Machine Translation

Session 1: Introduction & Evaluation

Sharid Loáiciga

Linguistics Department
University of Potsdam

April 22nd, 2020

Slides credits: Yves Scherrer,
who in turn credits Jörg Tiedemann, Philipp Koehn & Sara Stymmne

Practical Info

Course website:

`https://compling-potsdam.github.io/sose20-mt/`

How to contact me:

By email, `loaicigasanchez@uni-potsdam.de`

In my office, once we can go back on campus safely,

building 14, room 2.19/20

I have variable available hours for meetings, let me know by

email if you need to see me.

Course format

- Policies, news & updates will be posted in the **course website**.
- **Moodle** will be reserved for submitting assignments and distribution of copyright material.
- This is a hands-on course.
- We will alternate between classic lectures and in-class exercises. You may finish the exercises at home if more time is necessary.

How to pass this course?

- Option 1: Pass/Not-pass
  - Hands-on assignments (PW sessions)
  - Seminar presentation on special topic in MT
- Option 2: Module grade
  - Meet requirements for option 1
  - Project work
    - marked and graded
    - written report

Skills requirements

- Familiarity with unix command-line tools
  - Essential
- Programming skills
  - Not required, we'll trust the developers of existing toolkits (but it doesn't hurt to understand the code)
- Machine learning in general
  - Desirable but not essential
- Neural networks and maths
  - Not required, we aim for an intuitive understanding of the models

https://compling-potsdam.github.io/sose20-mt/

Machine translation today

A l'orée de ce débat télévisé inédit dans l'histoire de la Ve République, on attendait une forme de «Tous sur Macron» mais c'est la candidate du Front national qui s'est retrouvée au cœur des premières attaques de ses quatre adversaires d'un soir, favorisées par le premier thème abordé, les questions de société et donc de sécurité, d'immigration et de laïcité.

At the beginning of this televised debate, which was unheard of in the history of the Fifth Republic, a "Tous sur Macron" was expected, but it was the candidate of the National Front who found itself at the heart of the first attacks of its four Opponents of one evening, favored by the first theme tackled, the issues of society and thus security, immigration and secularism.

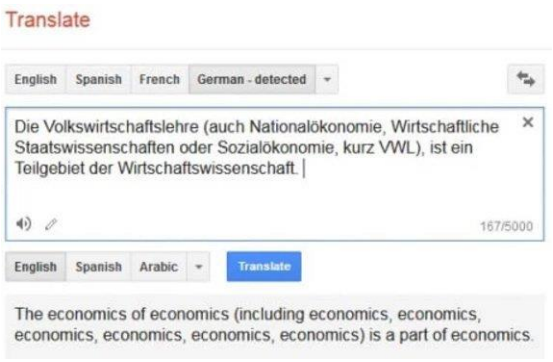
Machine translation today

记者从环保部了解到，《水十条》要求今年年底前直辖市、省会城市、计划单列市建成区基本解决黑臭水体。截至目前，全国224个地级及以上城市共排查确认黑臭水体2082个，其中34.9%完成整治，28.4%正在整治，22.8%正在开展项目前期。

Reporters learned from the Ministry of Environmental Protection, "Water 10" requirements before the end of this year before the municipality, the provincial capital city, plans to build a separate city to solve the basic black and black water. Up to now, the country's 224 prefecture-level and above cities were identified to confirm the black and white water 2082, of which 34.9% to complete the renovation, 28.4% is remediation, 22.8% is carrying out the project early.

Machine translation

- MT is a tough challenge (and fun)…
  - Doctor's office: Specialist in women and other diseases
  - Pub: Ladies are requested not to have children in the bar
  - Hotel: Please leave your values at the front desk



Machine translation

- MT is not a solved problem…
- Bit it is constantly improving…
  - Input: Vem vann Allsvenskan i fjol?
  - Google 2010: Who stole headlines last year?
  - Google 2013: Who won the Championship last year?
  - Google 2016: Who won the Olympics last year?
  - Google 2019: Who won the Allsvenskan last year?

Machine translation

- MT is a cool research topic
  - A complex but natural task
    - Natural input, natural output
    - Not fixed to particular linguistic theories or formalisms
  - Tied to intrinsic properties of language
    - What are the differences between languages?
    - How can we preserve meaning when translating?
  - Extraordinary progress in the last 5 years
    - MT is probably the language technology task that was impacted most by the “neural revolution” (2013-2017)

Machine translation

- MT is rather an art than a science
- This course is similar in spirit to “Building NLP applications” on BA level
- This command trains a neural machine translation system:

```
python3 train.py \
    -data data -save_model model \
    -encoder_type transformer -decoder_type transformer -position_encoding \
    -layers 6 -rnn_size 512 -word_vec_size 512 -transformer_ff 2048 -heads 8 \
    -train_steps 100000 -max_generator_batches 256 -dropout 0.1 \
    -batch_size 4096 -batch_type tokens -normalization tokens \
    -accum_count 2 -optim adam -adam_beta2 0.998 -decay_method noam \
    -warmup_steps 8000 -learning_rate 2 -max_grad_norm 0 \
    -param_init 0  -param_init_glorot -label_smoothing 0.1 \
    -valid_steps 10000 -valid_batch_size 16 -save_checkpoint_steps 10000 \
    -world_size 1 -gpu_ranks 0 -seed 1111
```

- What do these parameters mean and how do they influence translation quality?

