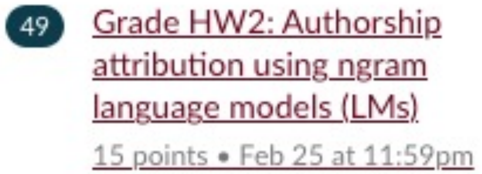
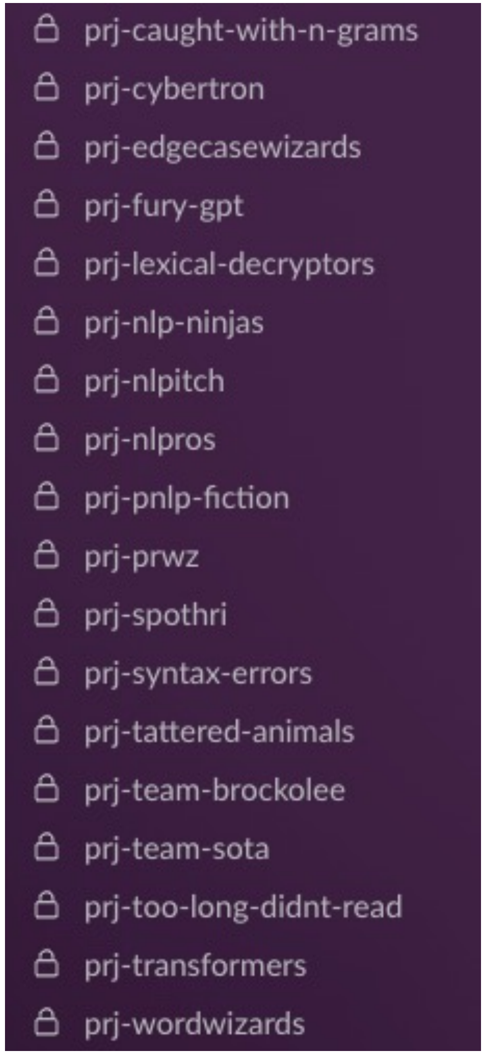
Announcement (0227)

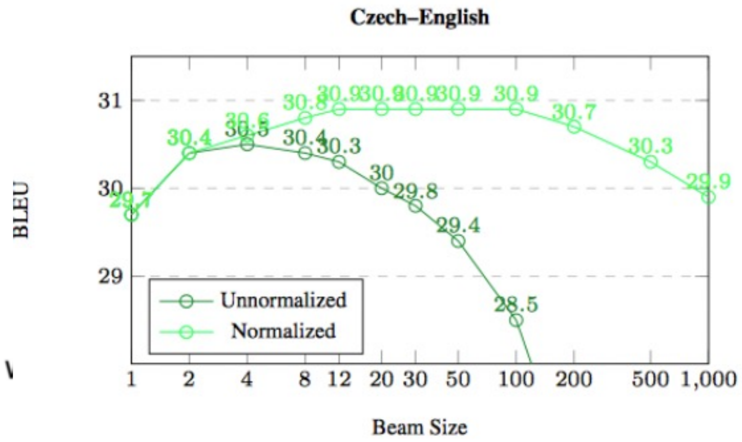


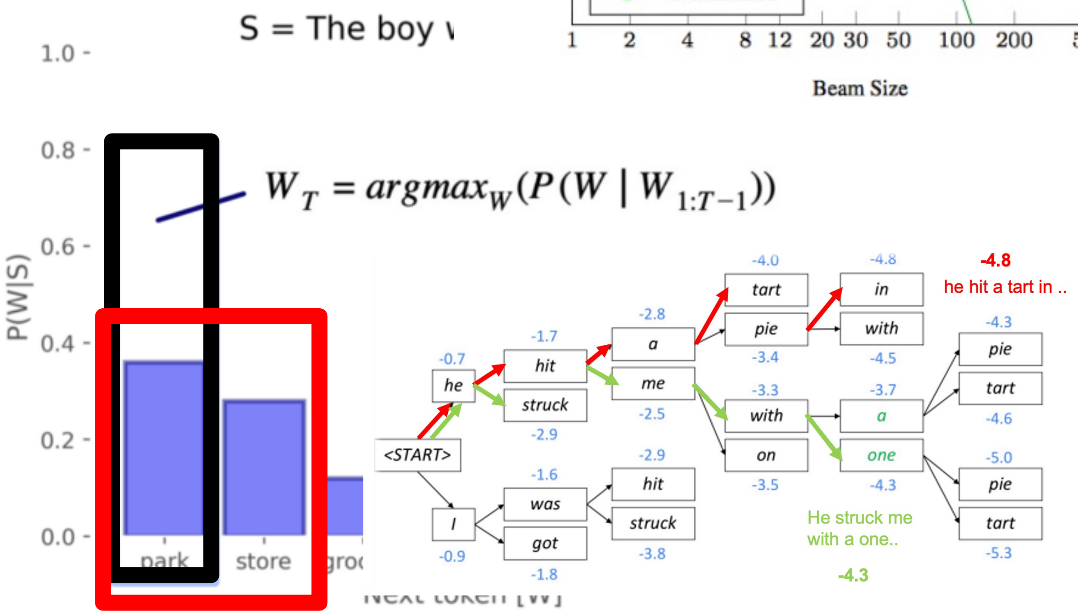
- ❑ Don’t give up your homework #2!
- ❑ Project brainstorming (1 point, due: Mar 1)
- ❑ Continue lecture on “LM: Search and decoding”
- ❑ New lecture on “LM: Evaluation and Applications”
- ❑ HW3 out



Sampling

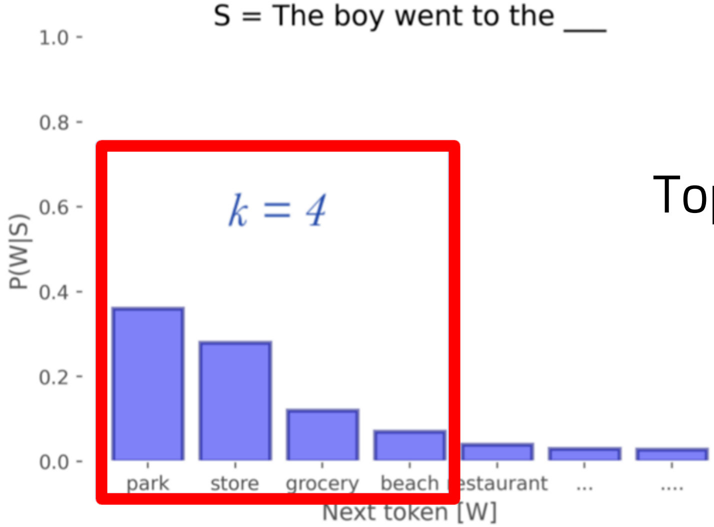
Recap



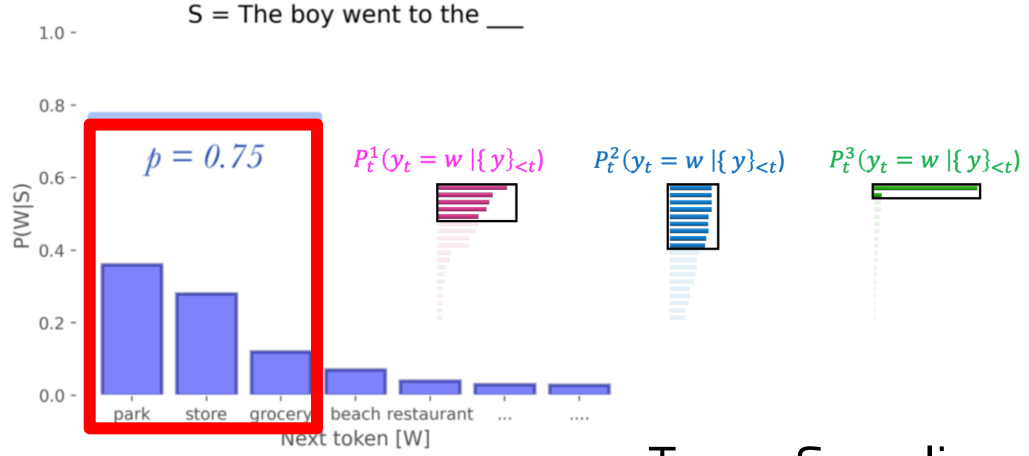


$$W_T = argmax_W(P(W \mid W_{1:T-1}))$$

Greedy/Beam-Search



Top-k Sampling



Top-p Sampling

Reward Estimation

$$\mathcal{L}_{RL} = -\sum_{t=1}^{T} (r(\hat{y}_t) - b) \log P(\ldots)$$

Recap

- How should we define a reward function? Just use your evaluation metric!
  - BLEU (machine translation; Ranzato et al., ICLR 2016; Wu et al., 2016)
  - ROUGE (summarization; Paulus et al., 2018; Celikyilmaz et al., 2018)
  - CIDEr (image captioning; Rennie et al., CVPR 2017)
  - SPIDEr (image captioning; Liu et al., ICCV 2017)
- Be careful about optimizing for the task as opposed to "gaming" the reward!
  - Evaluation metrics are merely proxies for generation quality!
  - "*even though RL refinement can achieve better BLEU scores, it barely improves the human impression of the translation quality*" – Wu et al., 2016

Reward Estimation

Recap

- What behaviors can we tie to rewards?
  - Sentence simplicity (Zhang and Lapata, EMNLP 2017)
  - Temporal Consistency (Bosselut et al., NAACL 2018)
  - Cross-modality consistency in image captioning (Ren et al., CVPR 2017)
  - Utterance Politeness (Tan et al., TACL 2018)
  - Paraphrasing (Li et al., EMNLP 2018)
  - Sentiment (Gong et al., NAACL 2019)
  - Formality (Gong et al., NAACL 2019)
- If you can formalize a behavior as a Python function (or train a neural network to approximate it!), you can train a text generation model to exhibit that behavior!

Search in Training: Takeaways

Recap

- Teacher forcing is still the main algorithm for training text generation models
- Diversity is an issue with sequences generated from teacher forced models
- Exposure bias causes text generation models to lose coherence easily
- Training with RL can allow models to learn behaviors that are challenging to formalize
    - But learning can be very unstable!
    - chatGPT: advanced RL algorithms (e.g., PPO) for better human alignment with human feedback

Other techniques not covered

- ❑ Decoding time control for controllable text generation (e.g., PPLM)
- ❑ Multi-attribute control using RL (will be covered)
- ❑ Unlikelihood training
- ❑ Data augmentation for reducing the exposure bias
- ❑ Retrieval-augmented Generation (RAG)
- ❑ Retrieval based generation (e.g., KNN Language Models)
- ❑ Instruction tuning and human feedback learning (will be covered)

