

DSC291: Advanced Statistical Natural Language Processing

Parsing
Text Generation

Zhiting Hu

Lecture 12, May 5, 2022

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

Logistics

- Paper presentation sign-up (see Piazza)

Outline

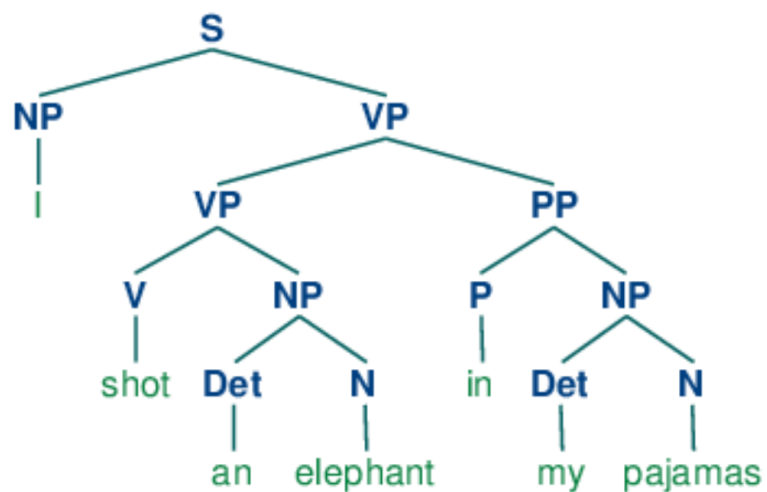
- Parsing
- Text Generation

Parsing

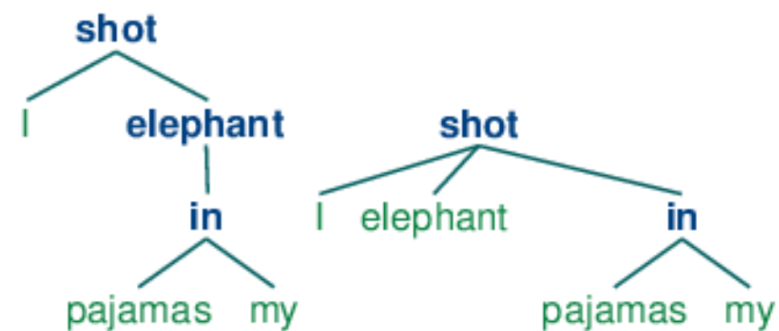
[Slides adapted from UW CSE 447 by Noah Smith; UCB Info 159/259 by David Bamman]

Formalisms

Phrase structure grammar
(Chomsky 1957)



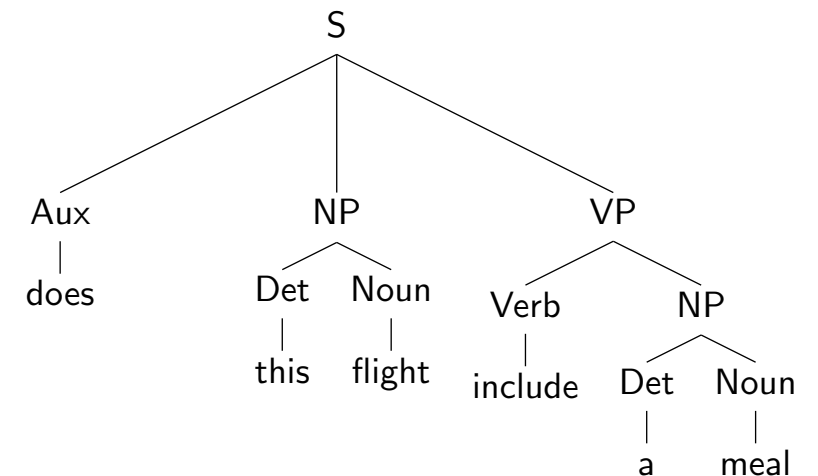
Dependency grammar
(Mel'čuk 1988; Tesnière 1959; Pāṇini)



Recap: Phrase Structure Grammar

- Constituents: groups of words behave as single units
- Context-Free Grammar (CFG)
 - A CFG gives a formal way to define a valid structure in a language

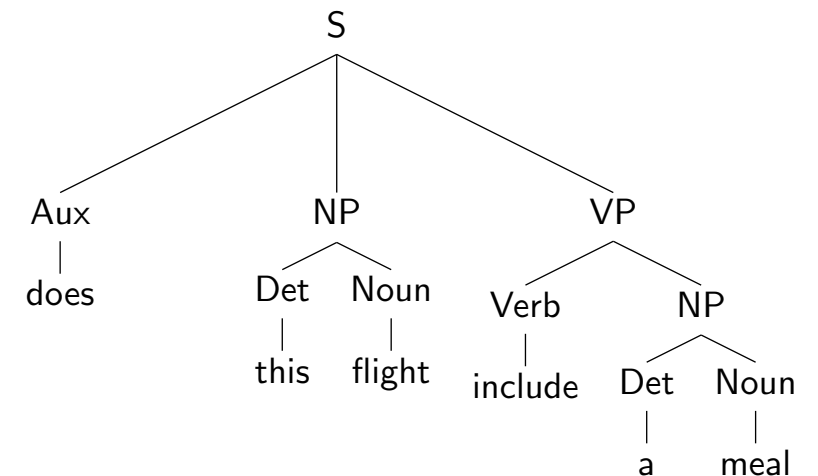
N	Finite set of non-terminal symbols	NP, VP, S
Σ	Finite alphabet of terminal symbols	the, dog, a
R	Set of production rules, each $A \rightarrow \beta$ $\beta \in (\Sigma, N)$	$S \rightarrow NP VP$ $Noun \rightarrow dog$
S	Start symbol	



Recap: Phrase Structure Grammar

- Constituents: groups of words behave as single units
- Context-Free Grammar (CFG)
 - A CFG gives a formal way to define a valid structure in a language
- Probabilistic Context-Free Grammar (PCFG)
 - Each production is also associated with a probability

N	Finite set of non-terminal symbols	NP, VP, S
Σ	Finite alphabet of terminal symbols	the, dog, a
R	Set of production rules, each $A \rightarrow \beta [p]$ $p = P(\beta A)$	$S \rightarrow NP VP$ Noun \rightarrow dog
S	Start symbol	



Recap: Phrase Structure Grammar

- Constituents: groups of words behave as single units
- Context-Free Grammar (CFG)
 - A CFG gives a formal way to define a valid structure in a language
- Probabilistic Context-Free Grammar (PCFG)
 - Each production is also associated with a probability
- Parsing:
 - Show one or more derivations for a sentence, using the grammar

Recap: PCFG Scores Trees

We can write the parsing problem as finding the best-scoring tree:

$$\hat{t} = \operatorname{argmax}_{t \in \mathcal{T}_x} \operatorname{Score}(t)$$

PCFGs view each tree t as a “bag of rules” (from \mathcal{R}), and define:

$$\begin{aligned} \operatorname{Score}(t) &= p(t) \\ &= \prod_{(N \rightarrow \alpha) \in \mathcal{R}} p(\alpha \mid N)^{\operatorname{count}(N \rightarrow \alpha; t)} \end{aligned}$$

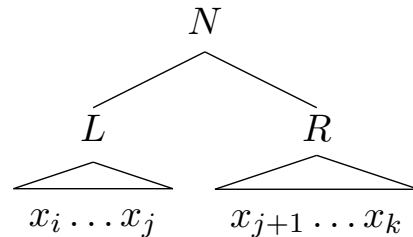
Recap: Probabilistic CKY

Base case: for $i \in \{1, \dots, n\}$ and for each $N \in \mathcal{N}$:

$$\heartsuit_{i:i}(N) = \log p(x_i \mid N)$$

For each i, k such that $1 \leq i < k \leq n$ and each $N \in \mathcal{N}$:

$$\heartsuit_{i:k}(N) = \max_{L, R \in \mathcal{N}, j \in \{i, \dots, k-1\}} \log p(L R \mid N) + \heartsuit_{i:j}(L) + \heartsuit_{(j+1):k}(R)$$

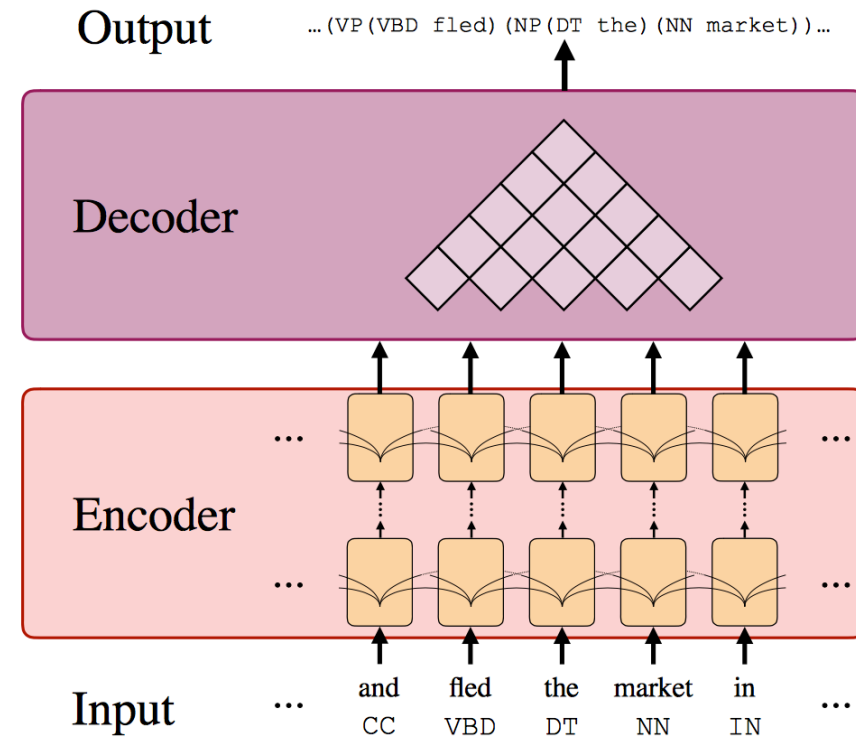


Solution:

$$\heartsuit_{1:n}(S) = \max_{t \in \mathcal{T}_x} \log p(t)$$

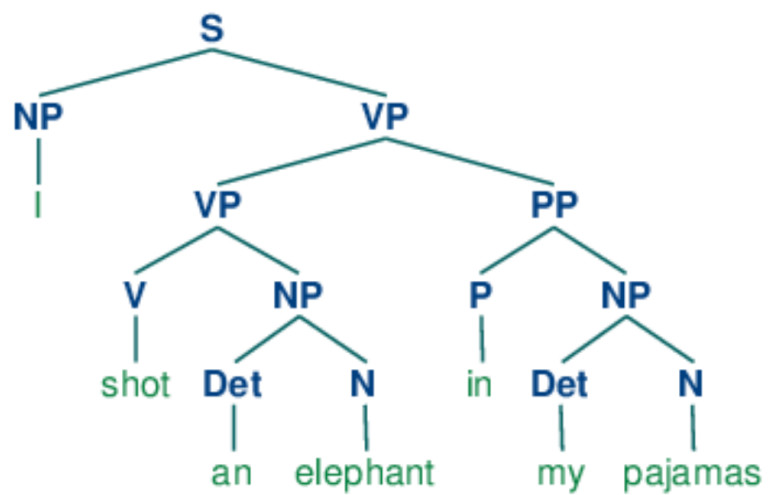
Neural Parsing

- Kitaev and Klein (2018), “Constituency Parsing with a Self-Attentive Encoder”
- Neural model (attention encoder) generates representations of each token in a sentence
- Learned scoring $s(i,j,k)$ function for each span from token i to token j with label k
- CKY for **decoding** to find the best tree through this space.

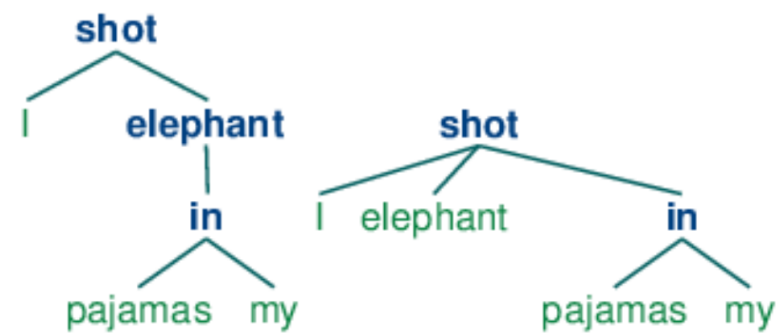


Formalisms

Phrase structure grammar
(Chomsky 1957)

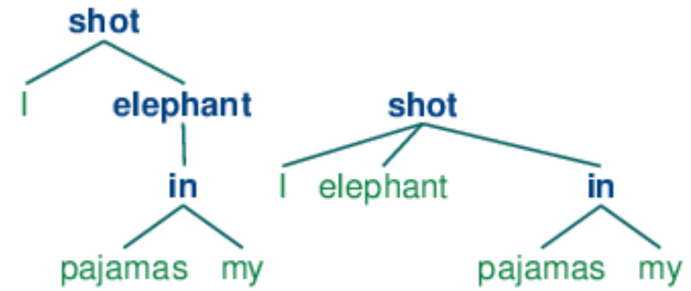
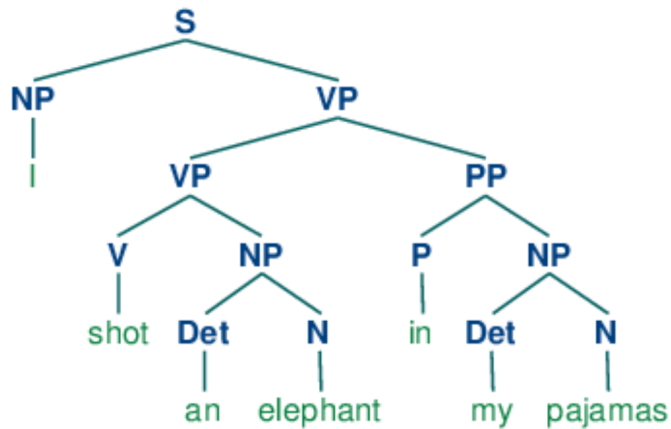


Dependency grammar
(Mel'čuk 1988; Tesnière 1959; Pāṇini)



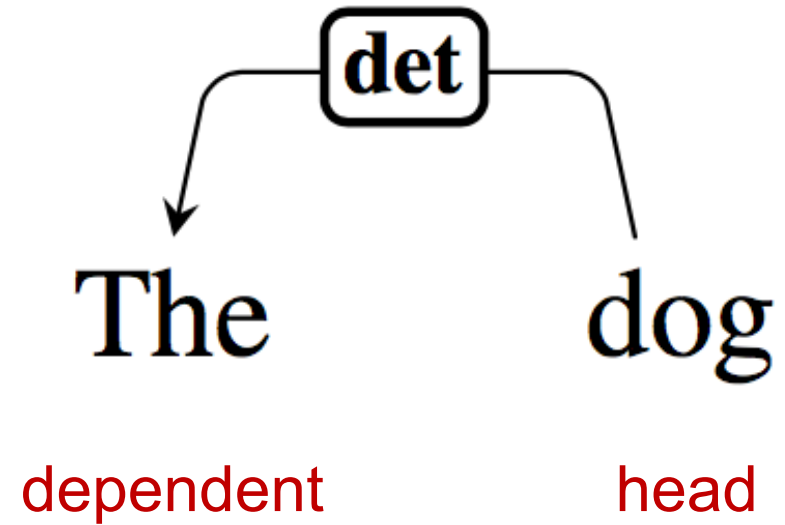
Dependency syntax

A different family of theories of syntax focuses on dependencies between words

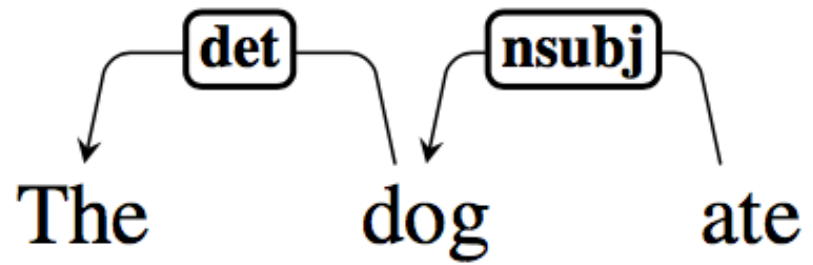


- Dependency syntax doesn't have non-terminal structure like a CFG; words are directly linked to each other.

