COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Machine Translation
History & Evaluation

CMSC 470

Marine Carpuat

Today's topics
Machine Translation

- Context: Historical Background
  - Machine Translation is an old idea

- Machine Translation Evaluation

1947

When I look at an article in Russian, I say to myself: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.

Warren Weaver

1950s-1960s

- 1954 Georgetown-IBM experiment
  - 250 words, 6 grammar rules
- 1966 ALPAC report
  - Skeptical in research progress
  - Led to decreased US government funding for MT

Rule based systems

- Approach
  - Build dictionaries
  - Write transformation rules
  - Refine, refine, refine
- Meteo system for weather forecasts (1976)
- Systran (1968), …

```
"have" :=

if
    subject(animate)
    and object(owned-by-subject)
then
    translate to "kade...  aahe"
if
    subject(animate)
    and object(kinship-with-subject)
then
    translate to "laa...  aahe"
if
    subject(inanimate)
then
            translate to "madhye...
aahe"
```

1988

A Statistical Approach to Machine Translation

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek,
John D. Lafferty, Robert L. Mercer, and Paul S. Roossin

IBM
Thomas J. Watson Research Center
Yorktown Heights, NY

In this paper, we present a statistical approach to machine translation. We describe the application of our approach to translation from French to English and give preliminary results.

The COLING Paper Review

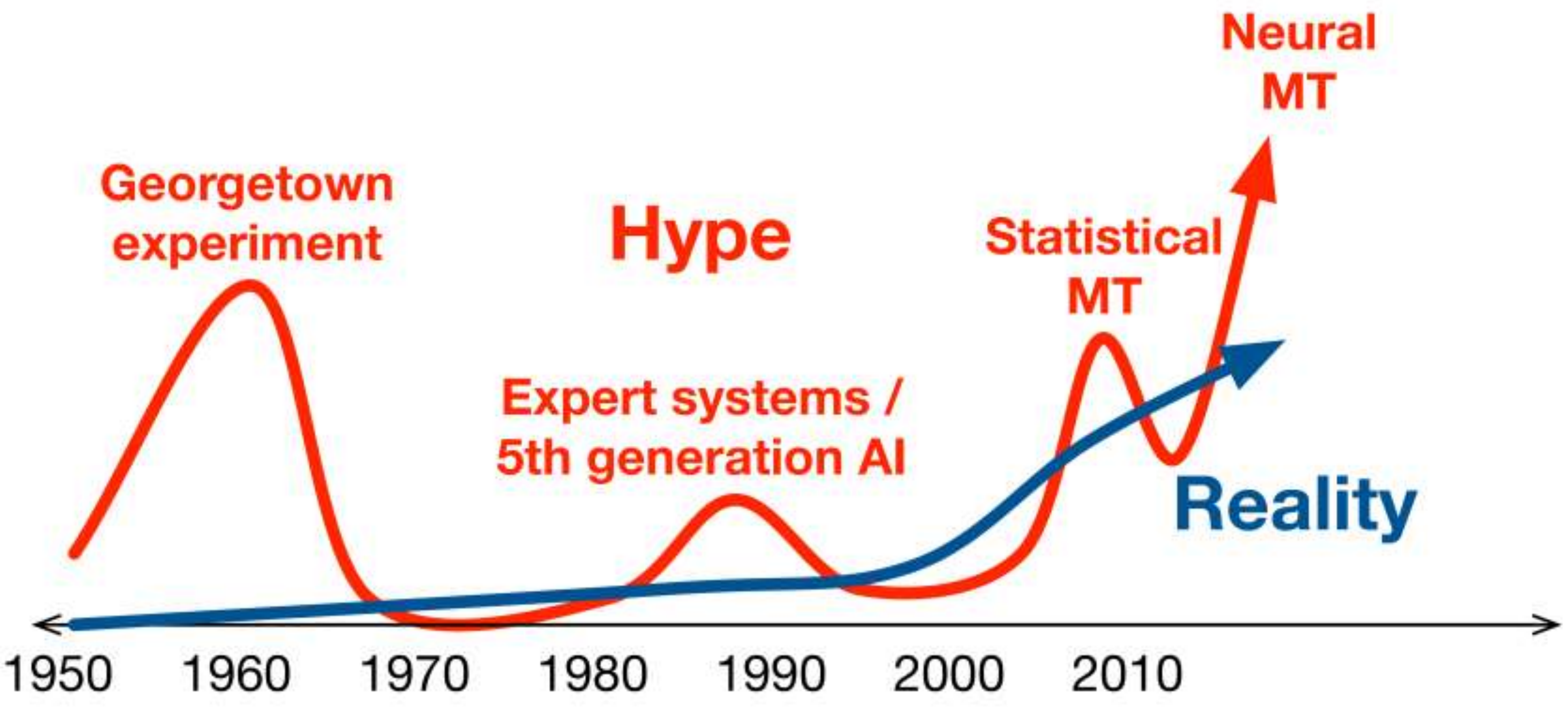
The validity of statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950. (cf. Hutchins, MT: Past, Present, Future, Ellis Horwood, 1986, pp. 30ff. and references therein) The crude force of computers is not science. The paper is simply beyond the scope of COLING.