…

CSCI 5541: Natural Language Processing

Lecture 8: Language Models: Evaluation and Applications

Dongyeop Kang (DK), University of Minnesota

dongyeop@umn.edu | twitter.com/dongyeopkang | dykang.github.io

computer science & engineering

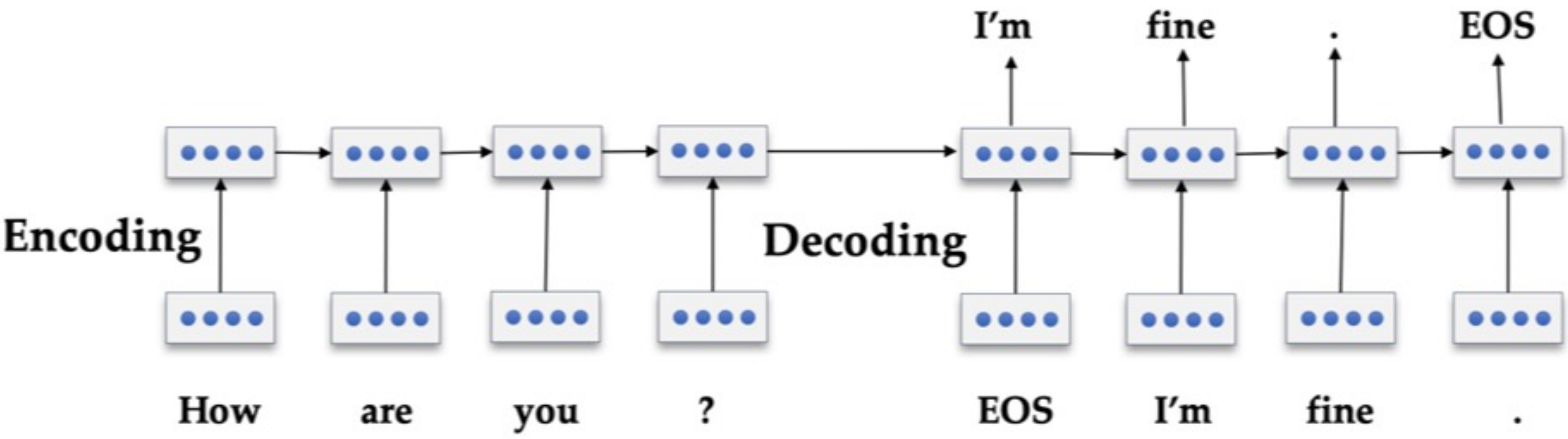
MINNESOTA NLP EST. 2021

University of Minnesota
Driven to Discover®

Applications

Dialogue Generation

Seq2Seq based chatbot



Knowledge/world grounding

Persuasion

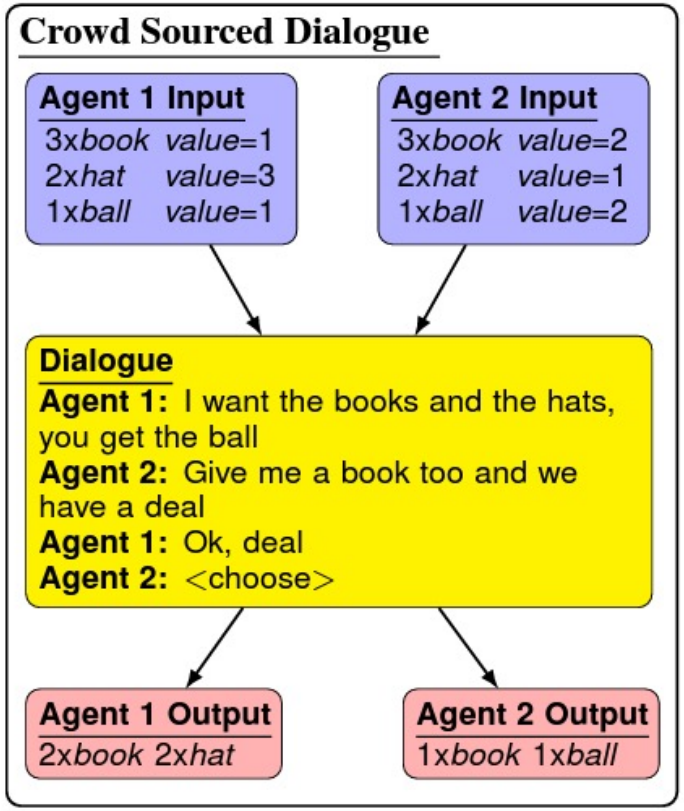
Negotiation

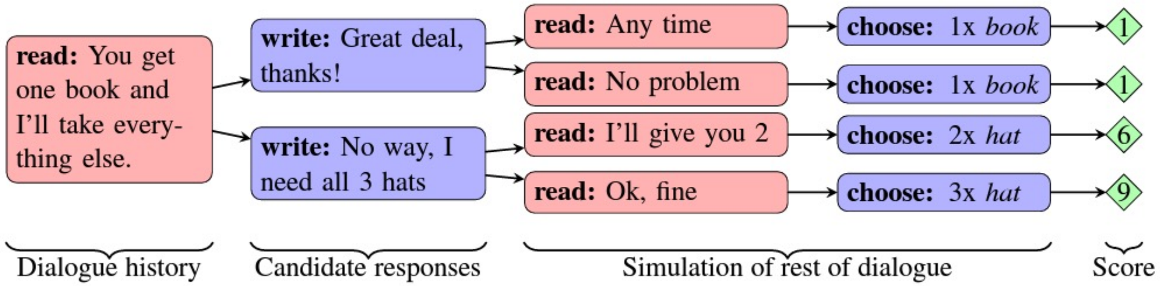
Collect human-human conversations with specific conditions/goals and computationally model their behaviors

Recommendation

Personalization

Knowledge





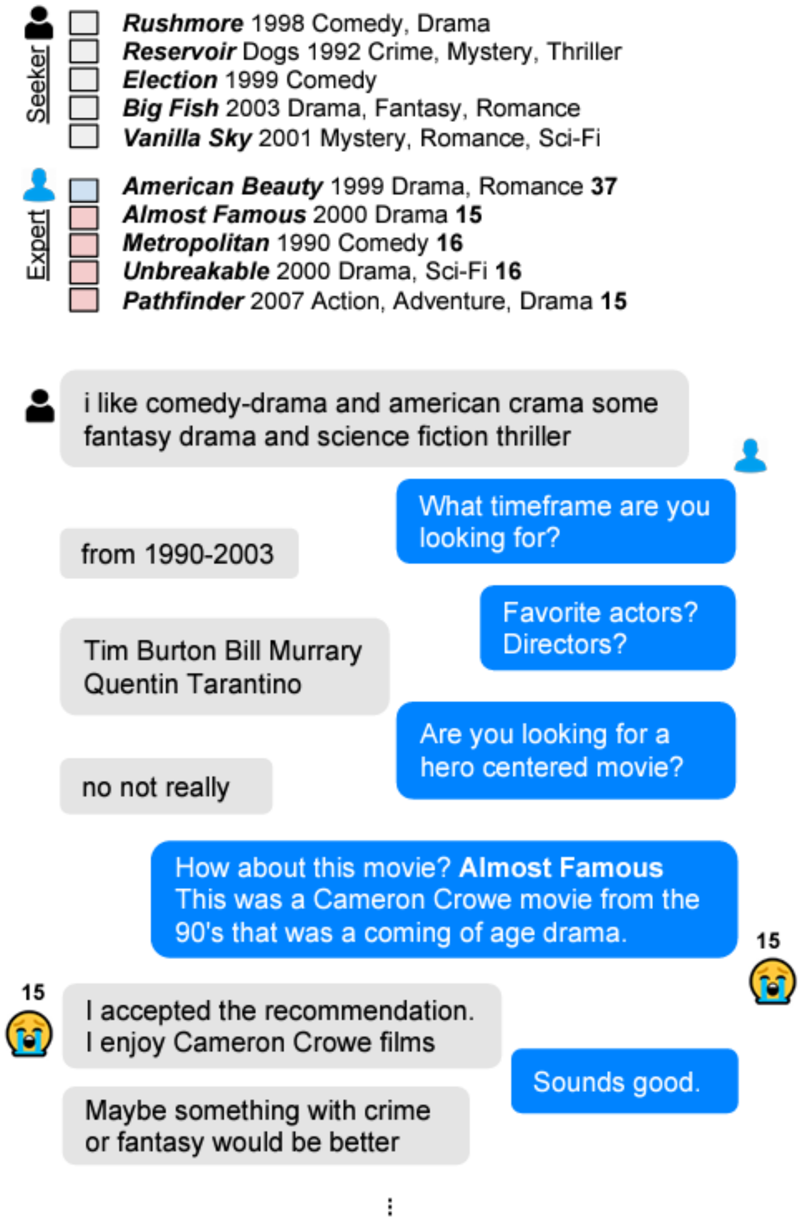
Deal or No Deal? End-to-End Learning for Negotiation Dialogues

Friends of agent A:

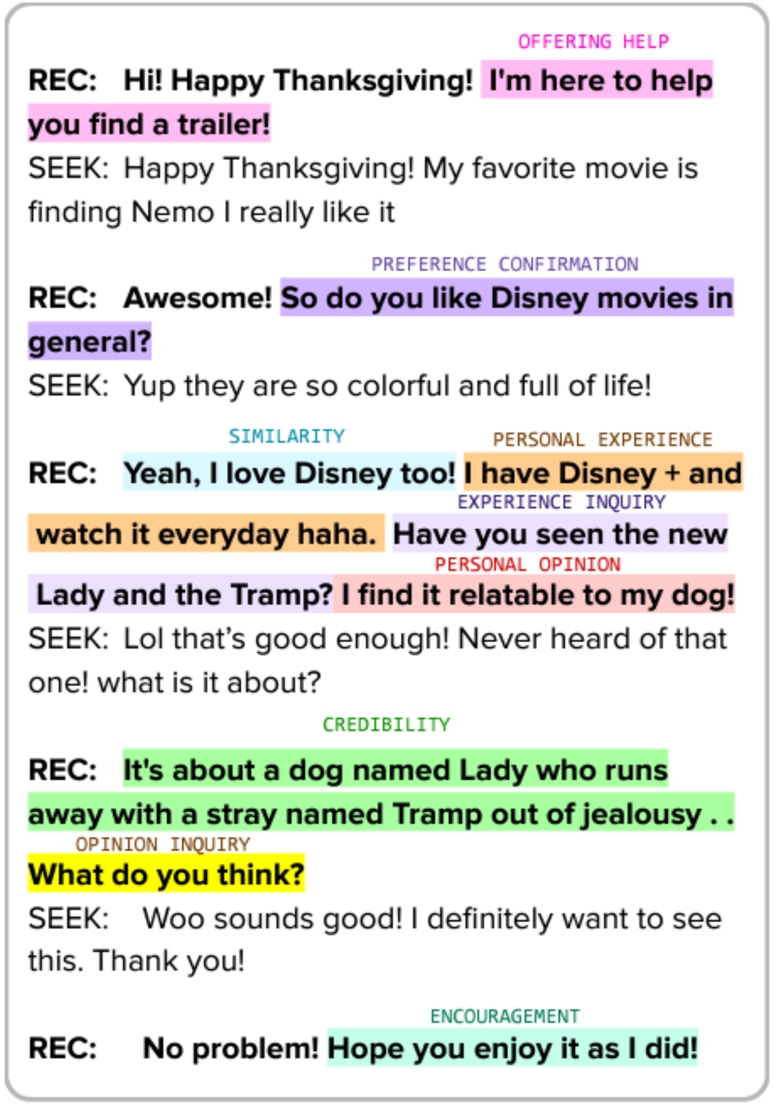
Name	School	Major	Company
Jessica	Columbia	Computer Science	Google
Josh	Columbia	Linguistics	Google
...	...	...	...

A: Hi! Most of my friends work for Google
B: do you have anyone who went to columbia?
A: *Hello?*
A: I have Jessica a friend of mine
A: and Josh, both went to columbia
B: *or anyone working at apple?*
B: SELECT (Jessica, Columbia, Computer Science, Google)
A: SELECT (Jessica, Columbia, Computer Science, Google)

Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings



Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue



INSPIRED: Toward Sociable Recommendation Dialog Systems

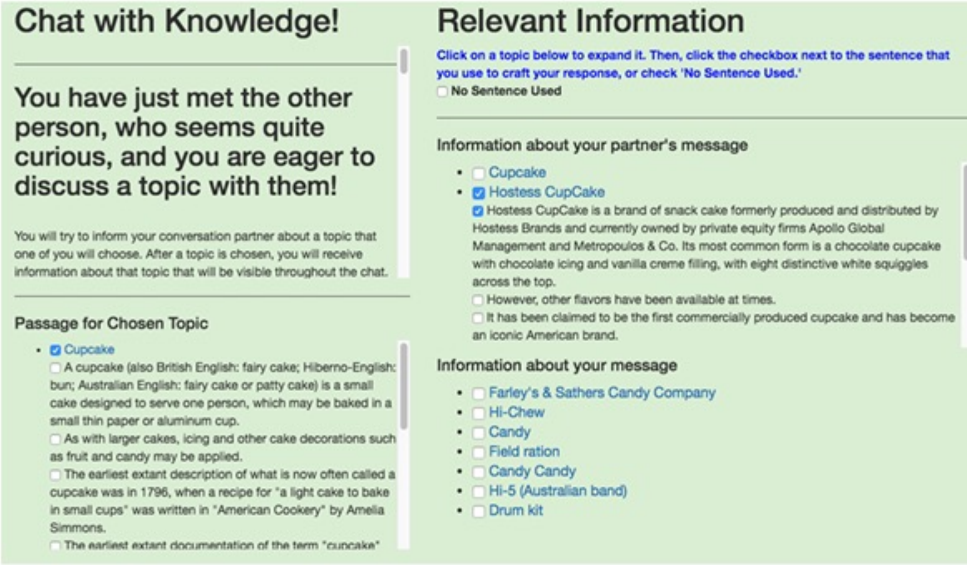
Role	Utterance	Annotation
ER	Hello, are you interested in protection of rights of children?	Source-related inquiry
EE	Yes, definitely. What do you have in mind?	
ER	There is an organisation called Save the Children and donations are essential to ensure children's rights to health, education and safety.	Credibility appeal
EE	Is this the same group where people used to "sponsor" a child?	
ER	Here is their website, https://www.savethechildren.org/. They help children all around the world. For instance, millions of Syrian children have grown up facing the daily threat of violence. In the first two months of 2018 alone, 1,000 children were reportedly killed or injured in intensifying violence.	Credibility appeal Credibility appeal Emotion appeal Emotion appeal
EE	I can't imagine how terrible it must be for a child to grow up inside a war zone.	
ER	As you mentioned, this organisation has different programs, and one of them is to "sponsor" child. You choose the location.	Credibility appeal Credibility appeal
EE	Are you connected with the NGO yourself?	
ER	No, but i want to donate some amount from this survey. Research team will send money to this organisation.	Self-modeling Donation information
EE	That sounds great. Does it come from our reward/bonuses?	
ER	Yes, the amount you want to donate is deducted from your reward.	Donation information
EE	What do you have in mind?	
ER	I know that my small donation is not enough, so i am asking you to also donate some small percentage from reward.	Proposition of donation
EE	I am willing to match your donation.	
ER	Well, if you go for full 0.30 i will have no moral right to donate less.	Self-modeling
EE	That is kind of you. My husband and I have a small NGO in Mindanao, Philippines, and it is amazing what a little bit of money can do to make things better.	
ER	Agree, small amount of money can mean a lot for people in third world countries. So agreed? We donate full reward each??	Foot-in-the-door Donation confirmation
EE	Yes, let's donate $0.30 each. That's a whole lot of rice and flour. Or a whole lot of bandages.	