Course goals

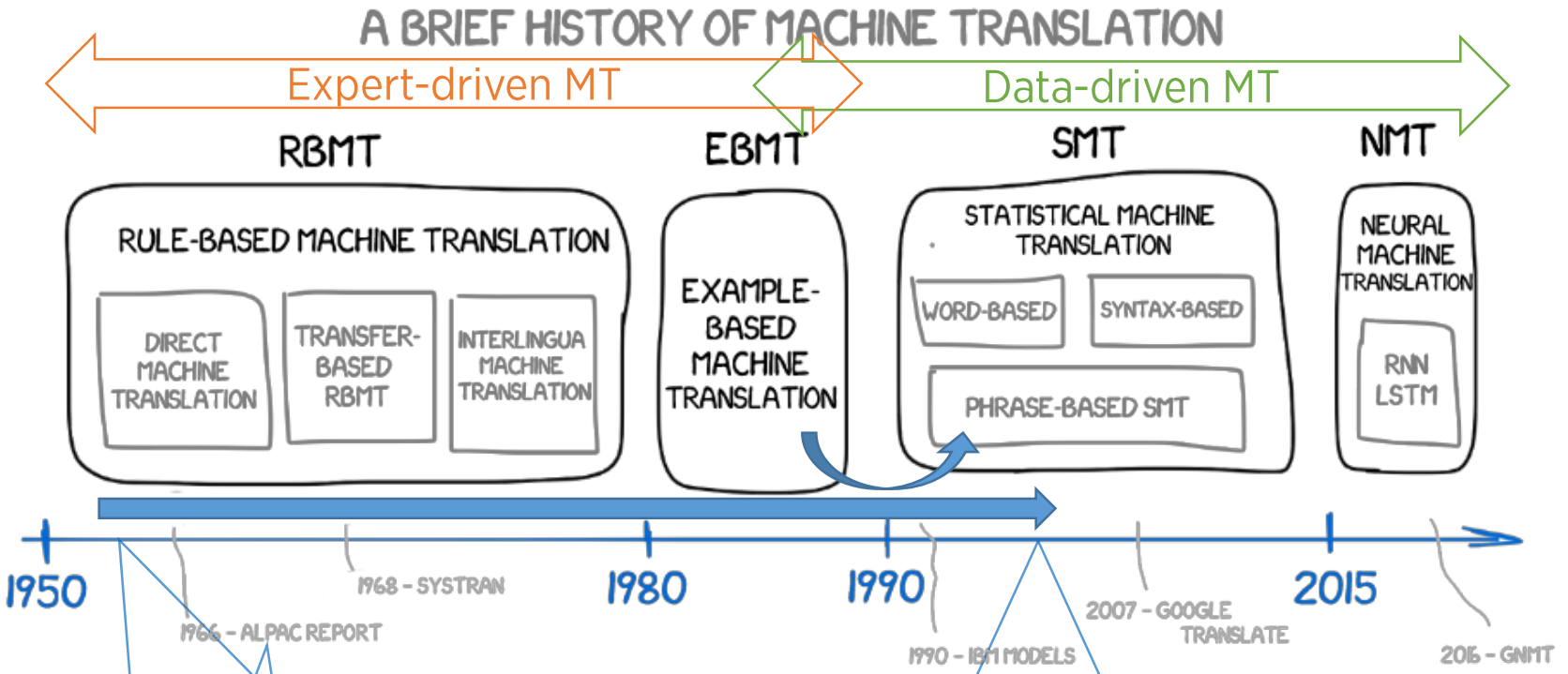
- Describe the differences between various machine translation paradigms (rule-based, statistical, neural machine translation)
- Explain the basics of neural machine translation and the most common model architectures
- Describe the issues related to machine translation evaluation
- Train, test and evaluate a neural machine translation system including the appropriate pre- and post-processing steps
- Read, understand and present scientific papers on current challenges in machine translation

A short history of machine translation

Figures from http://vas3k.com/blog/machine_translation/

MT is an old idea...

- In fact, one of the first language technology tasks.
- Warren Weaver (1947):
  - *When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode".*
  - Success of code-breaking in WWII
  - Beginning of cold war



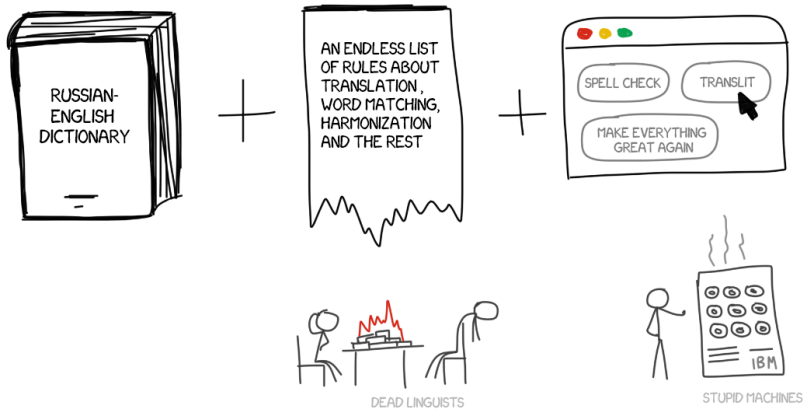
1954 – IBM-Georgetown experiment: a machine could translate 250 words and 6 grammar rules

1966 – ALPAC report:
MT is expensive, inaccurate, and unpromising

2002 – BLEU evaluation metric
2005 – Moses open-source system
2005 – Europarl corpus
2006 – First workshop on machine translation

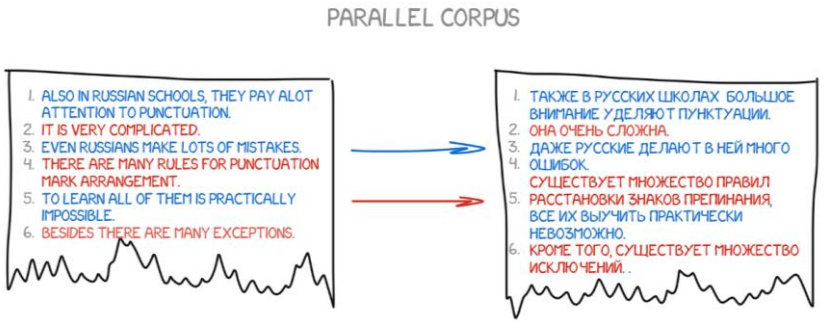
Rule-based MT

- Build dictionaries
- Write transformation rules
- Refine, refine, refine



- Commercial applications:
  - 1976: weather forecast translations French-English
  - 1968: Systran