More about the IBM story: [20 years of bitext workshop]

Statistical Machine Translation

- 1990s: increased research
- Mid 2000s: phrase-based MT
  - (Moses, Google Translate)
- Around 2010: commercial viability
- Since mid 2010s: neural network models

MT History: Hype vs. Reality



How Good is Machine Translation Today?

March 14 2018:

“Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English”

https://techcrunch.com/2018/03/14/microsoft-announces-breakthrough-in-chinese-to-english-machine-translation/

But also

https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427

How Good is Machine Translation Today? Output of Research Systems at WMT18

上周，**古装**剧《**美人私房菜**》临时停播，意外引发了关于国产剧收视率造假的热烈讨论。

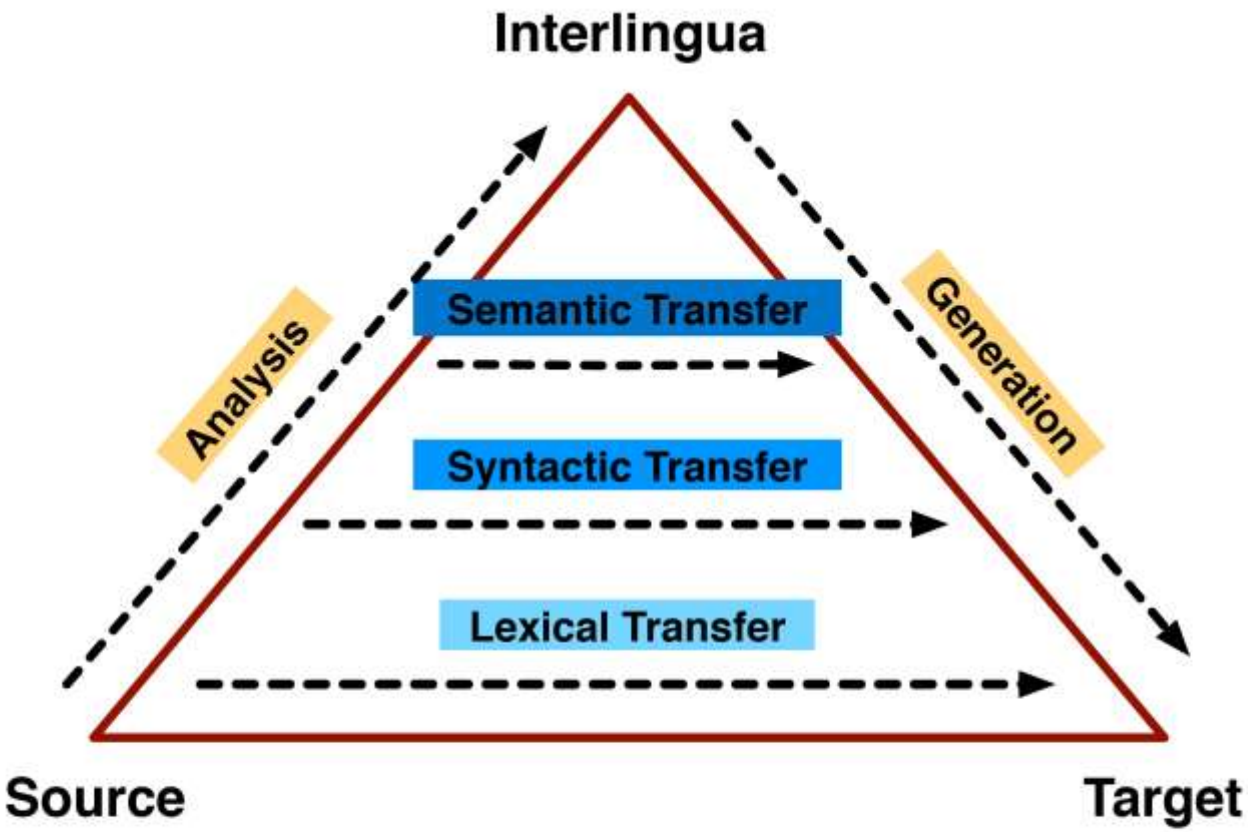
Last week, the vintage drama "Beauty private dishes" was temporarily suspended, accidentally sparking a heated discussion about the fake ratings of domestic dramas.