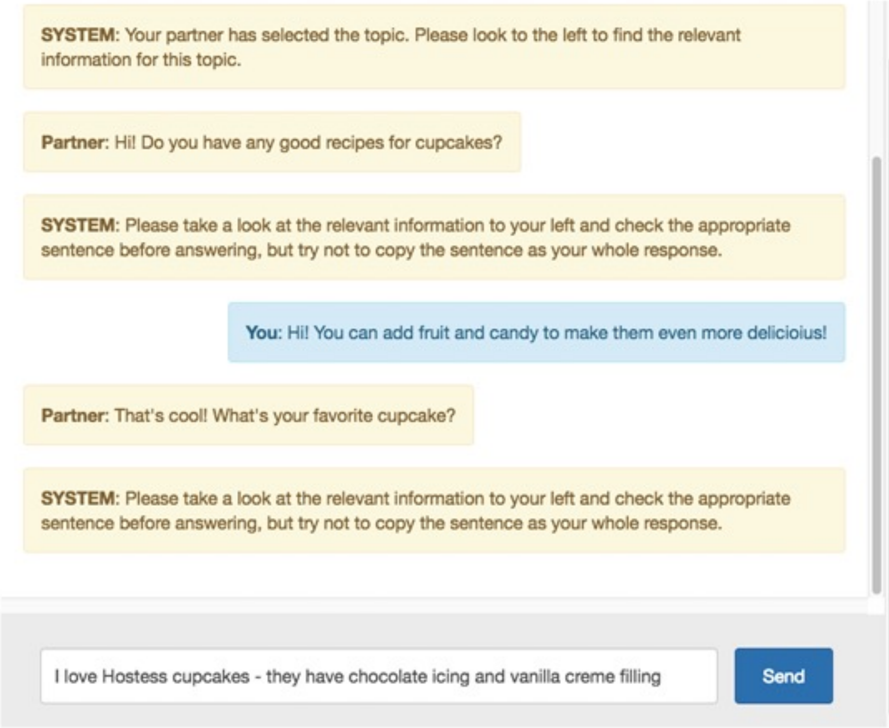
Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good

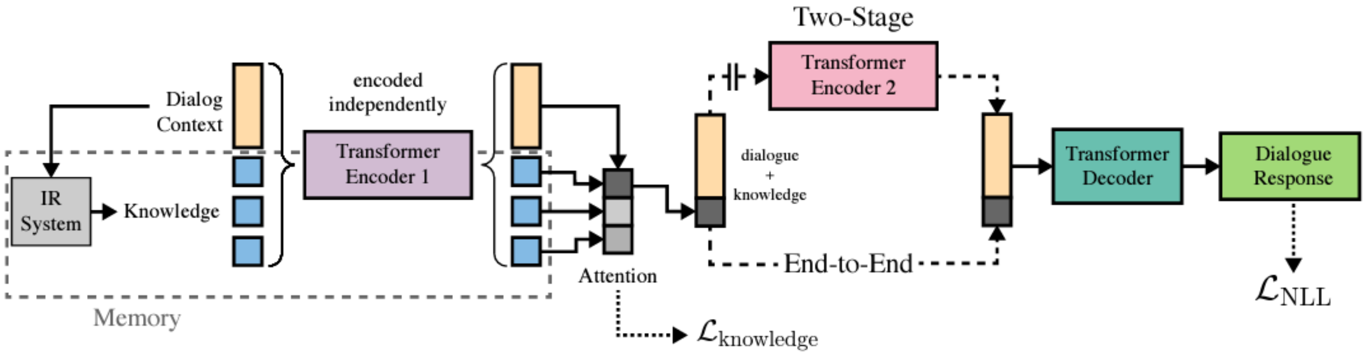
Persona 1	Persona 2
I like to ski My wife does not like me anymore I have went to Mexico 4 times this year I hate Mexican food I like to eat cheetos	I am an artist I have four children I recently got a cat I enjoy walking for exercise I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Personalizing Dialogue Agents: I have a dog, do you have pets too?







Topic:	Shower (*unseen*)
Human:	So I've been looking for a better showerhead with more water pressure for my shower.
Model:	I love pressure-versed water at nearly constant temperature to a shower
Human:	That sounds pretty great! Love those rainshower showers!
Model:	the ancient greeks were the first people to have showers.
Human:	Well I knew they were good for something.
Model:	the original showers were neither indoor structures nor man made but were common.
Human:	Ah that sounds intimidating!
Model:	it is required manual transport of both fresh and waste water.

Wizard of Wikipedia: Knowledge-Powered Conversational agents

Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.
Listener: I'm sorry to hear that. I wish I could help you figure it out

Label: Proud
Situation: Speaker felt this when...
"I finally got that promotion at work! I have tried so hard for so long to get it!"
Conversation:
Speaker: I finally got promoted today at work!
Listener: Congrats! That's great!
Speaker: Thank you! I've been trying to get it for a while now!
Listener: That is quite an accomplishment and you should be proud!

Emotion	Most-used speaker words	Most-used listener words	Training set emotion distrib
Surprised	got,shocked,really	that's,good,nice	5.1%
Excited	going,wait,i'm	that's,fun,like	3.8%
Angry	mad,someone,got	oh,would,that's	3.6%
Proud	got,happy,really	that's,great,good	3.5%
Sad	really,away,get	sorry,oh,hear	3.4%
Annoyed	get,work,really	that's,oh,get	3.4%
Grateful	really,thankful,i'm	that's,good,nice	3.3%
Lonely	alone,friends,i'm	i'm,sorry,that's	3.3%
Afraid	scared,i'm,night	oh,scary,that's	3.2%
Terrified	scared,night,i'm	oh,that's,would	3.2%
Guilty	bad,feel,felt	oh,that's,feel	3.2%
Impressed	really,good,got	that's,good,like	3.2%
Disgusted	gross,really,saw	oh,that's,would	3.2%
Hopeful	i'm,get,really	hope,good,that's	3.2%
Confident	going,i'm,really	good,that's,great	3.2%
Furious	mad,car,someone	oh,that's,get	3.1%
Anxious	i'm,nervous,going	oh,good,hope	3.1%
Anticipating	wait,i'm,going	sounds,good,hope	3.1%
Joyful	happy,got,i'm	that's,good,great	3.1%
Nostalgic	old,back,really	good,like,time	3.1%
Disappointed	get,really,work	oh,that's,sorry	3.1%
Prepared	ready,i'm,going	good,that's,like	3%
Jealous	friend,got,get	get,that's,oh	3%
Content	i'm,life,happy	good,that's,great	2.9%
Devastated	got,really,sad	sorry,oh,hear	2.9%
Embarrassed	day,work,got	oh,that's,i'm	2.9%
Caring	care,really,taking	that's,good,nice	2.7%
Sentimental	old,really,time	that's,oh,like	2.7%
Trusting	friend,trust,know	good,that's,like	2.6%
Ashamed	feel,bad,felt	oh,that's,i'm	2.5%
Apprehensive	i'm,nervous,really	oh,good,well	2.4%
Faithful	i'm,would,years	good,that's,like	1.9%

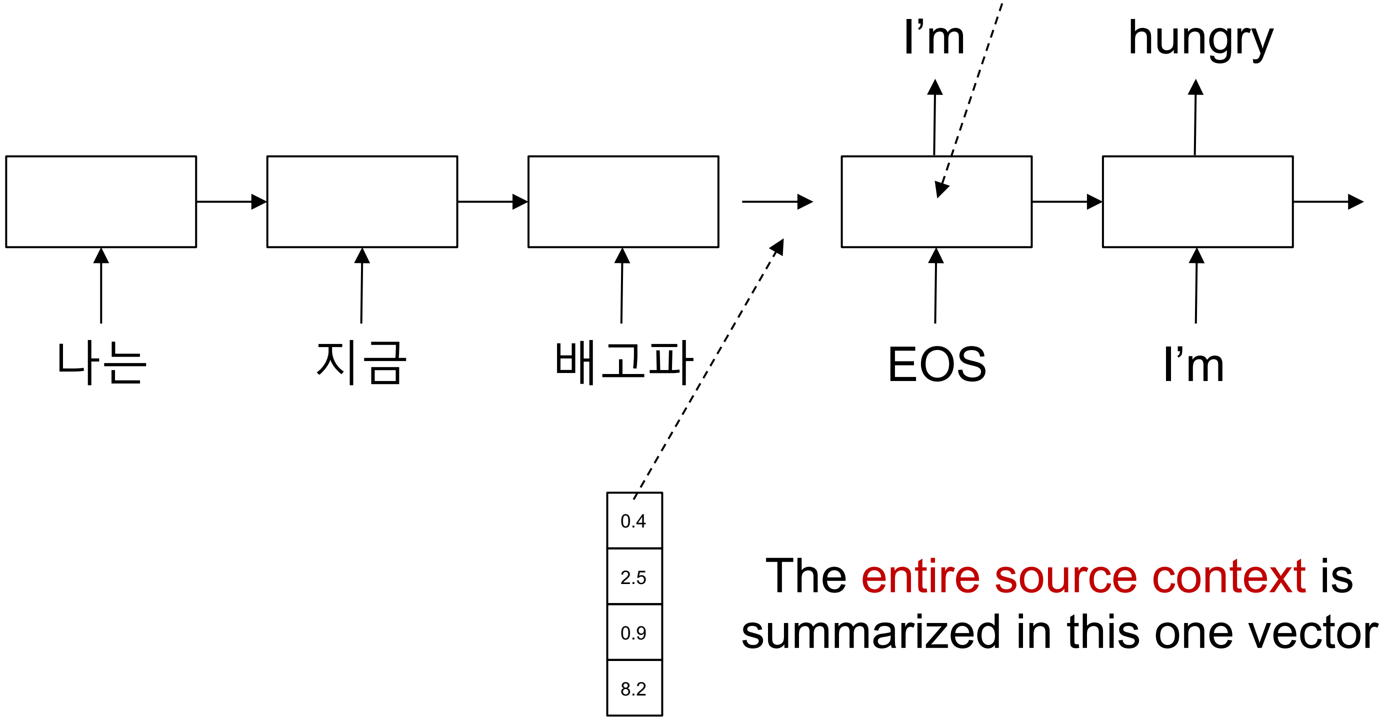
Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset

Machine Translation

Encoder-decoder

The decoder state depends just on the previous state and the previous output

$$s_i = f(s_{i-1}, y_{i-1})$$

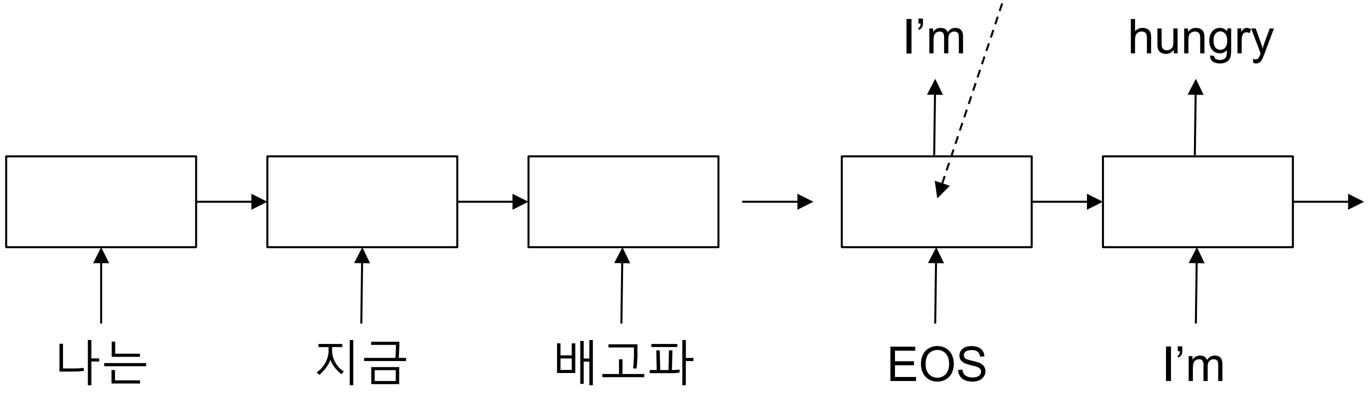


The entire source context is summarized in this one vector

Machine Translation with Attention

The decoder state depends just on the previous state, the previous output, and some context

