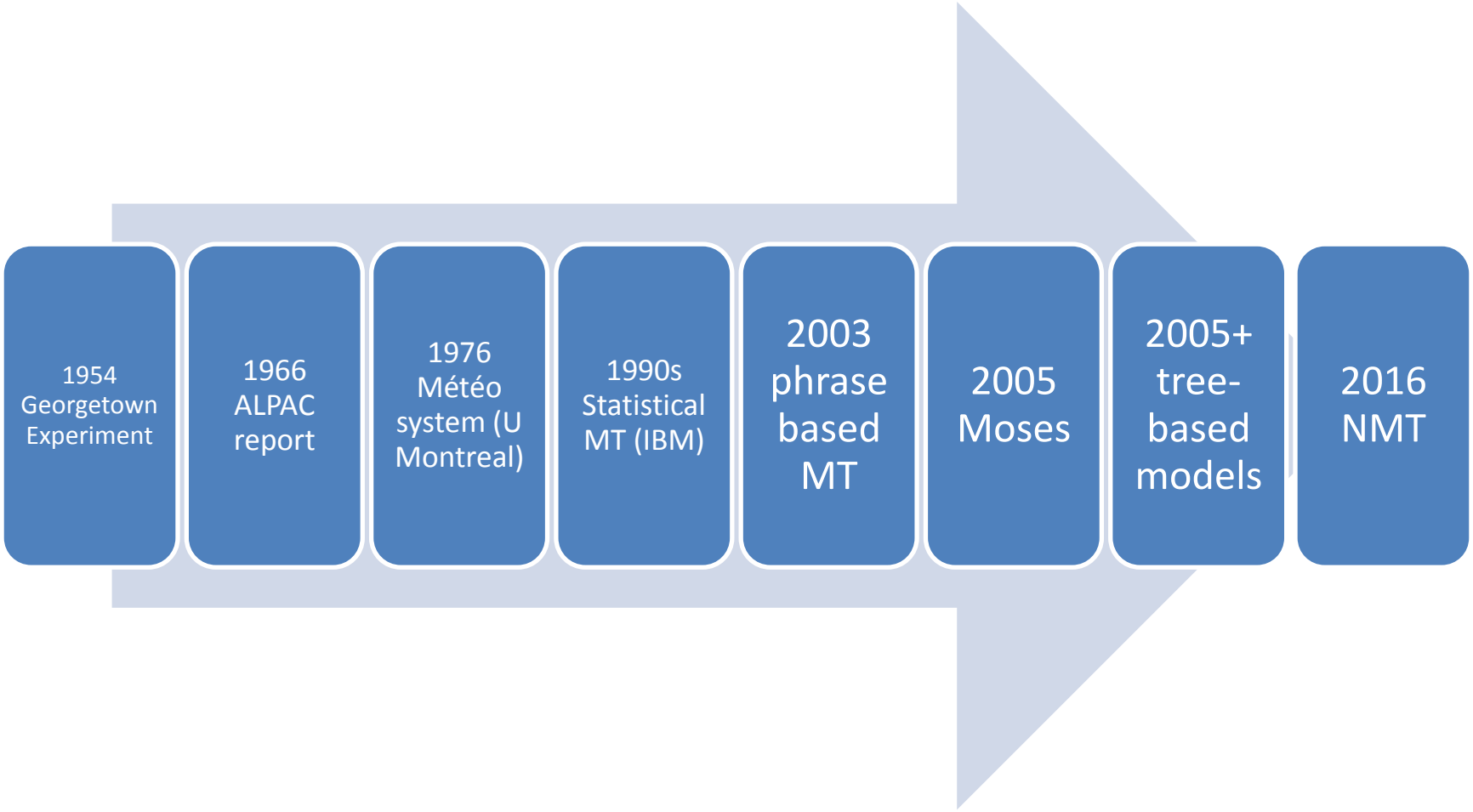
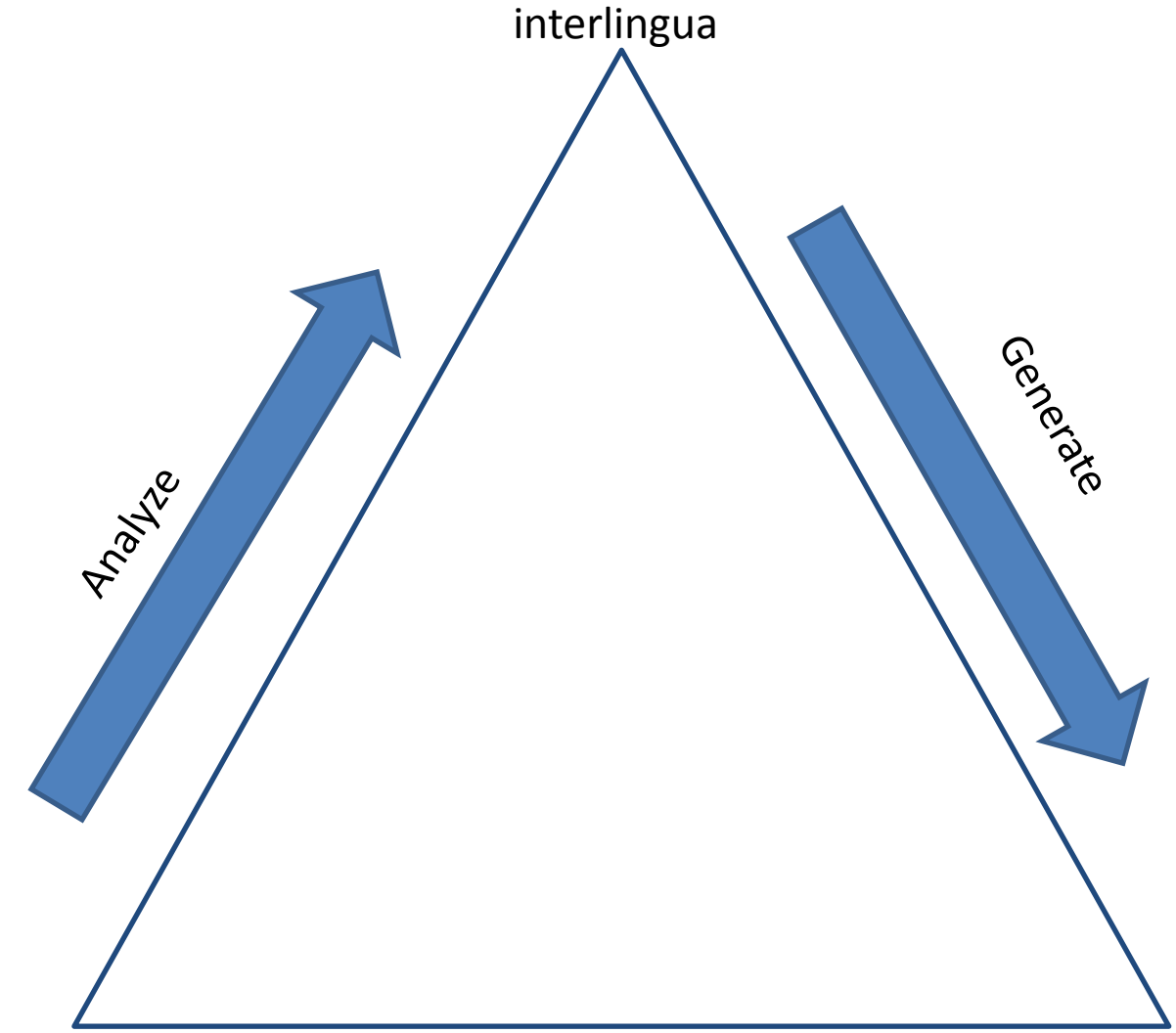
Lecture 4.1: Statistical machine translation, different approaches

