Approaches to Machine Translation: Rule-based, Statistical and Hybrid

MT Approaches

Approaches to Machine Translation

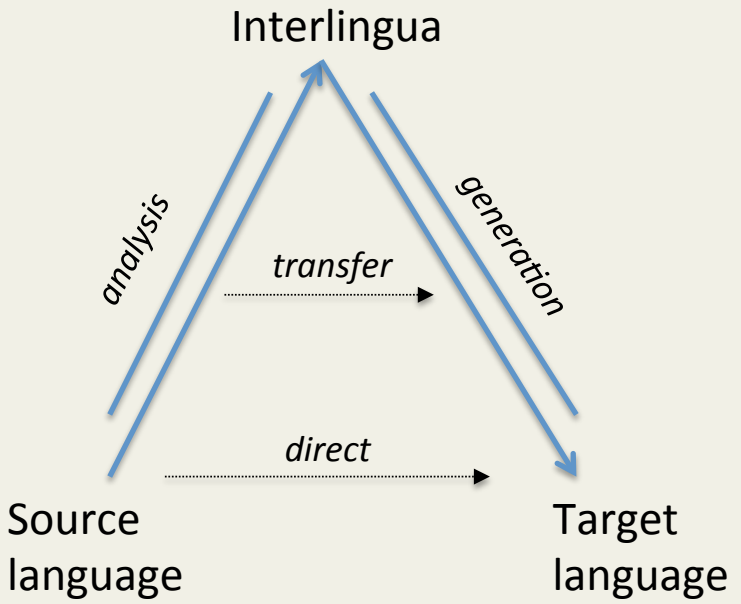
Sources of information

Rule-based: human written specific rules

Data-driven: use data to learn

Level of representation

Transfer-based MT has several depths of intermediary representation



Question

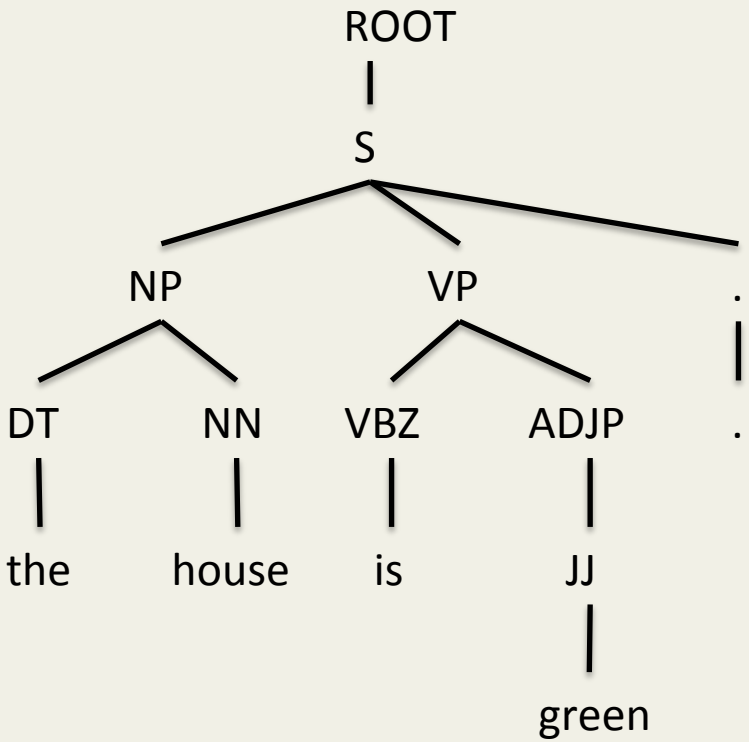
- Singapore has 4 official languages: English, Chinese, Malay and Tamil. The government wants to build MT systems in all directions. Choose all answers that match.
  - An interlingua approach would require to build 8 MT systems
  - A transfer-based approach would require to build 8 MT systems
  - A direct approach would require to build 12 MT systems

