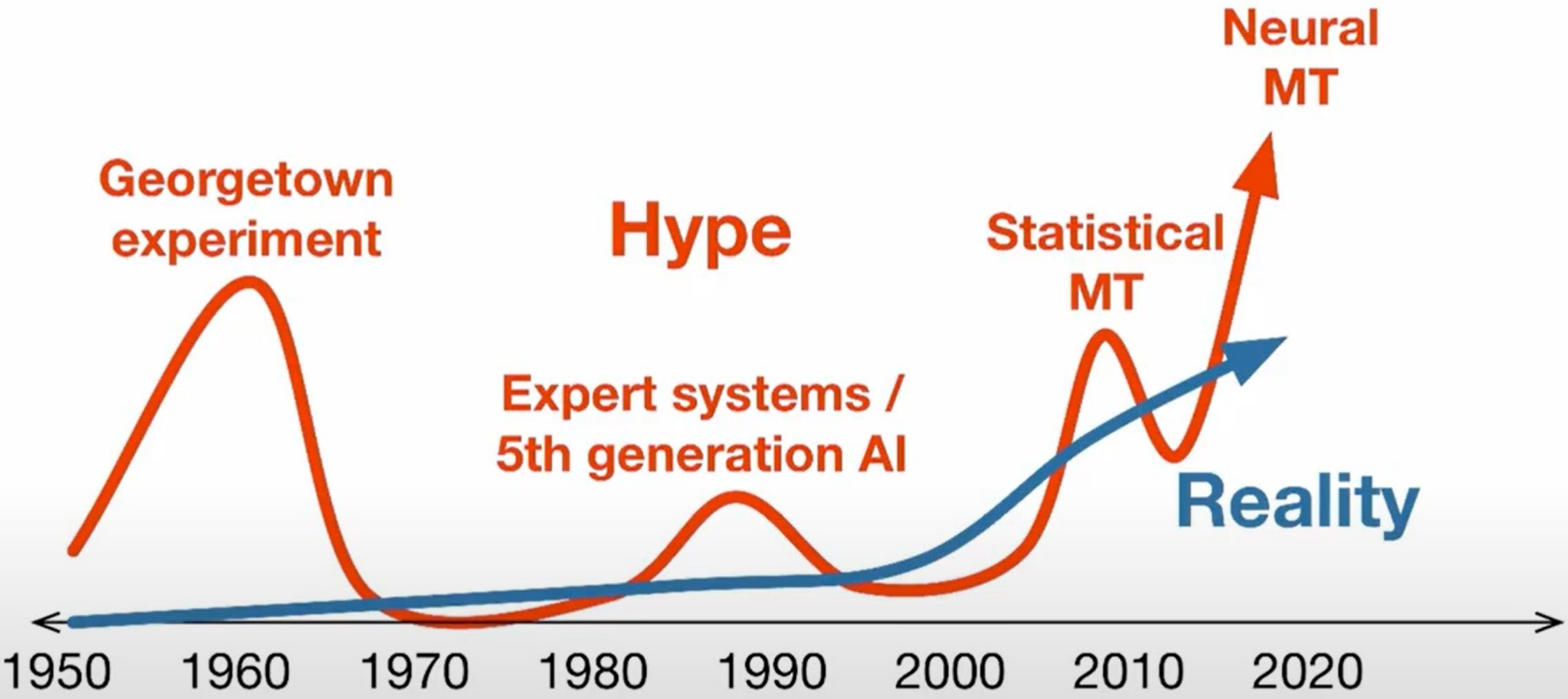
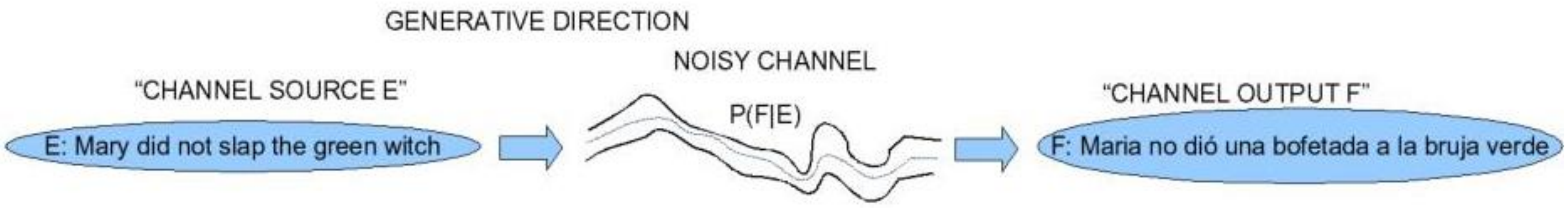