SMT: Introduction, history / various approaches

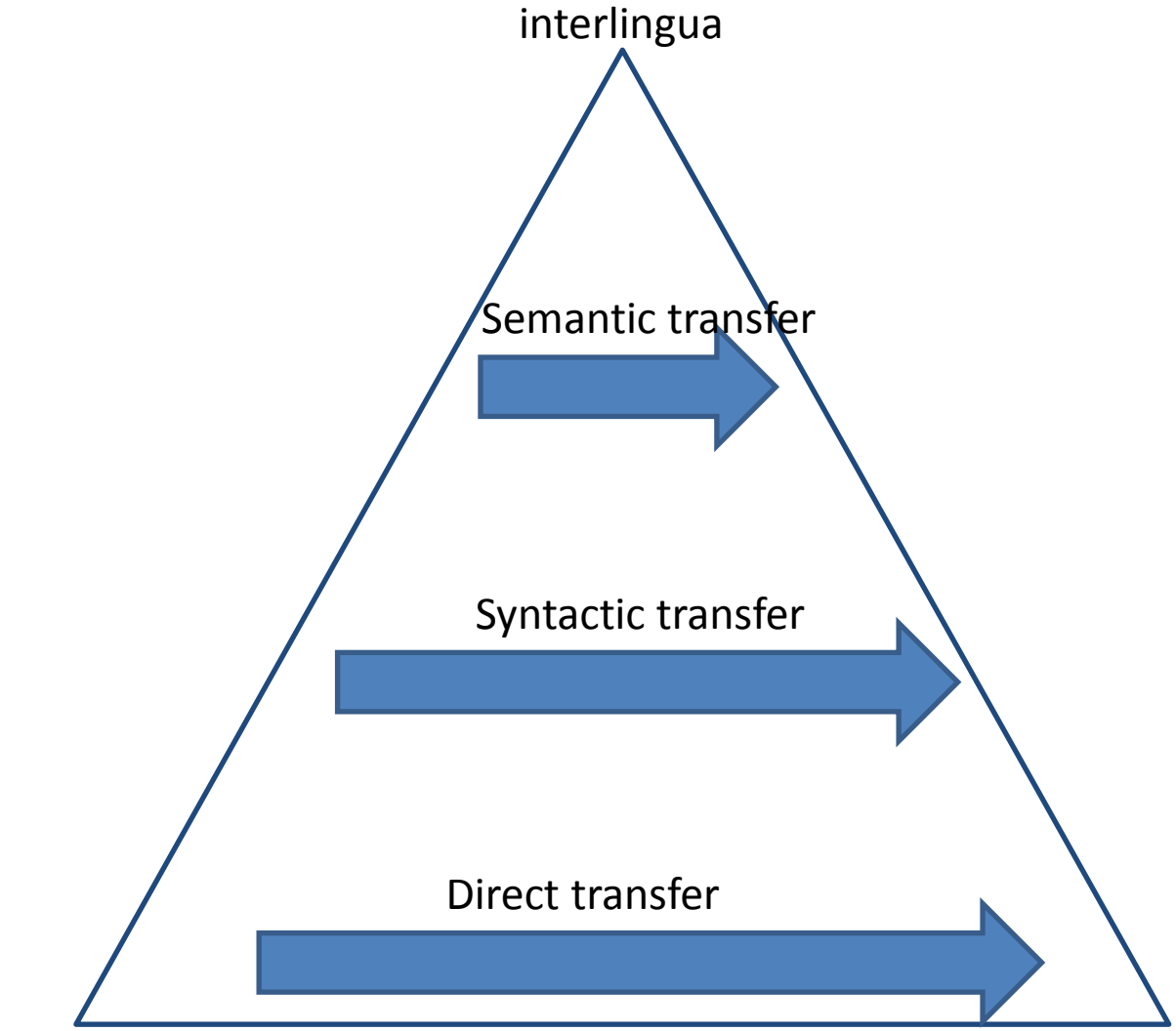
Machine translation history



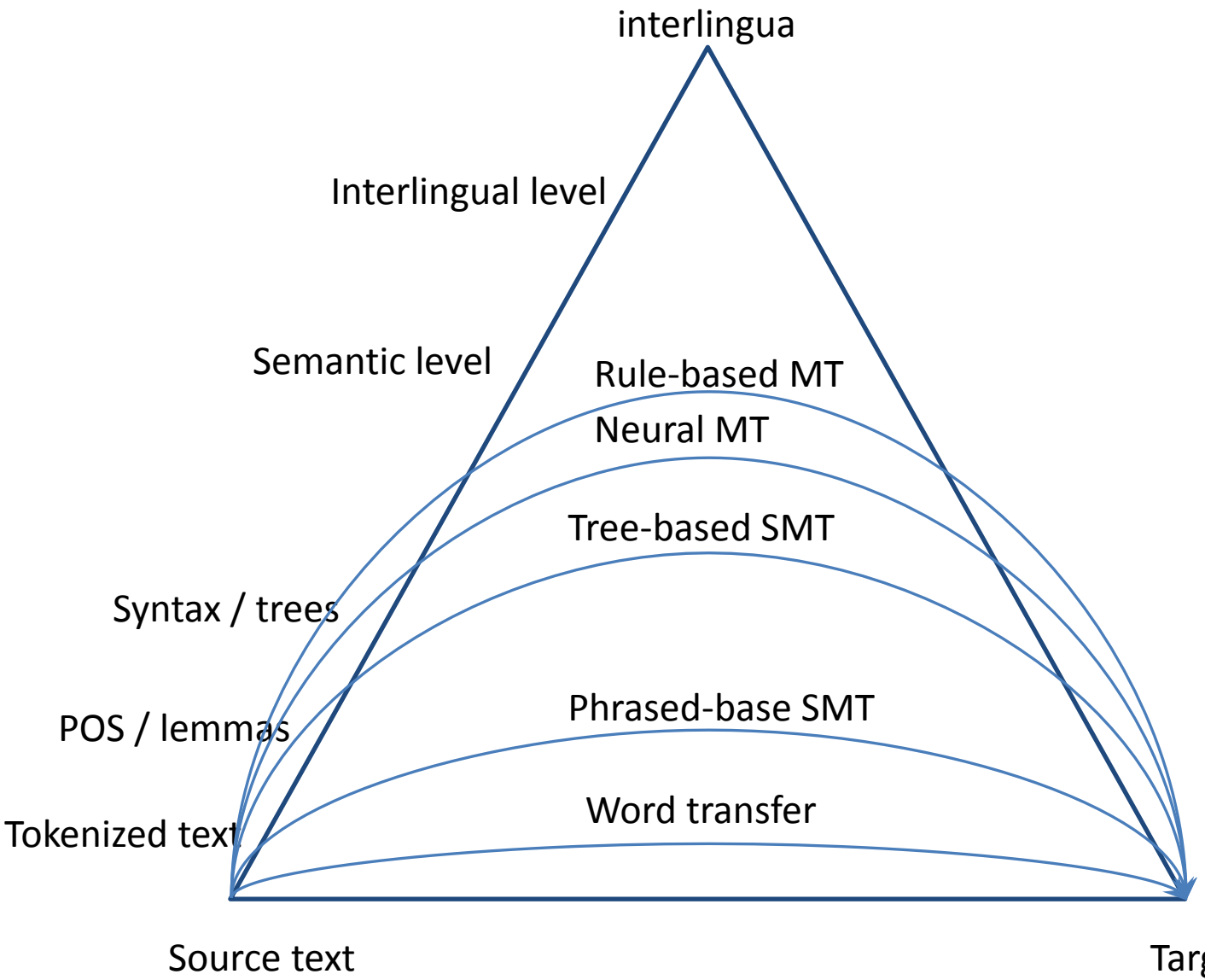
The dream of interlingua



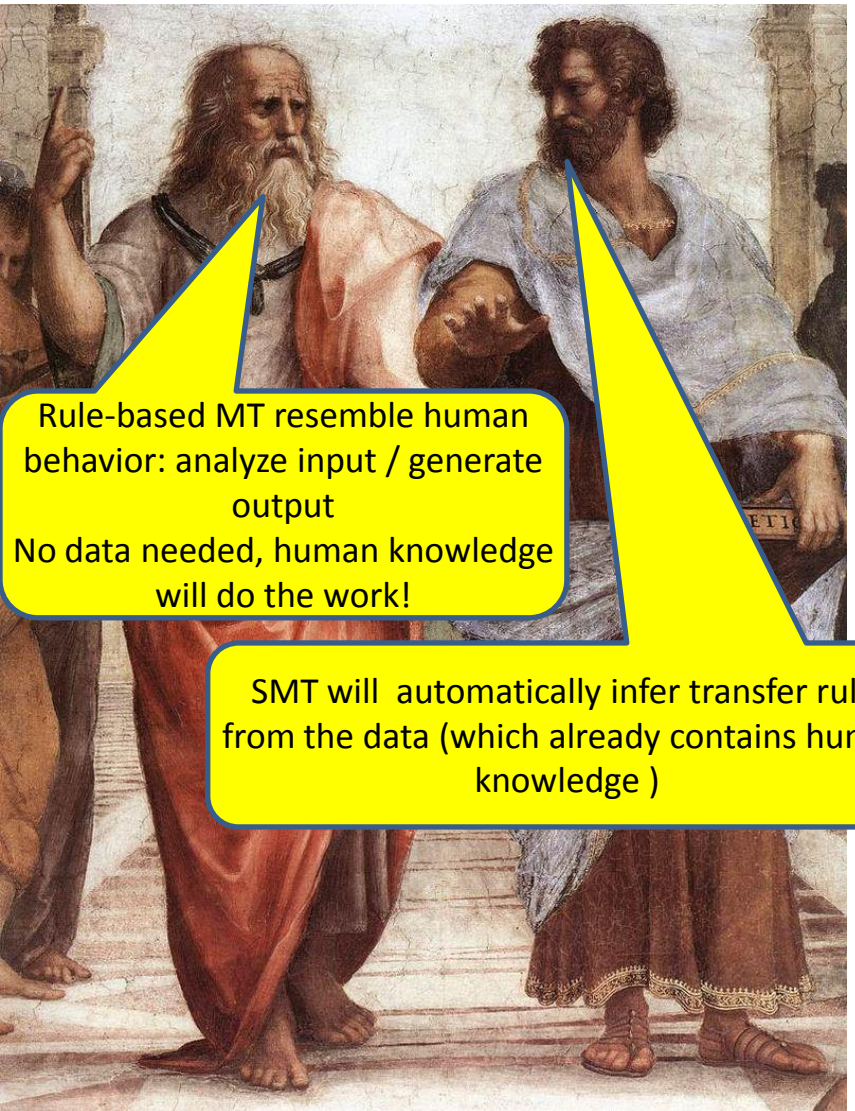