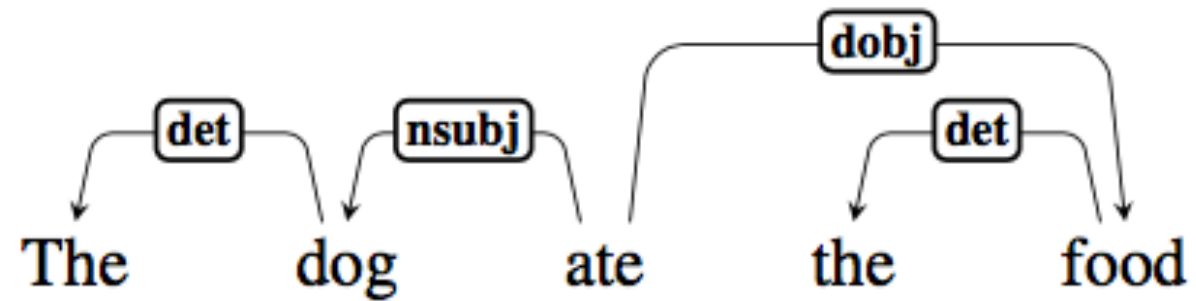
Dependency syntax



Dependency syntax

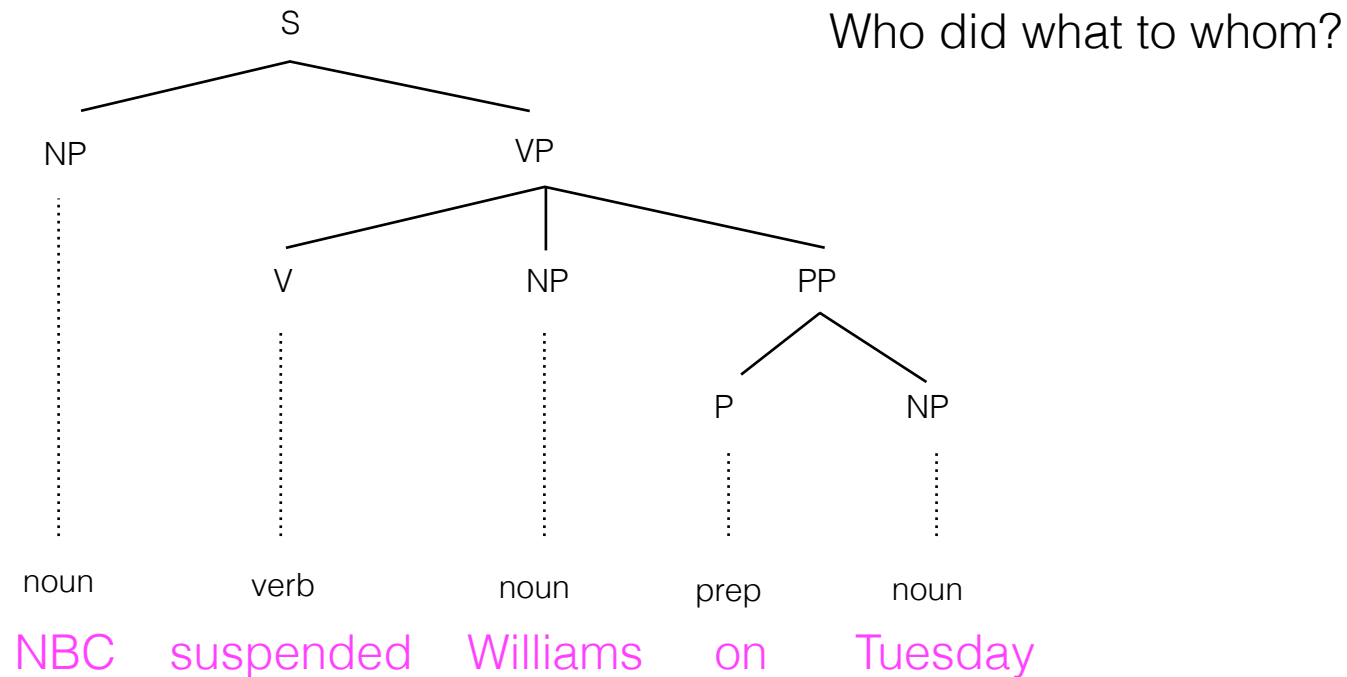


Dependency syntax



Dependencies vs constituents

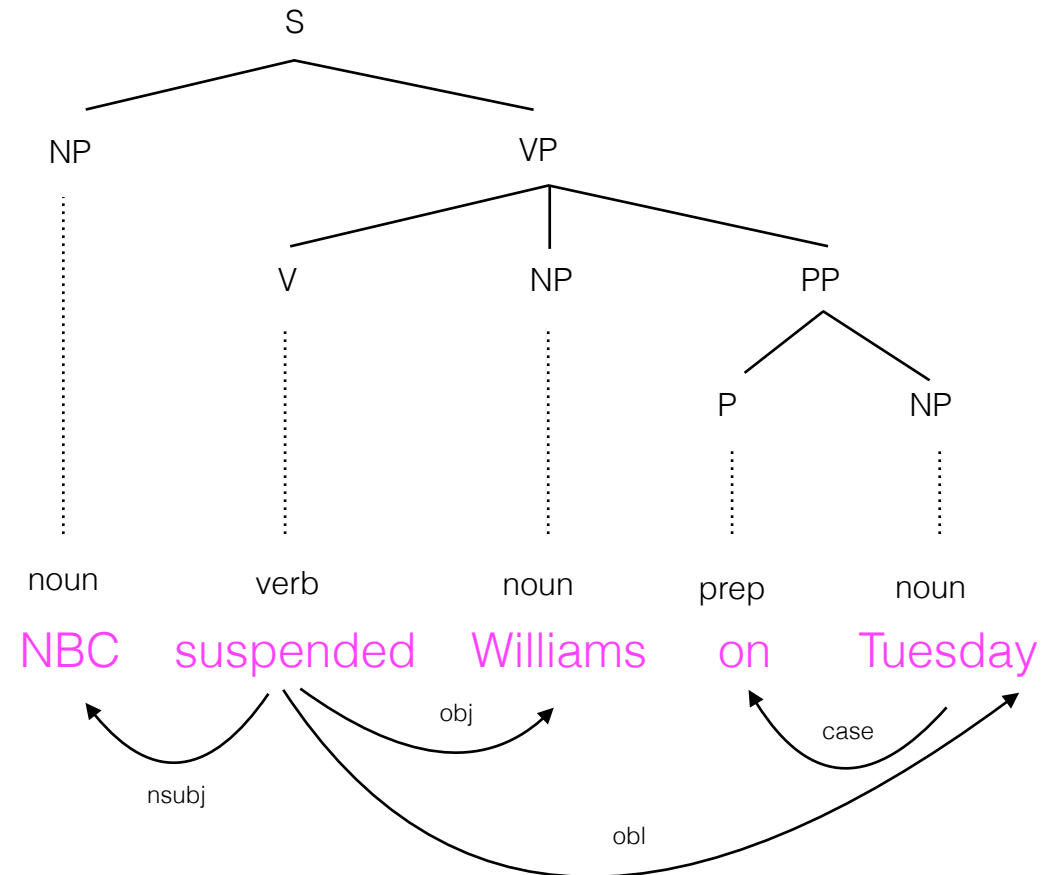
- Dependency links are closer to semantic relationships; no need to infer the relationships from the structure of a tree



subject: S → NP VP
direct object: S → NP (VP → ... NP ...)

Dependencies vs constituents

- Dependency links are closer to semantic relationships; no need to infer the relationships from the structure of a tree

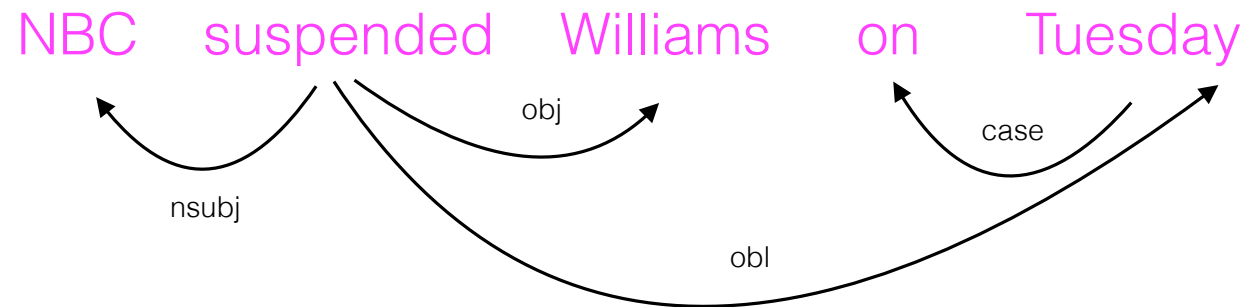


Dependencies vs constituents

- Dependency links are closer to semantic relationships; no need to infer the relationships from the structure of a tree

Captures binary relations between words

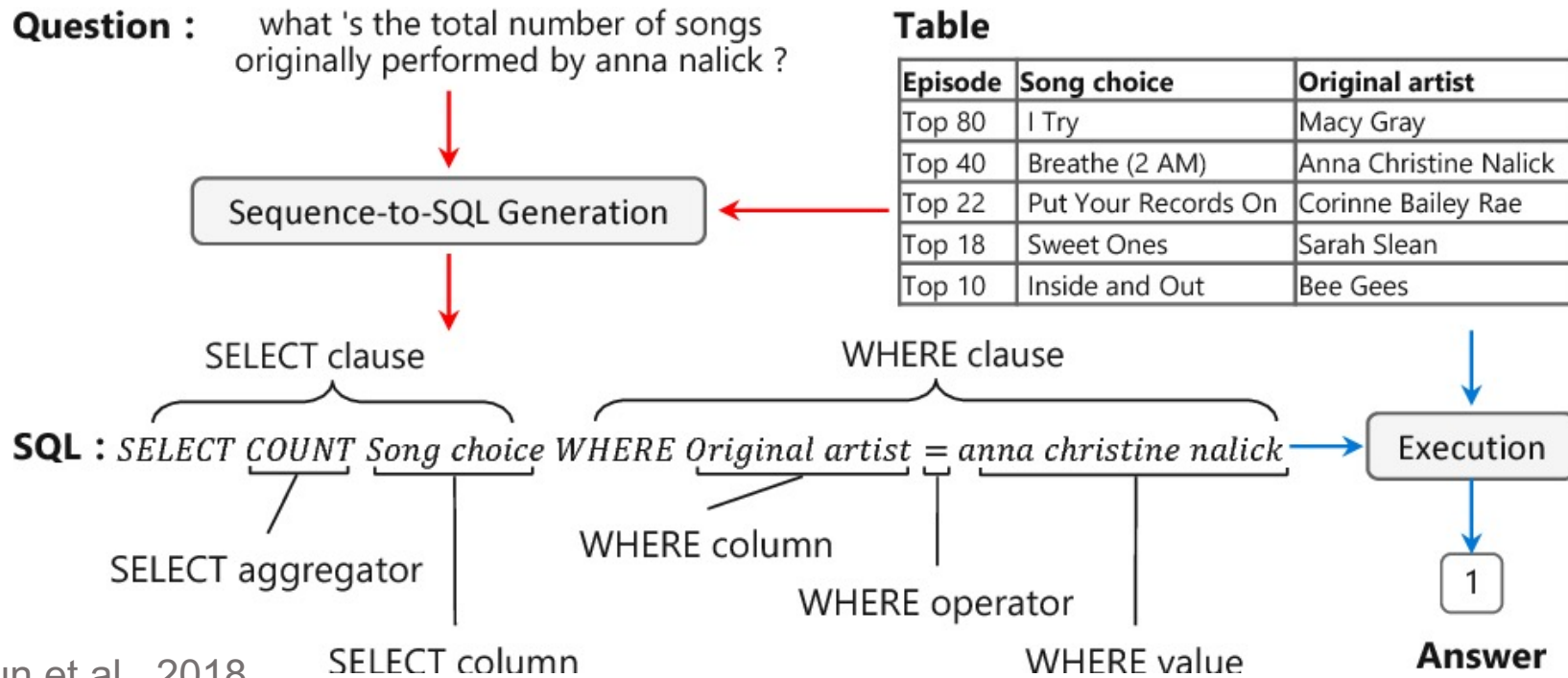
- nsubj(NBC, suspended)
- obj(Williams, suspended)



Semantic Parsing

Semantic parsing comprises a wide range of tasks where strings are mapped into meaning representation languages. Examples:

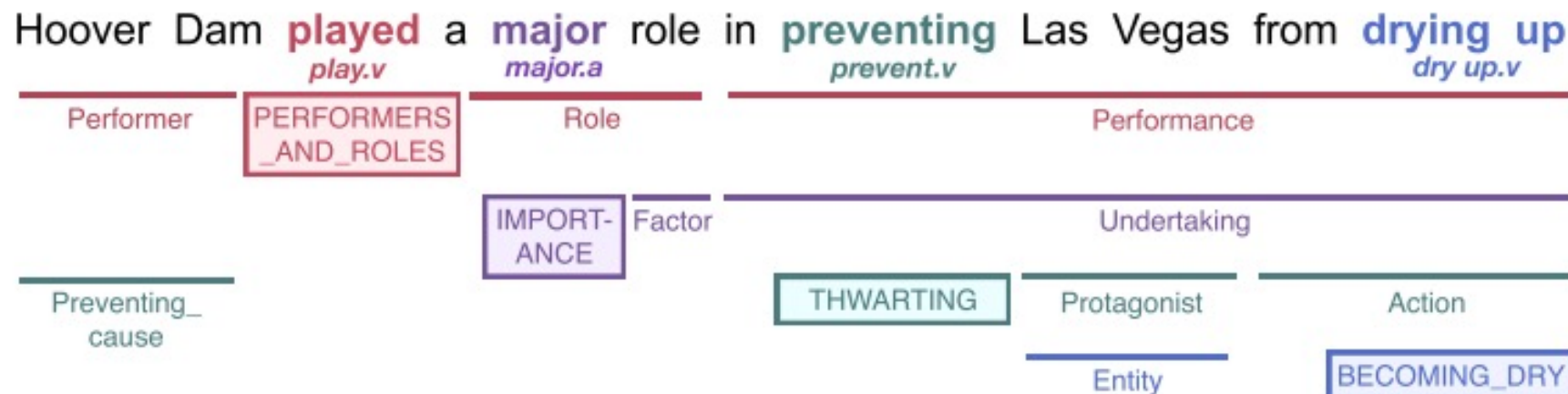
- ▶ Programming languages, especially query languages that can be used to answer questions using a database (Zettlemoyer and Collins, 2005, e.g.,)



Semantic Parsing

Semantic parsing comprises a wide range of tasks where strings are mapped into meaning representation languages. Examples:

- ▶ Programming languages, especially query languages that can be used to answer questions using a database (Zettlemoyer and Collins, 2005, e.g.,)
- ▶ Schemas designed around real-world event-types (called “frames”); trying to extract “who did what to whom?” (Baker et al., 1998; Palmer et al., 2005)



Other Examples of Linguistic Structure Prediction

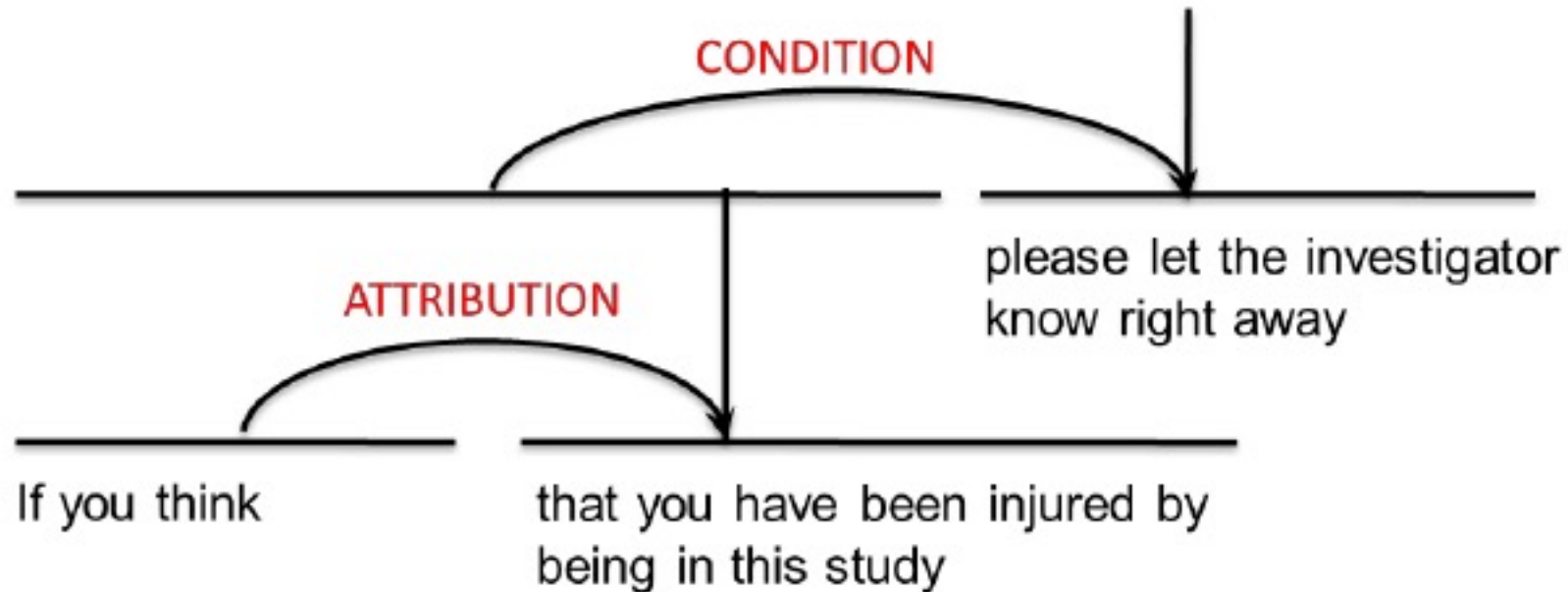
- Coreference resolution

*"I voted for Nader because he was most
aligned with my values," she said.*

The diagram shows three curved arrows indicating coreference relations. One arrow starts at the word 'I' in the first line and points to the word 'he' in the same line. A second arrow starts at the word 'she' in the second line and points to the word 'I' in the first line. A third arrow starts at the word 'she' in the second line and points to the word 'he' in the first line.