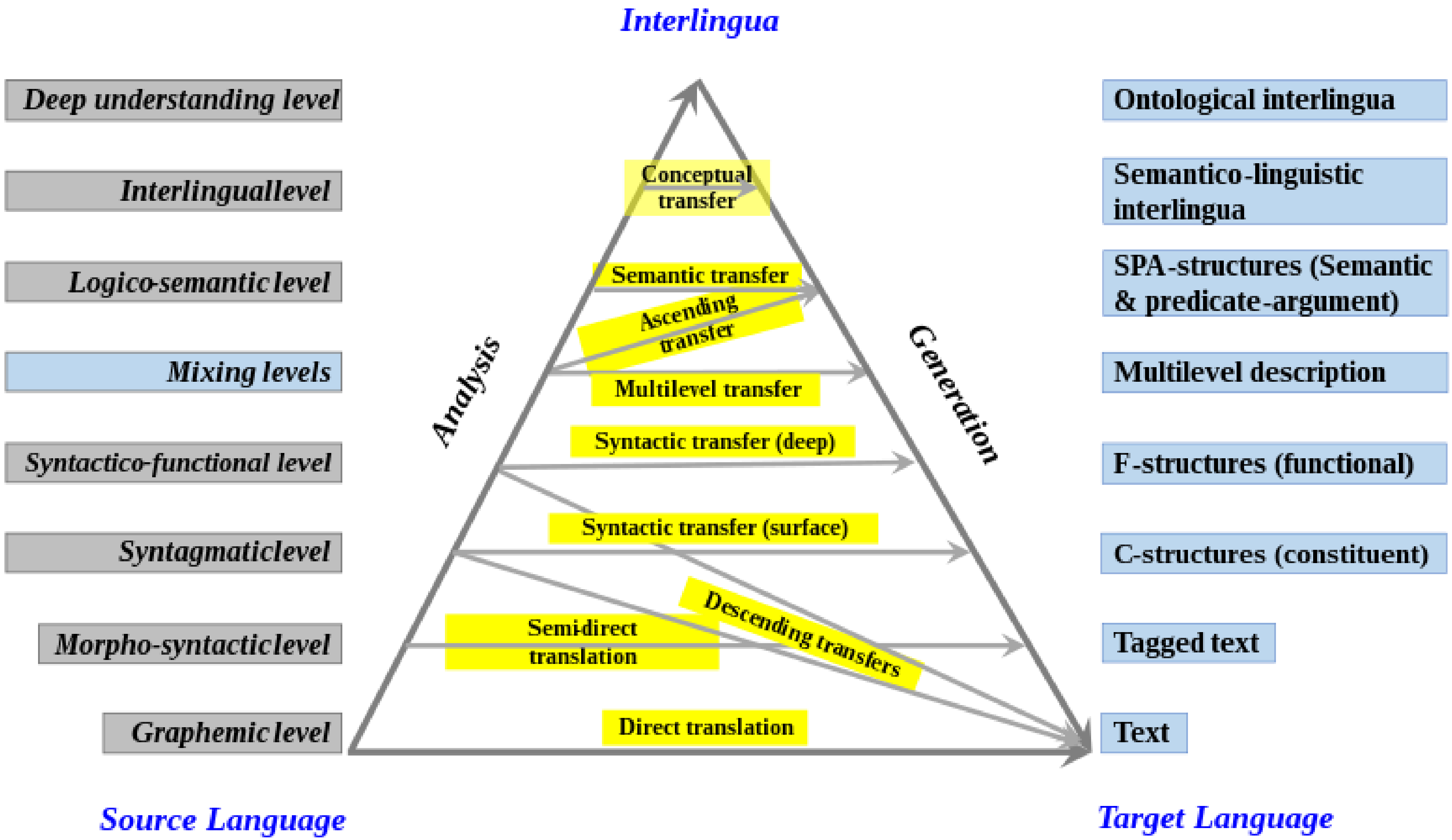
https://youtu.be/DuYkqCQEbpo

Noisy channel model



$$\begin{aligned} F^* &= \text{argmax}_F\, p(F|E) \\ &= \text{argmax}_F\, p(E|F)p(F)/p(E) \\ &= \text{argmax}_F\, \underbrace{p(E|F)}_{\text{Translation Model}}\underbrace{p(F)}_{\text{Language Model}} \end{aligned}$$