Statistical Machine Translation systems



Statistical Machine Translation systems



Two main families: rule-based & Statistical-based



Rule-based MT: the "classical" approach, translators/linguists build dictionaries and rules.

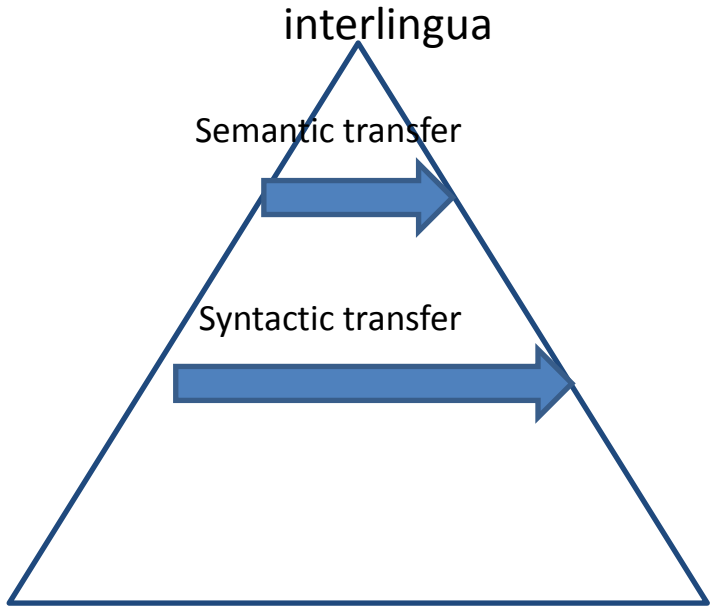
➔Top-down approach

Statistical-based MT: the "data-driven" approach, no rule, no dictionaries, only statistics on sentences.

