MT, Generation and Evaluation

CS 685, Fall 2025

Advanced Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

- Nov 20: Midterm #2

- Next two weeks are Thurs-only
  - Tu 11/4: no class for Election Day
  - Th 11/6: class like normal
  - Tu 11/11: no class for Veteran's Day
  - Th 11/13: class like normal

Quiz: which ones have errors?

Input: Tell me a bio of Bridget Moynahan.

Bridget Moynahan is a actress who were born in April 28, 1971 at Binghamton, New York. She become best known with her role as ...

Bridget Moynahan was born in Binghamton, New York. (...) Born in Springfield, Massachusetts, ...

Bridget Christie is an English stand-up comedian, actress and writer.

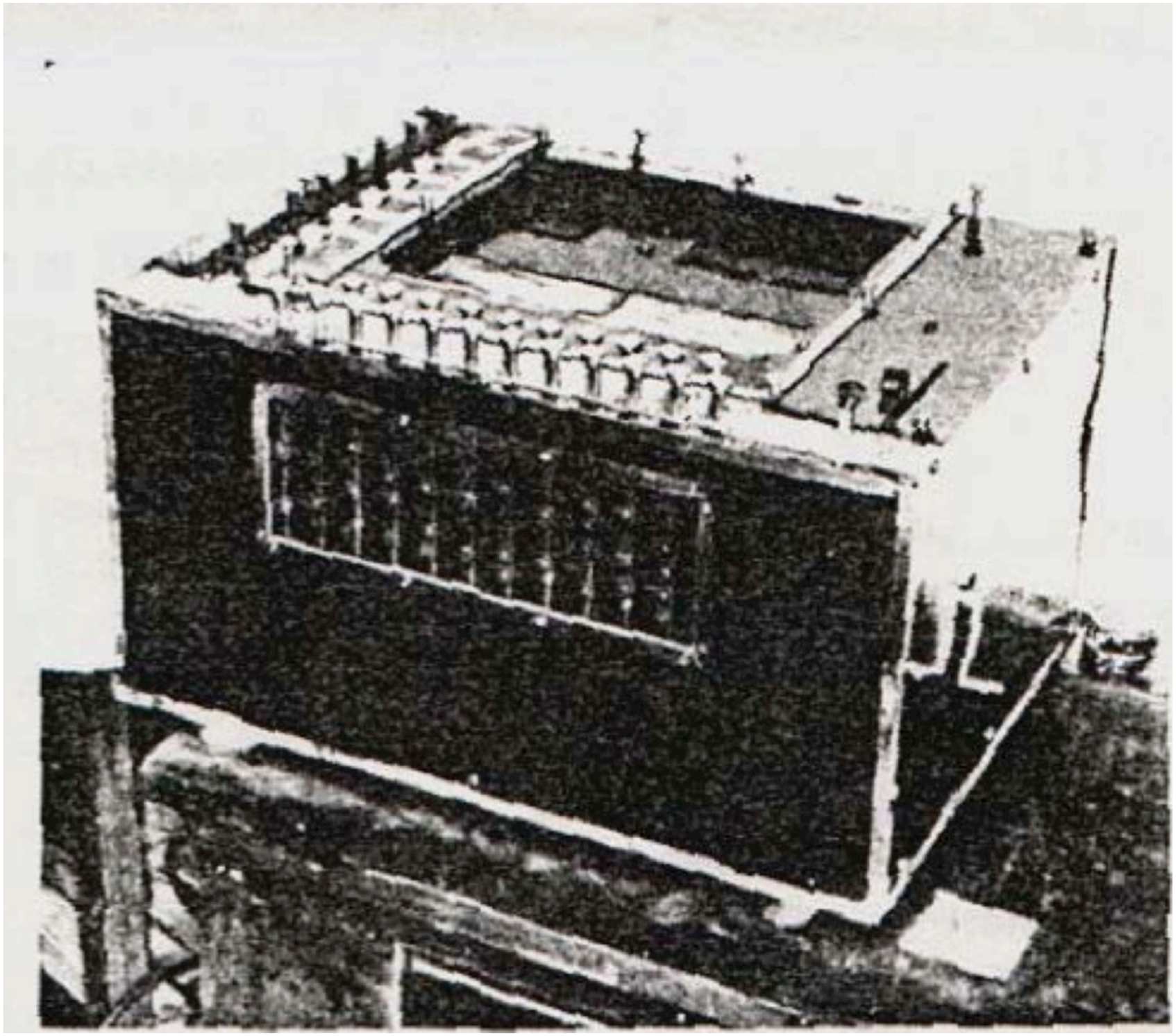
Bridget Moynahan is an American actress, model and producer, best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods.