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

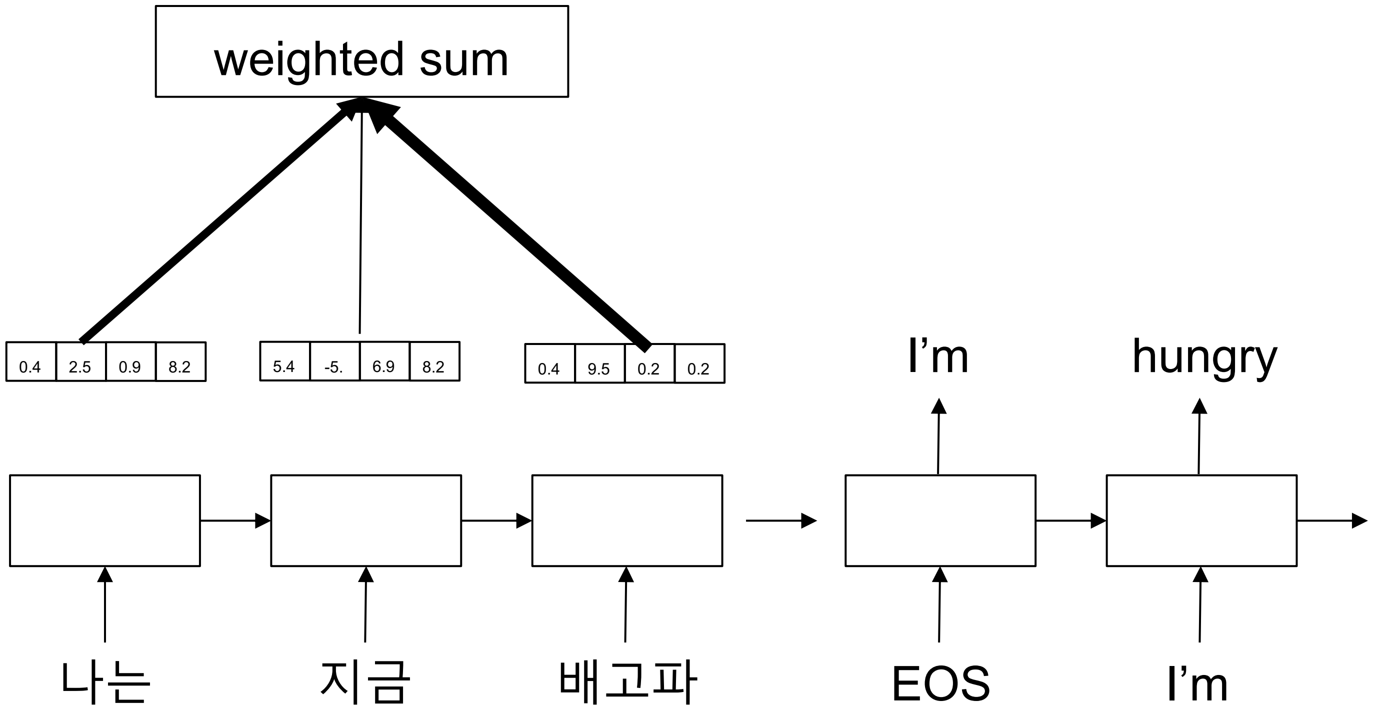


Machine Translation with Attention

The decoder state depends just on the previous state, the previous output, and some context

$$s_i = f\left(s_{i-1}, y_{i-1}, c_i\right)$$

$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$



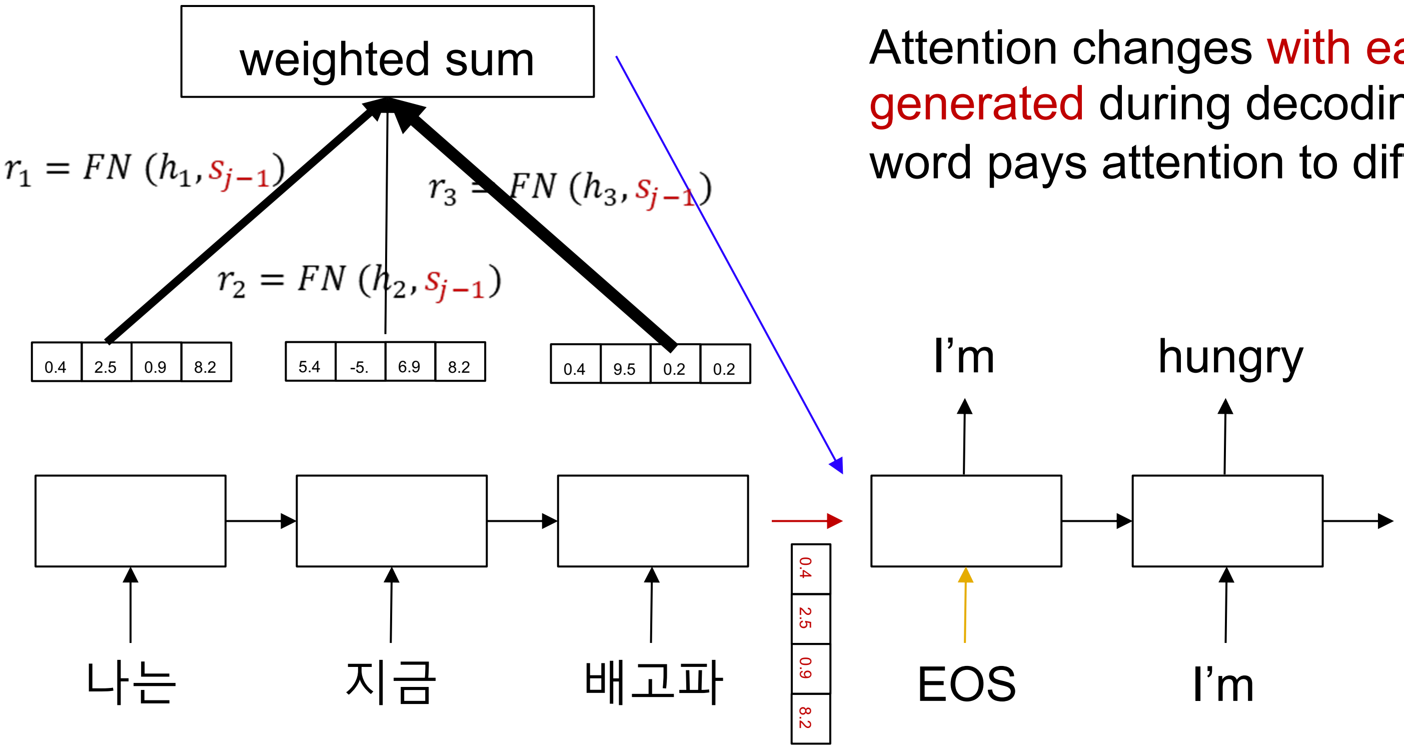
Machine Translation with Attention

$$s_i = f\,(s_{i-1}, y_{i-1}, c_i)$$

$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

$a = softmax(r)$

$r = [r_1, r_2, r_3]$



Attention changes with each word being generated during decoding. Each subsequent word pays attention to different parts of the input.