Other Examples of Linguistic Structure Prediction

- Coreference resolution
- Discourse parsing



Text Generation

Text Generation Tasks

- Generates natural language from input data or machine representations

Text Generation Tasks

- Generates natural language from input data or machine representations
- Spans a broad set of natural language processing (NLP) tasks:

<u>Task</u>	<u>Input X</u>	<u>Output Y (Text)</u>
Chatbot / Dialog System	Utterance	Response
Machine Translation	English	Chinese
Summarization	Document	Short paragraph
Description Generation	Structured data	Description
Captioning	Image/video	Description
Speech Recognition	Speech	Transcript

Two Central Goals

- Generating human-like, grammatical, and readable text
 - I.e., generating **natural** language
- Generating text that contains desired information inferred from inputs
 - Machine translation
 - Source sentence --> target sentence w/ the same meaning
 - Data description
 - Table --> data report describing the table
 - Attribute control
 - Sentiment: positive --> "I like this restaurant"
 - Conversation control
 - Control conversation strategy and topic

Text Generation Basics

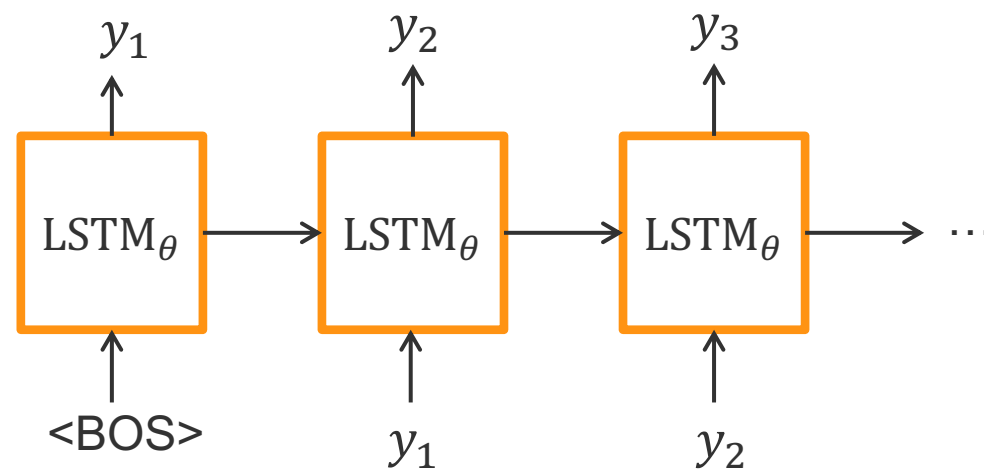
- Model
- Learning
- Inference (Decoding)
- Evaluation

Basic Building Block: Left-to-Right Language Model

- Calculates the probability of a sentence:
 - Sentence:

$$\mathbf{y} = (y_1, y_2 \dots, y_T)$$

$$p_{\theta}(\mathbf{y}) = \prod_t p_{\theta}(y_t | \mathbf{y}_{1:t-1})$$

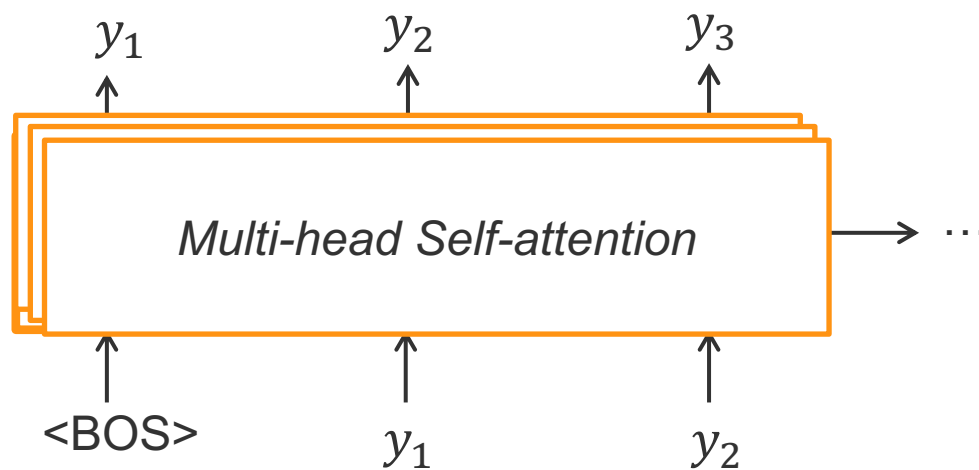


Basic Building Block: Left-to-Right Language Model

- Calculates the probability of a sentence:
 - Sentence:

$$\mathbf{y} = (y_1, y_2, \dots, y_T)$$

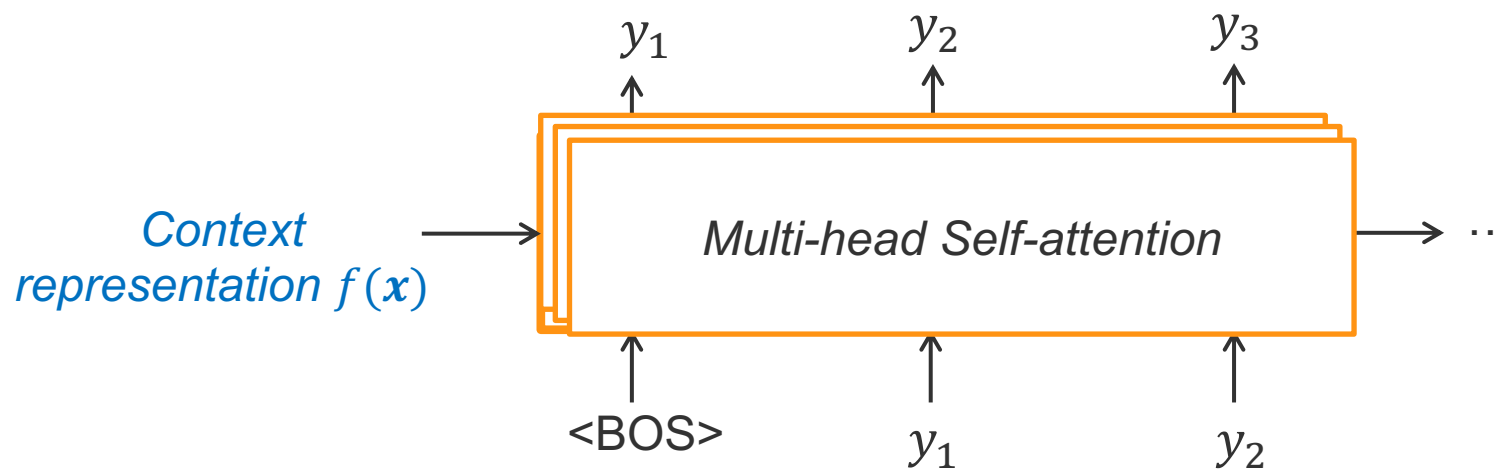
$$p_{\theta}(\mathbf{y}) = \prod_t p_{\theta}(y_t | \mathbf{y}_{1:t-1})$$



Basic Building Block: **Conditional** Language Model

- Calculates the probability of a sentence:
 - Sentence: $\mathbf{y} = (y_1, y_2, \dots, y_T)$, Context: \mathbf{x}

$$p_{\theta}(\mathbf{y} | \mathbf{x}) = \prod_t p_{\theta}(y_t | \mathbf{y}_{1:t-1}, \mathbf{x})$$

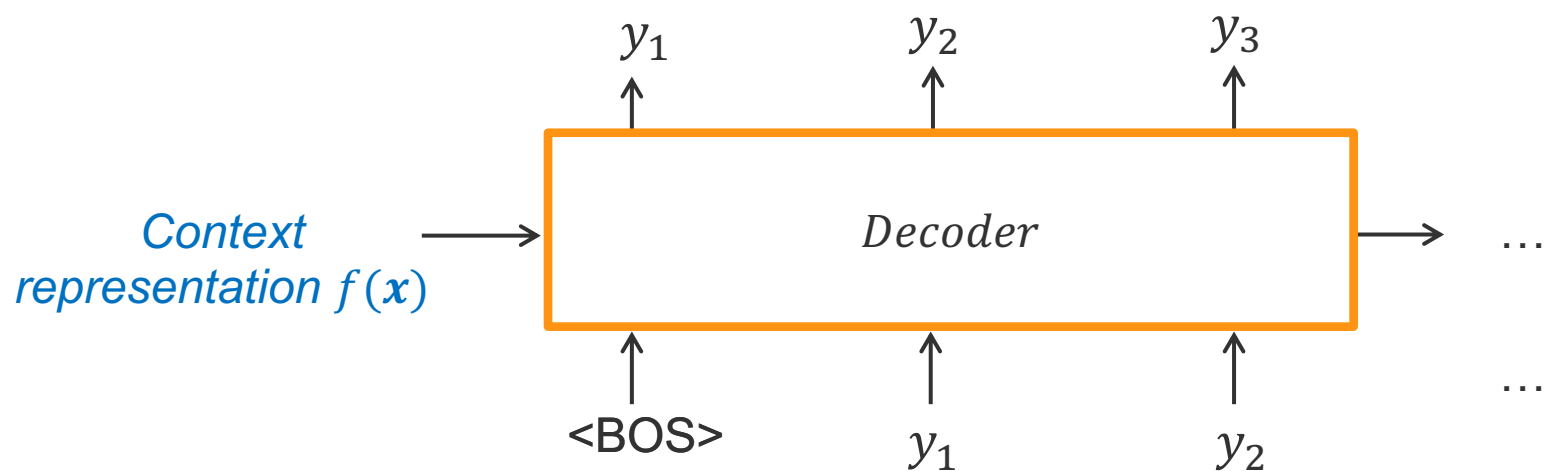


Basic Building Block: **Conditional** Language Model

- Calculates the probability of a sentence:
 - Sentence: $\mathbf{y} = (y_1, y_2 \dots, y_T)$, Context: \mathbf{x}

$$p_{\theta}(\mathbf{y} | \mathbf{x}) = \prod_t p_{\theta}(y_t | \mathbf{y}_{1:t-1}, \mathbf{x})$$

- Language model as a **decoder**

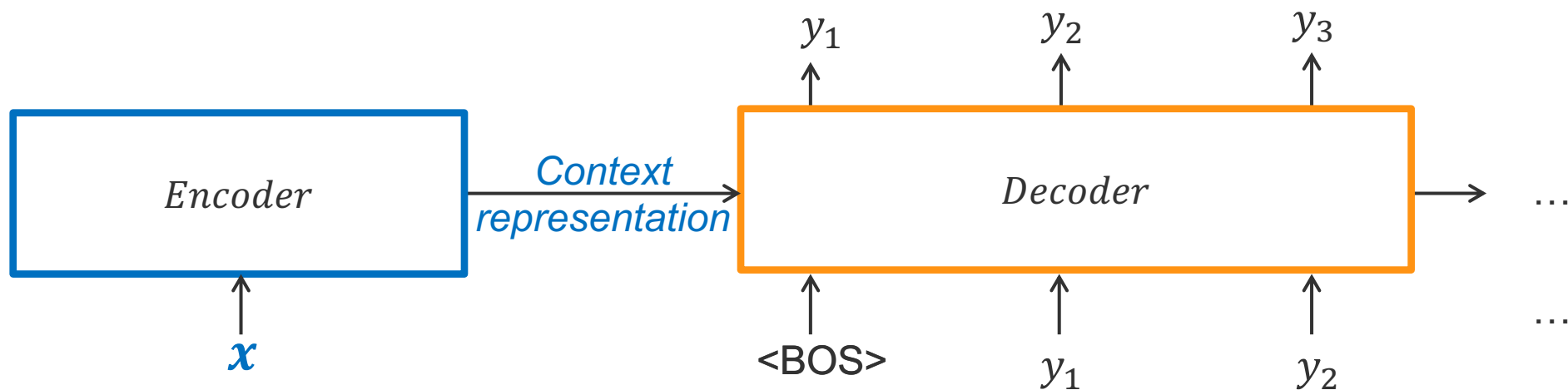


Encoder-Decoder Model

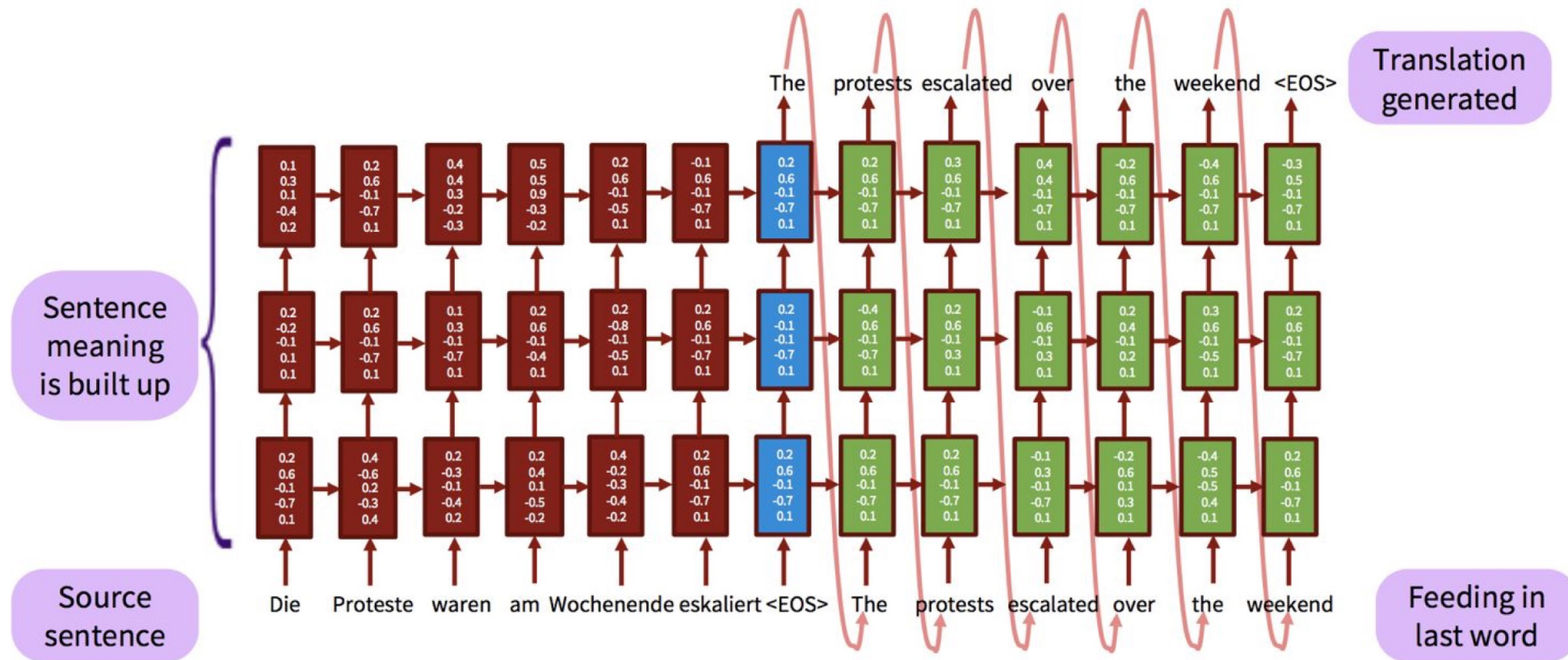
- Calculates the probability of a sentence:
 - Sentence: $\mathbf{y} = (y_1, y_2, \dots, y_T)$, Context: \mathbf{x}

$$p_{\theta}(\mathbf{y} | \mathbf{x}) = \prod_t p_{\theta}(y_t | \mathbf{y}_{1:t-1}, \mathbf{x})$$