Vauquois' Pyramid



Questions to Ponder

- What makes machine translation such a difficult problem?
- Are all language pairs equally difficult to translate? Why or why not?
- How do humans translate between languages (ie. what are the cognitive steps involved)?

Project
Checkpoint #1

- Decide and submit (through teams post) the following:
  - Your team members
  - L and a short note on why you chose L and its resource levels
  - The end-to-end application over translation that you are going to build.
    - What?
    - Why is it interesting?
    - How is it useful?
  - Any challenges you foresee.
- Deadline: 7th March

LECTURE 2

Evaluation & Training

3rd Mar 2023

Evaluation of Machine Translation Systems

- What are the Challenges?
- What are the features/dimensions?

HUMAN EVALUATION

5-point **comprehensibility** and fluency scale

Grade -1	No Output OR buffer clearance issue
Grade 0	Nonsense (If the sentence doesn't make any sense at all – it is like someone speaking to you in a language you don't know)
Grade 1	Some parts make sense but is not comprehensible over all (e.g., listening to a language which has lots of borrowed words from your language – you understand those words but nothing more)
Grade 2	Comprehensible but has quite a few errors (e.g., someone who can speak your language but would make lots of errors. However, you can make sense out of what is being said)
Grade 3	Comprehensible, occasional errors (e.g., someone speaking Hindi getting all its genders wrong)
Grade 4	Perfect (e.g., someone who knows the language)

Automatic Evaluation

BLEU = Bilingual Evaluation Understudy Score

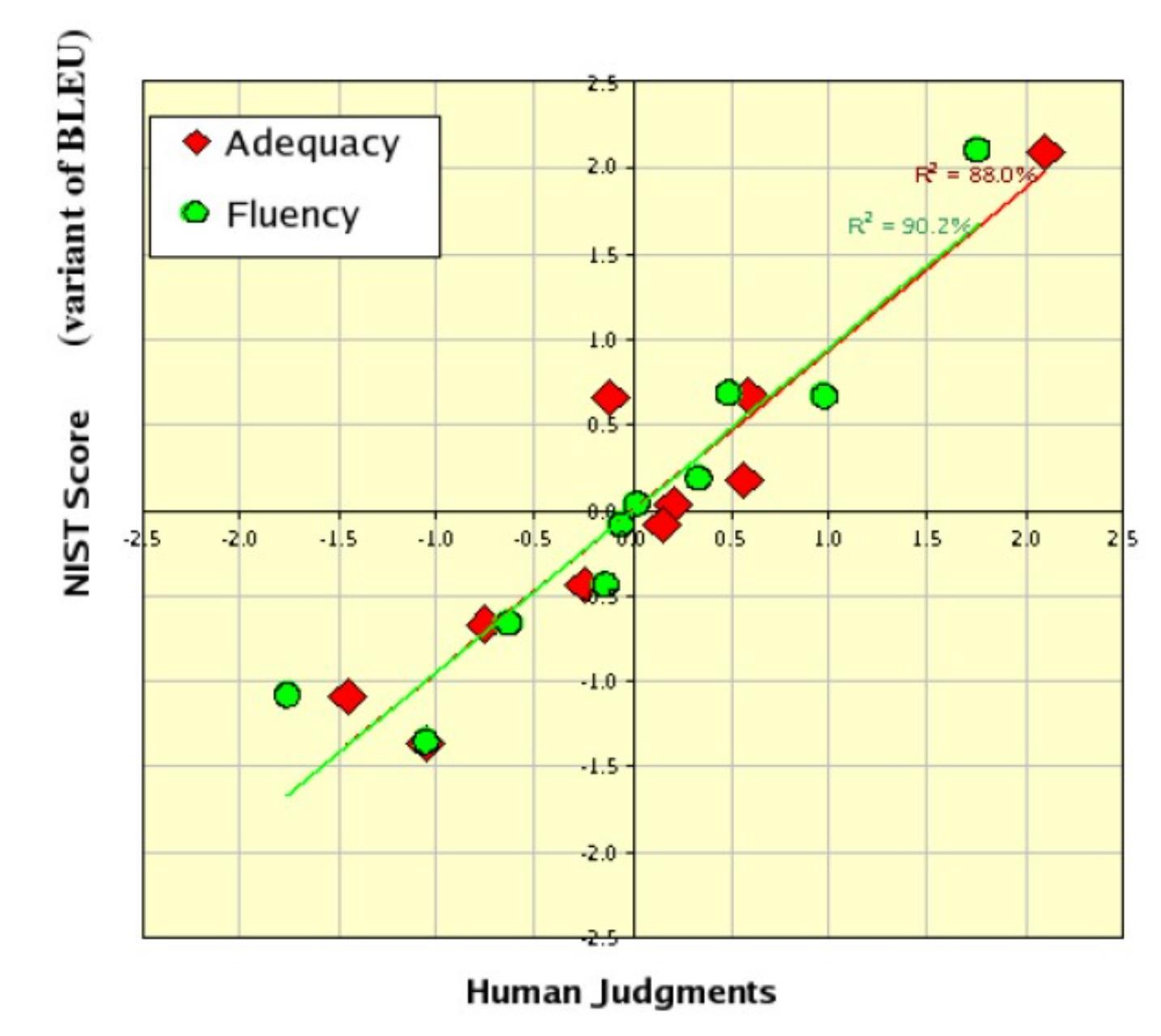
- N-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