[slide: Sewon Min]

- Generation as tasks
  - Machine Translation (text to text)
  - Summarization (text to shorter text)
  - Other open-ended (KB to text, etc.)
  - ...

- Evaluation challenge: no single answer!

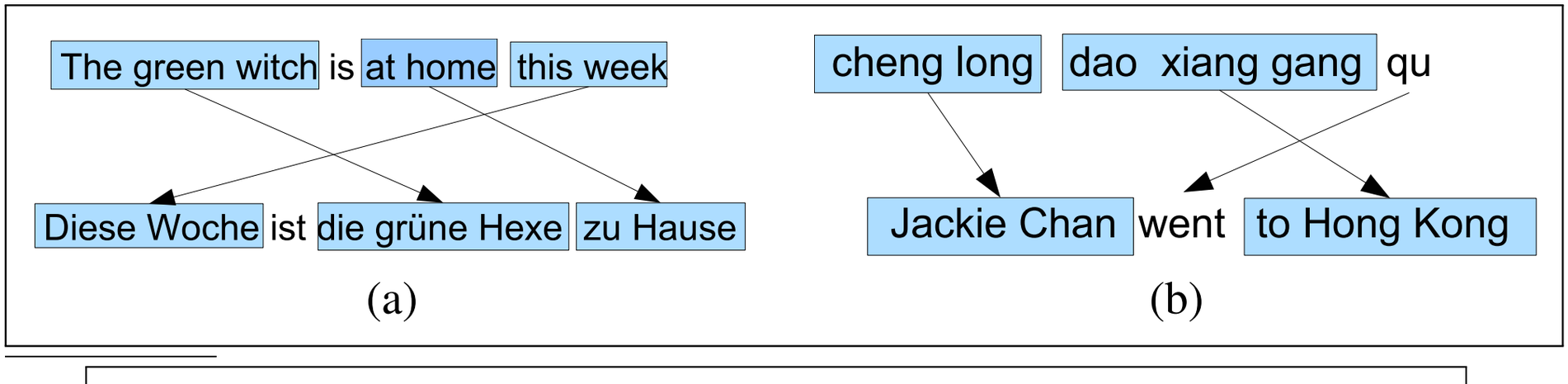
Machine Translation



Georges Artrouni's "mechanical brain", a translation device patented in France in 1933. (Image from Corbé by way of John Hutchins)