- Language model as a **decoder**
- Encodes context with an **encoder**



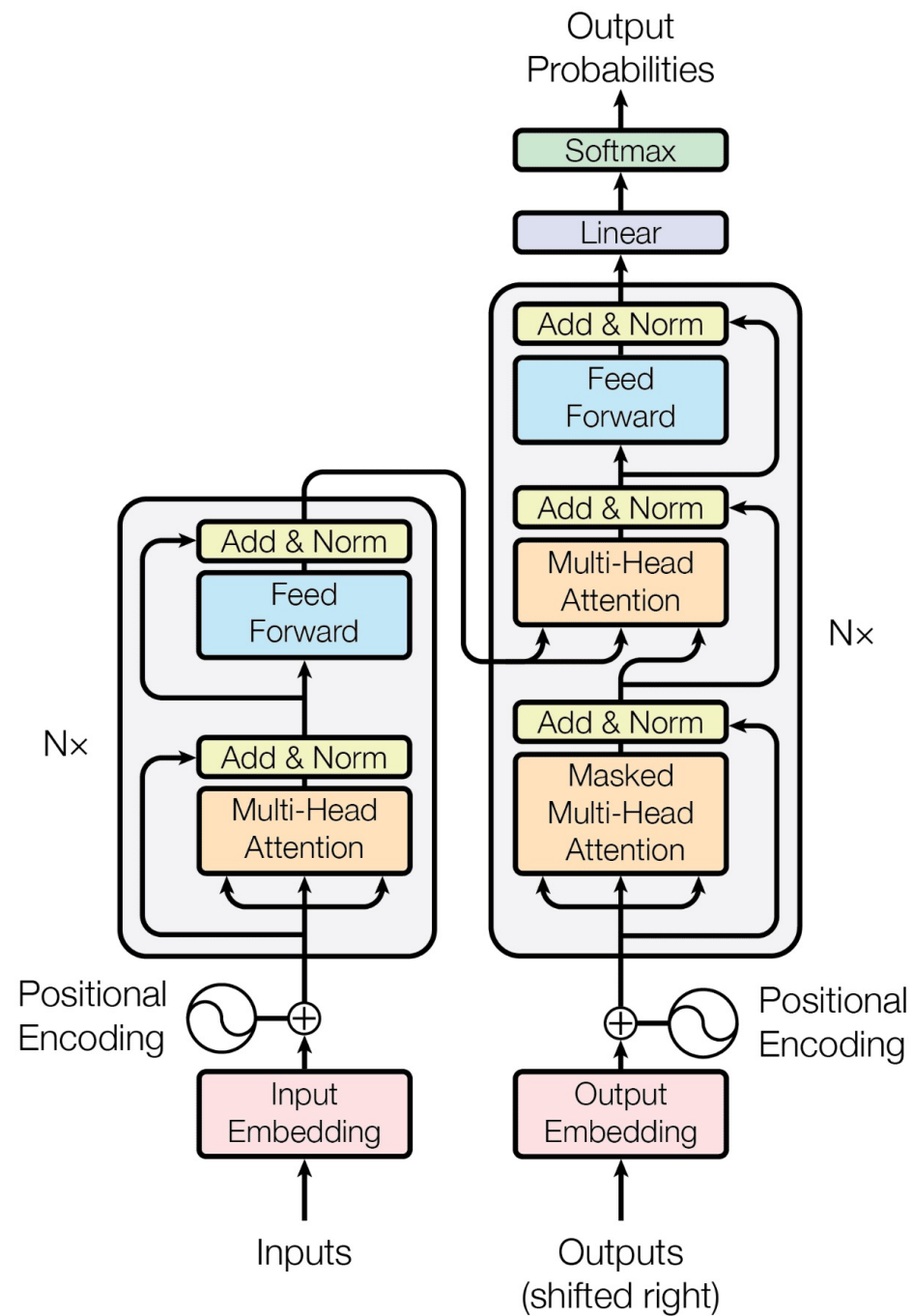
Encoder-Decoder Model



[Sutskever et al. 2014, Bahdanau et al. 2014]

Encoder-Decoder Model

Transformers encoder-decoder
(Lecture #3)



Text Generation Basics

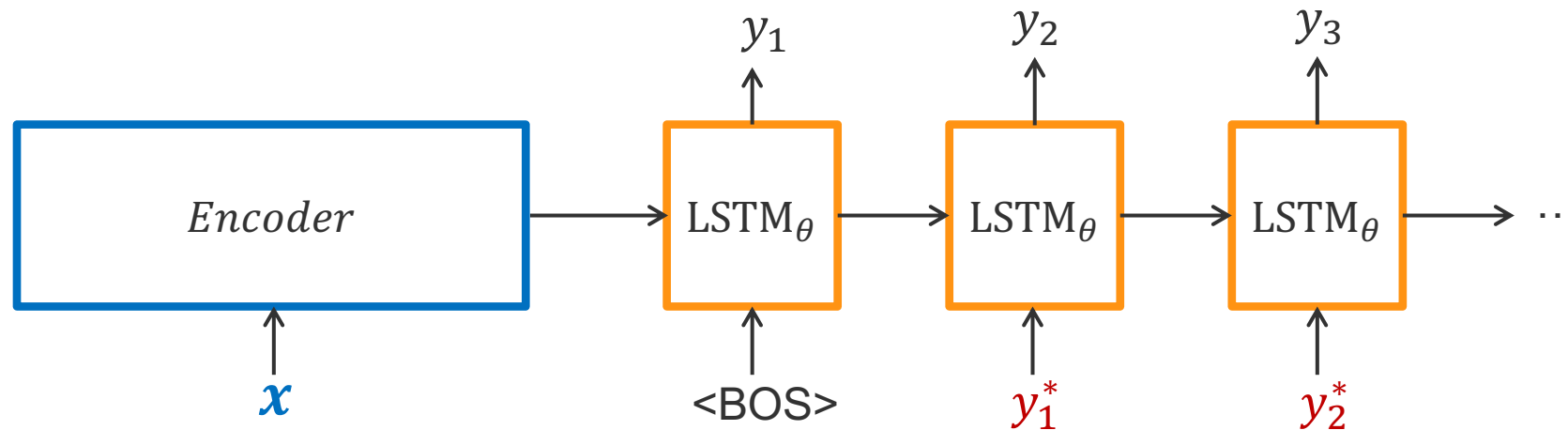
- Model
- Learning
- Inference (Decoding)
- Evaluation

Supervised Training

- Given data example $(\mathbf{x}^*, \mathbf{y}^*)$
- Minimizes negative log-likelihood of the data

$$\min_{\theta} \mathcal{L}_{\text{MLE}} = -\log p_{\theta}(\mathbf{y}^* | \mathbf{x}^*) = -\prod_{t=1}^T p_{\theta}(y_t^* | \mathbf{y}_{1:t-1}^*, \mathbf{x}^*)$$

- Sequence cross-entropy loss
- Inference: **teacher-forcing decoding**
 - For every step t , feed in the previous ground-truth tokens $\mathbf{y}_{1:t-1}^*$ to decode next step



Text Generation Basics

- Model
- Learning
- Inference (Decoding)
- Evaluation

Decoding

- Once the model is trained, we can apply different decoding methods to generate text sequence \mathbf{y}
- Popular basic decoding methods:
 - Beam-search decoding
 - Greedy decoding
 - Random sample decoding
 - Top-k decoding
 - Top-p decoding

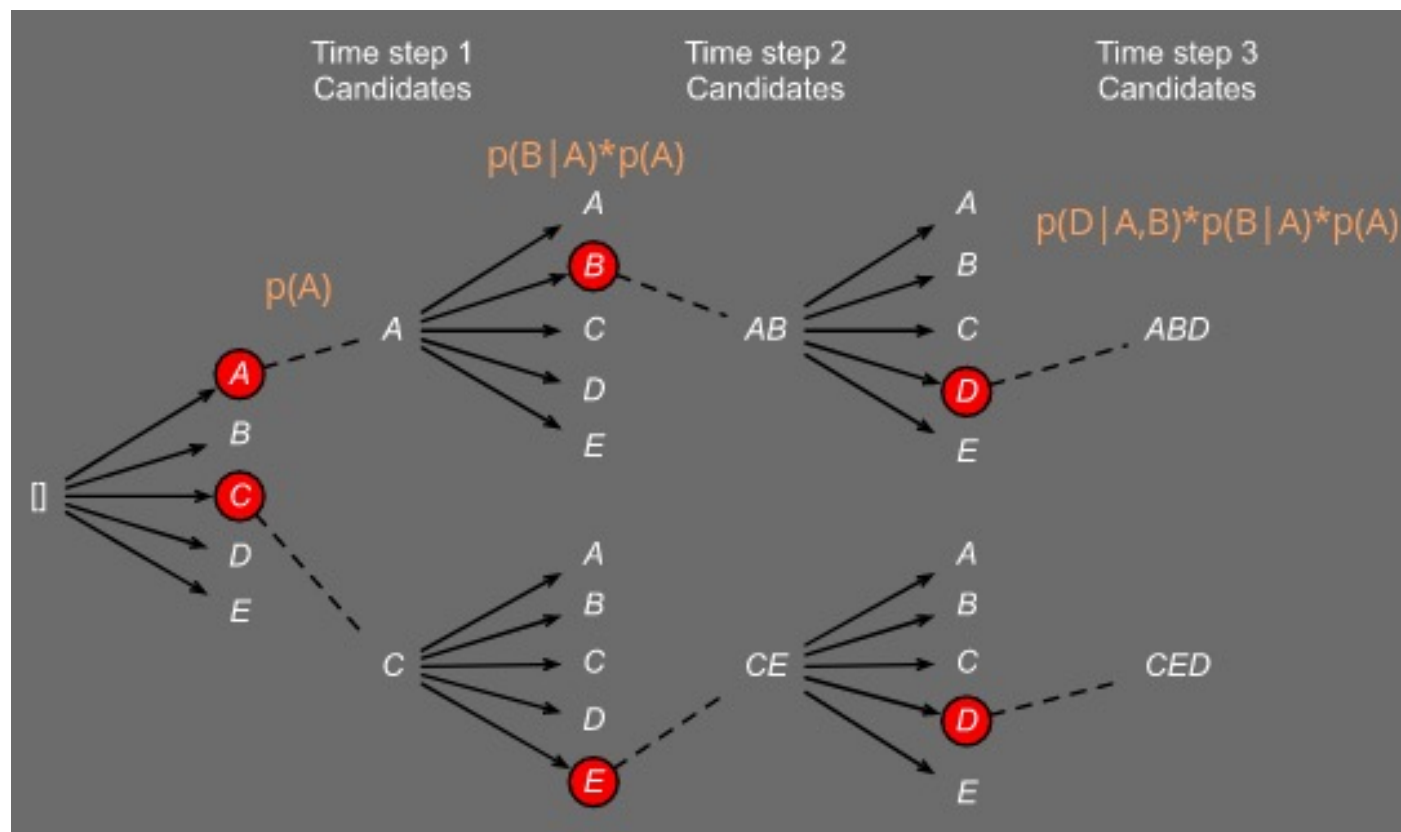
Decoding: Beam Search

- We want $\hat{y} = \operatorname{argmax}_{y \in \mathcal{L}^n} \operatorname{Score}(x, y; \theta)$

where $\operatorname{Score}(x, y; \theta) = p_{\theta}(y|x)$

- Beam Search approximately solves it

(Example: beam width = 2)

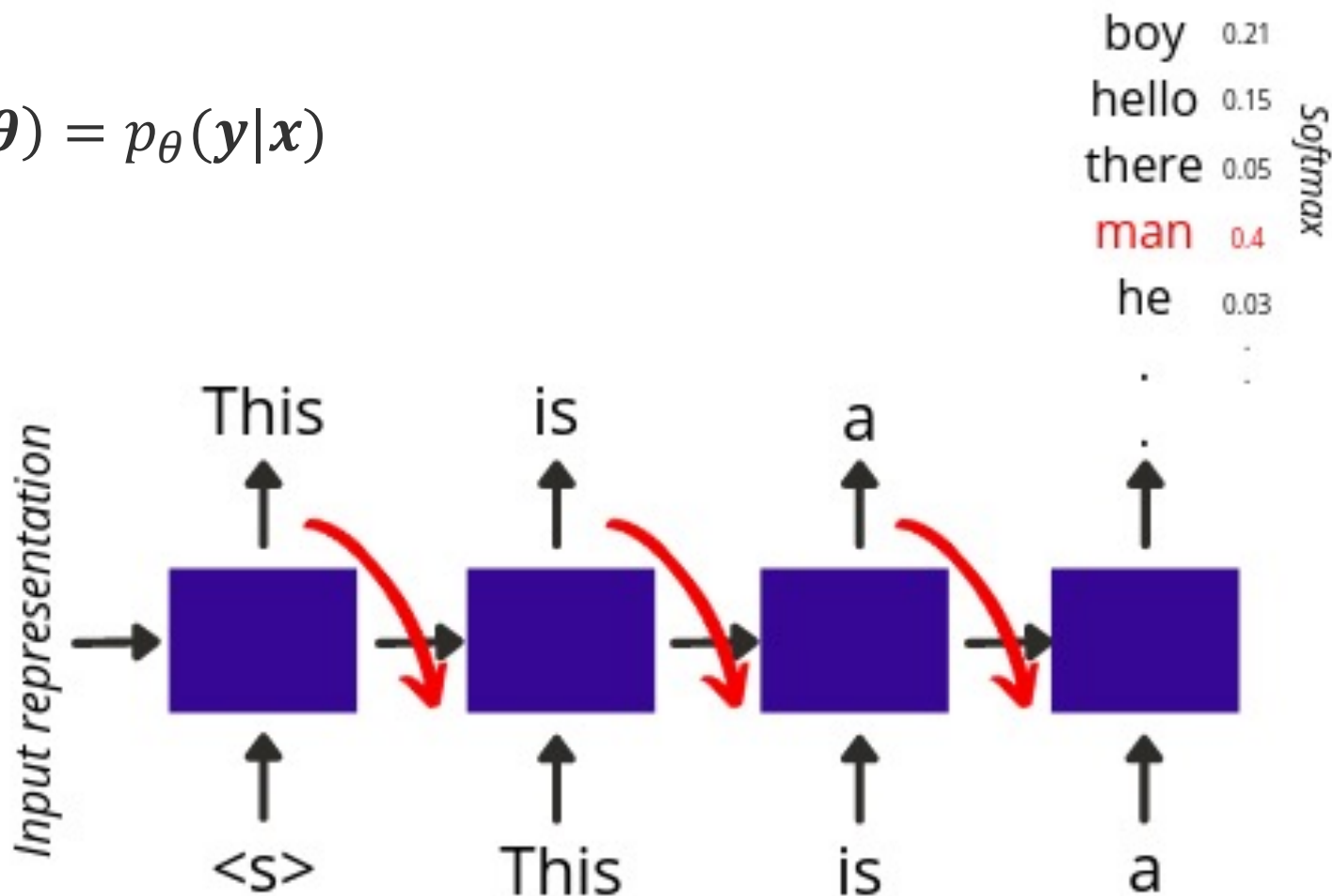


Decoding: Greedy

- We want $\hat{y} = \operatorname{argmax}_{y \in \mathcal{L}^n} \operatorname{Score}(x, y; \theta)$

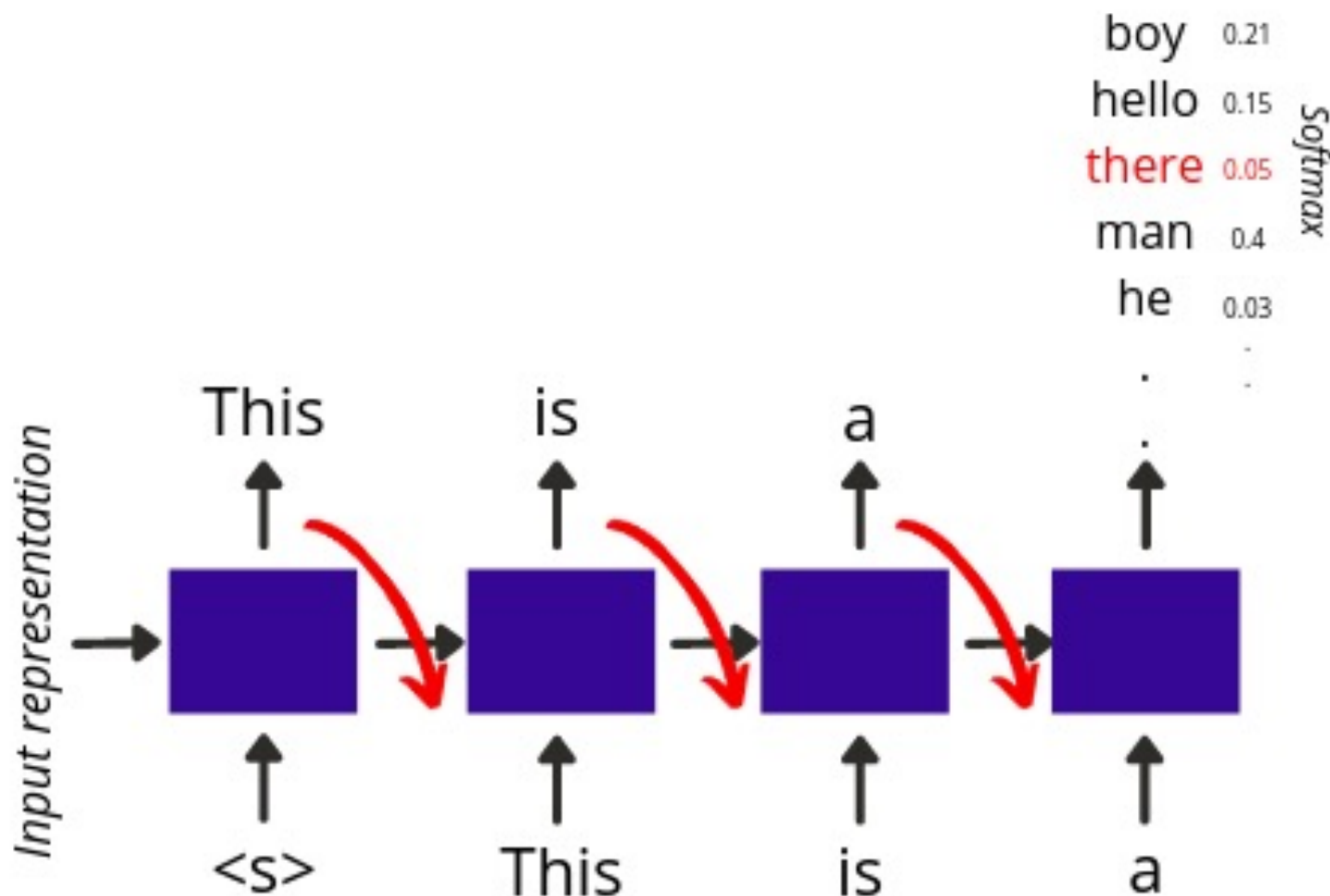
where $\operatorname{Score}(x, y; \theta) = p_{\theta}(y|x)$