民权团体针对密苏里州发出旅**行警告**

Civil rights groups issue travel warnings against Missouri

http://matrix.statmt.org

The Vauquois Triangle



Challenges: word translation ambiguity

- What is the best translation?

$$\text{Sicherheit} \rightarrow \text{security } 14{,}516$$
$$\text{Sicherheit} \rightarrow \text{safety } 10{,}015$$
$$\text{Sicherheit} \rightarrow \text{certainty } 334$$

- Solution intuition: use counts in parallel corpus (aka bitext)
  - Here European Parliament corpus

Challenges: word order

- Problem: different languages organize words in different order to express the same idea

  En: The red house

  Fr: La maison rouge

- Solution intuition: language modeling!

Challenges: output language fluency

- What is most fluent?

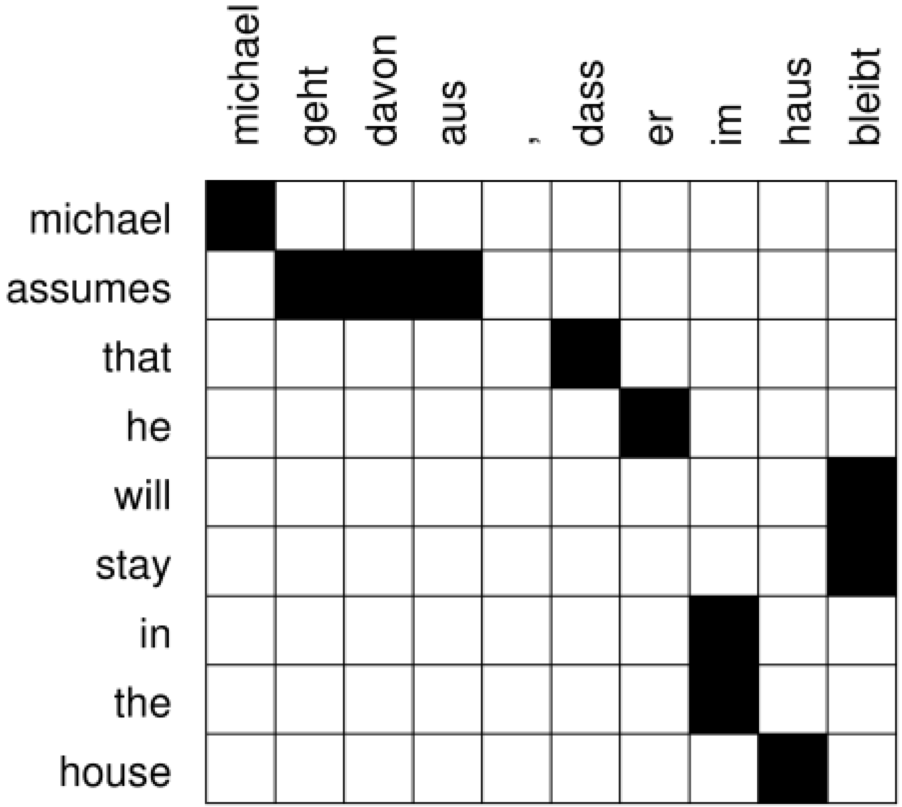
a problem for translation

a problem of translation

a problem in translation

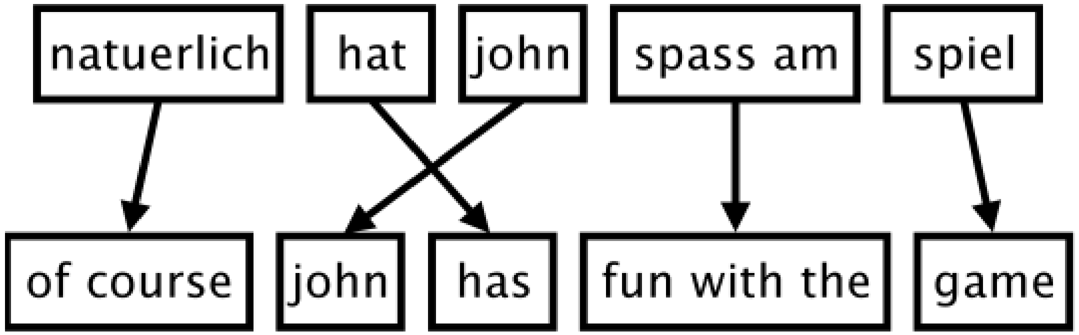
- Solution intuition: a language modeling problem!

Word Alignment

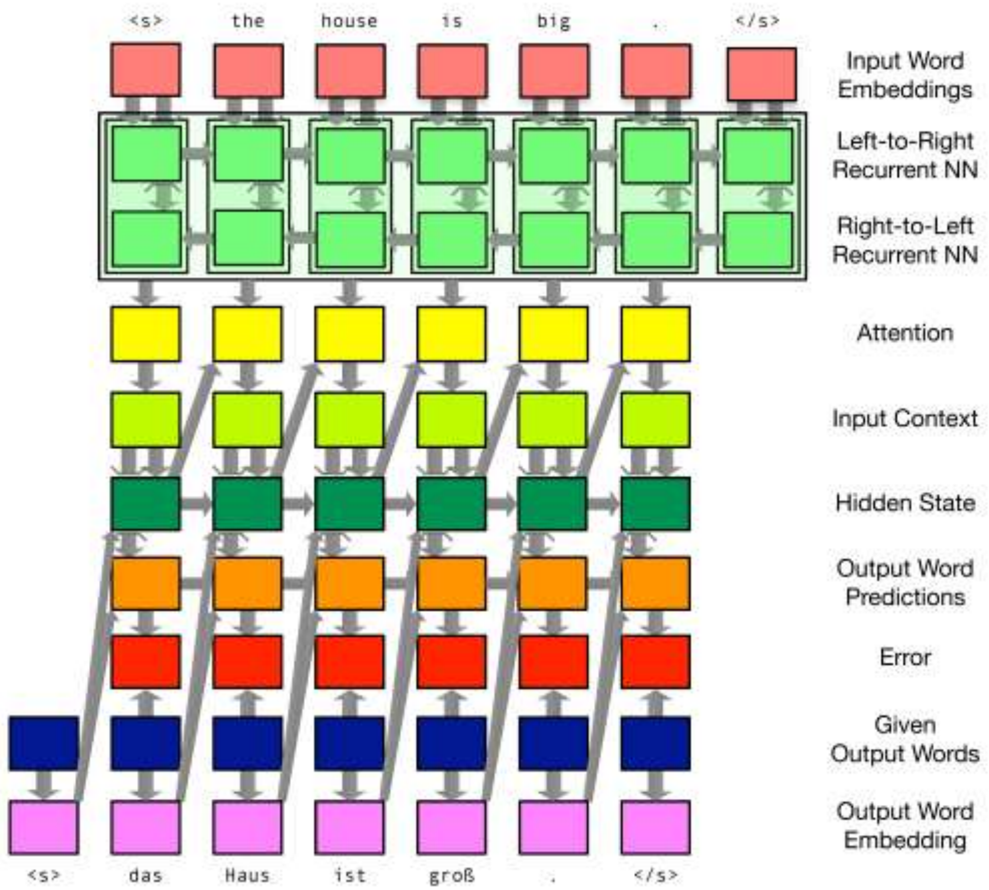


Phrase-based Models

- Input segmented in phrases
- Each phrase is translated in output language
- Phrases are reordered



Neural MT



<s>
the
house
is
big
.
</s>
Input Word Embeddings
Left-to-Right Recurrent NN
Right-to-Left Recurrent NN
Attention
Input Context
Hidden State
Output Word Predictions
Error
Given Output Words
Output Word Embedding
<s>
das
Haus
ist
groß
.
</s>