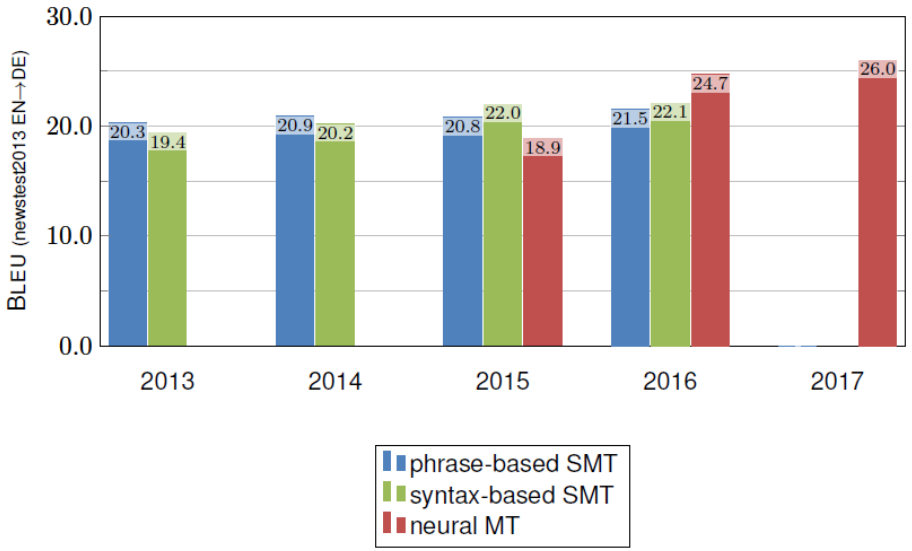
Statistical MT



- 1990: “IBM models”
  - Idea: learn everything from a parallel corpus
- Mid-2000s: phrase-based models
  - A lesson from EBMT: Translating each word separately is harder than it needs to be
  - Keep frequent word sequences (“phrases”) together and translate them as a whole
- Late 2000s: syntax-based models
- 2010: Commercial viability (Google Translate…)

Neural MT

- Late 2000s: successful use of neural models for computer vision
- 2012, 2013: first neural models for MT proposed
- Since 2016: NMT is the new state of the art



*NMT 2015 from U. Montréal: https://sites.google.com/site/acl16nmt/

Figure credit: Rico Sennrich

Hype and reality

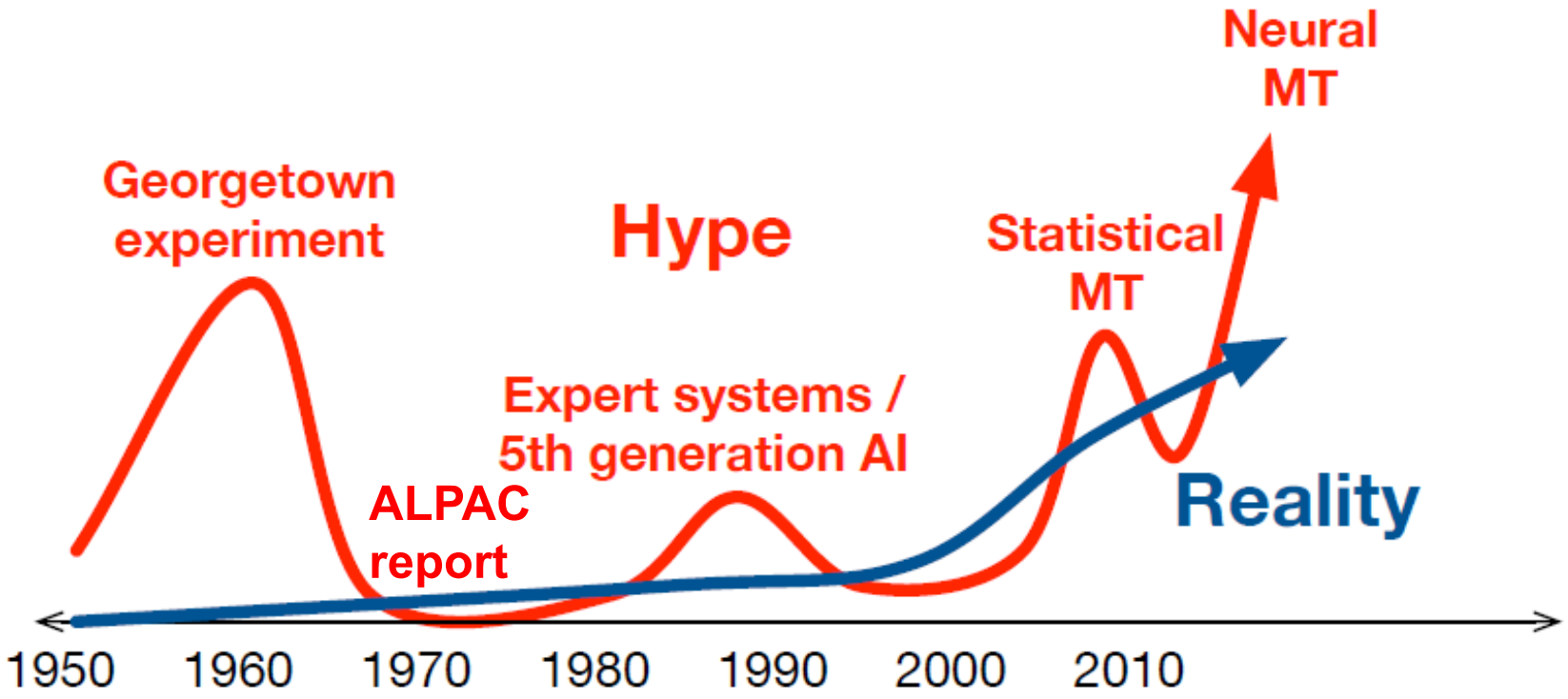
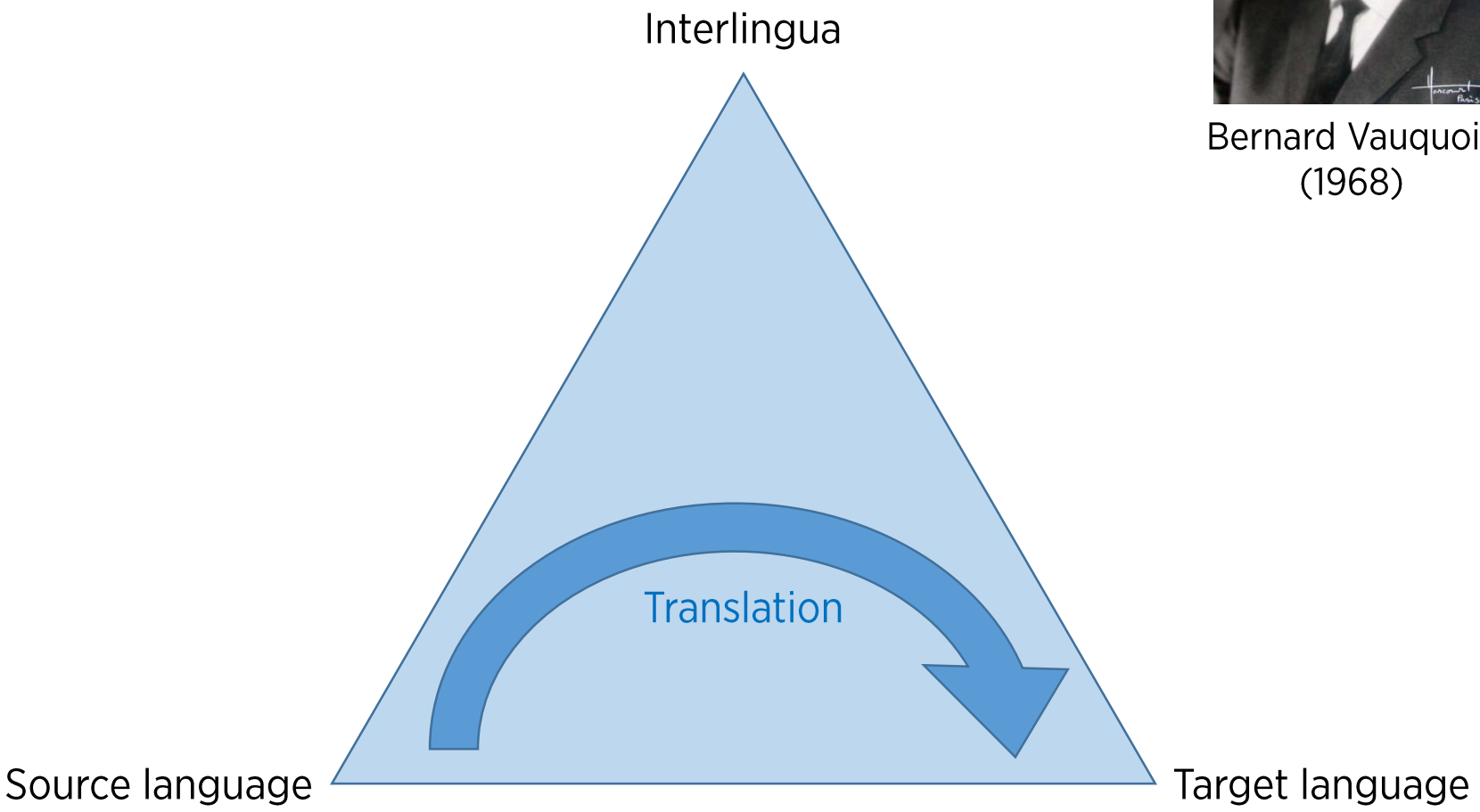