- Greedy decoding: beam width = 1



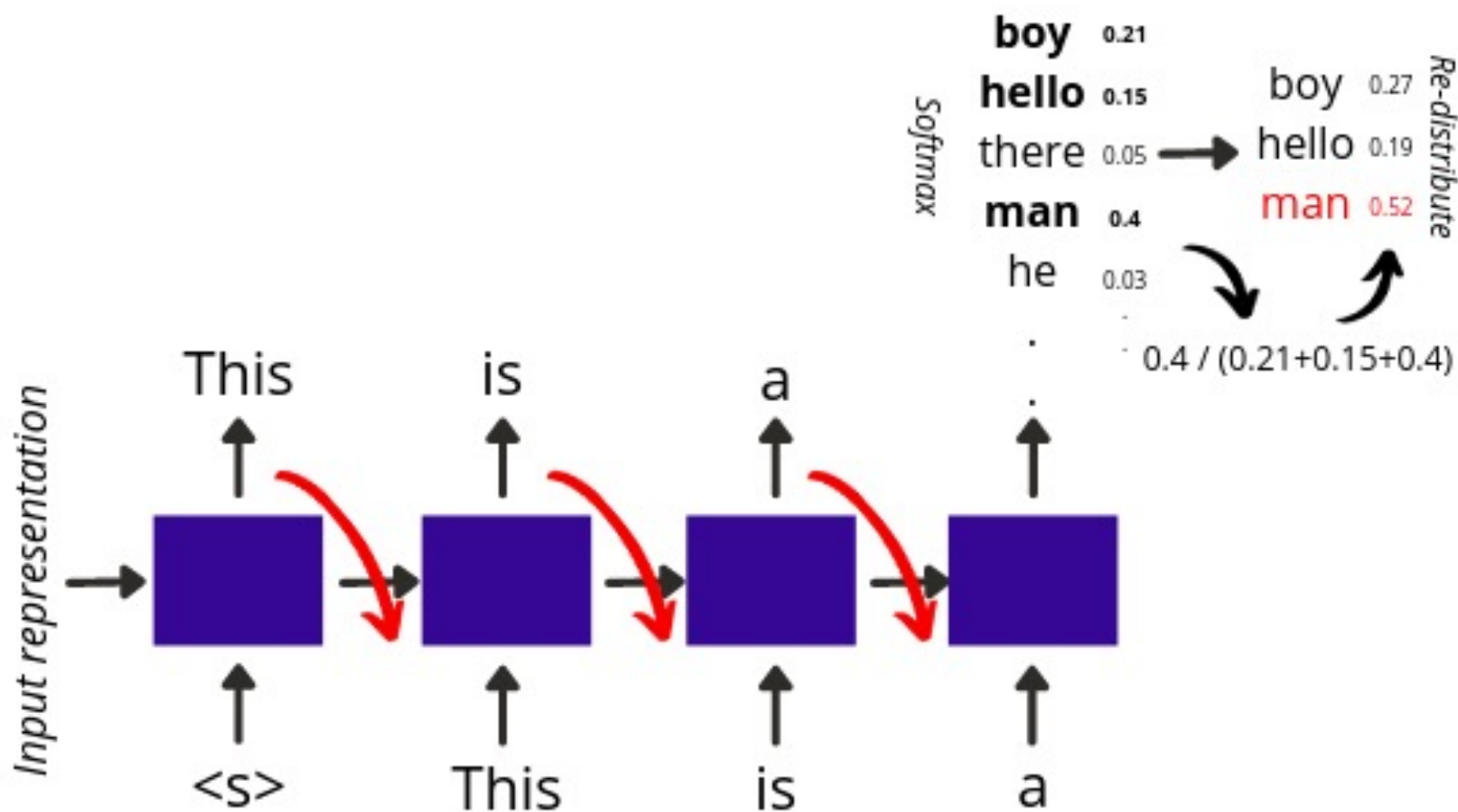
Decoding: Random Sample

- At each step t , sample a random word based on the conditional distribution $p_{\theta}(y_t \mid \mathbf{y}_{1:t-1}, \mathbf{x})$



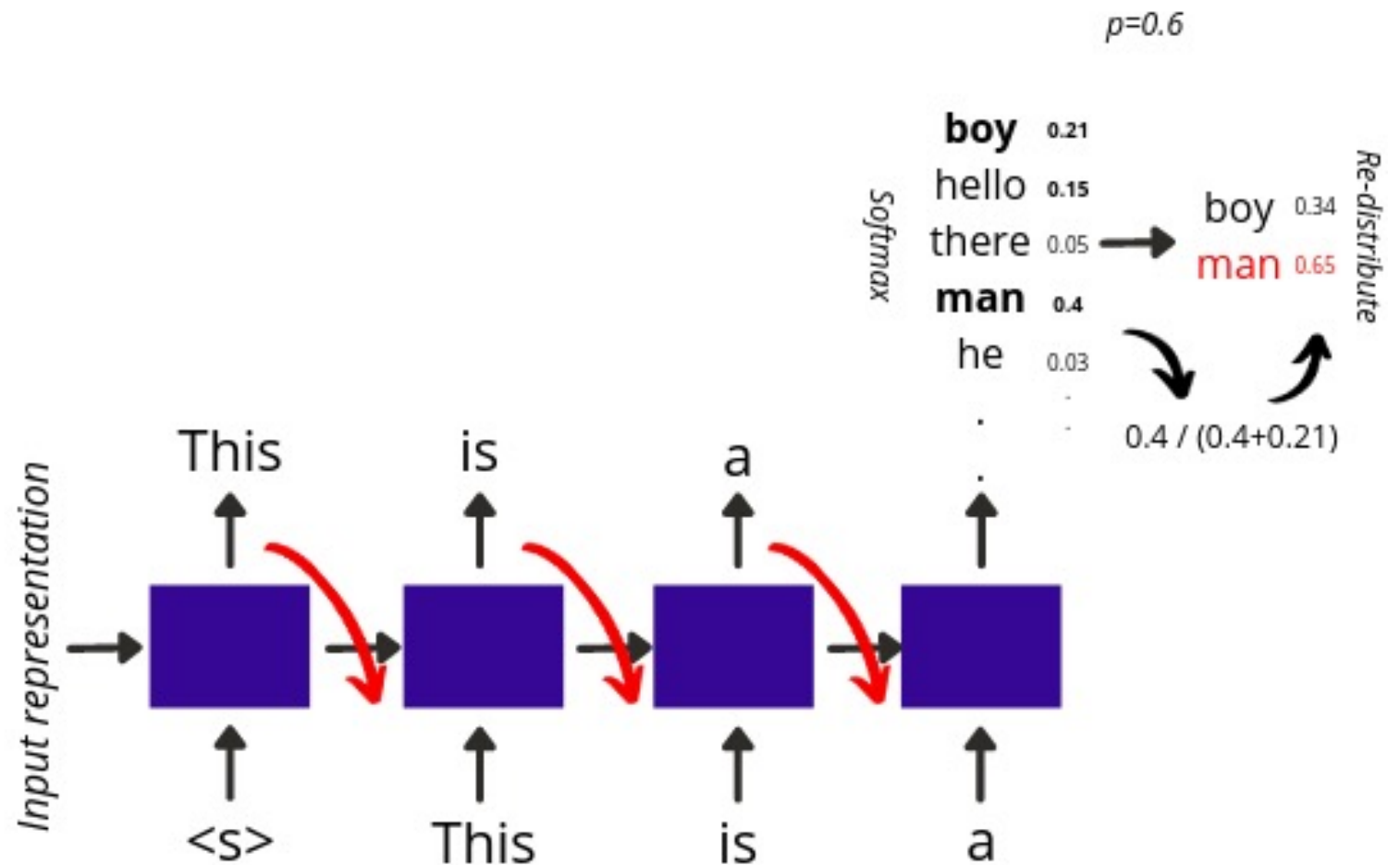
Decoding: Top-k

- At each step t , sample a word from the top-k most probable candidates based on the conditional distribution $p_{\theta}(y_t \mid \mathbf{y}_{1:t-1}, \mathbf{x})$



Decoding: Top- p

- At each step t , sample a word from the top candidates whose cumulative probability exceeds the probability mass p



Text Generation Basics

- Model
- Learning
- Inference (Decoding)
- Evaluation

Evaluation

- A big challenge in text generation research
- Many ways for automatic evaluation
 - E.g., comparing with human-written references
 - BLEU (Papineni et al., 2002) for machine translation
 - Weighted average of n-gram precision (across different n)
 - n-gram precision p_n

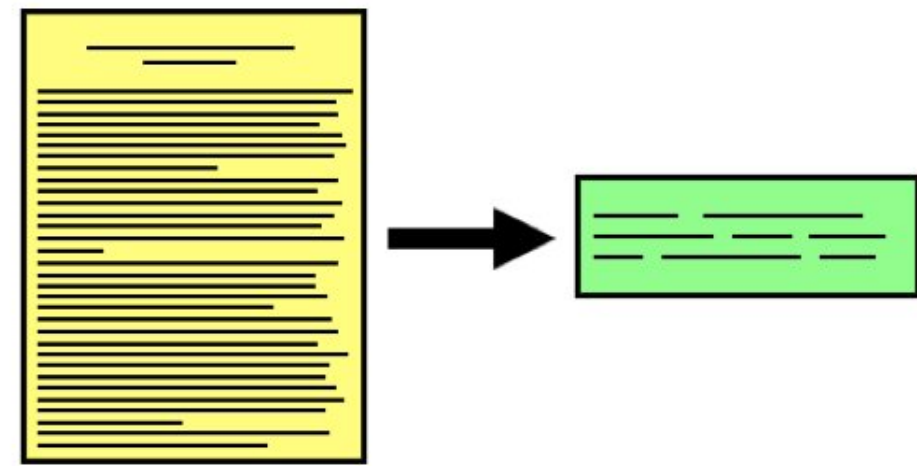
$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

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

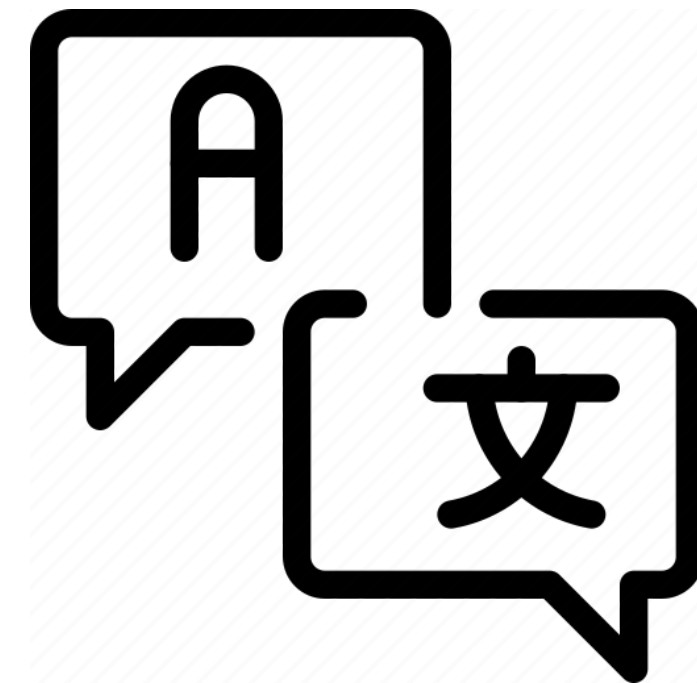
Then,

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) .$$

Natural language generation (NLG) tasks have diverse goals



Summarization



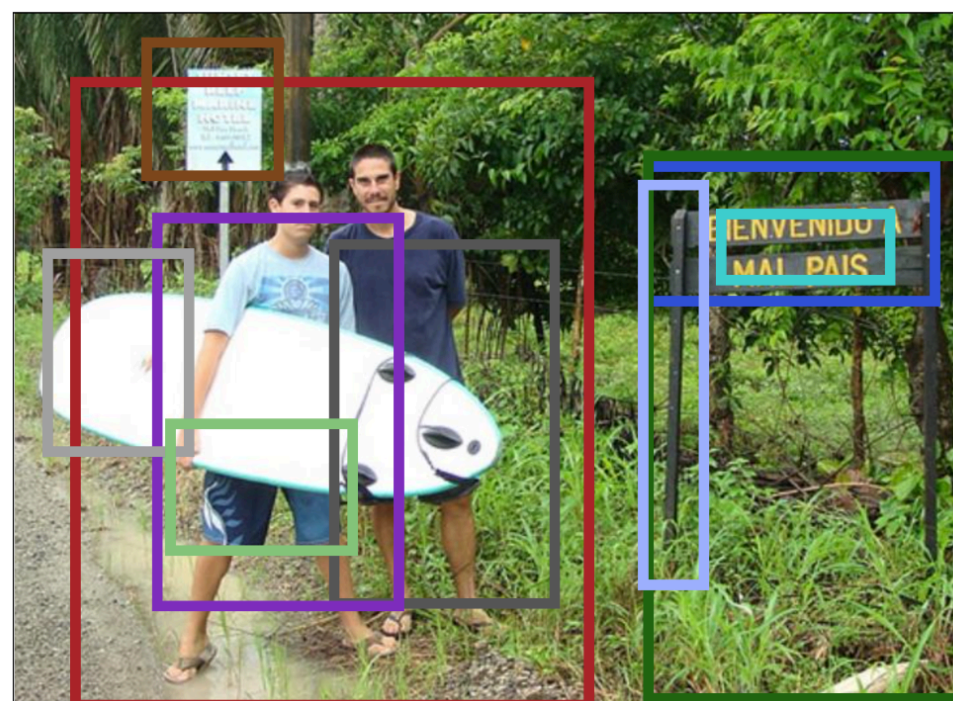
Translation



Dialog

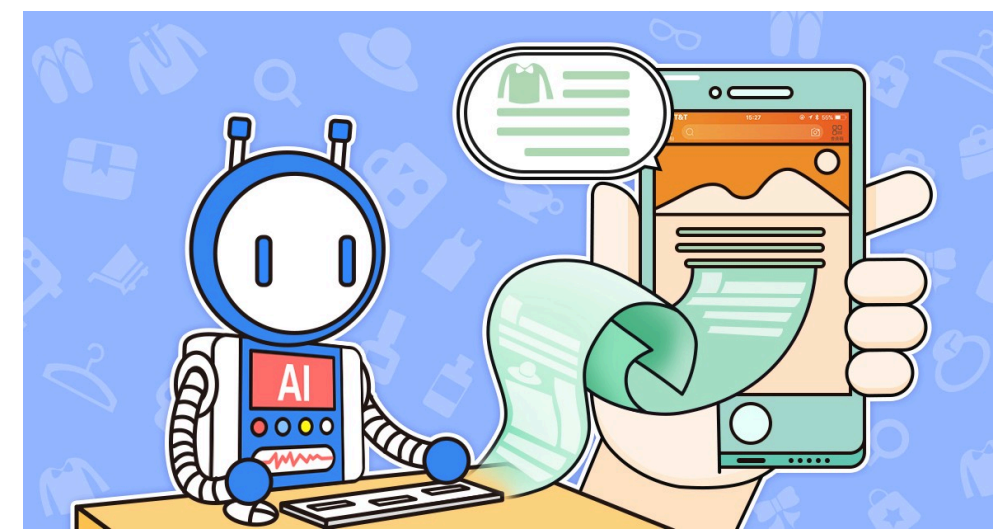


Poetry Generation



two men standing on the beach. the sign is black and white. a girl holding a frisbee. a wooden sign. white sign with black writing. man holding a white frisbee. white frisbee in the air. the shorts are blue. a metal pole holding a sign. the sign is yellow.

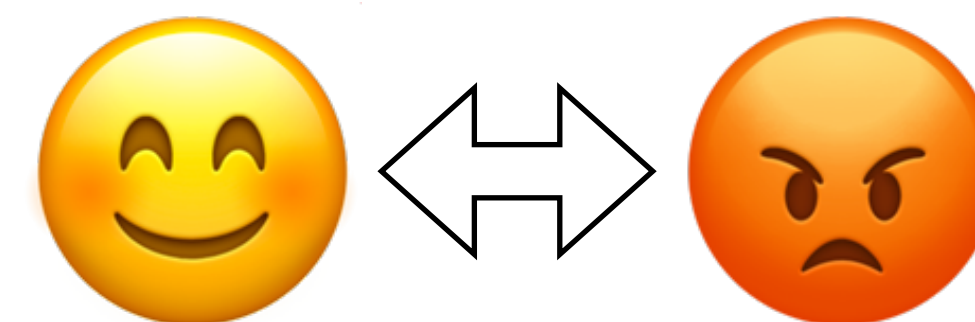
Image Captioning



Story Generation

<p>George Mikell</p>  <p>Mikell in March 2017</p> <p>Born Jurgis Mikelaitis 4 April 1929 (age 88) Elderslie, Lithuania</p> <p>Nationality Lithuanian, Australian</p> <p>Occupation Actor, writer</p> <p>Years active 1957–present</p> <p>Known for <i>The Guns of Navarone</i> <i>The Great Escape</i></p> <p>Height 6' 0" (1.83m)</p>	 <p>WIKIPEDIA The Free Encyclopedia</p> <p>→</p> <p>George Mikell (born Jurgis Mikelaitis; 4 April 1929) is a Lithuanian-Australian actor and writer best known for his performances in <i>The Guns of Navarone</i> (1961) and <i>The Great Escape</i> (1963).</p>
---	---

Data-to-Text



Sentiment Transfer

And the list is growing...

