Introduction to machine translation

Machine Translation

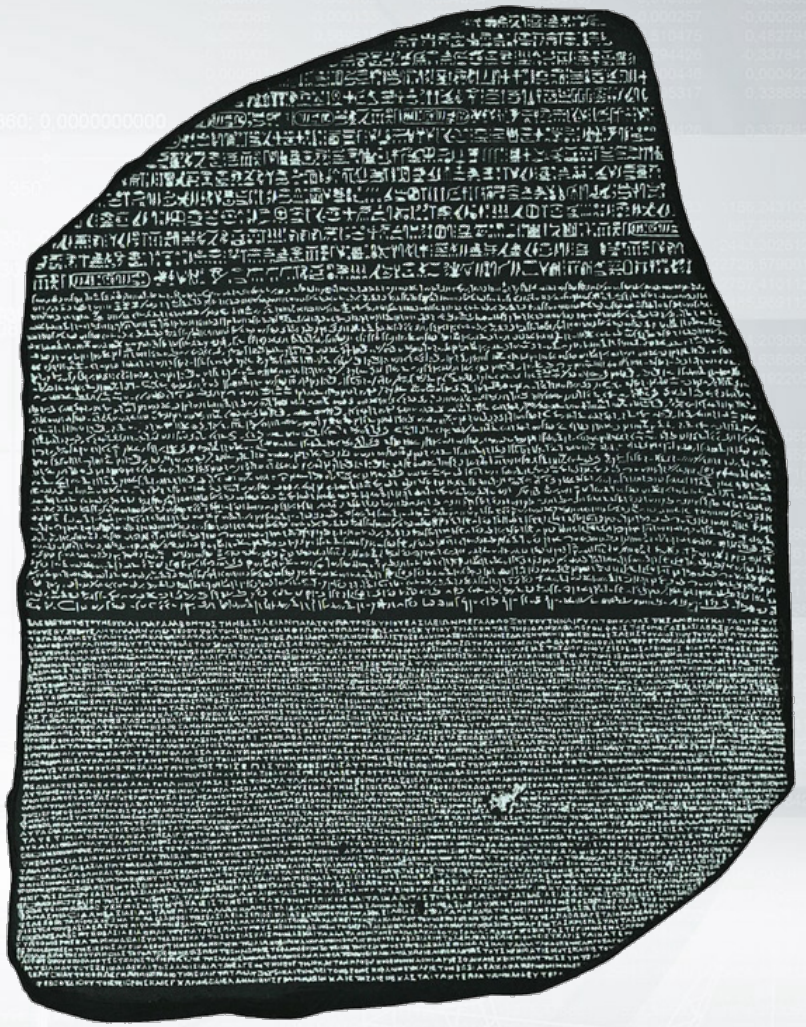
Parallel data

Parallel corpora:

- Europarl
- Movie subtitles
- Translated news, books
- Wikipedia (comparable)
- http://opus.lingfil.uu.se/

Lot's of problems with data:

- Noisy
- Specific domain
- Rare language pairs
- Not aligned, not enough



Evaluation

- How to compare two arbitrary translations?
- Low agreement rate even between reviewers
- BLEU score – a popular automatic technique

Evaluation

- How to compare two arbitrary translations?
- Low agreement rate even between reviewers
- BLEU score – a popular automatic technique

Reference: ***E-mail was sent on Tuesday.***

System output: ***The letter was sent on Tuesday.***

Evaluation

- How to compare two arbitrary translations?
- Low agreement rate even between reviewers
- BLEU score – a popular automatic technique

Reference: ***E-mail was sent on Tuesday.***

System output: ***The letter was sent on Tuesday.***

1-grams: 4 / 6

Evaluation

- How to compare two arbitrary translations?
- Low agreement rate even between reviewers
- BLEU score – a popular automatic technique

Reference: ***E-mail was sent on Tuesday.***

System output: ***The letter was sent on Tuesday.***

1-grams: 4 / 6

2-grams: 3 / 5

Evaluation

- How to compare two arbitrary translations?
- Low agreement rate even between reviewers
- BLEU score – a popular automatic technique

Reference: ***E-mail was sent on Tuesday.***

System output: ***The letter was sent on Tuesday.***

1-grams: 4 / 6

2-grams: 3 / 5

3-grams: 2 / 4

4-grams: 1 / 3

Evaluation

- How to compare two arbitrary translations?
- Low agreement rate even between reviewers
- BLEU score – a popular automatic technique