Today's topics
Machine Translation

- Context: Historical Background
  - Machine Translation is an old idea

- Machine Translation Evaluation

How good is a translation?
Problem: no single right answer

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

Evaluation

- How good is a given machine translation system?
- Many different translations acceptable
- Evaluation metrics
  - Subjective judgments by human evaluators
  - Automatic evaluation metrics
  - Task-based evaluation

Adequacy and Fluency

- Human judgment
  - Given: machine translation output
  - Given: input and/or reference translation
  - Task: assess quality of MT output

- Metrics
  - **Adequacy:** does the output convey the meaning of the input sentence? Is part of the message lost, added, or distorted?
  - **Fluency:** is the output fluent? Involves both grammatical correctness and idiomatic word choices.

Fluency and Adequacy: Scales

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Let's try:
rate fluency & adequacy on 1-5 scale

- Source:
  N'y aurait-il pas comme une vague hypocrisie de votre part ?
- Reference:
  Is there not an element of hypocrisy on your part?
- System1:
  Would it not as a wave of hypocrisy on your part?
- System2:
  Is there would be no hypocrisy like a wave of your hand?
- System3:
  Is there not as a wave of hypocrisy from you?

Challenges in MT evaluation

- No single correct answer

- Human evaluators disagree

Automatic Evaluation Metrics

- Goal: computer program that computes quality of translations
- Advantages: low cost, optimizable, consistent
- Basic strategy
  - Given: MT output
  - Given: human reference translation
  - Task: compute similarity between them

Precision and Recall of Words

SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

Precision

$$\frac{correct}{output\text{-}length} = \frac{3}{6} = 50\%$$

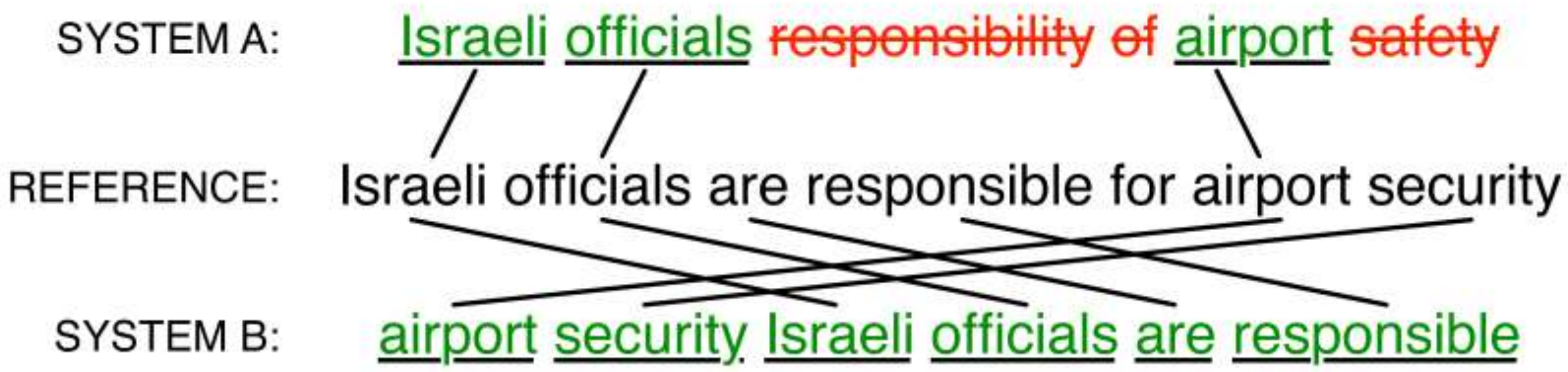
Recall

$$\frac{correct}{reference\text{-}length} = \frac{3}{7} = 43\%$$

F-measure

$$\frac{precision \times recall}{(precision + recall)/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall of Words



Metric	System A	System B
precision	50%	100%
recall	43%	100%
f-measure	46%	100%

flaw: no penalty for reordering

BLEU
Bilingual Evaluation Understudy

N-gram overlap between machine translation output and reference translation

Compute precision for n-grams of size 1 to 4

Add brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

Typically computed over the entire corpus, not single sentences

Multiple Reference Translations

To account for variability, use multiple reference translations

- n-grams may match in any of the references
- closest reference length used

Example

SYSTEM: Israeli officials responsibility of airport safety

2-GRAM MATCH 2-GRAM MATCH 1-GRAM

REFERENCES:

Israeli officials are responsible for airport security

Israel is in charge of the security at this airport

The security work for this airport is the responsibility of the Israel government

Israeli side was in charge of the security of this airport

BLEU examples

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

Some metrics use more linguistic insights in matching references and hypotheses

Partial credit for matching stems

SYSTEM	Jim went home
REFERENCE	Joe goes home

Partial credit for matching synonyms

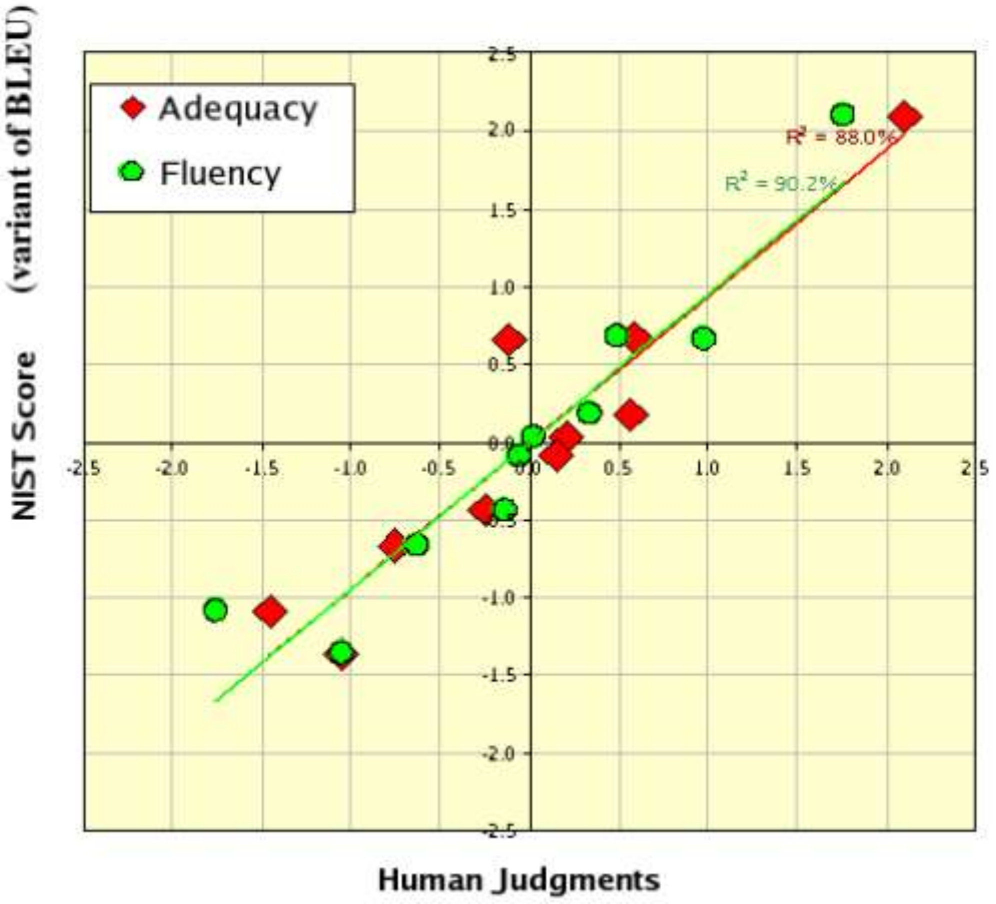
SYSTEM	Jim walks home
REFERENCE	Joe goes home