Comparing with reference is not enough

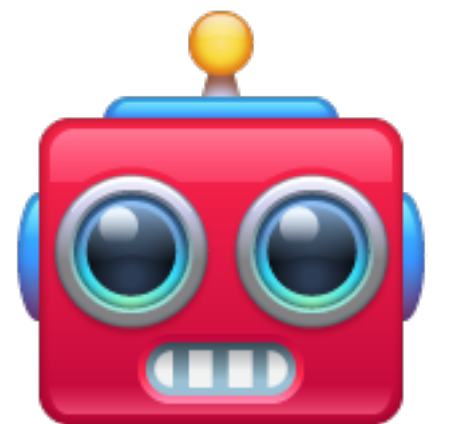


Article: McConaughey, 47, graduated from the university in 1993. He is an avid fan of its American football team...



Reference: McConaughey is a football fan

Summary 1: McConaughey is a **soccer** fan



Summary 2: McConaughey graduated from the university in 1993



Comparing with reference is not enough



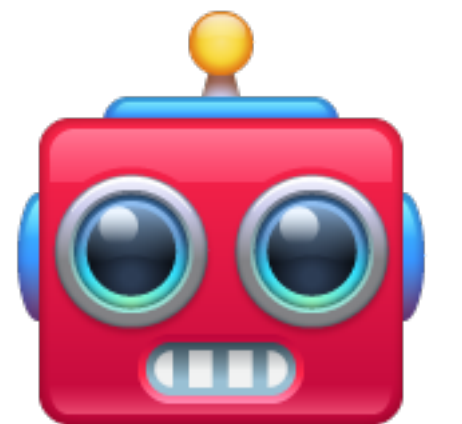
Article: McConaughey, 47, graduated from the university in 1993. He is an avid fan of its American football team...



Reference: McConaughey is a football fan

Factual Error

Summary 1: McConaughey is a **soccer** fan



Summary 2: McConaughey graduated from the university in 1993



Previous work on NLG evaluation

Grammaticality
Sequence Tagging
Interestingness
Redundancy
Persona Distinctiveness
Direct Assessment
Depth
Dullness
Informativeness
Content Selection
Pseudo Reference
Hallucination
Repetitiveness
Fluency
Pointwise Mutual Information
Human Score Regression
Clarity
Coverage
Knowledge Usage
Importance
Novelty
Linguistic Quality
Contradiction
Relevance
Entailment Classification
Factual Correctness
Lexical Matching
Perplexity
Semantic Similarity
Reference-Free
Engagingness
Shannon Game
Helpfulness
Consistency
Factuality
Naturalness
Word Mover Distance
Reference-Based
QA Metric
Embedding Matching
Coherence
Faithfulness
Knowledge Injection
Appropriateness
Automatic Turing Test
Diversity
Sensibleness

Previous work on NLG evaluation

- Need more common theoretical ground across tasks
- Need more unified guidance for new tasks/aspects

Grammaticality Sequence Tagging Interestingness Redundancy

Persona Distinctiveness Depth Dullness Informativeness

Content Selection Direct Assessment Repetitiveness

Fluency Pointwise Mutual Information Hallucination Clarity

Coverage Knowledge Importance Novelty

Linguistic Quality Entailment Classification

Factual Correctness Matching

Semantic Similarity Perplexity

Helpfulness Shannon Game

Word Mover Distance Naturalness

Embedding Matching Appropriateness

Coherence Faithfulness Knowledge Injection

Automatic Turing Test Diversity Sensibleness

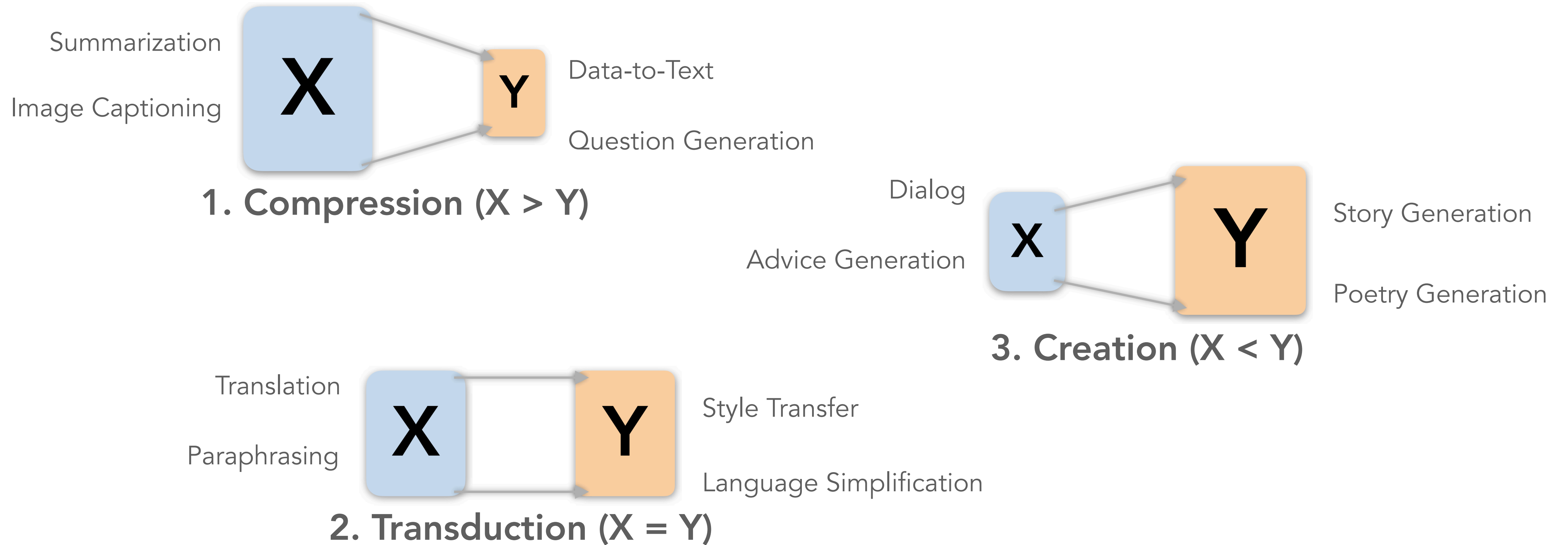
A More Unified Framework for NLG Evaluation

What to evaluate: based on NLG task category

Categorize based on information change from input (X) to output (Y)

What to evaluate: based on NLG task category

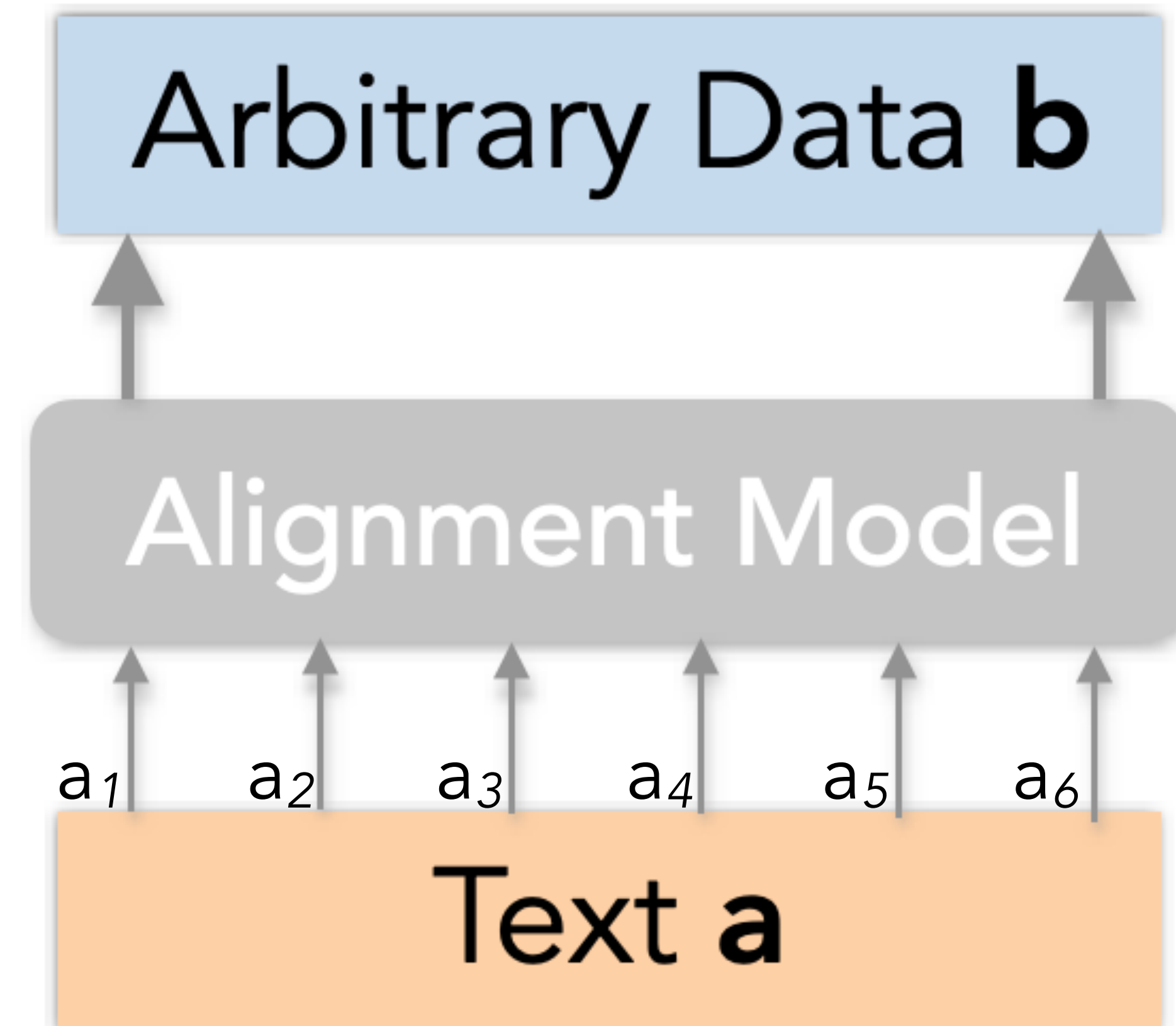
Categorize based on information change from input (X) to output (Y)



How to evaluate: unified information alignment

Definition: The *information alignment* from text **a** to arbitrary data **b** is

$$\text{align}(\mathbf{a} \rightarrow \mathbf{b}) = \langle \alpha_1, \alpha_2, \dots, \alpha_N \rangle$$

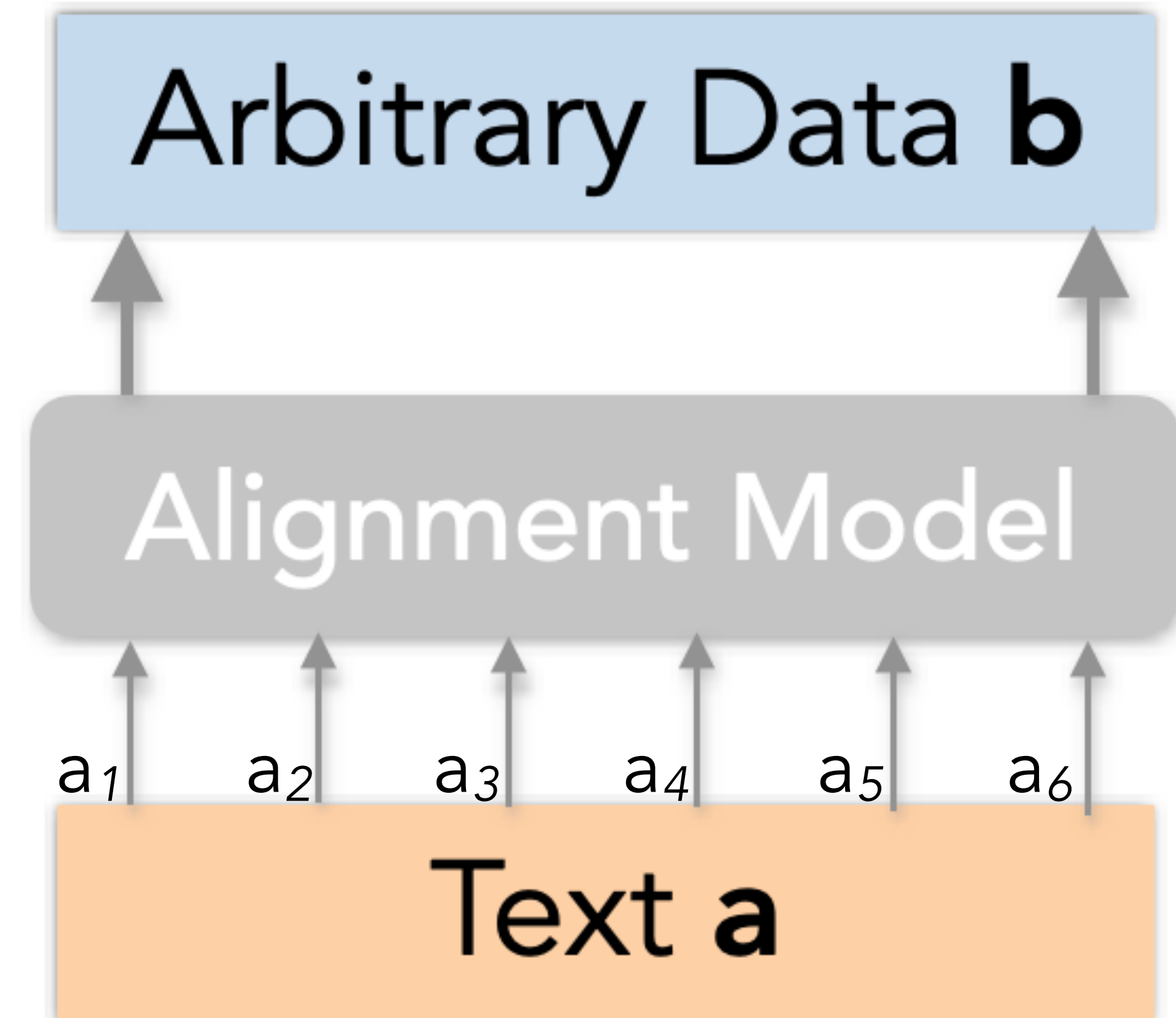


How to evaluate: unified information alignment

Definition: The *information alignment* from text **a** to arbitrary data **b** is

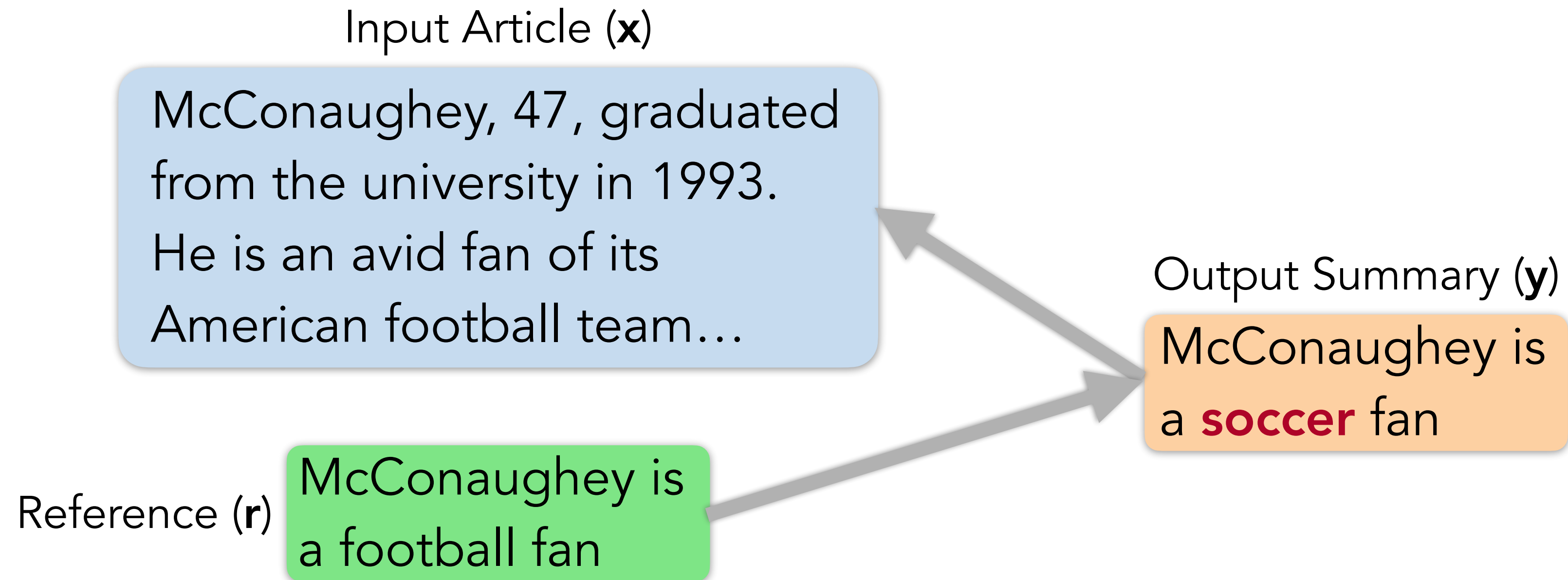
$$\text{align}(\mathbf{a} \rightarrow \mathbf{b}) = \langle \alpha_1, \alpha_2, \dots, \alpha_N \rangle$$