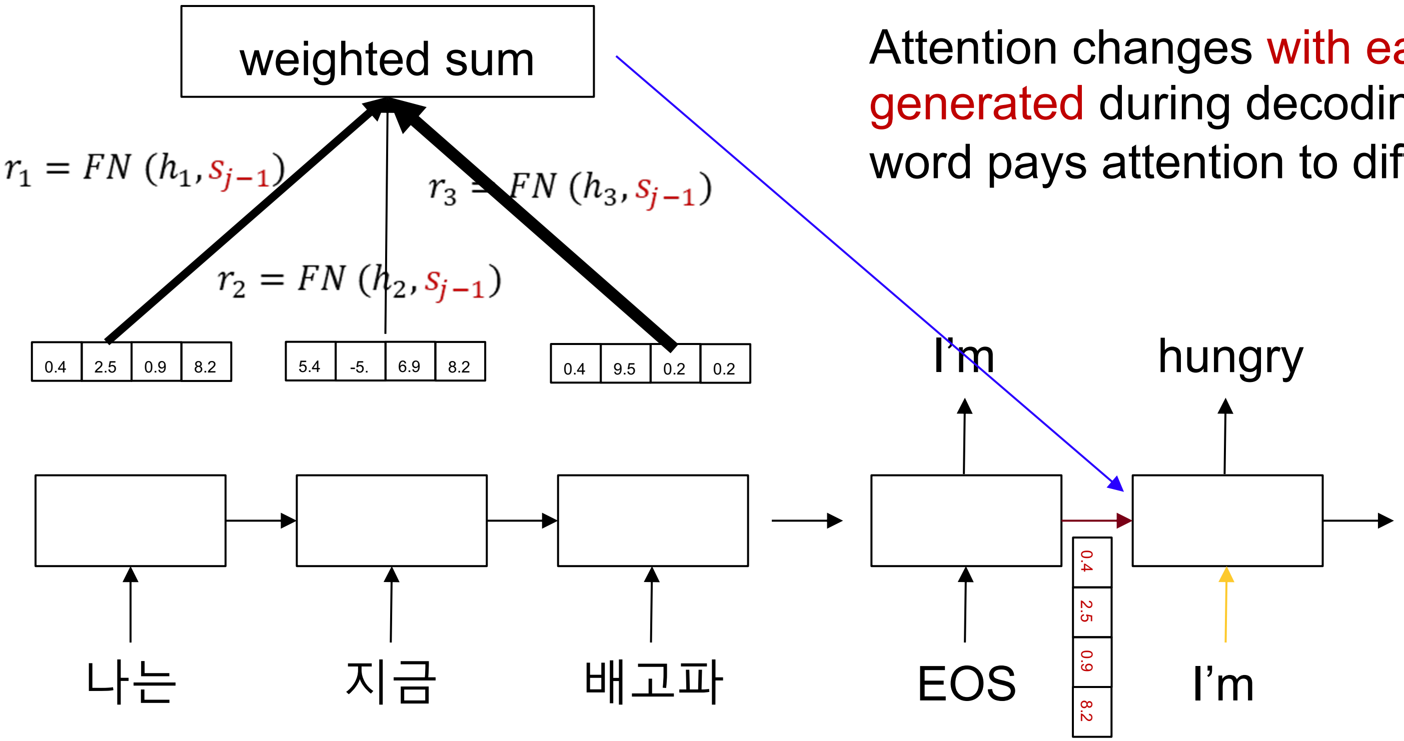
Machine Translation with Attention

$$s_i = f\,(s_{i-1}, y_{i-1}, c_i)$$

$$c = h_1 a_1 + h_2 a_2 + h_3 a_3$$

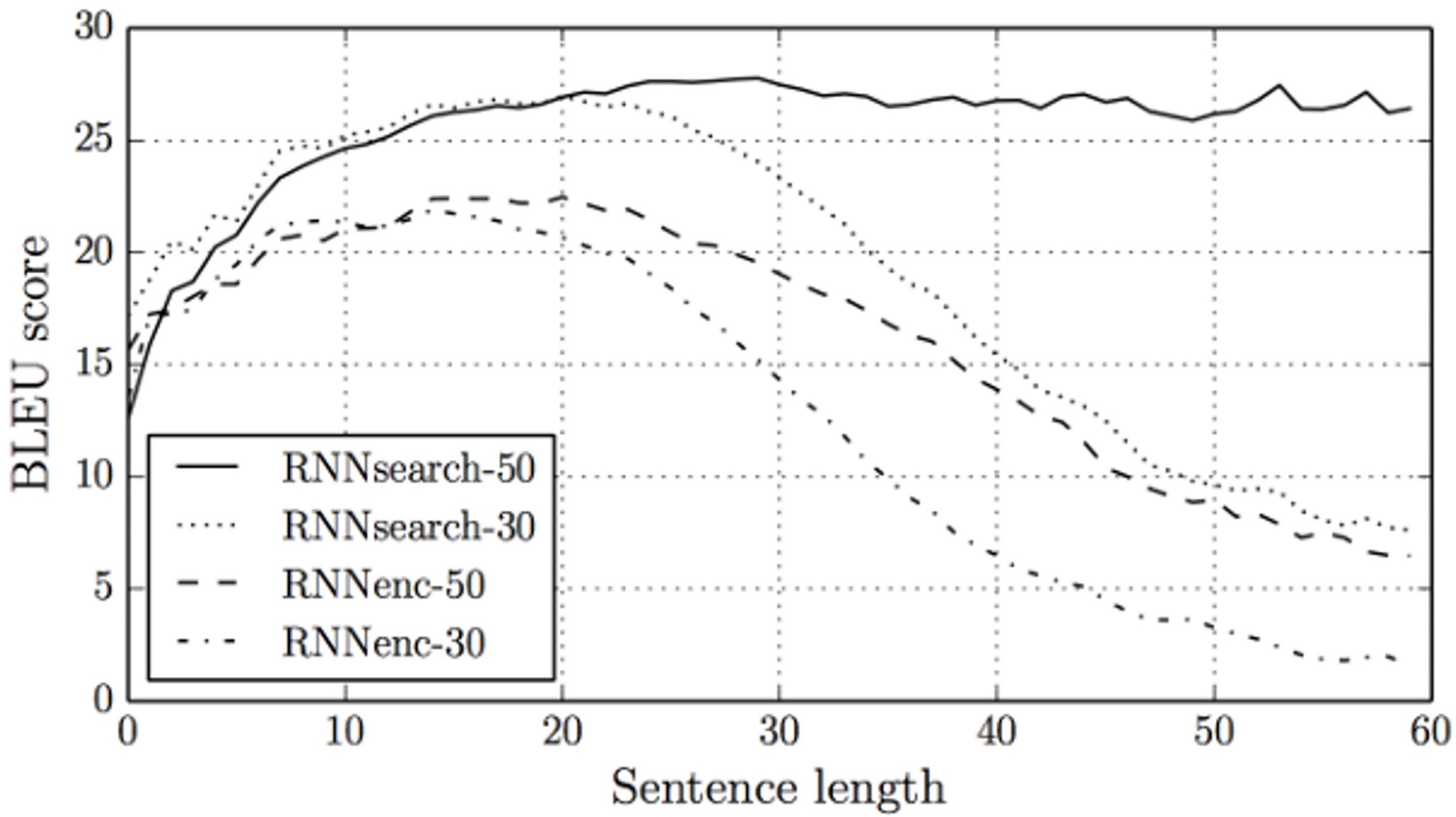
$$a = softmax(r)$$

$$r = [r_1, r_2, r_3]$$

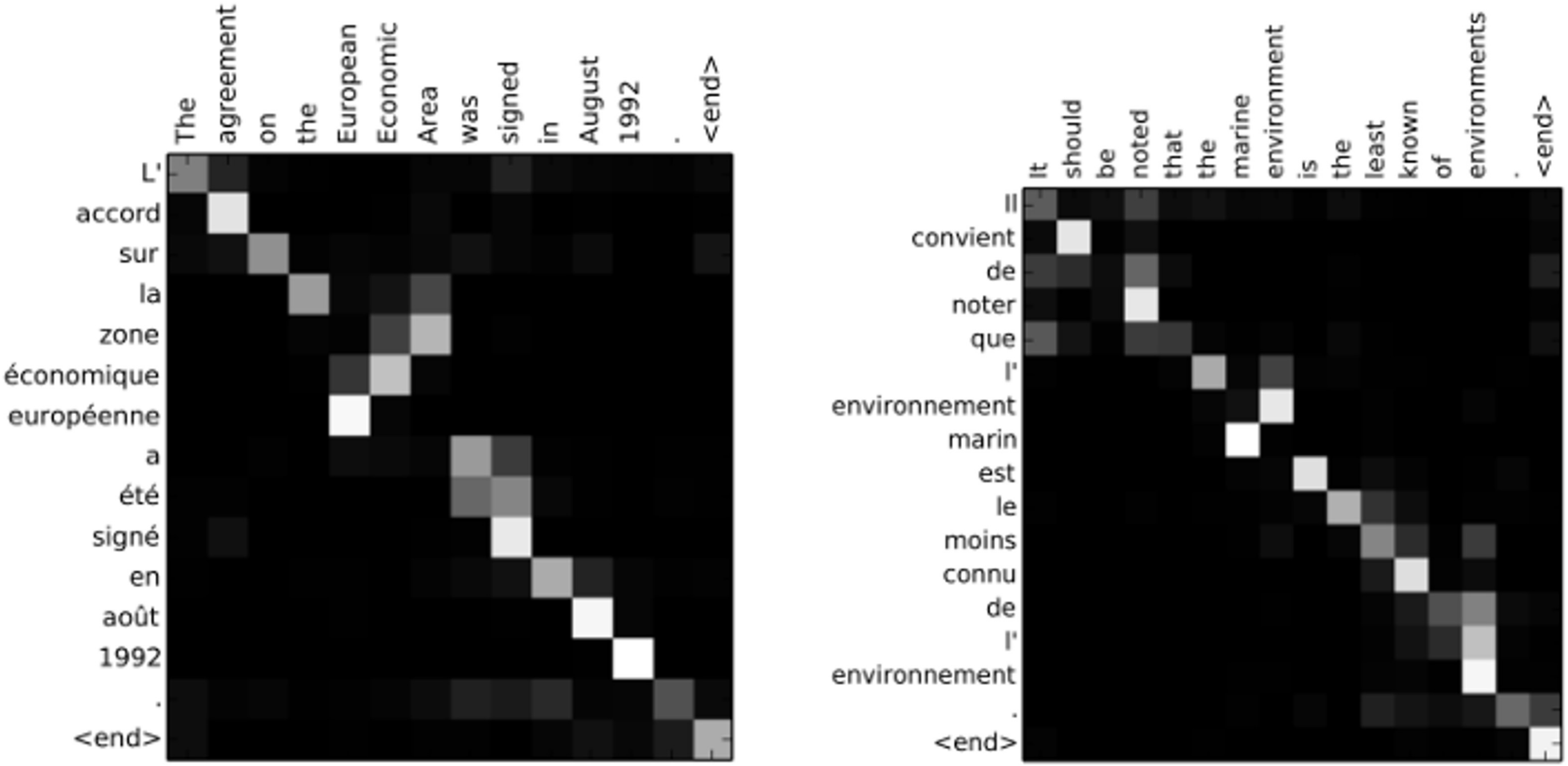


Attention changes with each word being generated during decoding. Each subsequent word pays attention to different parts of the input.

Better performance on long sentences

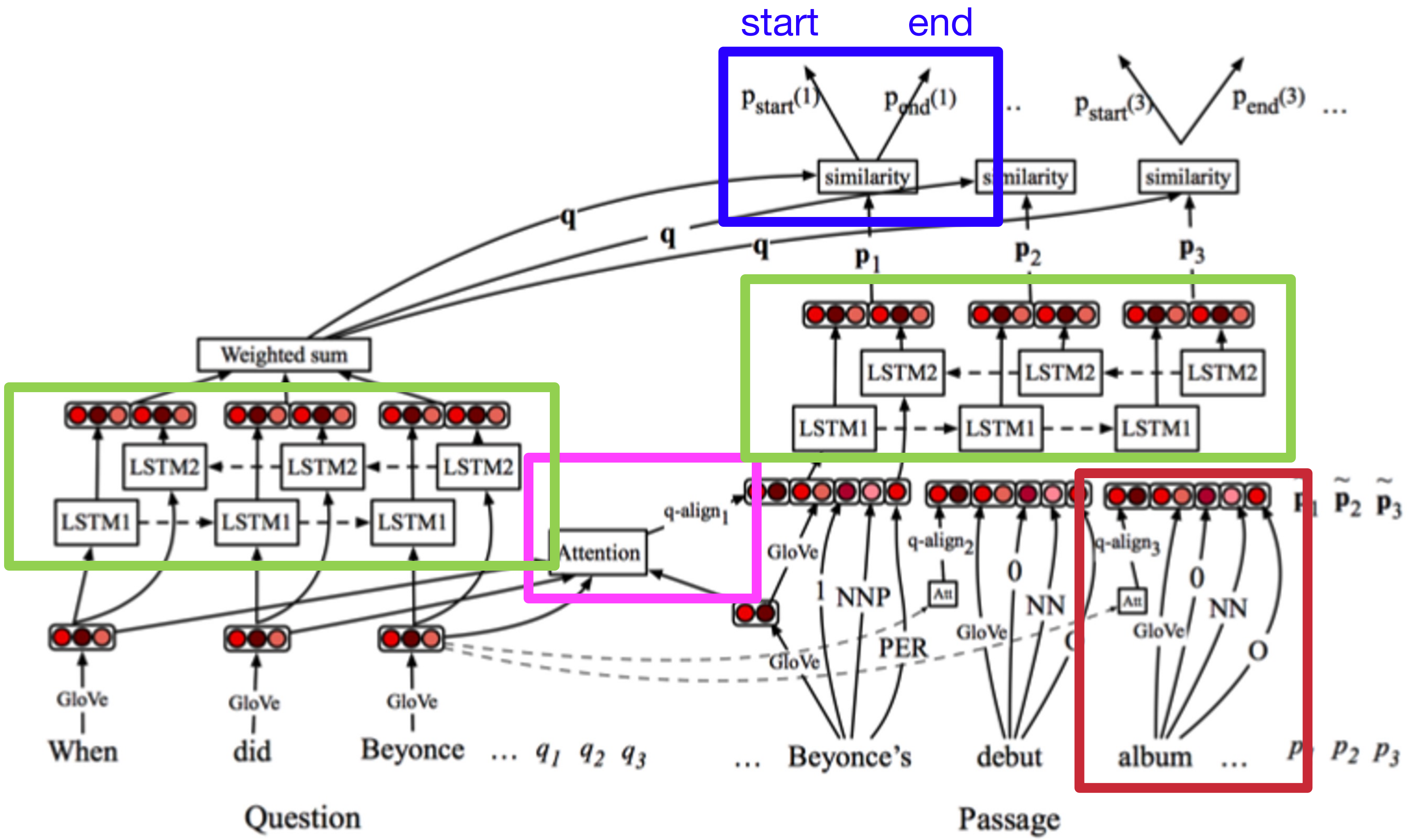


Bahdanau et al. (2016), “Neural Machine Translation by Jointly Learning to Align and Translate”

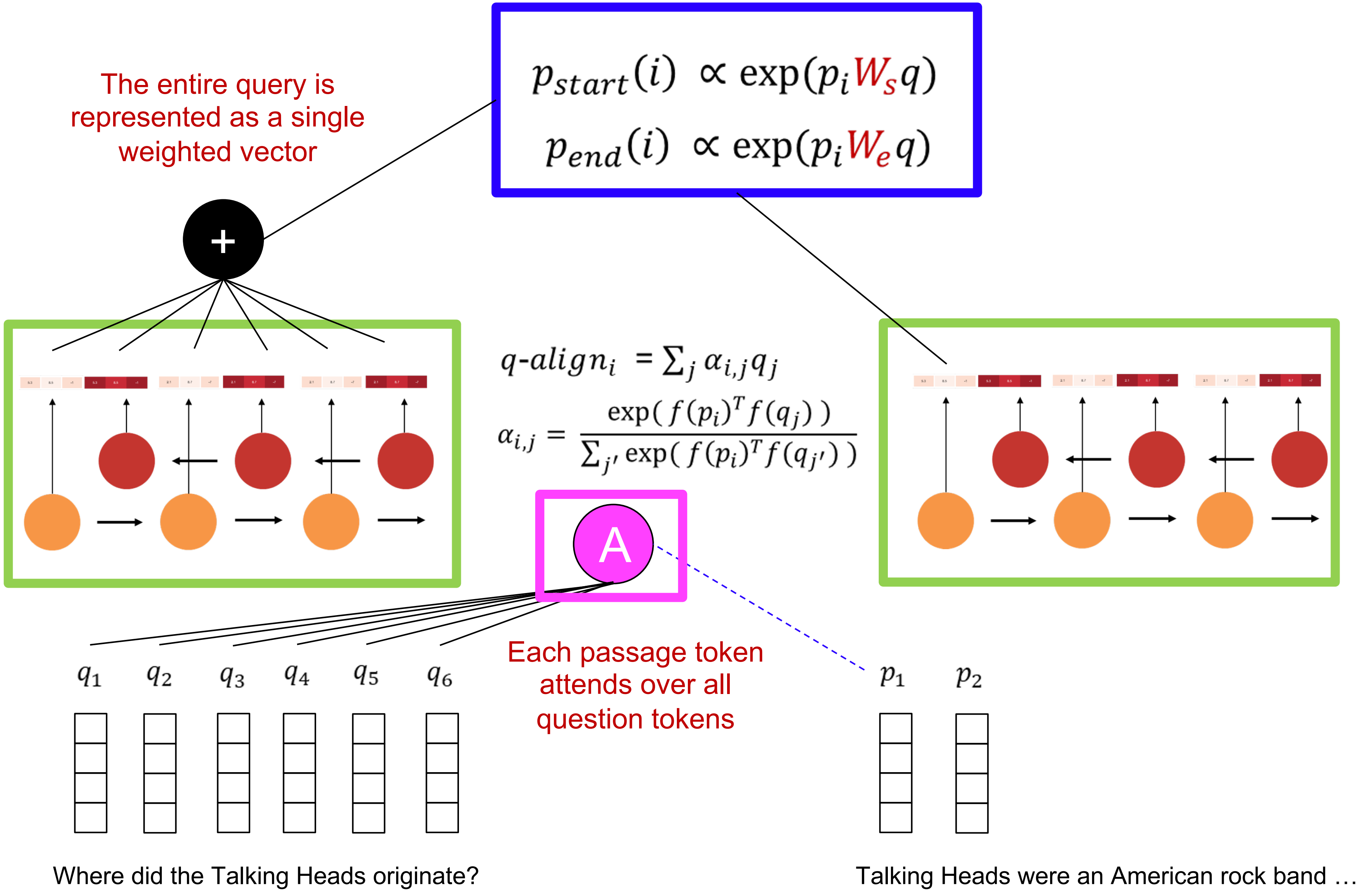


Bahdanau et al. (2016), "Neural Machine Translation by Jointly Learning to Align and Translate"