Question

- Singapore has 4 official languages: English, Chinese, Malay and Tamil. The government wants to build MT systems in all directions. Choose all answers that match.
  - **An interlingua approach would require to build 8 MT systems**
  - A transfer-based approach would require to build 8 MT systems
  - A direct approach would require to build 12 MT systems

Rule-based Machine Translation

- Resources:
  - Morphological dictionaries
  - Source analyzer
  - Translation lexicon
  - Transfer rules



Rule-based Free/Open Source Platform

Corpus-based MT are trained on parallel corpora

Collections of parallel texts at sentence level

English	Russian
This course is a thorough introduction to machine translation technology	Этот курс представляет собой интенсивное введение в технологию машинного перевода
We will describe all aspects of building a statistical machine translation system, from both formal and practical perspectives	Мы рассмотрим все аспекты построения системы статистического машинного перевода с теоретической и практической точки зрения

An early parallel text

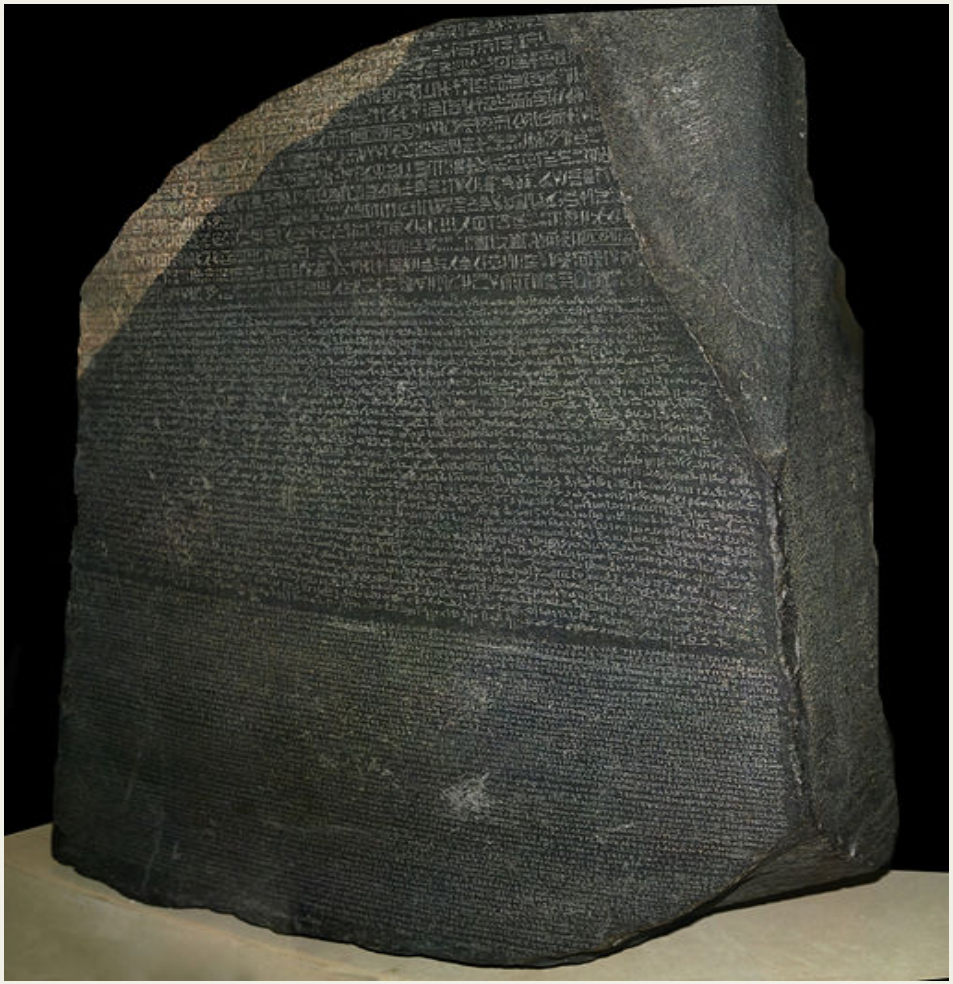
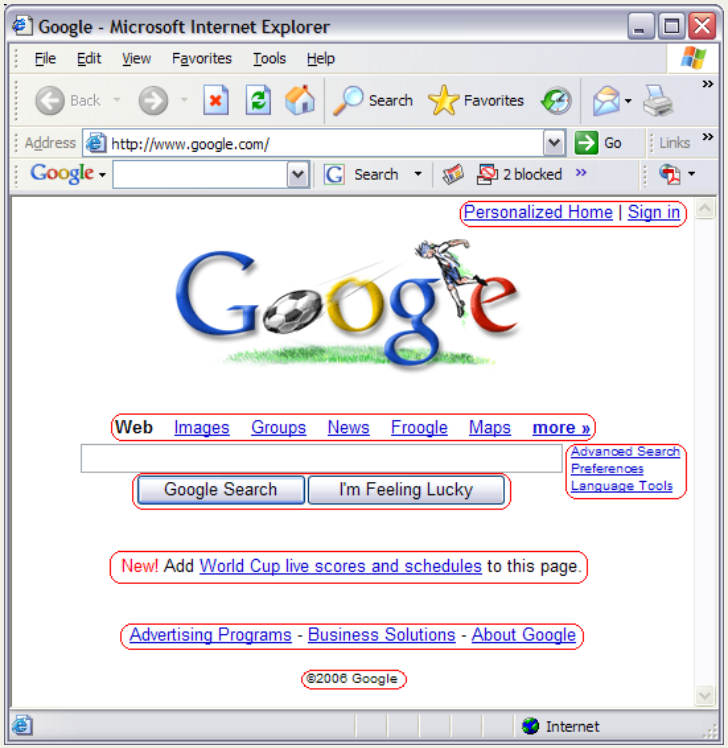