Machine Translation

- Hard problem! issues include word order and word meaning *[J&M ch. 12]*



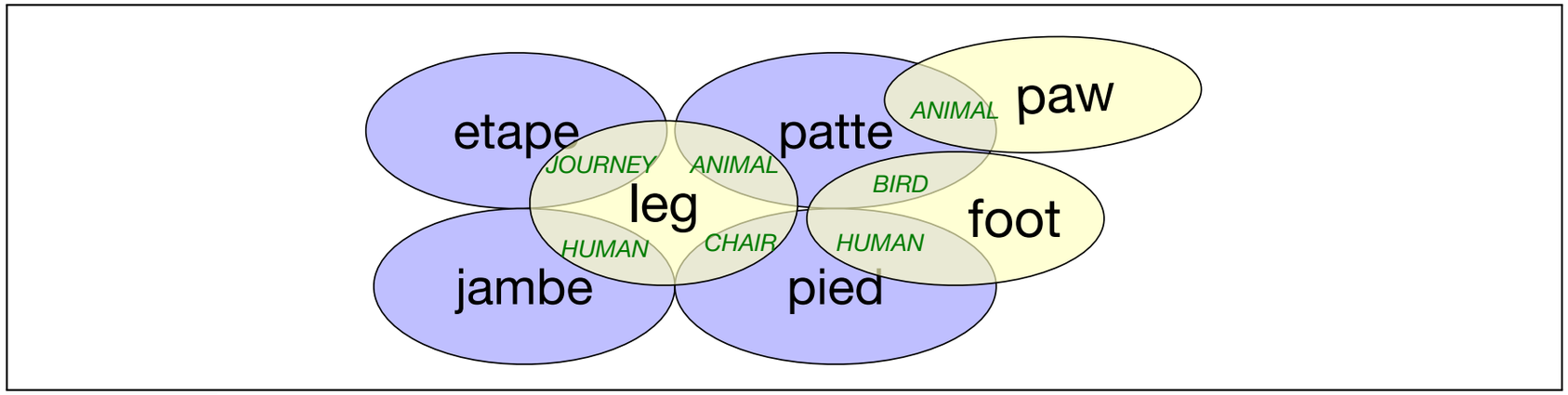


Figure 12.3 The complex overlap between English *leg*, *foot*, etc., and various French translations as discussed by Hutchins and Somers (1992).

- Exercise: Parallel data makes it possible to jointly learn alignments (attention) and cross-lingual word meanings (e.g. in the IBM Models, Brown et al. 1989)

1a. Garcia and associates.
1b. Garcia y asociados.

2a. Carlos Garcia has three associates.
2b. Carlos Garcia tiene tres asociados.

3a. his associates are not strong.
3b. sus asociados no son fuertes.

4a. Garcia has a company also.
4b. Garcia tambien tiene una empresa.

5a. its clients are angry.
5b. sus clientes están enfadados.

6a. the associates are also angry.
6b. los asociados tambien están enfadados.

7a. the clients and the associates are enemies.
7b. los clientes y los asociados son enemigos.

8a. the company has three groups.
8b. la empresa tiene tres grupos.

9a. its groups are in Europe.
9b. sus grupos están en Europa.

10a. the modern groups sell strong pharmaceuticals.
10b. los grupos modernos venden medicinas fuertes.

11a. the groups do not sell zanzanine.
11b. los grupos no venden zanzanina.

12a. the small groups are not modern.
12b. los grupos pequeños no son modernos.

exercise "solution"

Summarization

- Take in long document(s), output shorter summarization

- Major paradigms
  - Extractive summarization: select sentences, clauses, etc. to keep
  - Abstractive summarization: generate totally new text (*LLM default approach*)

- News article summarization is an often-studied version

NLG eval methods

- 1. Human (Manual) Evaluation

- 2. Reference-based Evaluation
  - (Need dataset with *reference* (gold-standard) outputs)
  - BLEU
  - ROUGE
  - BERTScore

- 3. LLM Evaluation
  - "LLM-as-judge"

Human evauation

- Have annotators directly read and evaluate system outputs

- For MT, attributes:
  - Fluency (in the target language)
  - Adequacy: does output convey meaning of input? (need bilingual annotator)

- Annotation type:
  - Likert scale (e.g. 1-5) for each attribute
  - Paired (forced-choice) comparisons

- Was done in NIST OpenMT and related regular evaluations

Reference-based metrics

- But you need something to evaluate faster on during development!