➔Bottom-up approach

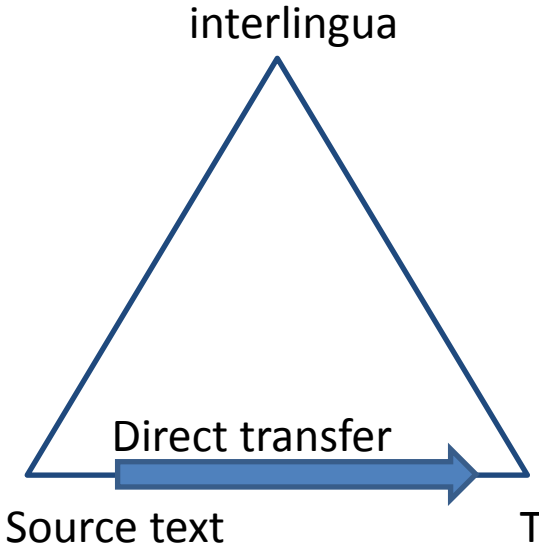
RBMT

- *Theory behind rule-based MT (RBMT)*
  - *An exhaustive bilingual dictionary*
  - *Built-in linguistic rules*



SMT

- *Theory behind statistical machine translation (SMT), empirical translation*
  - *Use a lot of "training" examples*
  - *Use machine learning (statistics / neural networks) to transfer from source to target*



SMT: various approaches

- Word-based (not used)
- Example-based (deprecated)
- Phrase-based
- Tree-based (syntax-based)
- Neural networks

Which method to choose?

Depends on many criteria

- Your background
- Your resources
- Your goal/scope

Choice depending on your background

- Linguists love rule-based (or syntax-based)
- Translators may prefer example-based
- Computer scientists may love parsing/interlingua/generation
- Statisticians love SMT
- Mathematicians love Neural MT

Depends on your resources

- Human resources:
  - Many employees to write rules, update dictionaries
  - … or to create parallel texts for SMT
- Data:
  - Is your language pair having a lot of parallel texts?
  - Do you have technical tools? Tokenizer / part-of-speech-tagger / parser etc.

Various corpora of parallel texts (JRC Acquis, Europarl, UN corpus, Indic corpora…)

Choice depends on your goal(s)