IMAGE SOURCE: WIKIPEDIA

Parallel vs Comaprable Text on the Web

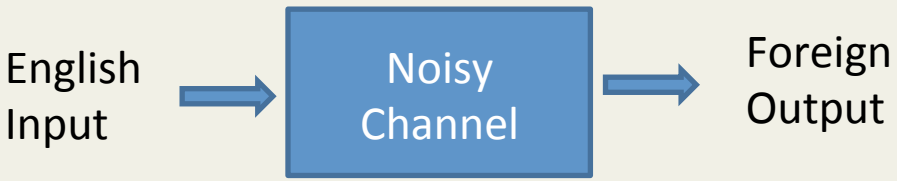


Corpus-based MT approaches

- Example-based:

  translation by analogy
- Statistical-based:

  translation generated on the basis of statistical models



Example-Based Machine Translation

- Simplest case
  - Sentence to be translated matches previously seen sentence
  - Same as 100% translation memory match
- Pattern recognition

English	Japanese
How much is that **red umbrella**?	Ano **akai kasa** wa ikura desu ka.
How much is that **small camera**?	Ano **chiisai kamera** wa ikura desu ka.

Question

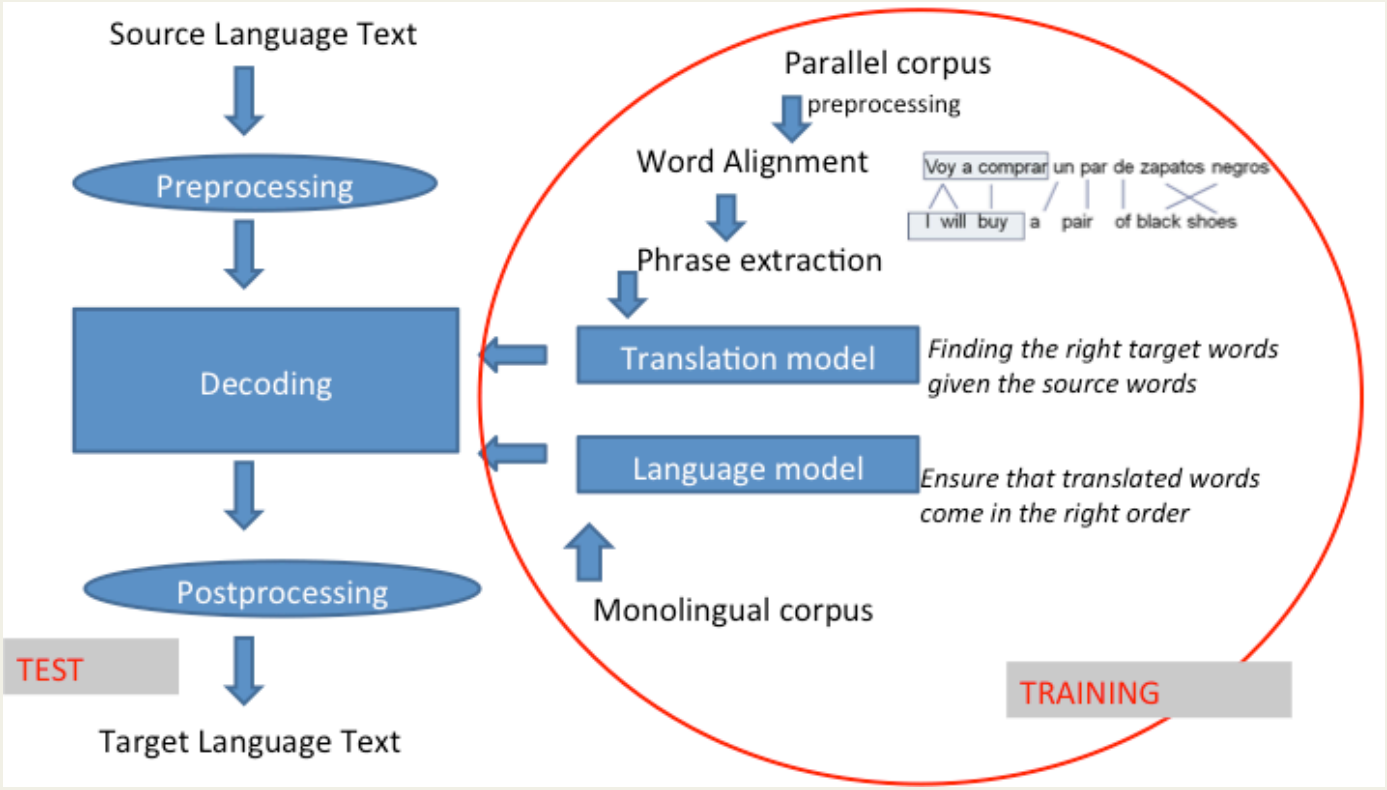
- Choose all properties that match to statistical machine translation
  - Language independent
  - No data needed
  - Difficult to deploy
  - Good knowledge of the language involved