Neural QA model



Chen et al. 2017


The entire query is represented as a single weighted vector
$p_{start}(i) \propto \exp(p_i W_s q)$
$p_{end}(i) \propto \exp(p_i W_e q)$
$q\text{-}align_i = \sum_j \alpha_{i,j} q_j$
$\alpha_{i,j} = \frac{\exp(f(p_i)^T f(q_j))}{\sum_{j'} \exp(f(p_i)^T f(q_{j'}))}$
A
Each passage token attends over all question tokens
q_1 q_2 q_3 q_4 q_5 q_6
p_1 p_2
Where did the Talking Heads originate?
Talking Heads were an American rock band …

Evaluation methods on generated text

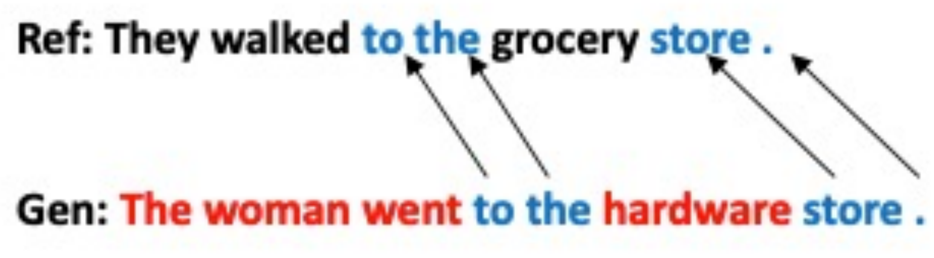
Types of evaluation methods in NLG

Content overlap metrics

Model-based metrics

Human evaluations

Content overlap metrics



- Compute a score that indicates the similarity between **generated** and **gold-standard** (human-written) text
- Fast, efficient and widely used
- Two broad categories:
  - **N-gram overlap metrics** (e.g., BLEU, ROUGE, METEOR)
  - **Semantic overlap metrics** (e.g., PYRAMID, SPICE)

N-gram overlap metrics

Word overlap–based metrics (BLEU, ROUGE, METEOR, CIDEr, etc.)

- ❑ They're not ideal for machine translation
- ❑ They get progressively much worse for tasks that are more open-ended than machine translation
  - ○ Worse for summarization, as longer output texts are harder to measure
  - ○ Much worse for dialogue, which is more open-ended than summarization
  - ○ Much, much worse for story generation, which is also open-ended, but whose sequence length can make it seem you're getting decent scores!

Bilingual Evaluation Understudy (BLEU)

- N-gram overlap between generated text and reference text
- Compute prevision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)
- Typically computed over the entire corpus, not single sentences

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

Bilingual Evaluation Understudy (BLEU)

BLEU (Papineni et al. 2002): what fraction of {1-4}-grams in the system translation appear in the reference translations?

$$P_n = \frac{\text{Number of ngrams in system and reference translations}}{\text{Number of ngrams in system translation}}$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{1-r/c} & if\ c \leq r \end{cases}$$

c = length of hypothesis translation
r = length of closest reference translation

$$BLEU = \text{BP} \quad \exp \frac{1}{N} \sum_{n=1}^{N} \log p_n$$

brevity penalty

Hypothesis/system translation

Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Appeared
calm
when
he
was
taken
to
the
American
plane
,
which
will
to
Miami
,
Florida
.

$$P_1 = \frac{15}{18} = 0.833$$

Reference translation

Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.

Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.

Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.

Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Ngrams appearing >1 time in the hypothesis can match up to the max number of times they appear in a single reference e.g., two commas in hypothesis but one max in any single reference.

Hypothesis/system translation

Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Appeared calm
calm when
when he
he was
was taken
taken to
to the
the American
American plane

plane ,
, which
which will
will to
to Miami
Miami ,
, Florida
Florida .

$$P_2 = \frac{10}{17} = 0.588$$

Reference translation

Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.

Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.

Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.

Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Recall Oriented Understudy for Gisting Evaluation (ROUGE)

- Overlap between generated text and reference text in terms of **recall**.
- Three types:
  - Rouge-N: the most prevalent form that detects n-gram overlap;
  - Rouge-L: identifies the Longest Common Subsequence
  - Rouge-S: concentrates on skip grams.

$$\frac{\text{number of n-grams found in model and reference}}{\text{number of n-grams in reference}}$$

The main difference between rouge and bleu is that bleu score is precision-focused whereas rouge score focuses on recall.

BLEU and ROUGE Examples

```python
from nltk.translate.bleu_score import sentence_bleu
reference = [['this', 'movie', 'was', 'awesome']]
candidate = ['this', 'movie', 'was', 'awesome', 'too']
score = sentence_bleu(reference, candidate)
print(score)
0.668740304976422
```

```python
from rouge import Rouge
reference = 'this movie was awesome'
candidate = 'this movie was awesome too'
rouge = Rouge()
scores = rouge.get_scores(candidate, reference)[0]
['rouge-2']
['f']
print(scores)
0.8571428522448981
```

https://arize.com/blog-course/generative-ai-metrics-bleu-score/

A simple failure case of BLEU

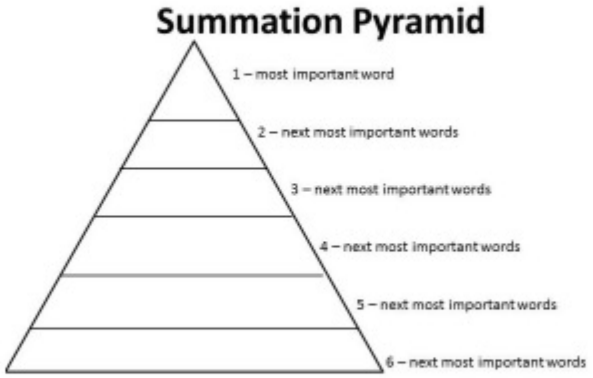
n-gram overlap metrics have no concept of semantic relatedness!

Are you enjoying your Homework #2 on ngram LM?

Heck Yes!

	BLEU = 0.61	Yes!
	BLEU = 0.25	You know it !
False Negative	BLEU = 0.0	Yup .
False Positive	BLEU = 0.67	Heck no !

Semantic overlap metrics



PYRAMID

(Nenkova et al., 2017)

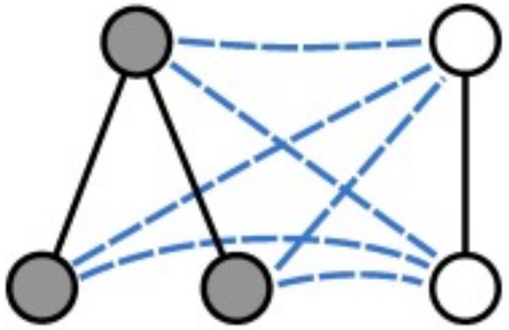
Incorporates **human content selection variation** in summarization evaluation.

Identifies **Summarization Content Units (SCU)**s to compare information content in summaries.

SPICE

(Anderson et al., 2016)

Semantic propositional image caption evaluation is an image captioning metric that initially parses the reference text to derive an abstract scene graph representation.



SPIDER

(Liu et al., 2017)

A combination of semantic graph similarity (**SPICE**) and n-gram similarity measure (**CIDER**), the SPICE metric yields a more complete quality evaluation metric.

Types of evaluation methods in NLG

Content overlap metrics

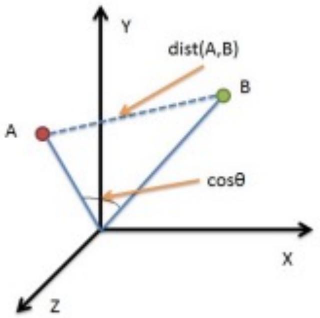
Model-based metrics

Human evaluations

Model-based metrics

- Use learned representations of words and sentences to compute semantic similarity between generated and reference texts
- No more n-gram bottleneck because text units are represented as embeddings
- Even though embeddings are pretrained, distance metrics used to measure the similarity can be fixed

Model-based metrics: Word distance functions



Vector Similarity

Embedding based similarity for semantic distance between text.

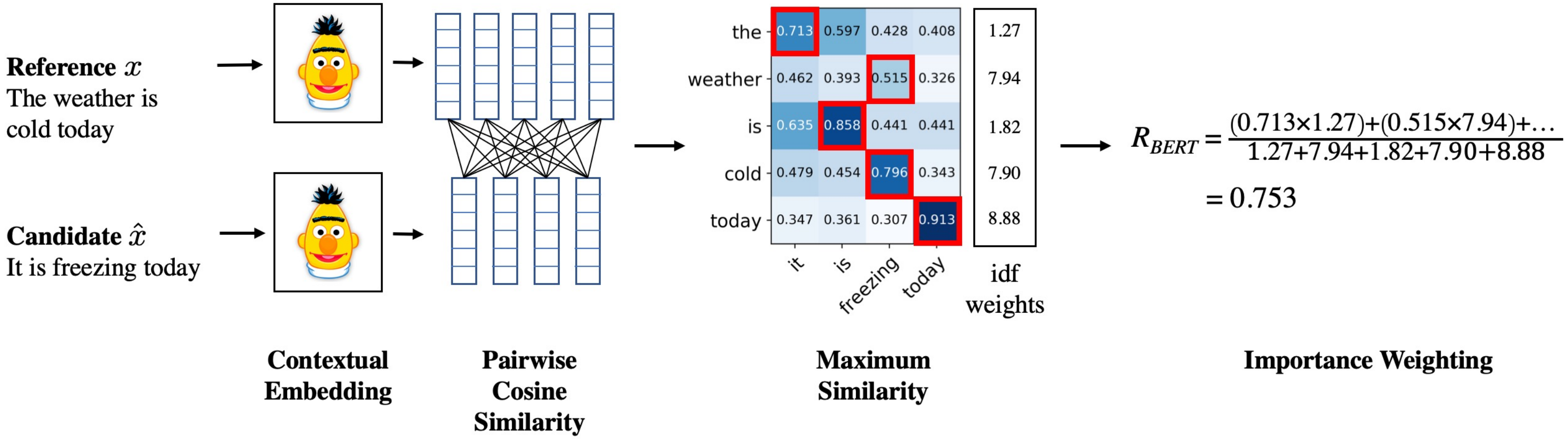
- ❑ Embedding Average (Liu et al., 2016)
- ❑ Vector Extrema (Liu et al., 2016)
- ❑ MEANT (Lo, 2017)
- ❑ YISI (Lo, 2019)

Word Mover's Distance

Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching. (Kusner et.al., 2015; Zhao et al., 2019)

BERTScore

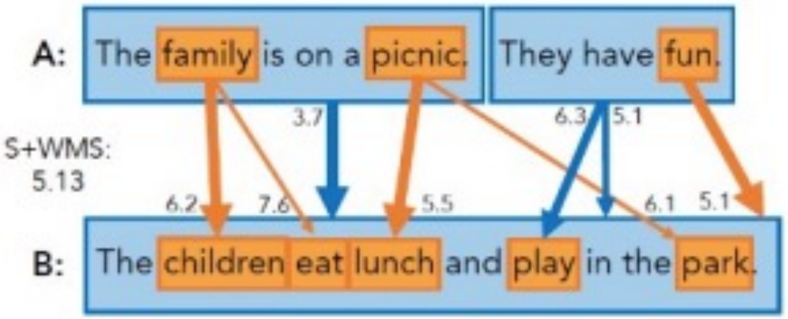
Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. (Zhang et.al. 2020)