- Idea: compare predicted output to reference (gold-standard) output
  - Need some way to allow soft matching

- BLEU score: n-gram overlap
- BERTScore: contextual embedding alignment

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to match same portion of reference translation twice at a certain n-gram level (two MT words *airport* are only correct if two reference words *airport;* can't cheat by typing out "the the the the the")
  - Do count unigrams also in a bigram for unigram precision, etc.
- Brevity Penalty
  - Can't just type out single word "the" (precision 1.0!)
- It was thought quite hard to "game" the system (i.e., to find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- BLEU is a weighted geometric mean, with a brevity penalty factor added.
  - Note that it's precision-oriented
- BLEU4 formula

  (counts n-grams up to length 4)

exp (1.0 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Note: only works at corpus level (zeroes kill it); there's a smoothed variant for sentence-level

BLEU in Action

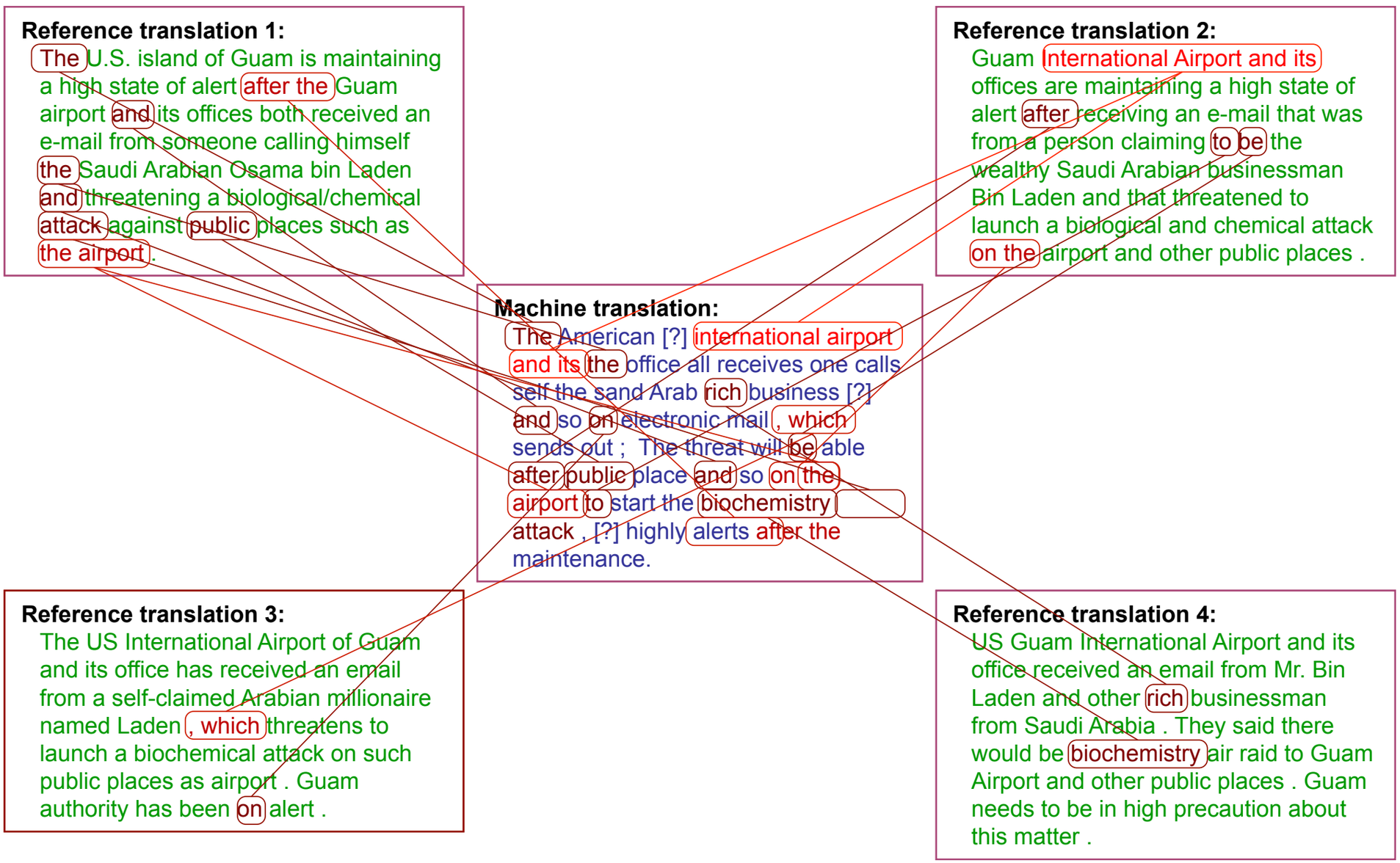
枪手被警方击毙。 **(Foreign Original)**

the gunman was shot to death by the police . **(Reference Translation)**

the gunman was police kill .	**#1**
wounded police jaya of	**#2**
the gunman was shot dead by the police .	**#3**
the gunman arrested by police kill .	**#4**
the gunmen were killed .	**#5**
the gunman was shot to death by the police .	**#6**
gunmen were killed by police ?SUB>0 ?SUB>0	**#7**
al by the police .	**#8**
the ringer is killed by the police .	**#9**
police killed the gunman .	**#10**

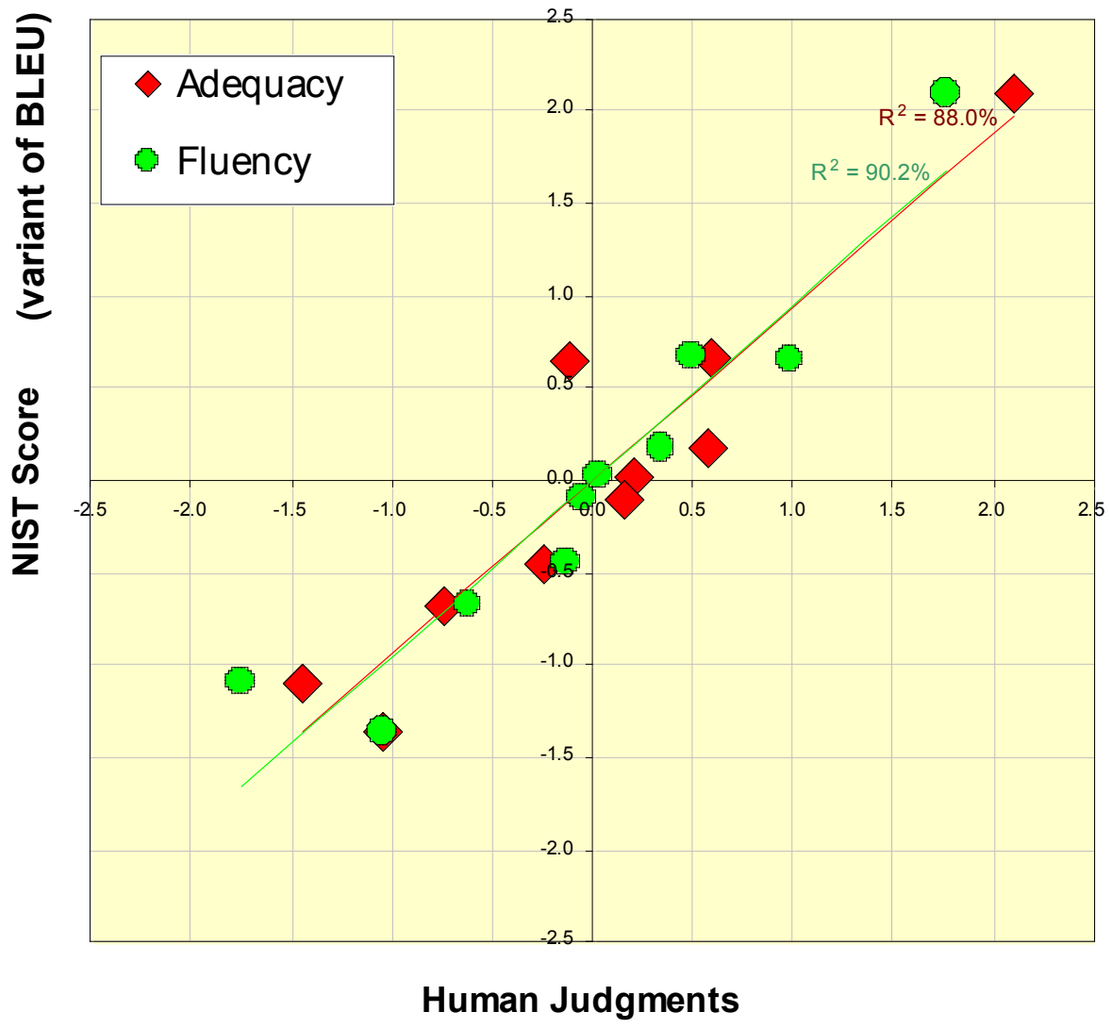
green = 4-gram match (good!)
red = word not matched (bad!)