Question

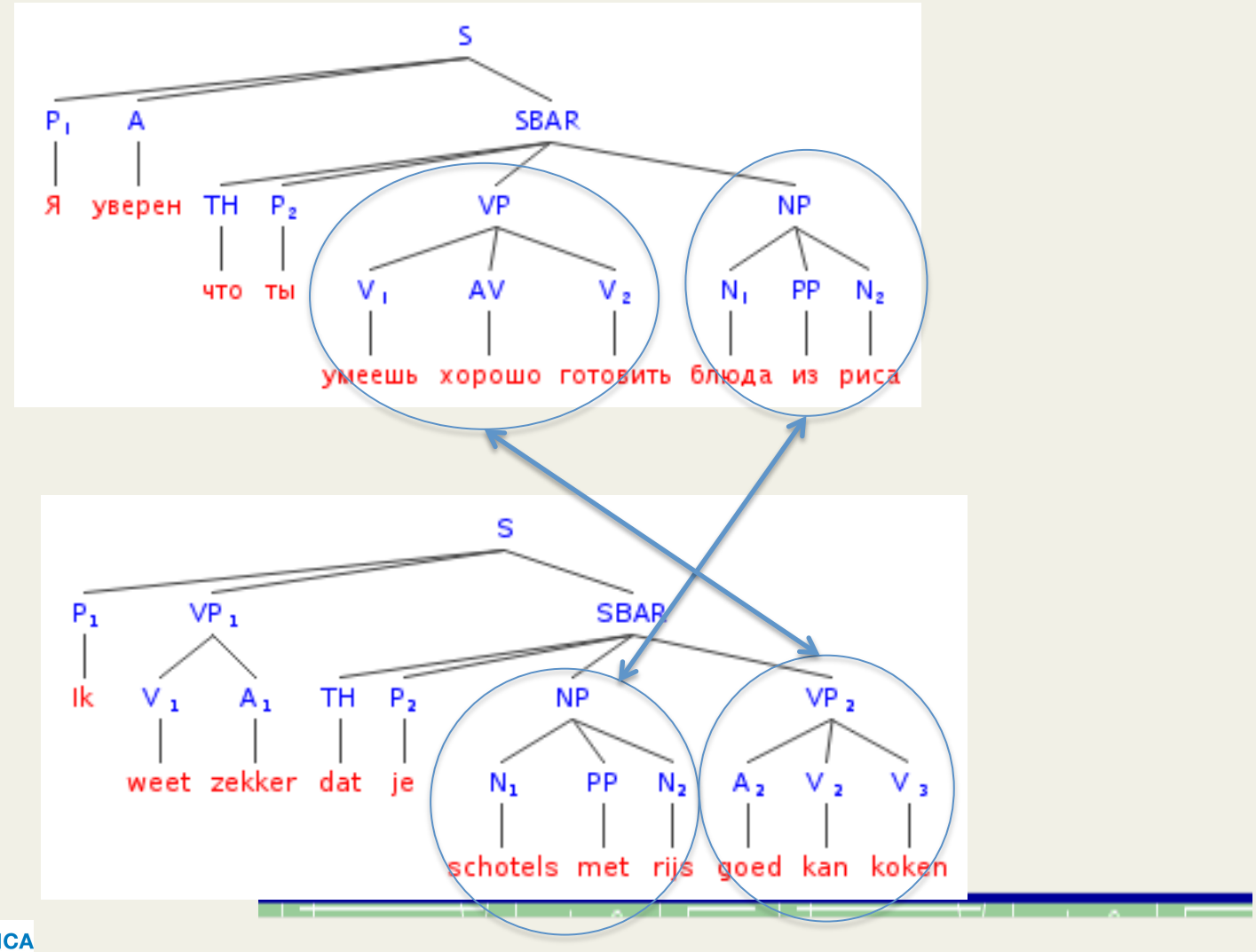
- Choose all properties that match to statistical machine translation
  - **Language independent**
  - No data needed
  - Difficult to deploy
  - Good knowledge of the language involved

https://www.youtube.com/watch?feature=player_detailpage&v=_GdSC1Z1Kzs

A picture is worth a million equations



Syntax Augmented by parse trees



Hierarchical-based introduce hierarchical rules

- Hierarchical rules allow for hierarchical phrases that can contain other phrases
  - [Я] [уверен] [что] [ты] [хорошо умеешь готовить] [блюда с рисом]
  - [Ik] [weet zeker] [dat] [je] [schotels met rijst] [goed kan koken]
  - [ты][X][блюда с рисом] - > [je][schotels met rijst][X]

The parallel copus is the main required resource for SMT

Free parallel corpus:

- EPPS
- JRC-Acquis
- UN data
- Canadian Hansards

Others:

TAUS data

LDC data

Popular SMT software

- Word alignment GIZA++, Berkeley
- Language modeling: SRILM, IRSTLM
- Phrase extraction: THOT, Moses
- Decoder: Moses

Advantages of SMT

- Data driven
- Language independent
- No need for staff of language experts
- Can prototype a new system quickly and at a very low cost
- High coverage and flexibility of matching heuristics

Example of online SMT system: GOOGLE, YANDEX

Next

SMT CHALLENGES