Model-based metrics: Beyond word matching

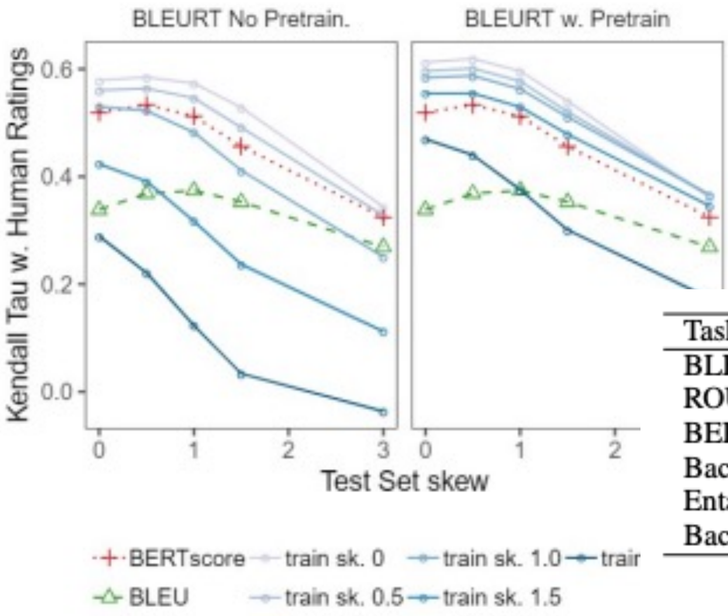
Sentence Movers Similarity

Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations. (Clark et.al., 2019)



BLEURT

A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text. (Sellam et.al. 2020)



Task Type	Pre-training Signals	Loss Type
BLEU	τ_{BLEU}	Regression
ROUGE	$\tau_{\text{ROUGE}} = (\tau_{\text{ROUGE-P}}, \tau_{\text{ROUGE-R}}, \tau_{\text{ROUGE-F}})$	Regression
BERTscore	$\tau_{\text{BERTscore}} = (\tau_{\text{BERTscore-P}}, \tau_{\text{BERTscore-R}}, \tau_{\text{BERTscore-F}})$	Regression
Backtrans. likelihood	$\tau_{\text{en-fr},\boldsymbol{z}\mid\tilde{\boldsymbol{z}}}, \tau_{\text{en-fr},\tilde{\boldsymbol{z}}\mid\boldsymbol{z}}, \tau_{\text{en-de},\boldsymbol{z}\mid\tilde{\boldsymbol{z}}}, \tau_{\text{en-de},\tilde{\boldsymbol{z}}\mid\boldsymbol{z}}$	Regression
Entailment	$\tau_{\text{entail}} = (\tau_{\text{Entail}}, \tau_{\text{Contradict}}, \tau_{\text{Neutral}})$	Multiclass
Backtrans. flag	$\tau_{\text{backtran-flag}}$	Multiclass

Table 1: Our pre-training signals.

```python
import torch
from bert_score import score


# reference and generated texts
ref_text = "The quick brown fox jumps over the lazy dog."
gen_text = "A fast brown fox leaps over a lazy hound."

# compute Bert score
P, R, F1 = score([gen_text], [ref_text], lang="en", model_type="bert-base-uncased")

# print results
print(f"Bert score: P={P.item():.4f} R={R.item():.4f} F1={F1.item():.4f}")
```

Automatic metrics in general don't really work

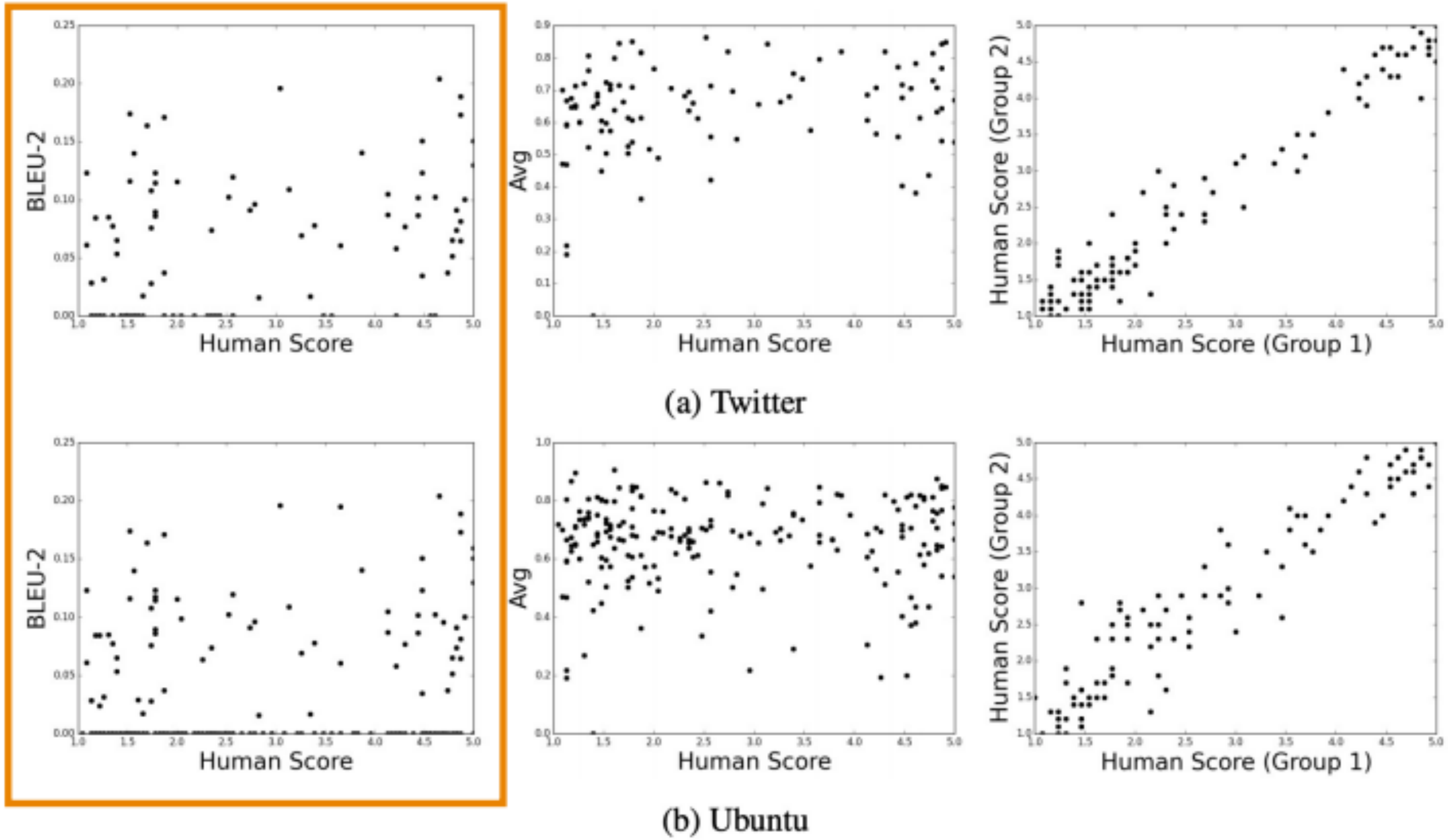


Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

(Liu et.al., 2016)

What if there is no reference text?

Types of evaluation methods in NLG

Content overlap metrics

Model-based metrics

Human evaluations

Human Evaluations

- Automatic metrics fall short of matching human decisions
- Human evaluation is most important form of evaluation for text generation systems
  - >75% generation papers at ACL 2019 included human evaluations
- Gold standard in developing new automatic metrics
  - New automated metrics must correlate well with human evaluations!

Human Evaluations

- Ask humans to evaluate the quality of generated text

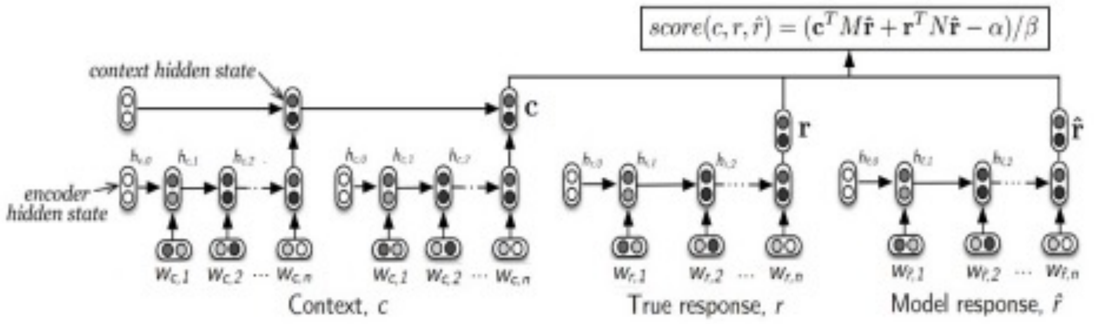
- Overall or along some specific dimension:
  - fluency
  - coherence / consistency
  - factuality and correctness
  - commonsense
  - style / formality
  - grammaticality
  - typicality
  - redundancy

Note: Don't compare human evaluation scores across differently conducted studies Even if they claim to evaluate the same dimensions!

Human evaluation: Issues

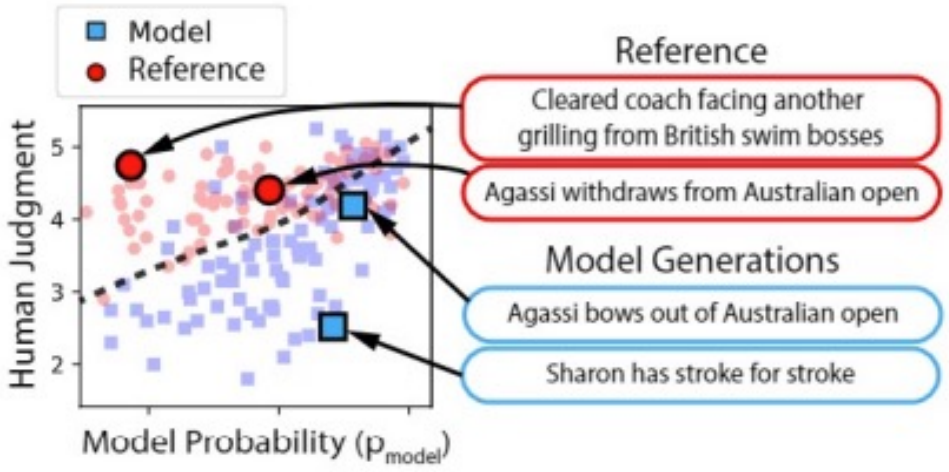
- Human judgments are regarded as the gold standard
- Of course, we know that human eval is slow and expensive
- Conducting human evaluation effectively is very difficult
  - Humans are *are inconsistent*
    can be illogical
    lose concentration
    misinterpret your question
    can't always explain why they feel the way they do

