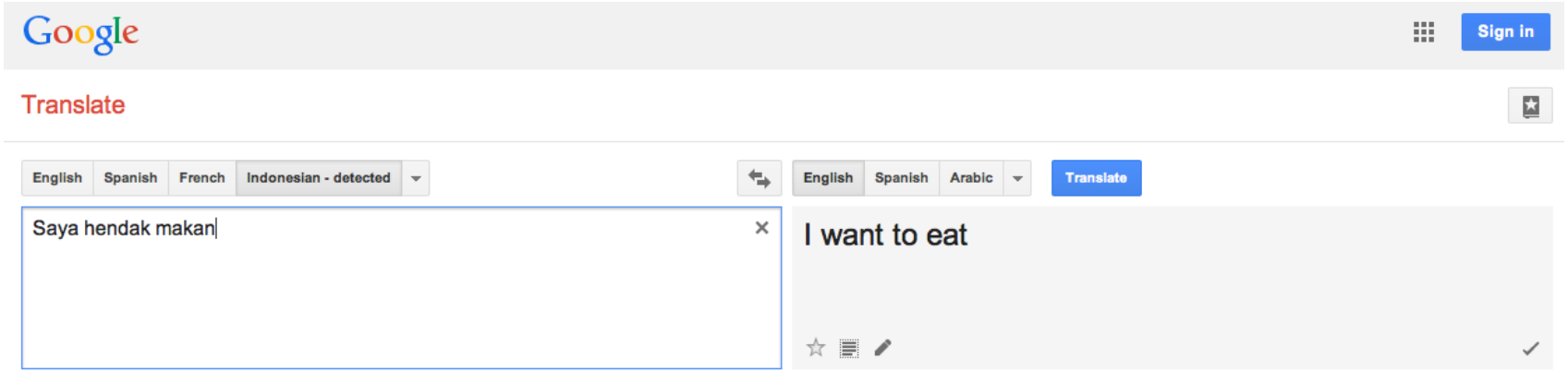
- Broad coverage systems already available via the Web
- Fast turnaround, acceptable error rate for gisting
- Dominant systems are now statistical (e.g. Google's)
- Higher accuracy can be achieved by carefully domain-targetted systems

New Directions

- Spoken language translation for general-purpose
- Translation for minority and under-resourced languages
- Systems for monolinguals: from unknown source language to unknown target language
- Improvement expectations (particularly PC commercial and Internet systems)
- Reusability of resources (particularly dictionaries and translation memories)
- Integration
  - MT as option in word processing packages, on Web pages
  - MT as option with summarisation, information extraction, information retrieval, data retrieval, question-answering, Internet search tools

Tool: Google Translate

- A free online translator
- Developed by GOOGLE
- Supports 80 languages world wide
- Google Translate is able to translate words, sentences and web pages
- You can try ▶ Google



Tool: SiSTeC–ebmt

- SiSTeC stands for SiStem Terjemahan berasaskan SSTC (SSTC-based Translation System)
- SiSTeC-ebmt is an example-based machine translation (EBMT) engine
- SiSTeC-ebmt ables to translate language pairs include English-Malay (bidirectional) and English-Chinese
- For further information, go to SiSTeC-ebmt

References

Jurafsky, D., Martin, J. H. (2008). Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence). Prentice Hall. ISBN: 0131873210.

Al-Adhaileh, M., Tang E.K, & Zaharin, Y. (2002). A Synchronization Structure of SSTC and Its Applications in Machine Translation. The COLING 2002 Post-Conference Workshop on Machine Translation in Asia, Taipei, Taiwan.

Callison-BurchKoehn, C. & P. et al. (2008). Introduction to Statistical Machine Translation. MT Marathon, EuroMatrix project, 2008.

A. Eisele, (2007). Hybrid machine translation translation: Combining rule rule-based and statistical MT systems. 1st MT Marathon, Edinburgh

Tiedemann, J. (2009). Machine Translation - Rule-based MT and MT evaluation.