- Vector of scores for each **a** token
- Score α_i : confidence token a_i is grounded in **b**



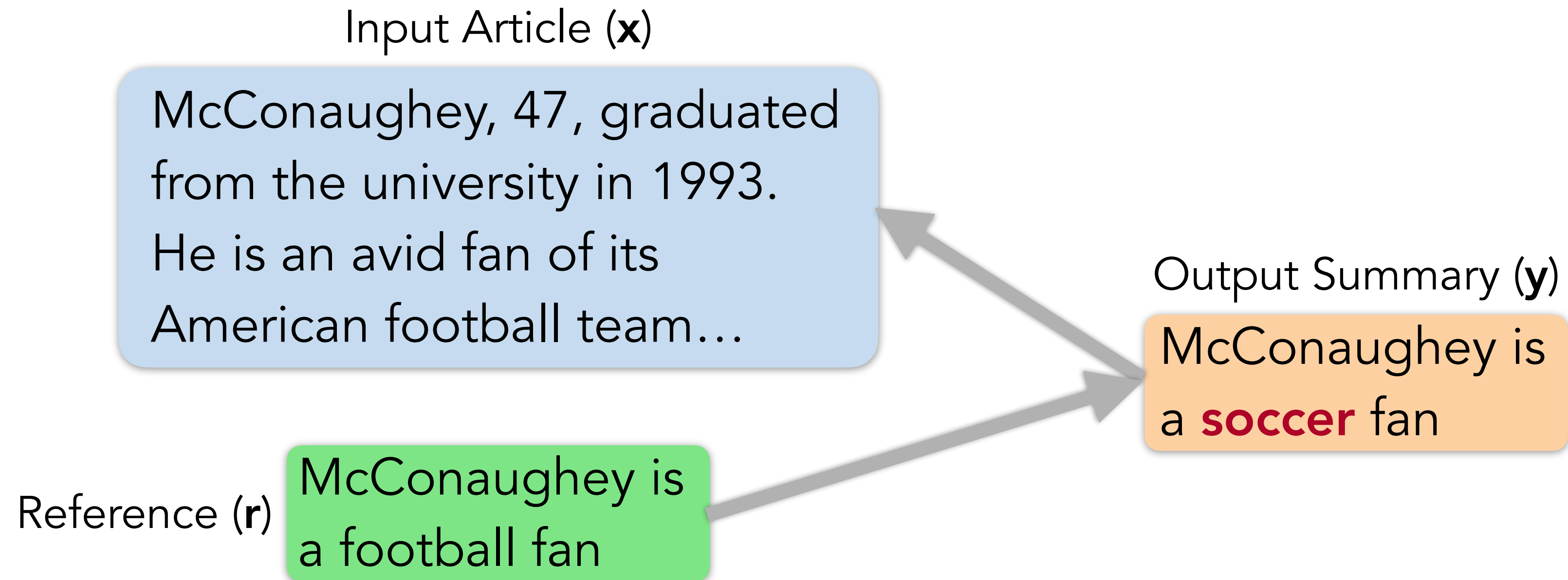
Evaluation of *compression* tasks

e.g. summarization



Evaluation of *compression* tasks

e.g. summarization

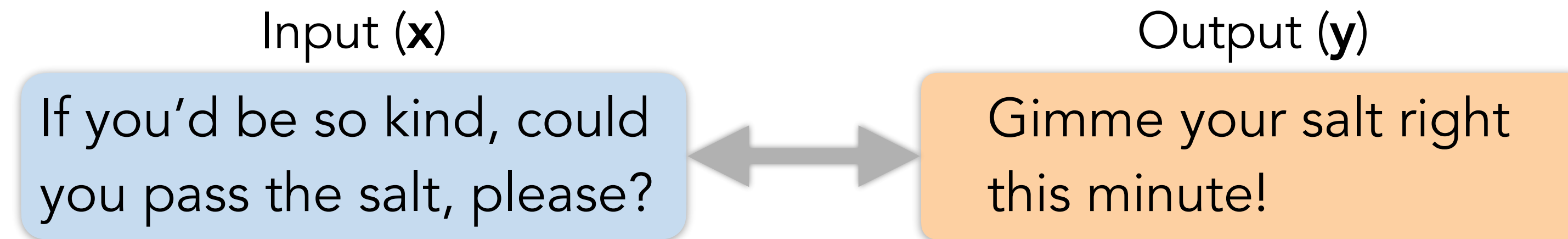


$$\text{CONSISTENCY}(\mathbf{y}, \mathbf{x}) = \text{mean}(\text{align}(\mathbf{y} \rightarrow \mathbf{x}))$$

$$\text{RELEVANCE}(\mathbf{y}, \mathbf{x}, \mathbf{r}) = \text{mean}(\text{align}(\mathbf{r} \rightarrow \mathbf{y})) \times \text{mean}(\text{align}(\mathbf{y} \rightarrow \mathbf{x}))$$

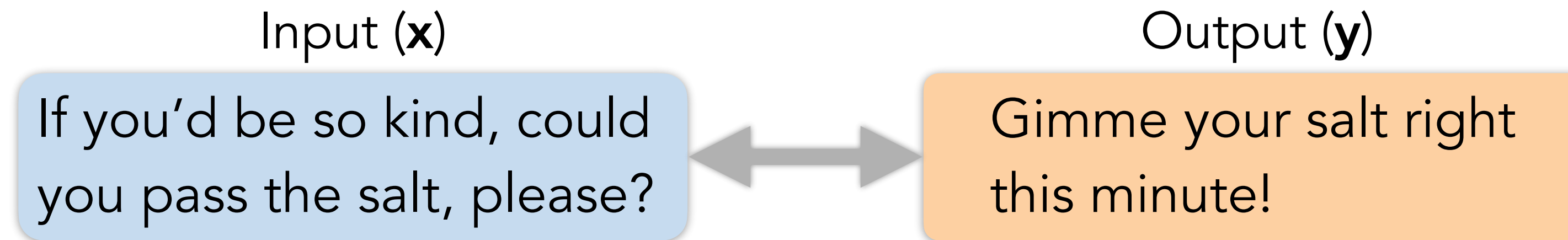
Evaluation of *transduction* tasks

e.g. style transfer



Evaluation of *transduction* tasks

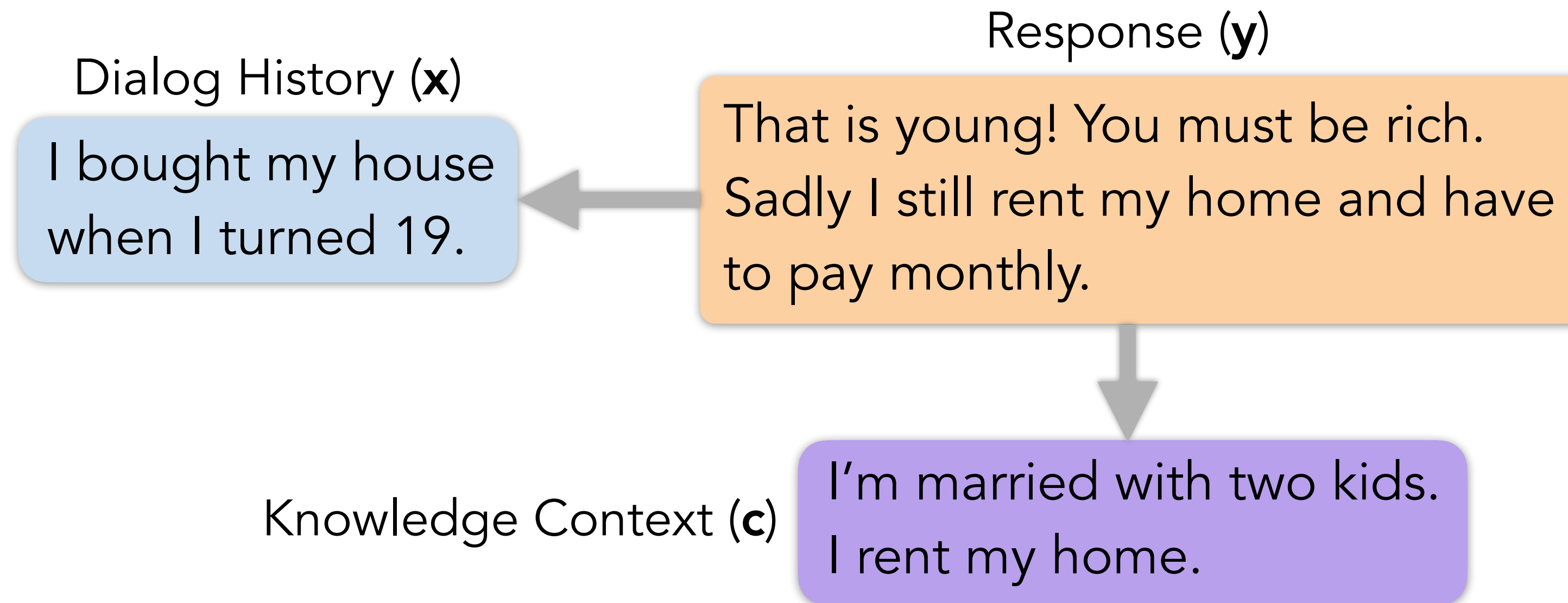
e.g. style transfer



$$\text{PRESERVATION}(\mathbf{y}, \mathbf{x}) = \frac{\text{mean}(\text{align}(\mathbf{y} \rightarrow \mathbf{x})) \times \text{mean}(\text{align}(\mathbf{x} \rightarrow \mathbf{y}))}{\text{mean}(\text{align}(\mathbf{y} \rightarrow \mathbf{x})) + \text{mean}(\text{align}(\mathbf{x} \rightarrow \mathbf{y}))}$$

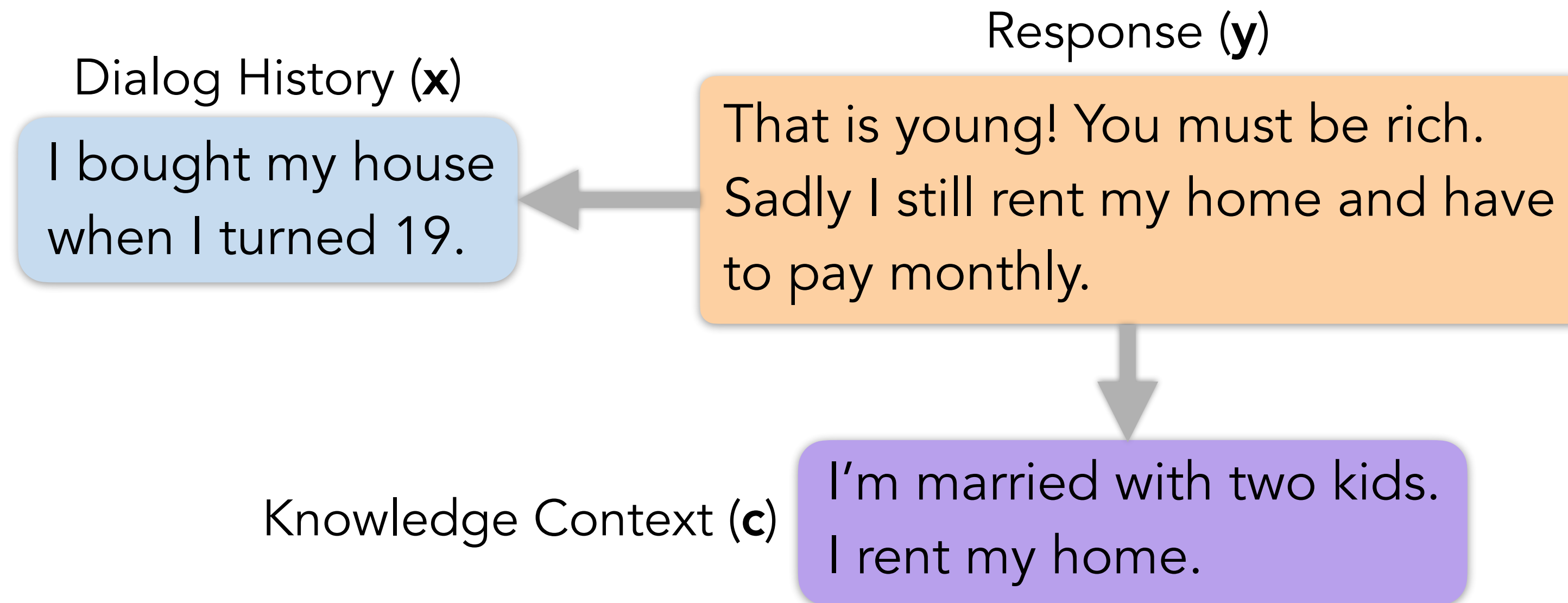
Evaluation of *creation* tasks

e.g. knowledge-based dialog



Evaluation of *creation* tasks

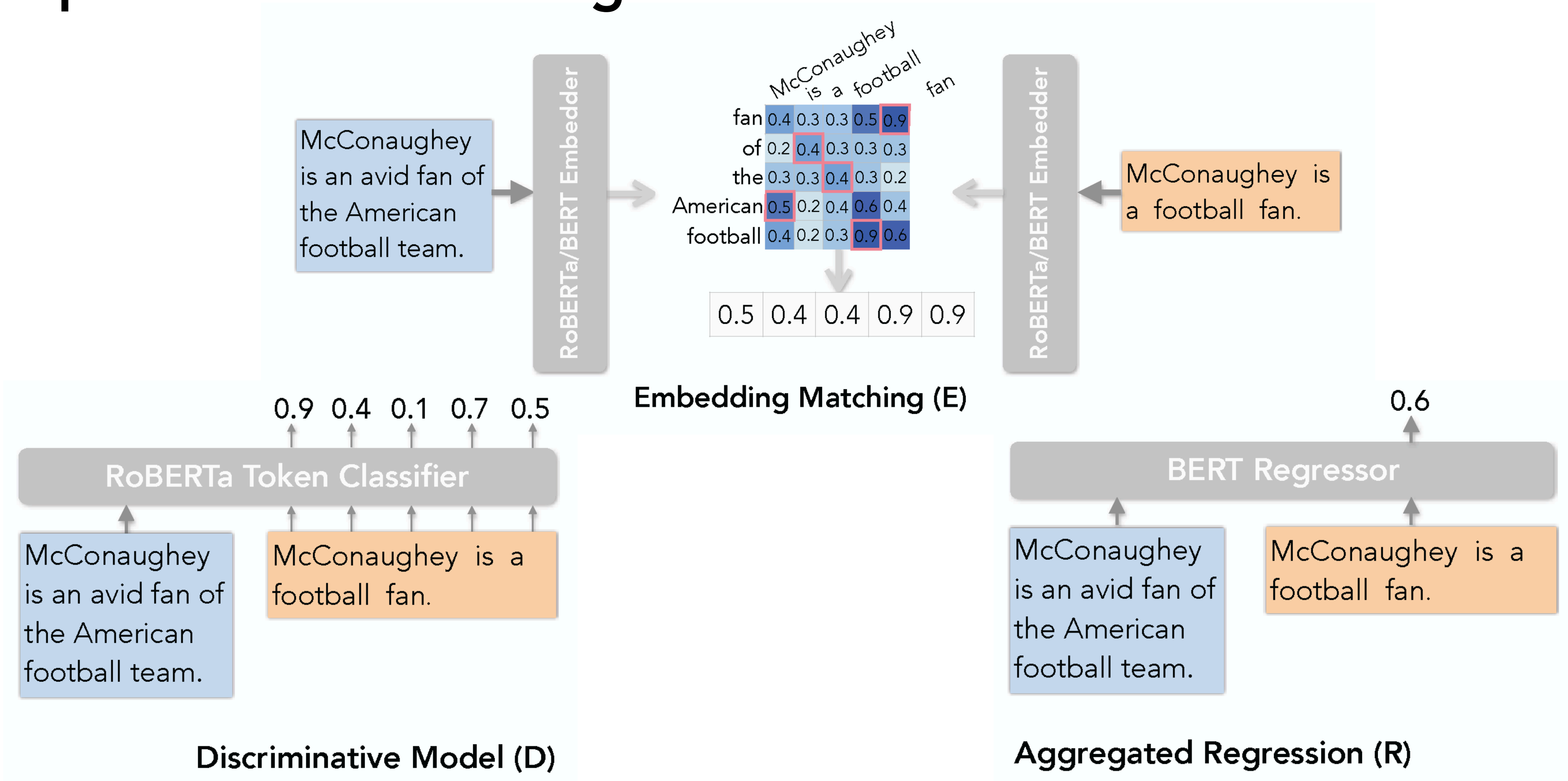
e.g. knowledge-based dialog



$$\text{ENGAGINGNESS}(\mathbf{y}, \mathbf{x}, \mathbf{c}) = \text{sum}(\text{align}(\mathbf{y} \rightarrow [\mathbf{x}, \mathbf{c}])))$$

$$\text{GROUNDEDNESS}(\mathbf{y}, \mathbf{c}) = \text{sum}(\text{align}(\mathbf{y} \rightarrow \mathbf{c}))$$