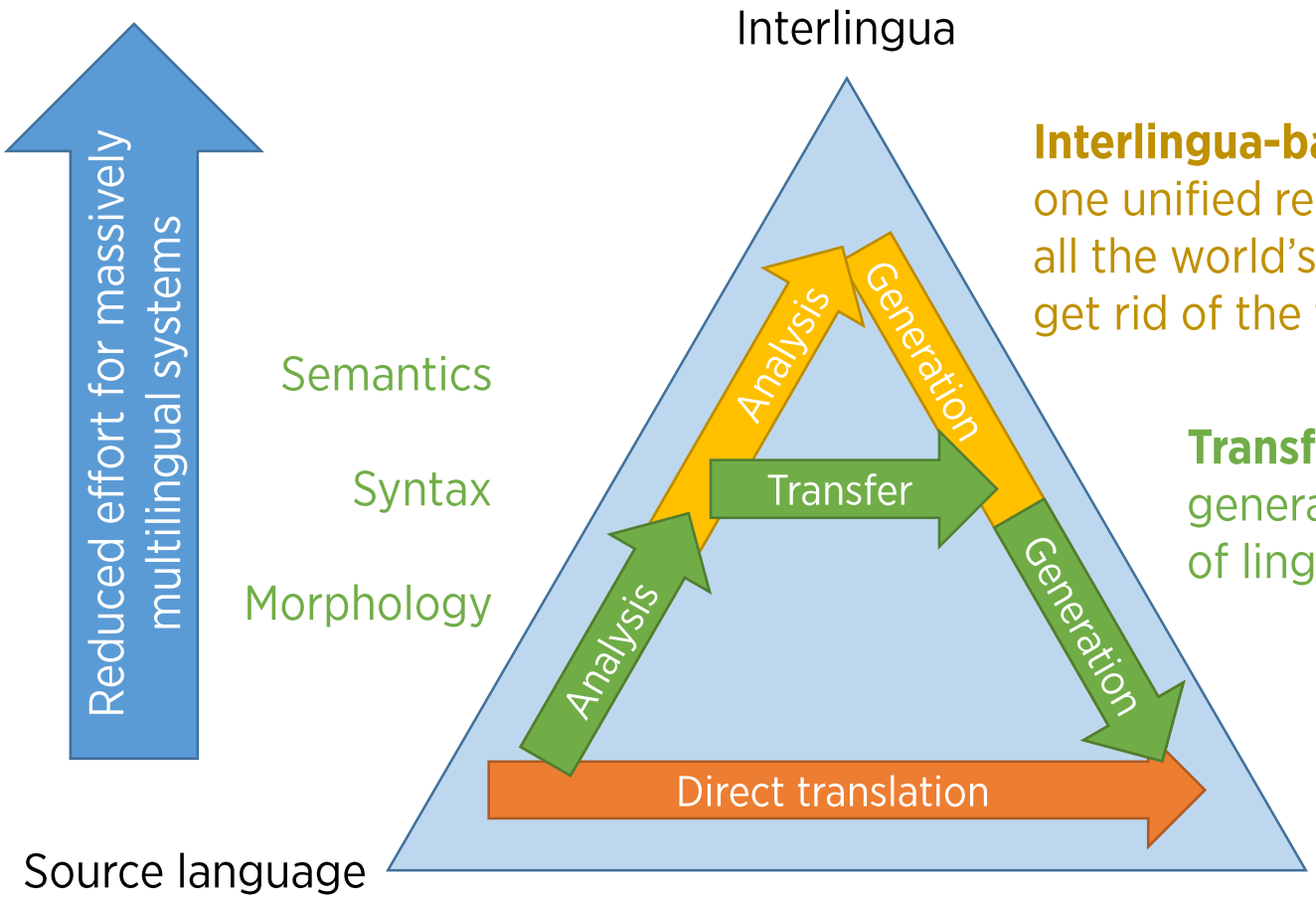
Figure credit: Philipp Koehn

The Vauquois triangle

Bernard Vauquois
(1968)