Multiple Reference Translations



This helps alleviate a significant issue in BLEU

Initial results showed that BLEU predicts human judgments well



slide from G. Doddington (NIST)

but uneven more broadly

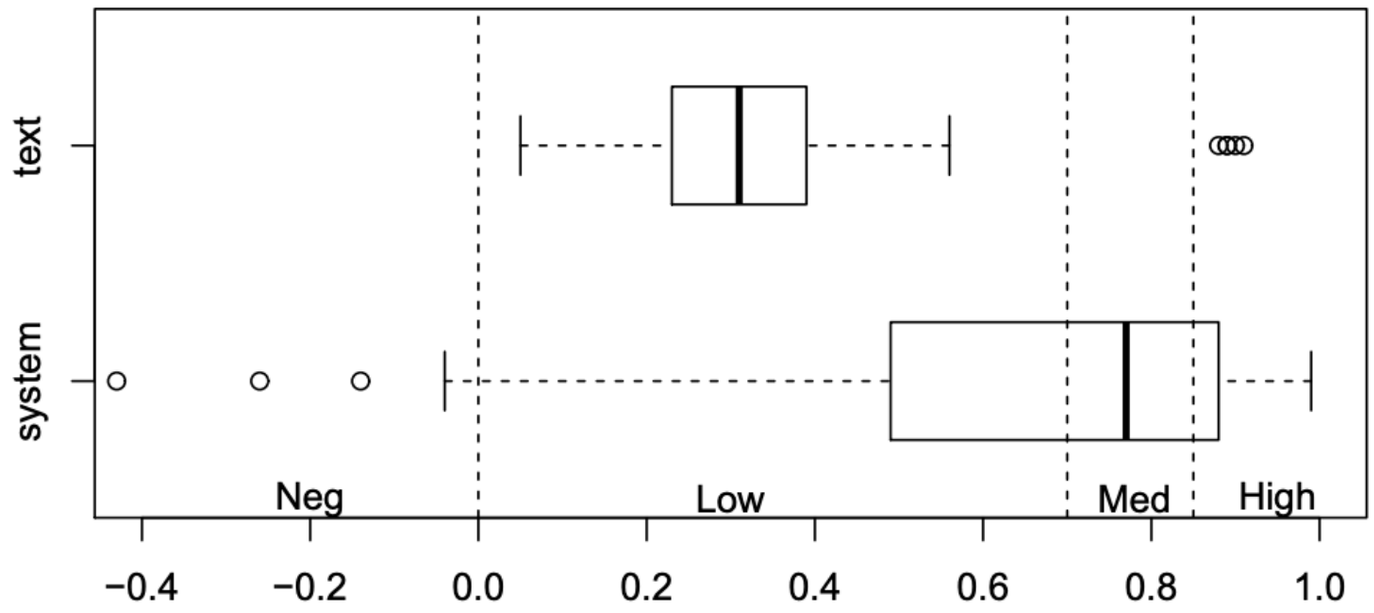


Figure 5
Box plot of BLEU–human correlations for MT, at system and text granularities.