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

- Typically computed over the entire corpus, not single sentences

Autimatic Evaluation

BLEU = Bilingual Evaluation Understudy Score



http://webcast.in2p3.fr/videos-how_can_we_measure_machine_translation_quality

What is a Better Translation? Reflections on Six Years of Running Evaluation Campaigns

Philipp Koehn

Automatic Evaluation

Adaptations of BLEU

- NIST: Like BLEU but higher weights to rare n-grams
- METEOR (Metric for Evaluation of Translation with Explicit ORdering): Considers recall, precision, word order and synonyms.

CharacTER

Translation Edit-distance

(in post-editing machine translation)

How many edit operations are required to convert the translator output to a satisfactory form?

Two paradigms of machine translation:

- Information preserving non-critical
- Information+style preserving critical

CheckList: Behavioral Testing of NLP Models
(ACL 2020 best paper)

- Usually, evaluation of NLP systems results in a single number. However, it is important to know which linguistic (and other) capabilities the system has, and which it doesn’t.

- Therefore, instead of a functional testing, can we resort to behavioral testing of NLP models?

Case-Study I: Sentiment Analysis

Labels: positive, negative, or neutral; INV: same pred. (INV) after removals/ additions; DIR: sentiment should not decrease (↑) or increase (↓)

	Test *TYPE* and Description	Failure Rate (%)					Example test cases & expected behavior
						RoB	
Vocab.+POS	***MFT:*** Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is Australian. neutral That is a private aircraft. neutral
	MFT: Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is extraordinary. pos I despised that aircraft. neg
	INV: Replace neutral words with other neutral words	9.4	16.2	12.4	10.2	10.2	@Virgin should I be concerned that → when I'm about to fly ... INV @united the → our nightmare continues... INV
	DIR: Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir Great trip on 2672 yesterday... You are extraordinary. ↑ @AmericanAir AA45 ... JFK to LAS. You are brilliant. ↑
	DIR: Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways your service sucks. You are lame. ↓ @JetBlue all day. I abhor you. ↓
Robust.	***INV:*** Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that selfie was extreme. @pi9QDK INV @united stuck because staff took a break? Not happy 1K.... https://t.co/PWK1jb INV
	INV: Swap one character with its neighbor (typo)	5.6	10.2	10.4	5.2	3.8	@JetBlue → @JeBtlue I cri INV @SouthwestAir no thanks → thakns INV
NER	***INV:*** Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba → Canada... INV @VirginAmerica I miss the #nerdbird in San Jose → Denver INV
	INV: Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon → Erin was your saviour INV @united 8602947, Jon → Sean at http://t.co/58tuTgli0D, thanks. INV
Temporal	***MFT:*** Sentiment change over time, present should prevail	41.0	36.6	42.2	18.8	11.0	I used to hate this airline, although now I like it. pos In the past I thought this airline was perfect, now I think it is creepy. neg
Negation	***MFT:*** Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not poor. pos or neutral It isn't a lousy customer service. pos or neutral
	MFT: Negated neutral should still be neutral	40.4	39.6	74.2	98.4	95.4	This aircraft is not private. neutral This is not an international flight. neutral
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plane would be awful, but it wasn't. pos or neutral I thought I would dislike that plane, but I didn't. pos or neutral
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, given it's a Tuesday, that this pilot was great. neg I don't think, given my history with airplanes, that this is an amazing staff. neg
SRL	***MFT:*** Author sentiment is more important than of others	45.4	62.4	68.0	38.8	30.0	Some people think you are excellent, but I think you are nasty. neg Some people hate you, but I think you are exceptional. pos
	MFT: Parsing sentiment in (question, "yes") form	9.0	57.6	20.8	3.6	3.0	Do I think that airline was exceptional? Yes. neg Do I think that is an awkward customer service? Yes. neg
	MFT: Parsing sentiment in (question, "no") form	96.8	90.8	81.6	55.4	54.8	Do I think the pilot was fantastic? No. neg Do I think this company is bad? No. pos or neutral