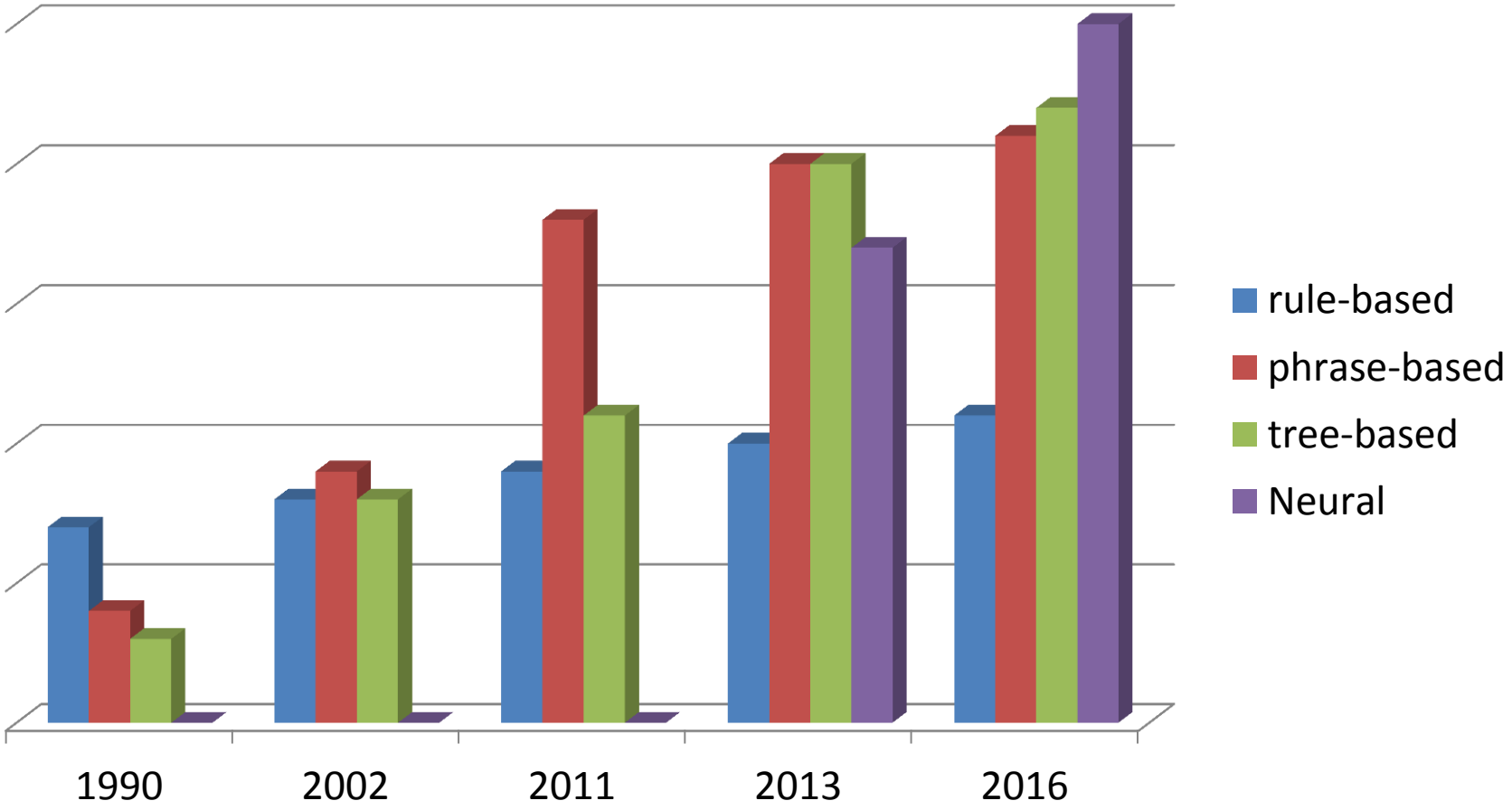
- You plan to build MT for:
  - Assimilation? (gist translation)
  - Dissemination? (MT as a support for translators to produce publishable quality)
  - One language pair only? More languages?
  - On a general or specific domain?

Pros and cons

MT	Advantages	Disadvantages
Rule-based(*)	• Based on linguistic theories • Adequate for languages with limited resources • Does not require many computational resources • Easy to perform error analysis	• Requires linguistic rules and dictionaries • Human Language Inconsistency (i.e. exceptions) • Disambiguation problems • Local translations • Language dependent • Expensive to maintain and extend
PBSMT(*)	• No linguistic knowledge required • Reduces the human resources cost • Easy to build • Easy to maintain • Trained with human translations Independent from the pair of languages	• Requires parallel text • Requires quite high computational resources • Difficult to perform error analysis • Problems languages pairs with different morphology/order • No linguistic background • Model size can be huge (100Gb)
Neural MT	• No linguistic knowledge required • Reduces the human resources cost • Easy to build *(with GPUs)* • Easy to maintain and to incrementally extend • Handles long-distant reordering & word agreement • Trained with human translations Independent from the pair of languages • Model size is small (100Mb) • Active research area (e.g. models from/to many languages)	• Requires parallel text • Requires high and specific computational resources (GPU) to train • Almost impossible to perform error analysis • No linguistic background • Problems with unknown words

(*) Based on [Costa-Jussa et al. 2012]

General trends about MT



Please note: it highly depends on the available resources and the language pairs.
E.g. Japanese-English rule-based system are often better than phrase-based, while English-Spanish are better with PBSMT
For languages that do not have enough training data: RBSMT is still a better option

Combining various approaches?