Learning from human reference



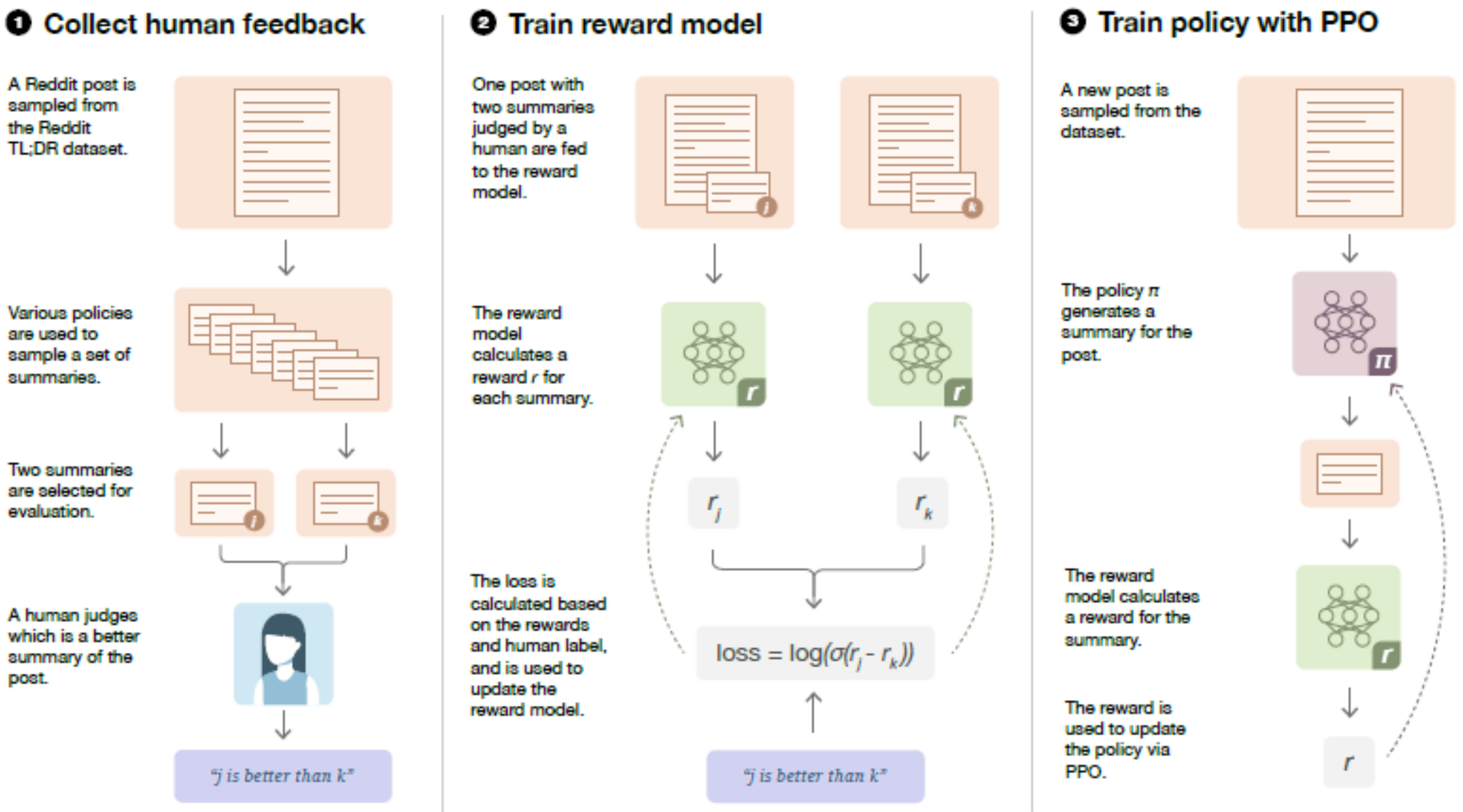
ADEM

A learned metric from human judgments for dialog system evaluation in a chatbot setting. (Lowe et.al., 2017)

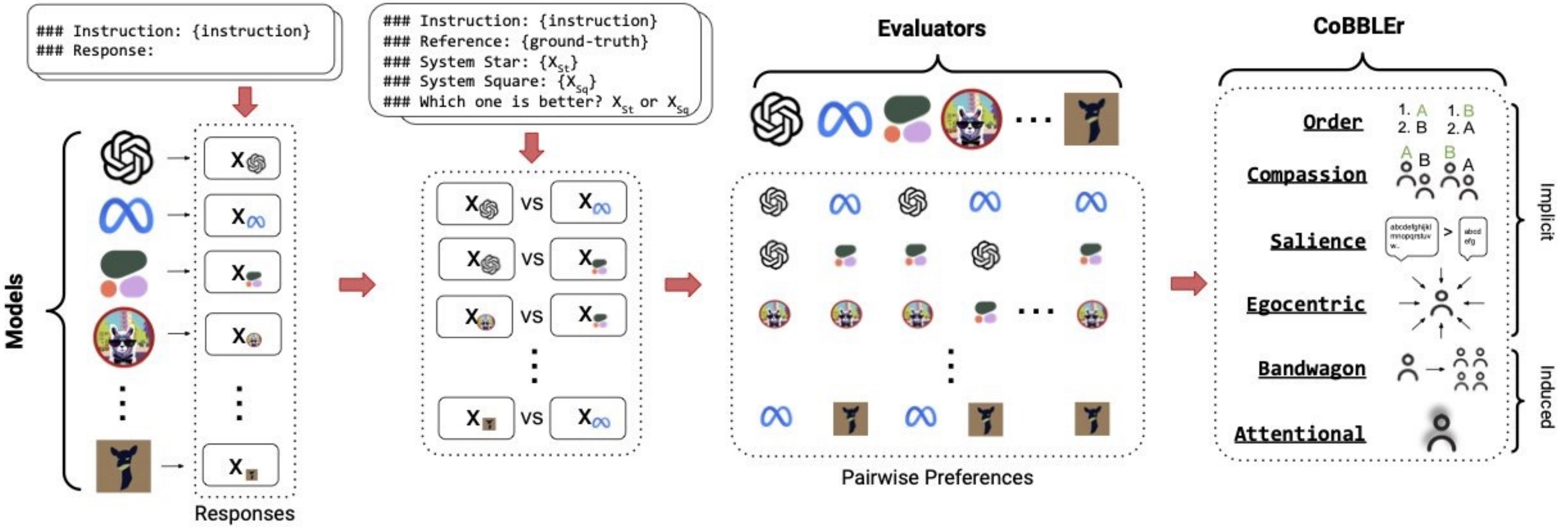


HUSE

Human Unified with Statistical Evaluation (HUSE), determines the similarity of the output distribution and a human reference distribution. (Hashimoto et.al. 2019)



[2009.01325] Learning to summarize from human feedback



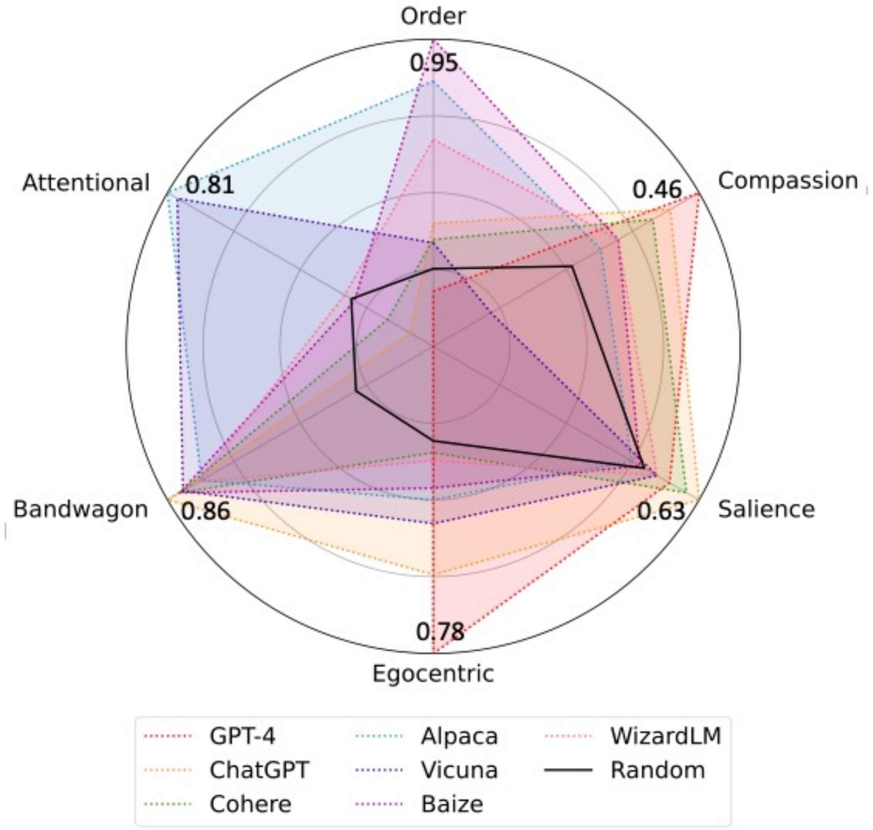
[Benchmarking Cognitive Biases in Large Language Models as Evaluators arxiv.org/abs/2309.17012](https://arxiv.org/abs/2309.17012)

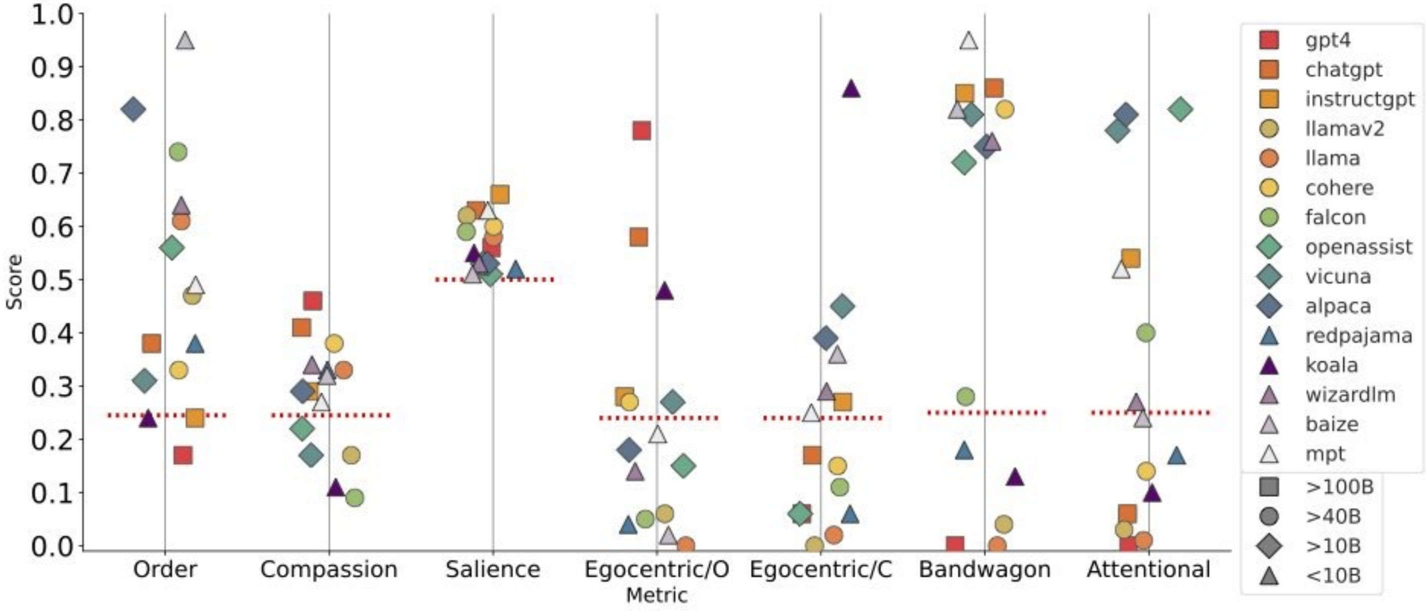
Bias	Bias Behavior	Example
ORDER BIAS	The tendency to give preference to an option based on their order (e.g. first, second, or last).	**`System Star`**: x **`System Square`**: y `System Square`: y `System Star`: x
COMPASSION FADE	The tendency to observe different behaviors when given recognizable names as opposed to anonymized aliases.	`Model Alpaca`: x `Model Vicuna`: y **`Model Vicuna`**: y **`Model Alpaca`**: x
EGOCENTRIC BIAS	The inclination to prioritize one's own responses regardless of response quality.	**`Model Star (You)`**: x `Model Square`: y
SALIENCE BIAS	The tendency to prefer responses based on the length of the response (i.e., more often preferring longer responses over shorter ones).	**`System Star`**: The quick brown fox jumps over the lazy dog. `System Square`: The fox jumped.
BANDWAGON EFFECT	The tendency to prefer majority belief without critical evaluation.	**85%** believe that `System Star` is better.
ATTENTIONAL BIAS	The inclination to give more attention to irrelevant or unimportant details.	`System Square` **likes to eat oranges and apples**

Table 1: We display the characteristic format for each bias and bold answers that indicate behavior influenced by the bias. For example, in COMPASSION FADE (recognizable names) `Model Alpaca` and `Model Vicuna` are associated with `System Star` and `System Square` respectively, in which the preferred response (bolded) is inconsistent with the preferred response from ORDER (anonymized names).

Benchmarking Cognitive Biases in Large Language Models as Evaluators arxiv.org/abs/2309.17012

Benchmarking Cognitive Biases in Large Language Models as Evaluators





[Benchmarking Cognitive Biases in Large Language Models as Evaluators arxiv.org/abs/2309.17012](https://arxiv.org/abs/2309.17012)

Evaluation: Takeaways

- Content overlap metrics provide a good starting point for evaluating the quality of generated text. You will need to use one but they're not good enough on their own.
- Model-based metrics can be more correlated with human judgment, but behavior is not interpretable
- Human judgments are critical
  - Only thing that can directly evaluate factuality, but humans are inconsistent!
- In many cases, the best judge of output quality is YOU!
  - Look at your model generations. Don't just rely on numbers!
  - Don't cherry pick! Publicly release large samples of the output of systems that you create!

Conclusion

- Interacting with natural language generation systems quickly shows their limitations
- Even in tasks with more progress, there are still many improvements ahead
- Evaluation remains a huge challenge.
  - We need better ways of automatically evaluating performance of NLG systems
- One of the most exciting and fun areas of NLP to work in!

Questions?