Reference: ***E-mail was sent on Tuesday.***

System output: ***The letter was sent on Tuesday.***

1-grams: 4 / 6

2-grams: 3 / 5

3-grams: 2 / 4

4-grams: 1 / 3

Brevity: min(1, 6/5)

Evaluation

- How to compare two arbitrary translations?
- Low agreement rate even between reviewers
- BLEU score – a popular automatic technique

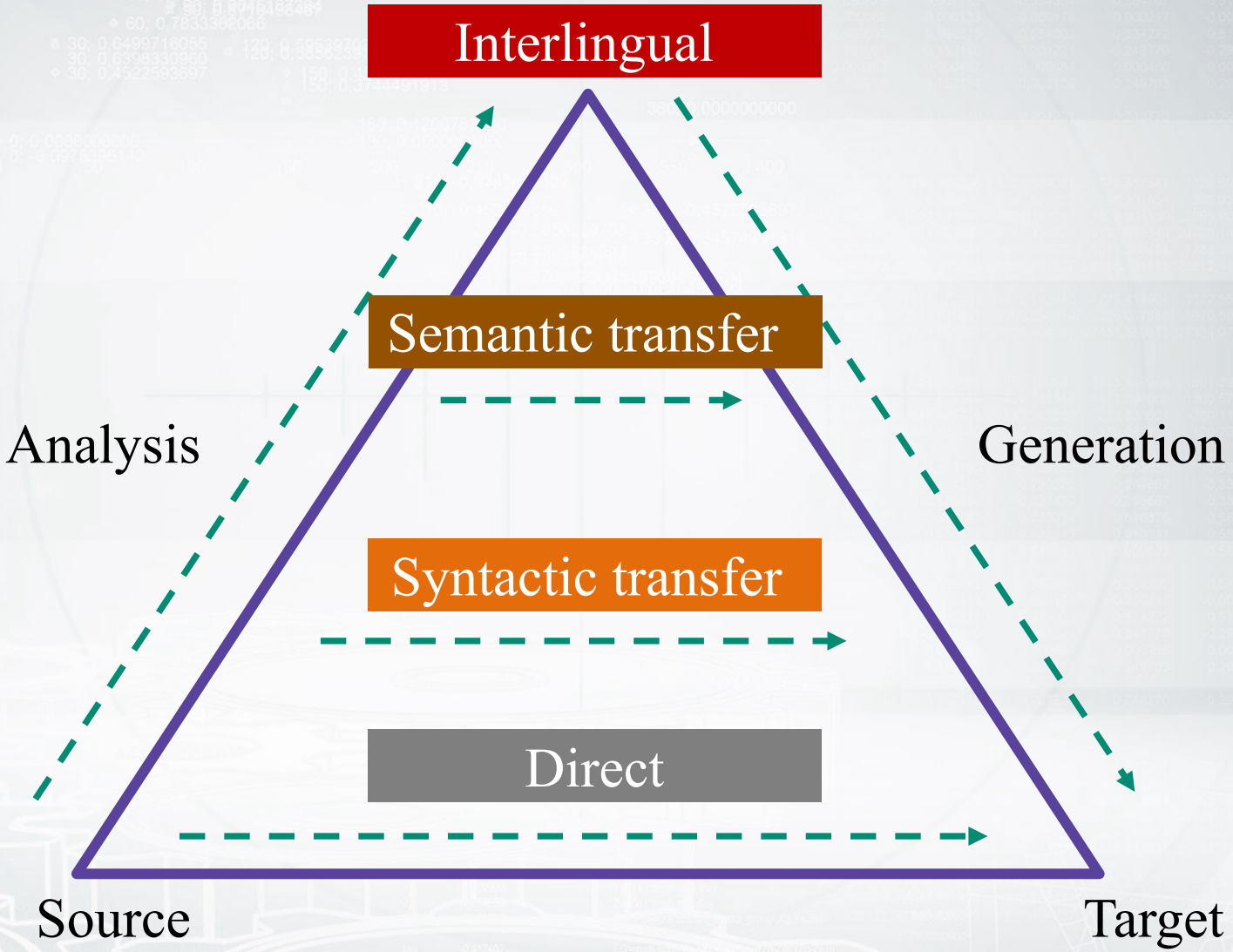
Reference: ***E-mail was sent on Tuesday.***