Table 1: A selection of tests for sentiment analysis. All examples (right) are failures of at least one model.

CheckList

Capability			
Vocabulary			
NER			
Negation			
…			

CheckList

Capability	Minimum Functionality Test	Invariance	Directional
Vocabulary			
NER			
Negation			
…			

CheckList

Capability	Minimum Functionality Test	Invariance	Directional
Vocabulary			
NER			
Negation			
…			

Template I: I <**NEGATION**> <**POS_VERB**> the <**THING**>
Test1: *I did not like the acting.*
Test2: *I can't say I recommend the book.*
Test3: …
Template II: I thought <**POS_STMT**>, but <**PRON** ><**VERB**> not.
Test1: *I thought the movie was great, but it was not.*
Test2: *I thought I will like the book, but I did not.*

CheckList

Capability	Minimum Functionality Test	Invariance	Directional
Vocabulary			
NER			
Negation			
…			

Same prediction after changing the named entity.

Test1: *Thanks to the staff, we were put in another flight to Delhi → Bangalore*

Test2: *@SouthWestern Great trip on 2374 → 7753 yesterday.*

Test3: …

Same prediction after addition of a random URL

Test1: *The movie is certainly worthy of watching. aka.ms\gluecos*

Test2: *You won't regret visiting this hotel. Github.io/microsoft*

CheckList

Capability	Minimum Functionality Test	Invariance	Directional
Vocabulary			
NER			
Negation			
…			

Adding intensifiers should not change sentiment polarity

Test1: *I absolutely love this restaurant.*

Test2: *I couldn't move past the third chapter. It was very boring.*

Test3: …

Adding negative/positive phrases at the end should not make predictions more positive/negative.

Test1: *Service wasn't great. You are lame*.

Test2: *You won't regret visiting this hotel. It is fantastic*.

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral			
`Template: I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos / neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral / neg	X
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR*** Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg / neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg / neutral	X
...			
Failure rate = 34.6%			

Figure 1: CheckListing a commercial sentiment analysis model (G). Tests are structured as a conceptual matrix with capabilities as rows and test types as columns (examples of each type in A, B and C).

How CheckList works

- Decide on capabilities.
- Come up with test templates
- Generate test cases with the help of tools
  - Available online
  - Variety of abstractions and support provided, including MLM type predictions to support semi-automatic test case generation.
- Run the system and calculate accuracy
- Report the findings in the table

Problem to ponder

- Can you use LLMs like ChatGPT to evaluate translations?
- Does Automatic metrics work equally well for all languages and language pairs?

Further Reading

- https://slator.com/machine-translation/a-quick-primer-on-edit-distance-a-key-metric-in-post-editing-machine-translation/
- https://www.topbots.com/evaluation-metrics-for-dialog-systems/#:~:text=Evaluation%20is%20a%20crucial%20part%20of%20the%20dialog,rely%20on%20automatic%20metrics%20when%20developing%20dialog%20systems.
-