The Vauquois triangle

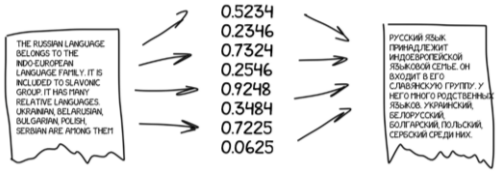


Interlingua-based translation:
one unified representation for all the world's languages to get rid of the transfer process

Transfer-based translation:
generalize through the use of linguistic analysis

Direct translation:
"word by word"

The Vauquois triangle



Do high-dimensional vectors of numbers represent some kind of interlingua?

Interlingua

RBMT

RULE-BASED MACHINE TRANSLATION

DIRECT MACHINE TRANSLATION

TRANSFER-BASED RBMT

INTERLINGUA MACHINE TRANSLATION

SMT

STATISTICAL MACHINE TRANSLATION

WORD-BASED

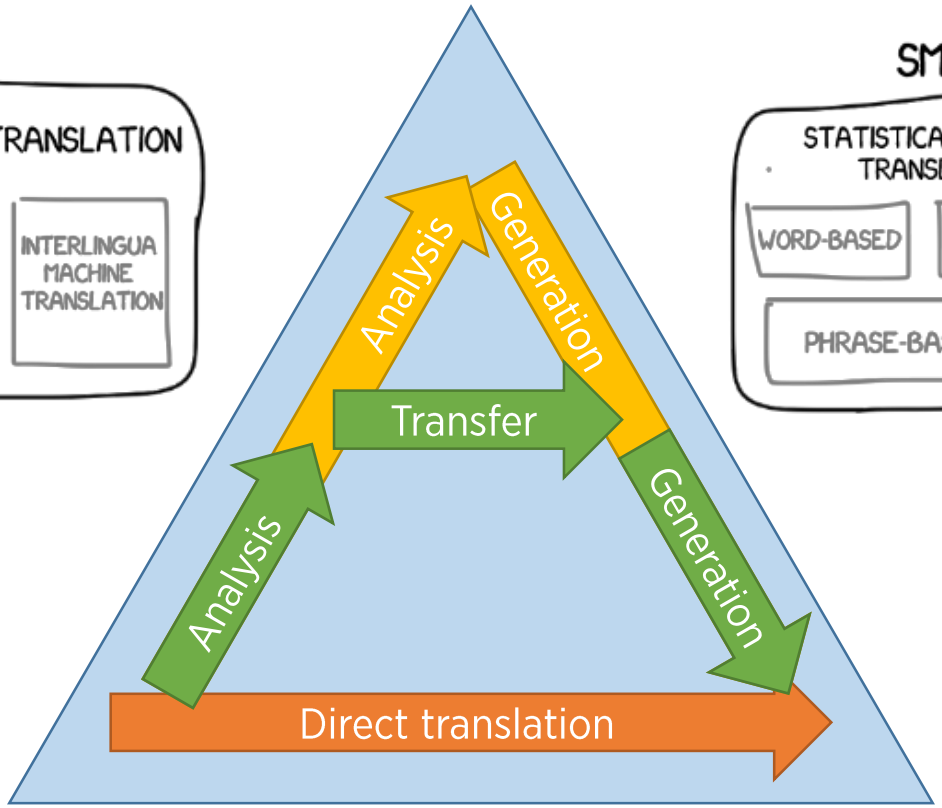
SYNTAX-BASED

PHRASE-BASED SMT

NMT

NEURAL MACHINE TRANSLATION

RNN LSTM



Source language

Target language

Evaluation

What is a good translation?

What is a good machine translation system?

Ten translations of a Chinese sentence

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.

(An example from the 2001 NIST evaluation set)

Why do we need evaluation?

- How good is a given system?
  - general evaluation and assessment
- Which one is the best system for a specific purpose?
  - task-specific evaluation
- How much did we improve our system?
  - system development
- How can we tune our system to become better?
  - parameter estimation

Evaluation methods

- Human evaluation:
  - Given: source + machine translation output
  - Given: reference translation + machine translation output
- Automatic evaluation:
  - Given: reference translation + machine translation output
- Extrinsic/task-based evaluation:
  - How much post-editing effort?
  - How well is the key information conveyed?
  - Test suites focusing on particular linguistic phenomena

Evaluation methods

- Human evaluation is...
  - ultimately what we are interested in
  - very time consuming (and boring)
  - not re-usable
  - subjective
- Automatic translation is...
  - cheap and re-usable
  - not necessarily reliable

Human evaluation

- Evaluation measures:
  - Adequacy and fluency *vs* general translation quality
- Evaluation setup:
  - Direct assessment *vs* ranking
- Required language skills of evaluators:
  - Source-based *vs* reference-based evaluation
  - Needs bilingual evaluators *vs* can be done by one translator + monolingual evaluators
- Task-based evaluation measures:
  - Post-editing effort
  - Content understanding tests

Adequacy and fluency

- Adequacy:
  - Does the output convey the same meaning as the input sentence?
  - Is part of the message lost, added, or distorted?
- Scale:

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

- Fluency:
  - Is the output good fluent English?
  - This involves both grammatical correctness and idiomatic word choices.
- Scale:

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

Example

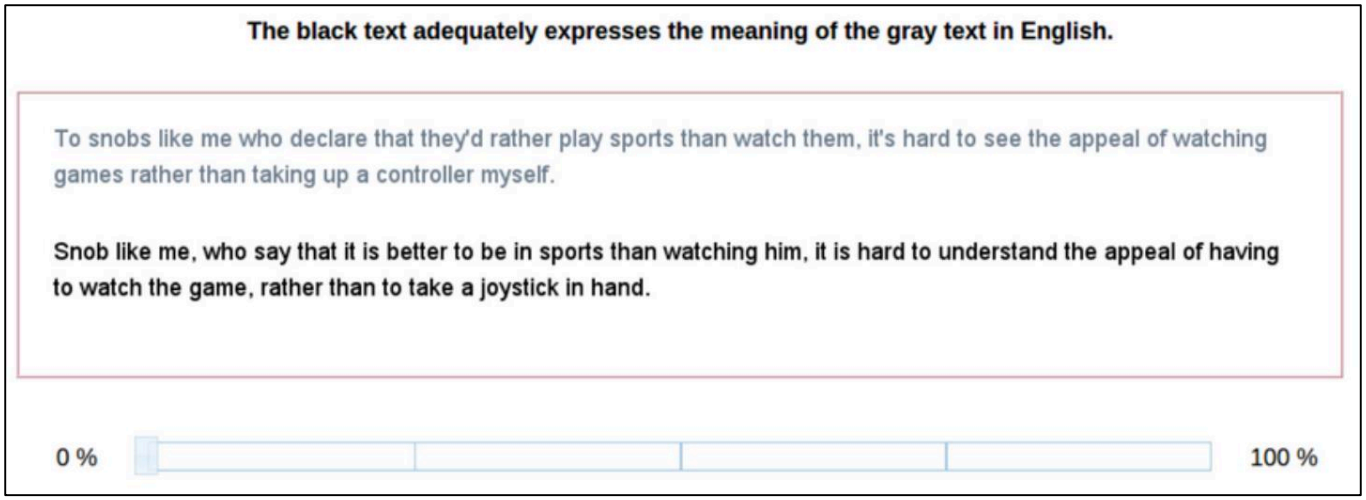
- Judge adequacy and fluency on a 1-5 scale:
  - Source:
    L'affaire NSA souligne l'absence totale de débat sur le renseignement
  - Reference:
    NSA Affair Emphasizes Complete Lack of Debate on Intelligence
  - System 1:
    The NSA case underscores the total lack of debate on intelligence
  - System 2:
    The case highlights the NSA total absence of debate on intelligence
  - System 3:
    The matter NSA underlines the total absence of debates on the piece of information

Evaluation setup

- Direct assessment:
  - The evaluator sees one translated sentence and needs to give a score
- Ranking:
  - The evaluator sees several translations (typically, the outputs from different systems) and has to rank them.
  - Evaluators are more consistent on ranking tasks than on absolute scoring tasks

Single-score reference-based direct assessment

- It is sometimes difficult to distinguish between fluency errors and adequacy errors
  - Use a single quality scale (1-5 or 0-100%)

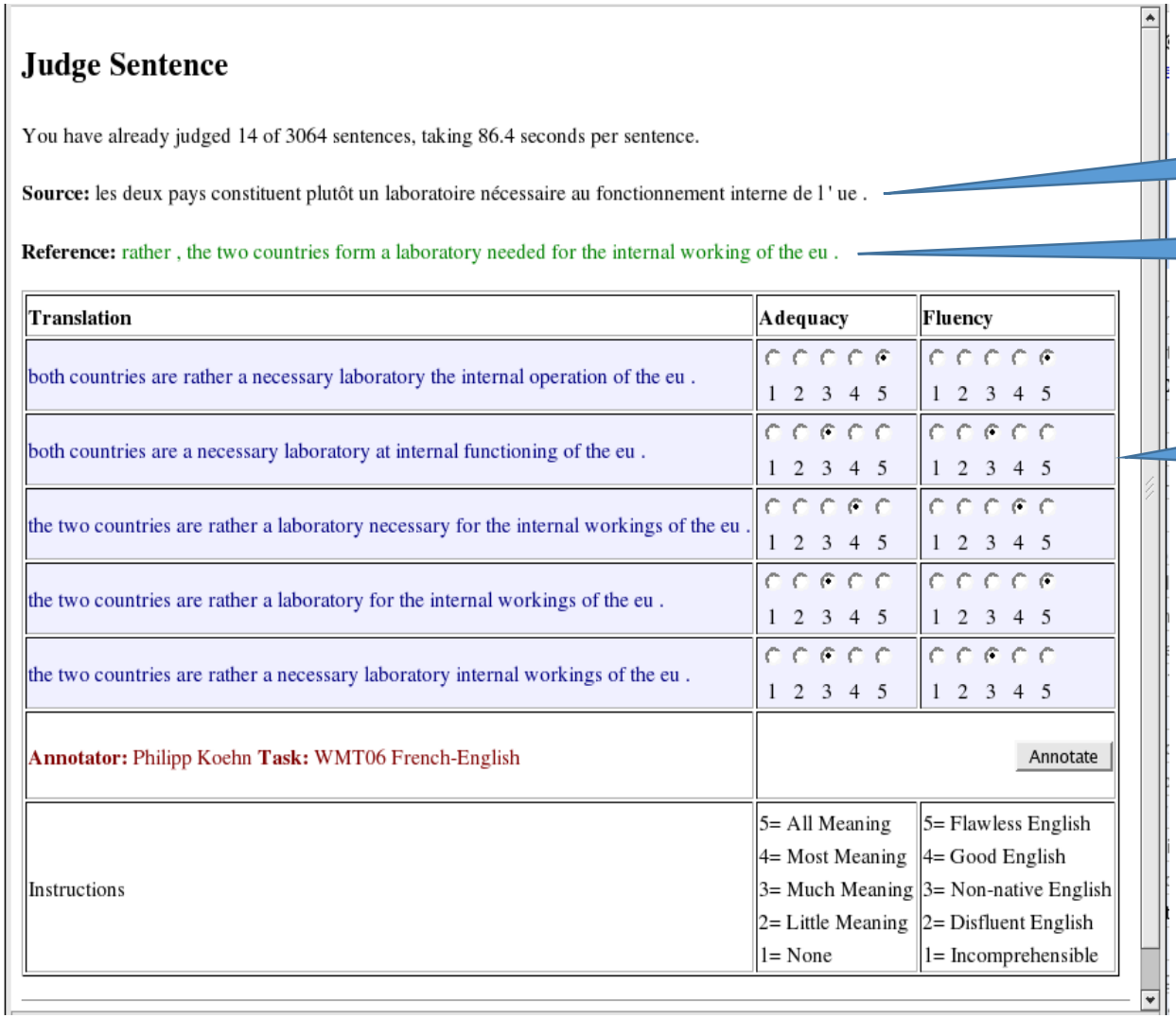


Single-score reference-based ranking

- Task: Is translation X better / worse / equal than translation Y?