Use of paraphrases

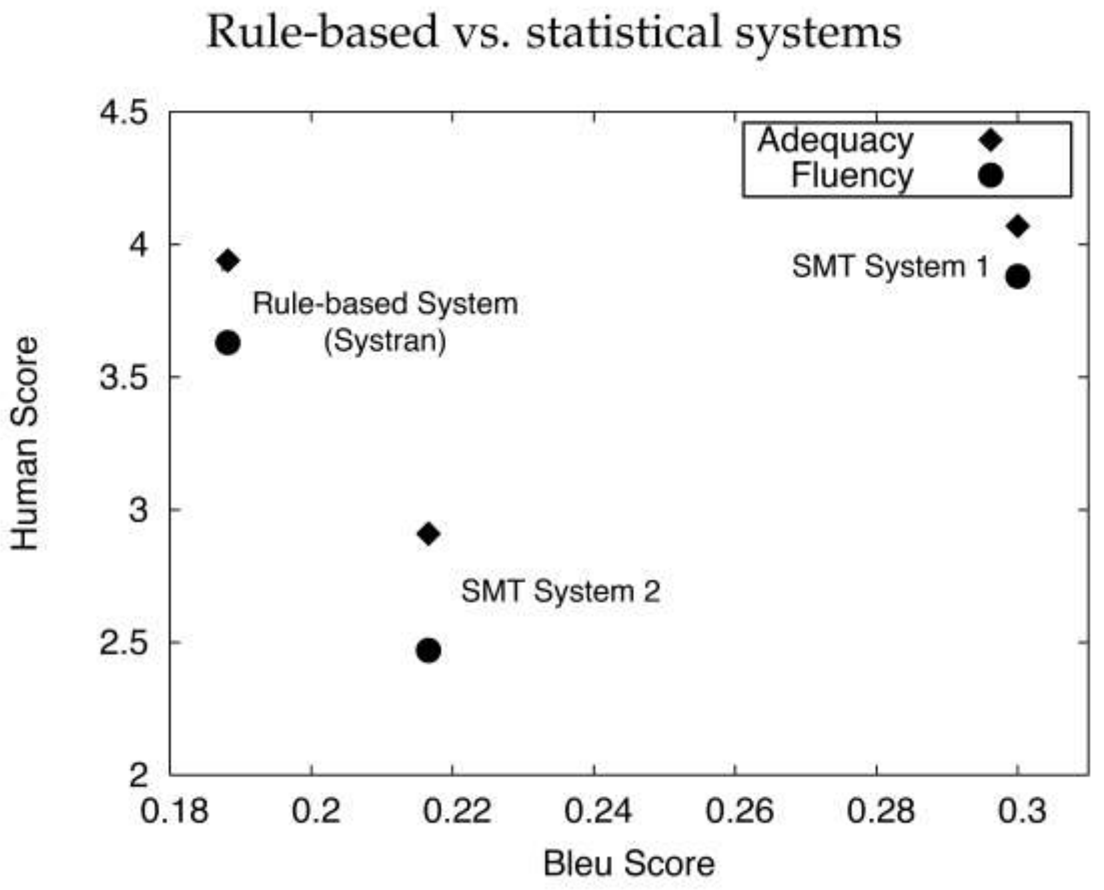
Drawbacks of Automatic Metrics

- All words are treated as equally relevant
- Operate on local level
- Scores are meaningless (absolute value not informative)
- Human translators score low on BLEU

Yet automatic metrics such as BLEU correlate with human judgement



Caveats: bias toward statistical systems



Automatic metrics

- Essential tool for system development
- Use with caution: not suited to rank systems of different types
- Still an open area of research
  - Connects with semantic analysis

Task-Based Evaluation
Post-Editing Machine Translation

Measuring time spent on producing translations

- baseline: translation from scratch
- post-editing machine translation

But: time consuming, depend on skills of translator and post-editor

Metrics inspired by this task

- TER: based on number of editing steps
  Levenshtein operations (insertion, deletion, substitution) plus movement
- HTER: manually construct reference translation for output, apply TER
  (very time consuming, used in DARPA GALE program 2005-2011)

Task-Based Evaluation
Content Understanding Tests

Given machine translation output, can monolingual target side speaker answer questions about it?

1. basic facts: who? where? when? names, numbers, and dates
2. actors and events: relationships, temporal and causal order
3. nuance and author intent: emphasis and subtext

Very hard to devise questions

Sentence editing task (WMT 2009–2010)

- person A edits the translation to make it fluent
  (with no access to source or reference)
- person B checks if edit is correct
  $\rightarrow$ did person A **understand** the translation correctly?

Today’s topics
Machine Translation

- Historical Background
  - Machine Translation is an old idea

- Machine Translation Today
  - Use cases and method

- Machine Translation Evaluation

What you should know

- Context: Historical Background
  - Machine Translation is an old idea
  - Difference between hype and reality!

- Machine Translation Evaluation
  - What are adequacy and fluency
  - Pros and cons of human vs automatic evaluation
  - How to compute automatic scores: Precision/Recall and BLEU