[Reiter 2018]

- Data: from WMT07 to WMT16
- Over-optimization for the metric?
- Comparing system variations
  vs. comparing between major paradigms
- BLEU score still used; e.g. SacreBLEU software (Post, 2018)

ROUGE scores

- "Recall-Oriented ... Evaluation"
- Used in summarization.
  - Key assumption: the reference summary is the approximate length you want
- N-gram version:
  - Recall of reference's n-grams
  - ROUGE-2: bigram overlap
- ROUGE-L: longest common subsequence

BERTScore

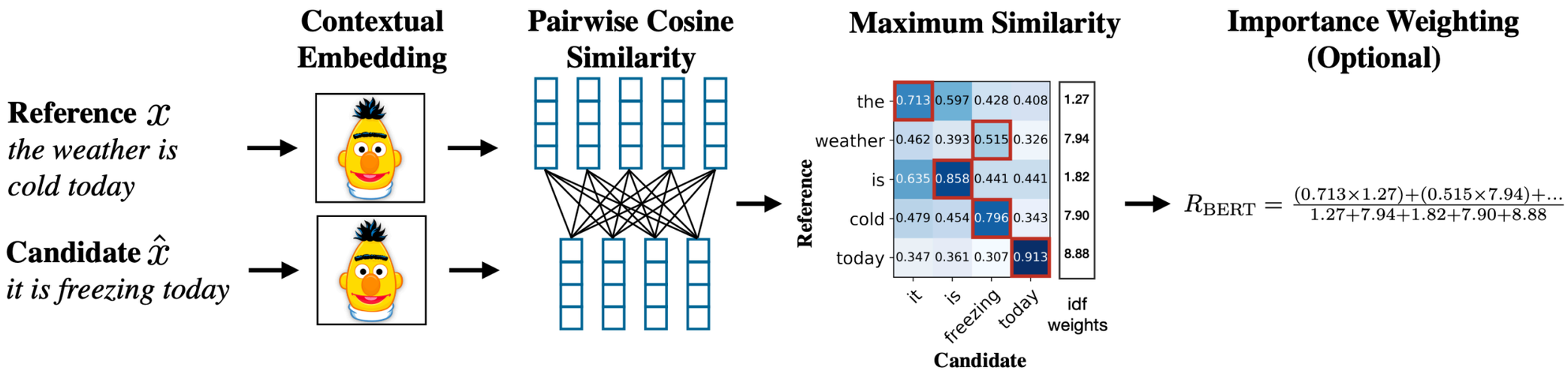


Figure 1: Illustration of the computation of the recall metric R_{BERT}. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red, and include the optional idf importance weighting.

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad , \quad F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad .$$

Reference vs Human Eval

- Data: News summarization
- From: "News Summarization and Evaluation in the Era of GPT-3" (~2023)

CNN article: https://www.cnn.com/2022/09/09/politics/judge-throws-out-trumps-rico-lawsuit-against-hillary-clinton-and-democrats/index.html

Length-constrained prompt

Summarize the above article in 2 sentences.

GPT-3: Donald Trump's lawsuit against Hillary Clinton, the Democratic National Committee, and several other individuals and entities has been dismissed by a federal judge. The judge ruled that Trump's claims are unsupported by any legal authority and that many of his specific factual assertions are implausible or unsupported.

Keyword-constrained prompt

Summarize the above article briefly focusing on Alina Habba.

GPT-3: Alina Habba is an attorney for Donald Trump. She stated that Trump will "immediately move to appeal this decision." Habba also said that the judge's opinion is "rife with erroneous applications of the law" and that it "disregards the numerous independent governmental investigations which substantiate our claim that the defendants conspired to falsely implicate our client and undermine the 2016 Presidential election."