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst

Ranking with adequacy and fluency



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

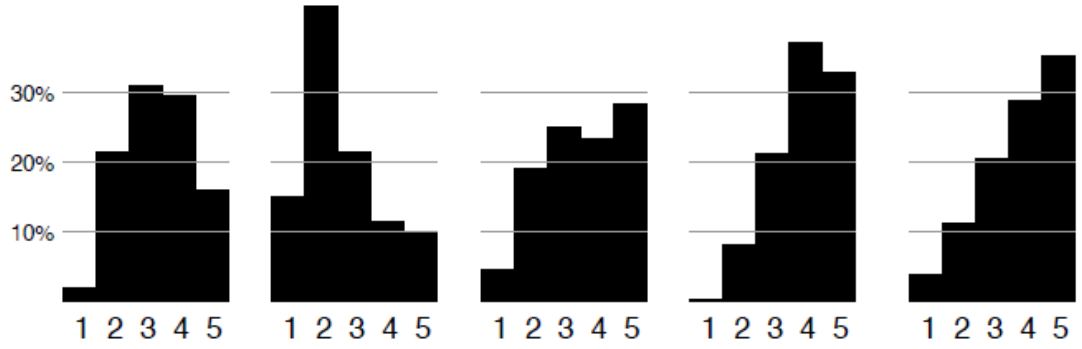
Source sentence

Reference translation

Various MT systems

Problems of human evaluation

- Source-based evaluation requires bilingual evaluators, reference-based evaluation does not
  - Results may differ
- Annotators disagree
  - Histogram of adequacy judgements by different human evaluators:



(from WMT 2006 evaluation)

Post-editing effort

- How much does a machine translated text need to be changed (post-edited) in order to produce a high-quality translation?
  - Time spent on post-editing
  - Number of edit operations (HTER)

Content understanding tests

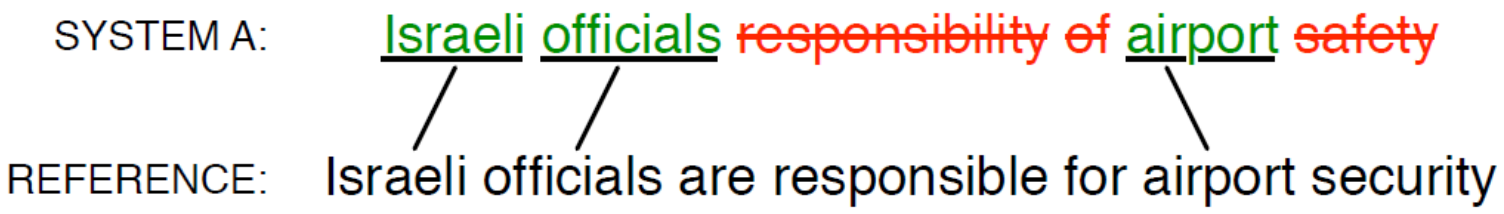
- Given machine translation output, can a monolingual target language speaker answer questions about it?
  - Basic facts:
    who? where? when? names, numbers, and dates
  - Actors and events:
    relationships, temporal and causal order
  - Nuance and author intent:
    emphasis and subtext
- Very hard to devise questions
- Just as time-consuming (or even more so) than direct assessment

Automatic evaluation

- Basic strategy:
  - provide a human reference translation
  - compute similarity between reference translation and machine translation output

- Different similarity measures:
  - Precision/recall
  - WER/TER
  - BLEU
  - METEOR

A first try: precision and recall



- Precision

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

- Recall

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

- F-measure

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

A first try: precision and recall

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

flaw: no penalty for reordering

Word error rate (WER)
Translation edit rate (TER)

- Minimum number of editing steps to transform output to reference (Levenshtein distance)
  - **match:** words match, no cost
  - **substitution:** replace one word with another
  - **insertion:** add word
  - **deletion:** drop word
- $\text{WER} = \frac{\text{substitutions+insertions+deletions}}{\text{reference—length}}$
- TER introduces an additional operation:
  - **shift:** move a word sequence

Word error rate (WER)

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
security	7	6	5	5	5	4	4

		airport	security	Israeli	officials	are	responsible
	0	1	2	3	4	5	6
Israeli	1	1	2	2	3	4	5
officials	2	2	2	3	2	3	4
are	3	3	3	3	3	2	3
responsible	4	4	4	4	4	3	2
for	5	5	5	5	5	4	3
airport	6	5	6	6	6	5	4
security	7	6	5	6	7	6	5

Metric	System A	System B
word error rate (WER)	57%	71%

Error rates: lower is better!

BLEU
BiLingual Evaluation Understudy

- N-gram overlap between reference translation and MT output
- Compute precision for n-grams of size 1 to 4 (geometric mean)
- Add a brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right) \cdot \sqrt[4]{\prod_{i=0}^{4} \text{precision}_i}$$

- Typically computed over the entire corpus, not single sentences

BLEU

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

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (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%