- Usually a good idea… but how?
- Multi-engine (launch 2 engines in parallel)
  - Decide the best

  or
  - Combine results
- Multi-pass (cascade various MT)
  - Guided by RBMT

  or
  - Guided by corpus-based MT (SMT or example-based)

Challenges in MT

1) Word order (e.g. translating between SVO to SOV)

2) Word sense ambiguity (Hindi "कल-kal")

3) Anaphora (zero-pronoun anaphora, pro-drop)

4) Handling agreement (subject-verb / adjective-noun) – long distance dependencies

5) idiomatic expressions

Hindi word "kal" कल is yesterday or tomorrow?

kal main dukaan jaegaa (कल मैं दुकान जाएगा)
[tomorrow] [I] [shop] [will-go]
kal main dukaan gaya tha (कल मैं दुकान गया था)
[yesterday] [I] [shop] [go] [was]

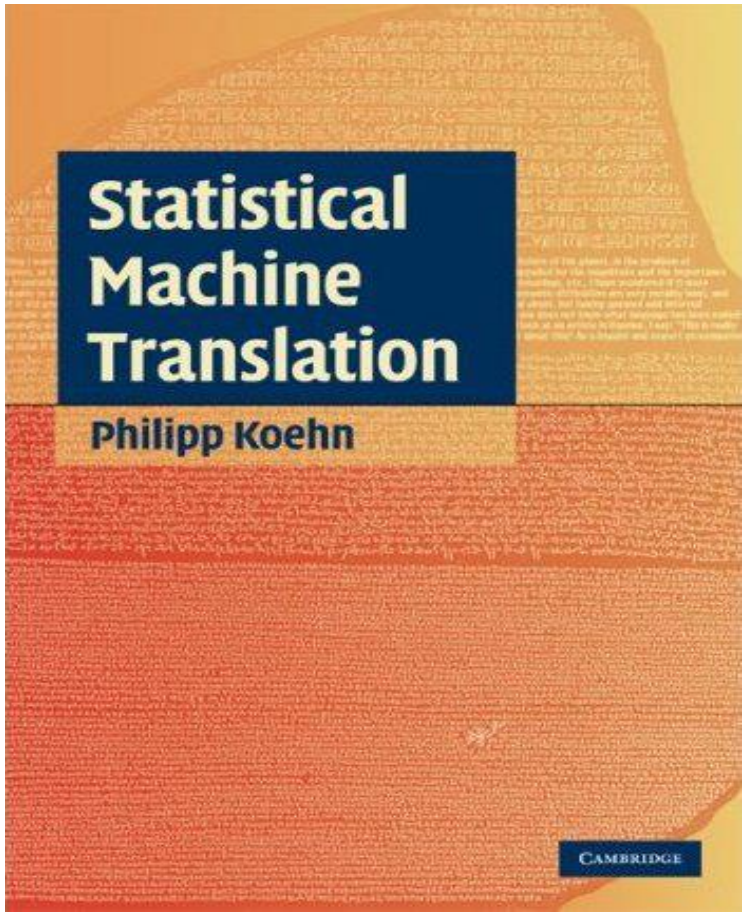
A glass on a table with water in it. It falls down.
Fr: Un verre (masc) sur la table (fem) avec de l'eau dans **celui/celle**-ci. **Il/Elle** tomba.
Es: Un vaso (masc) sobre una mesa (fem) con agua en **ella/él**. Se cayó.

Challenges in the usage of MT

news.bbc.co.uk/2/hi/7702913.stm

Irish says “I am not in the office at the moment. Send any work to be translated.”

More about SMT…



http://www.statmt.org/book/