Article: https://www.cnn.com/2022/03/01/africa/africa-condemns-racism-ukraine-intl/index.html

Prompt: Summarize the article in N sentences.

N The three African nations on the UN Security Council condemned reports of discrimination against African citizens at the Ukrainian border during a meeting at the UN HQ in New York City Monday.

N = 2 The United Nations Security Council condemned the reports of discrimination against African citizens at the Ukrainian border. The African Union has said it is "disturbed" by the reports of segregation against Africans in Ukraine, which it described as "shockingly racist."

N = 3 The article discusses the reports of discrimination against African citizens at the Ukrainian border. The representatives from the three African nations on the UN Security Council condemned the reports and called for the mistreatment of African peoples on Europe's borders to cease immediately. Foreign students attempting to flee Ukraine after Russia invaded the country told CNN that they experienced racial discrimination at the Ukrainian border.

Reference vs Human Eval

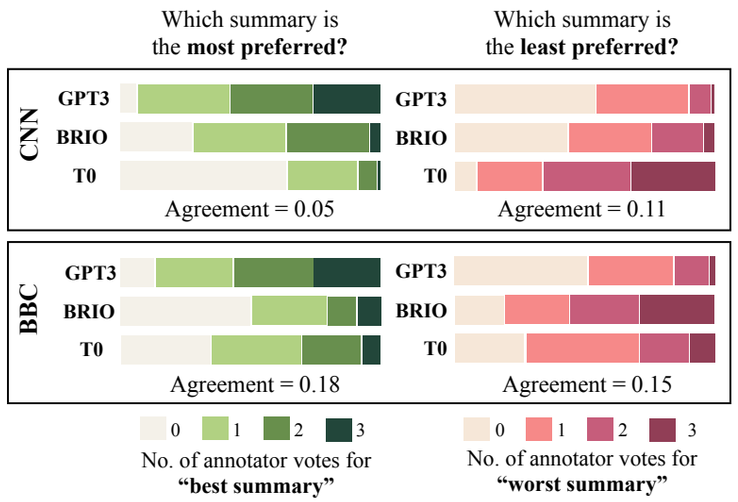


Figure 5: Annotator vote distribution for best and worst summaries across all datasets and models. Although GPT3-D2 is the clear winner according to majority vote, this choice is unanimous for less than 30% of the articles. This demonstrates the inherent variance in different annotators' definitions of "best summary", especially when comparing high-quality summaries from strong models.

Dataset	Model	Overlap-Based ROUGE(1/2/L)	METEOR	BLEU	Similarity-Based BERTScore	MoverScore
CNN	PEGASUS	34.85/14.62/28.23	.24	7.1	.858	.229
	BRIO	38.49/17.08/31.44	.31	6.6	.864	.261
	T0	35.06/13.84/28.46	.25	5.9	.859	.238
	GPT3-D2	31.86/11.31/24.71	.25	3.8	.858	.216
DailyMail	PEGASUS	45.77/23.00/36.65	.33	12.2	.865	.308
	BRIO	49.27/24.76/39.21	.37	11.7	.871	.331
	T0	42.97/19.04/33.95	.28	8.9	.863	.290
	GPT3-D2	38.68/14.24/28.08	.26	6.6	.859	.248
XSum	PEGASUS	47.97/24.82/39.63	.36	9.8	.901	.362
	BRIO	49.66/25.97/41.04	.39	10.6	.901	.372
	T0	44.20/20.72/35.84	.34	8.0	.896	.340
	GPT3-D2	28.78/7.64/20.60	.19	2.2	.869	.197
Newsroom	PEGASUS	39.21/27.73/35.68	.39	.14	.873	.272
	BRIO	-	-	-	-	-
	T0	25.64/9.49/21.41	.20	.04	.849	.145
	GPT3-D2	27.44/10.67/22.18	.22	.05	.859	.159