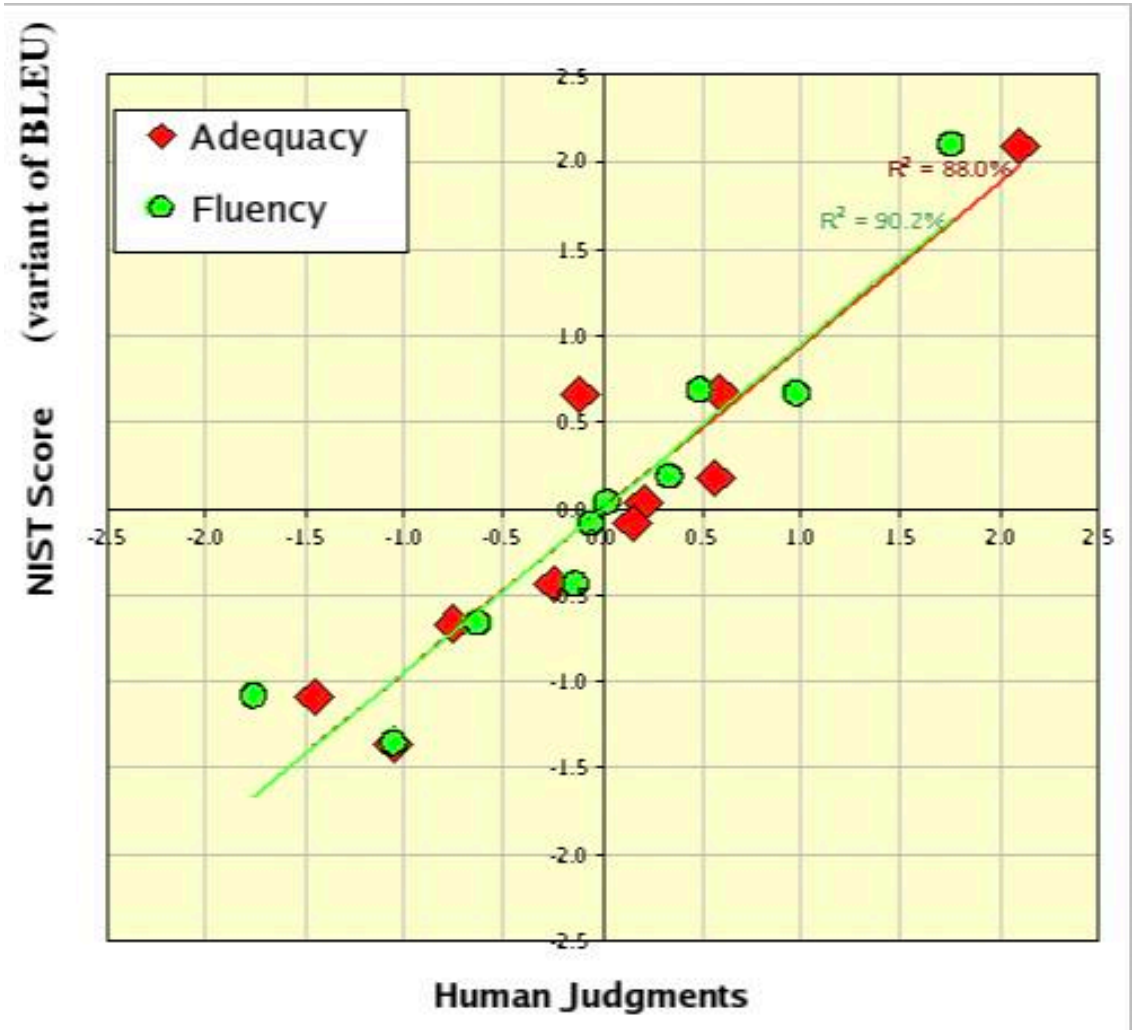
BLEU

- Use multiple reference translations to account for variability
  - n-grams may match in any of the references
  - use closest reference length

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

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

Does BLEU correlate with human judgements?



Typical BLEU scores

- Koehn (2005):

%	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	31.2	32.1	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	30.1	31.1	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	30.5	-	40.2	12.5	32.3	21.4	35.9	23.9
fr	23.7	18.5	26.1	30.0	38.4	-	12.6	32.4	21.1	35.3	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	34.0	36.0	11.0	-	20.0	31.2	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	30.1	37.9	39.0	11.9	32.0	20.2	-	21.9
sv	30.3	18.9	22.8	30.2	28.6	29.7	15.3	23.9	21.9	25.9	-

- 2019: add about 10 BLEU points...

METEOR: Flexible matching

- Partial credits for matching stems
  - System: Jim went home
  - Reference: Joe goes home
- Partial credits for matching synonyms
  - System: Jim walks home
  - Reference: Joe goes home
- Use of paraphrases

- Requires language-specific tools
  - Stemmer/lemmatizer, synonym dictionary, ...

Metrics research

This may be a topic for your presentation…

- Active development of new metrics
  - syntactic similarity
  - semantic equivalence or entailment
  - metrics targeted at reordering
  - trainable metrics
  - etc.
- Evaluation campaigns that rank metrics

Test suites

This may be a topic for your presentation…

- An old idea (from RBMT-times), but had a recent revival (since 2017):
  1. Create a test set targeting a specific linguistic phenomenon you want to evaluate
     - Pronouns
     - Ambiguous words
     - Morphology
  2. Translate it using MT systems of interest
  3. Evaluate how well your specific phenomenon is translated
     - Automatically (easily reuseable)
     - By humans

Critique of automatic evaluation metrics

- They ignore the relevance of words
  - Names and core concepts should more important than determiners and punctuation
- They operate on local level
  - Do not consider overall grammaticality of the sentence or sentence meaning
- Scores are meaningless
  - Scores are very test-set specific, their absolute value is not informative
- Human translators score low on BLEU
  - Possibly because of higher variability, different word choices

Purposes of machine translation

- **Assimilation:** reader initiates translation, wants to know content
  - user is tolerant of inferior quality
  - focus of majority of research (GALE program, etc.)
- **Communication:** participants don't speak same language, rely on translation
  - users can ask questions when something is unclear
  - chat room translations, hand-held devices
  - often combined with text-to-speech, speech-to-text
- **Dissemination:** publisher wants to make content available in other languages
  - high demands for quality
  - currently almost exclusively done by human translators

Current state of the art

HTER	Assessment	Examples	Language pairs, domains
0%	publishable	Seamless bridging of language divide	French-English restricted domain (e.g. technical document localization)
10%		Automatic publication of official announcements	French-English news stories
20%	editable	Access to official publications; Multilingual communication (chat, social networks)	German-English news stories
30%	gistable	Information gathering; Trend spotting	Swahili-English news stories
40%	triageable	Identifying relevant documents	Uyghur-English news stories

(Ph. Koehn's informal rough estimates)

Readings

- History of MT:
  - Vasily Zubarev's blog post:
  - http://vas3k.com/blog/machine_translation/
  - https://medium.freecodecamp.org/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5
- Evaluation:
  - Philipp Koehn, *Statistical machine translation*, ch. 9
  - PDF available on Moodle
- Pre-assignment for Friday:
  - Make sure you have access to Taito