System output: ***The letter was sent on Tuesday.***

1-grams: 4 / 6
2-grams: 3 / 5
3-grams: 2 / 4
4-grams: 1 / 3
Brevity: min(1, 6/5)

$$\mathrm{BLEU} = 1 \cdot \sqrt[4]{\frac{4}{6} \cdot \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3}}$$

The mandatory slide



Henry S. Thompson, https://www.inf.ed.ac.uk/teaching/courses/anlp/lectures/30/

Roller-coaster of machine translation

1954 Georgetown IBM experiment Russian to English:

- Claimed that MT would be solved **within 3-5 years.**

1966 ALPAC report:

- Concluded that MT was **too expensive and ineffective.**

Two main paradigms

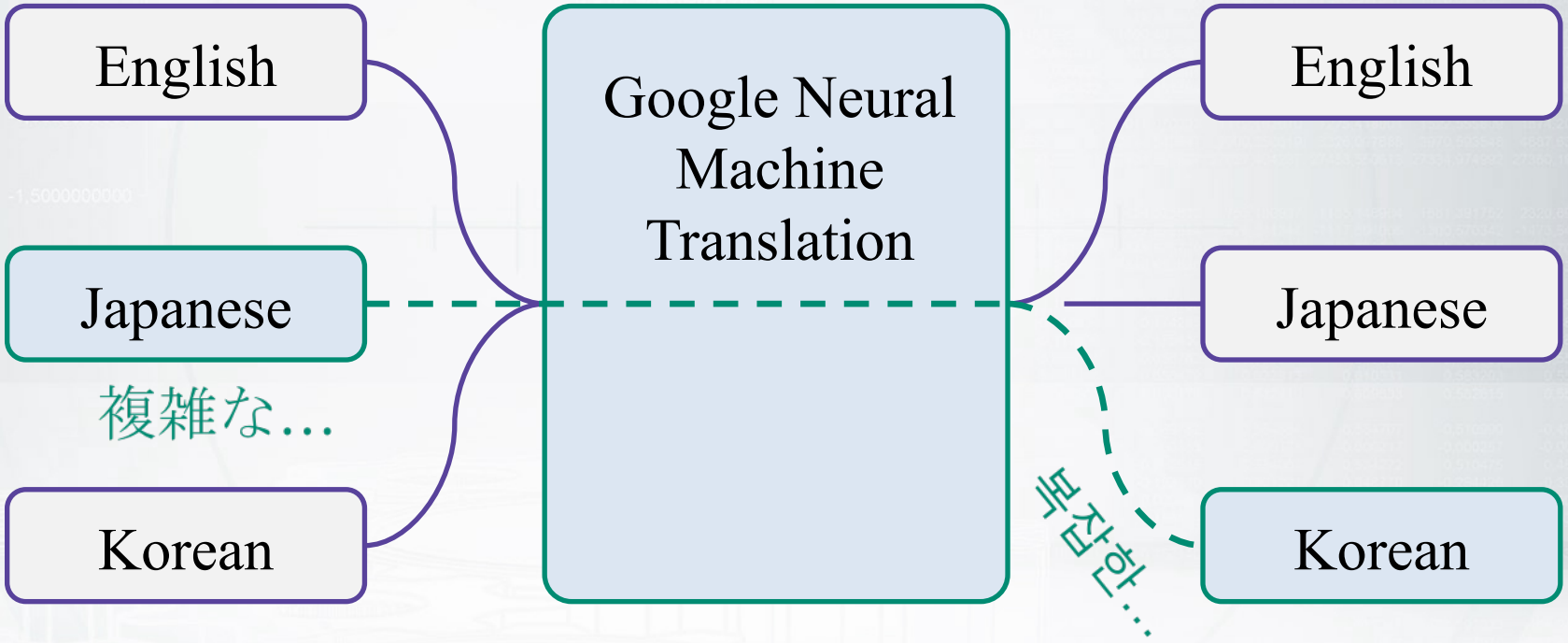
Statistical Machine Translation (SMT):

- 1988 – Word-based models (IBM models)
- 2003 – Phrase-based models (Philip Koehn)
- 2006 – Google Translate (and Moses, next year)

Neural Machine Translation (NMT):

- 2013 – First papers on pure NMT
- 2015 – NMT enters shared tasks (WMT, IWSLT)
- 2016 – Launched in production in companies

Zero-shot translation



https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html