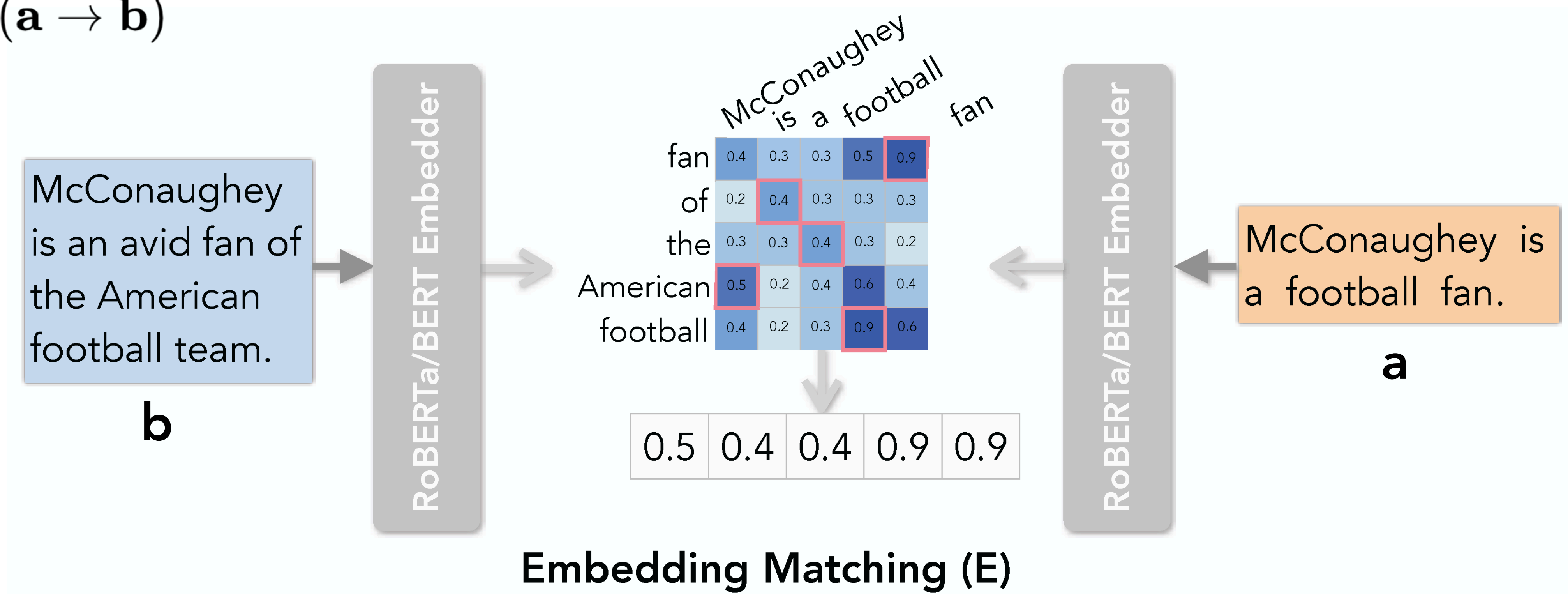
Implementations for alignment models



Illustrations depict alignment from output (in orange) to input (in blue)

Implementations for alignment models (1)

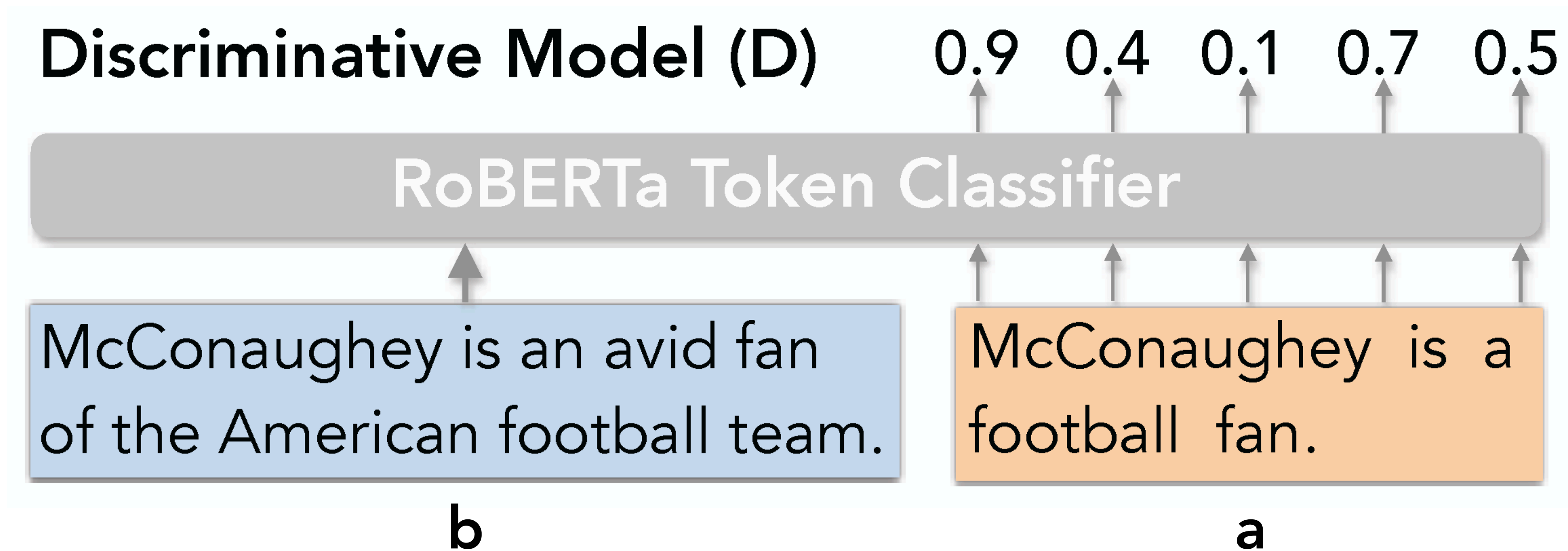
$align(\mathbf{a} \rightarrow \mathbf{b})$



- Compute the contextual representations of tokens in **a** (orange) and **b** (blue) with BERT or RoBERTa
- For each token in **a**, take the maximum cosine similarity with tokens in **b** as the alignment score

Implementations for alignment models (2)

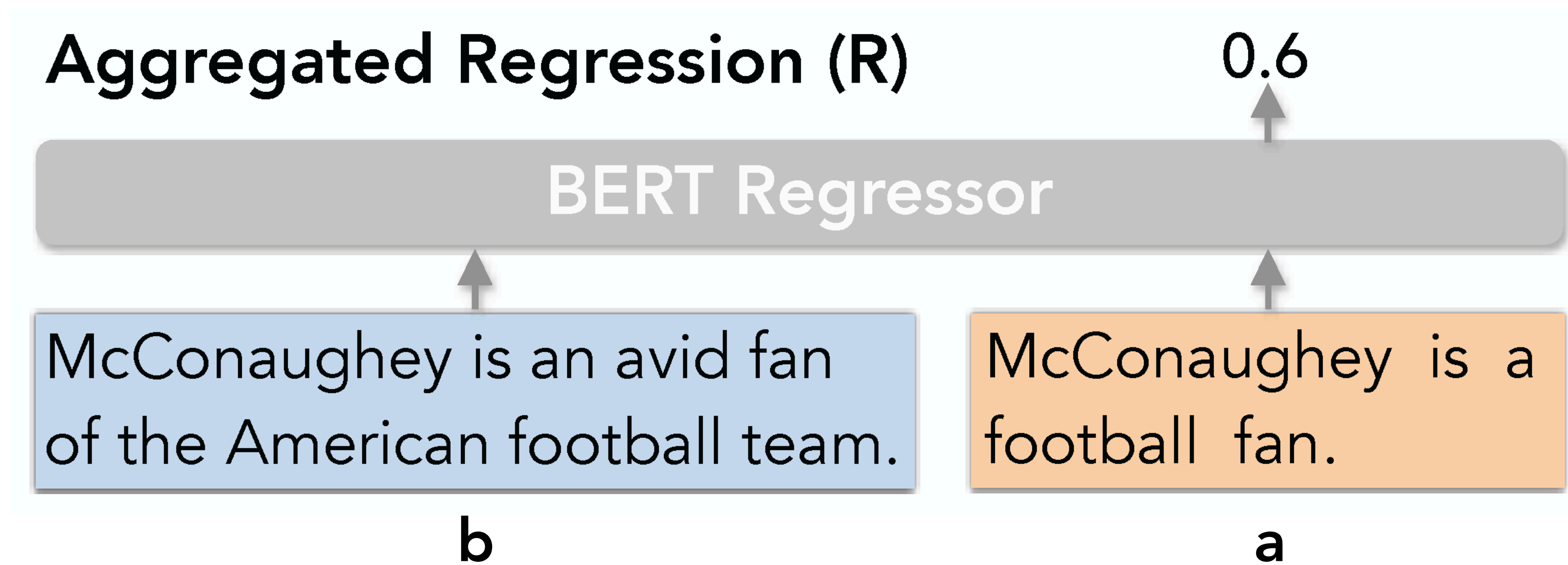
$align(\mathbf{a} \rightarrow \mathbf{b})$



- Train a token classifier to predict alignment with weakly-supervised data
- For each token in **a**, the predicted probability of alignment is the alignment score

Implementations for alignment models (3)

align(**a** → **b**)



- Train a regression model on the aggregated scores from weakly-supervised data
- The prediction is the aggregated alignment score for the entire text

Experiments

- **Setting:** Commonly used human annotation datasets in the following tasks
 - *Compression:* Summarization
 - *Transduction:* Style transfer
 - *Creation:* Knowledge-based dialog

Experiments

- **Setting:** Commonly used human annotation datasets in the following tasks

- *Compression:* Summarization
- *Transduction:* Style transfer
- *Creation:* Knowledge-based dialog

- **Evaluation Criteria:** Sample-level Pearson and Spearman correlations with human judgments

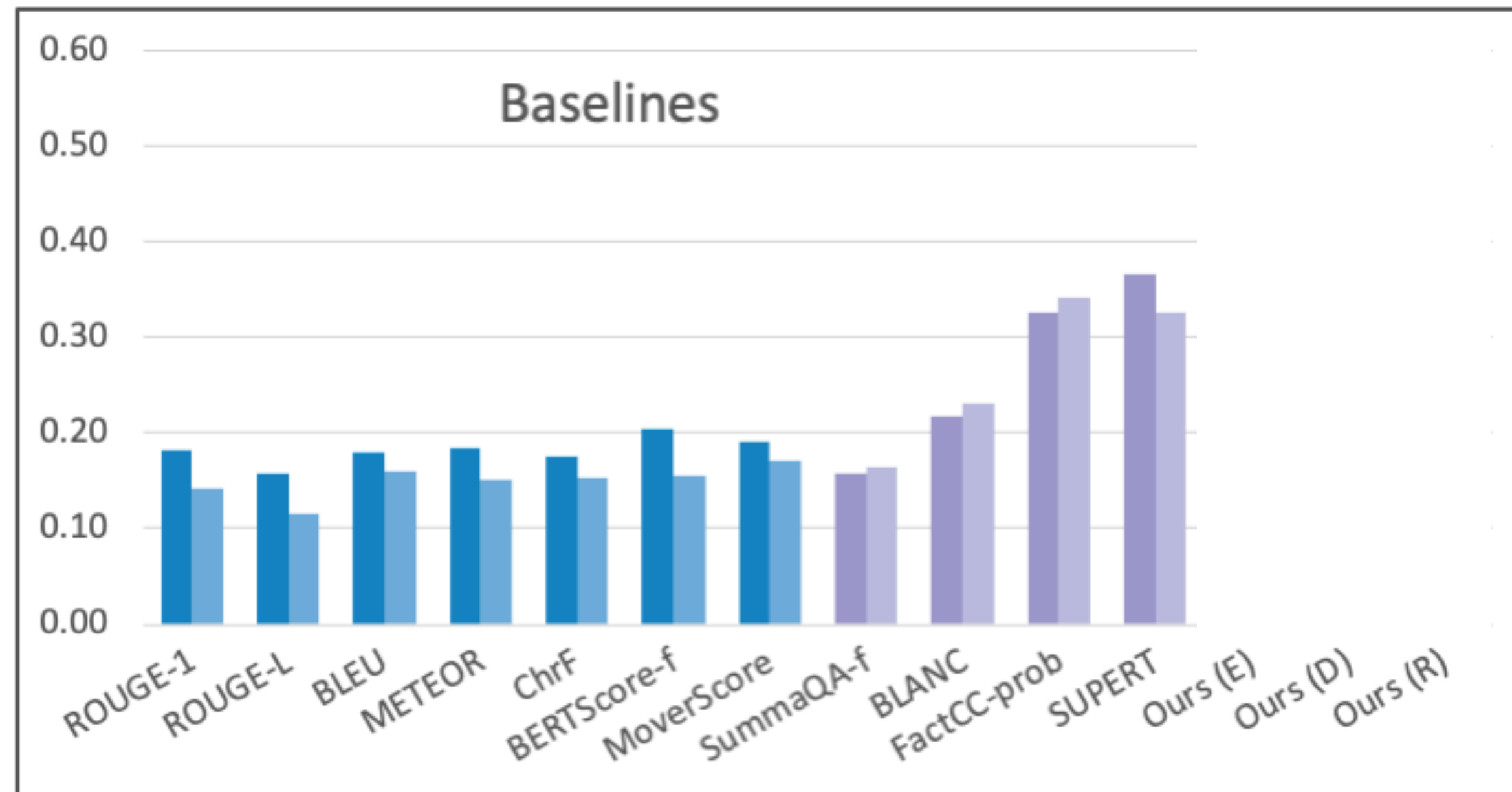
(More results available in paper appendix)

Compression metrics - consistency results

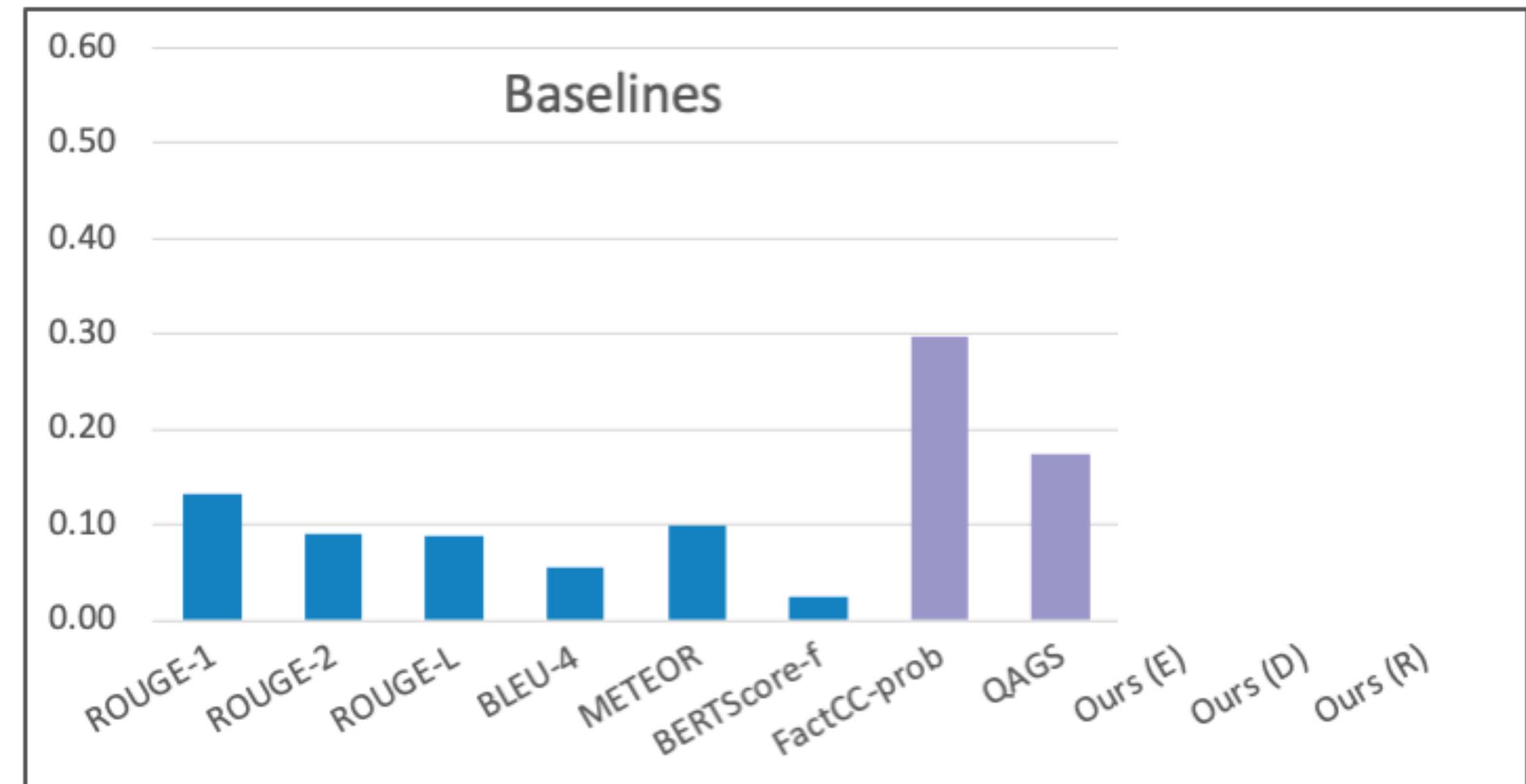
•**Dataset:** 1) SummEval on CNN/DM summarization dataset; 2) QAGS on XSUM

•**Results:**

Consistency (CNN/DM – SummEval)



Consistency (XSUM – QAGS)

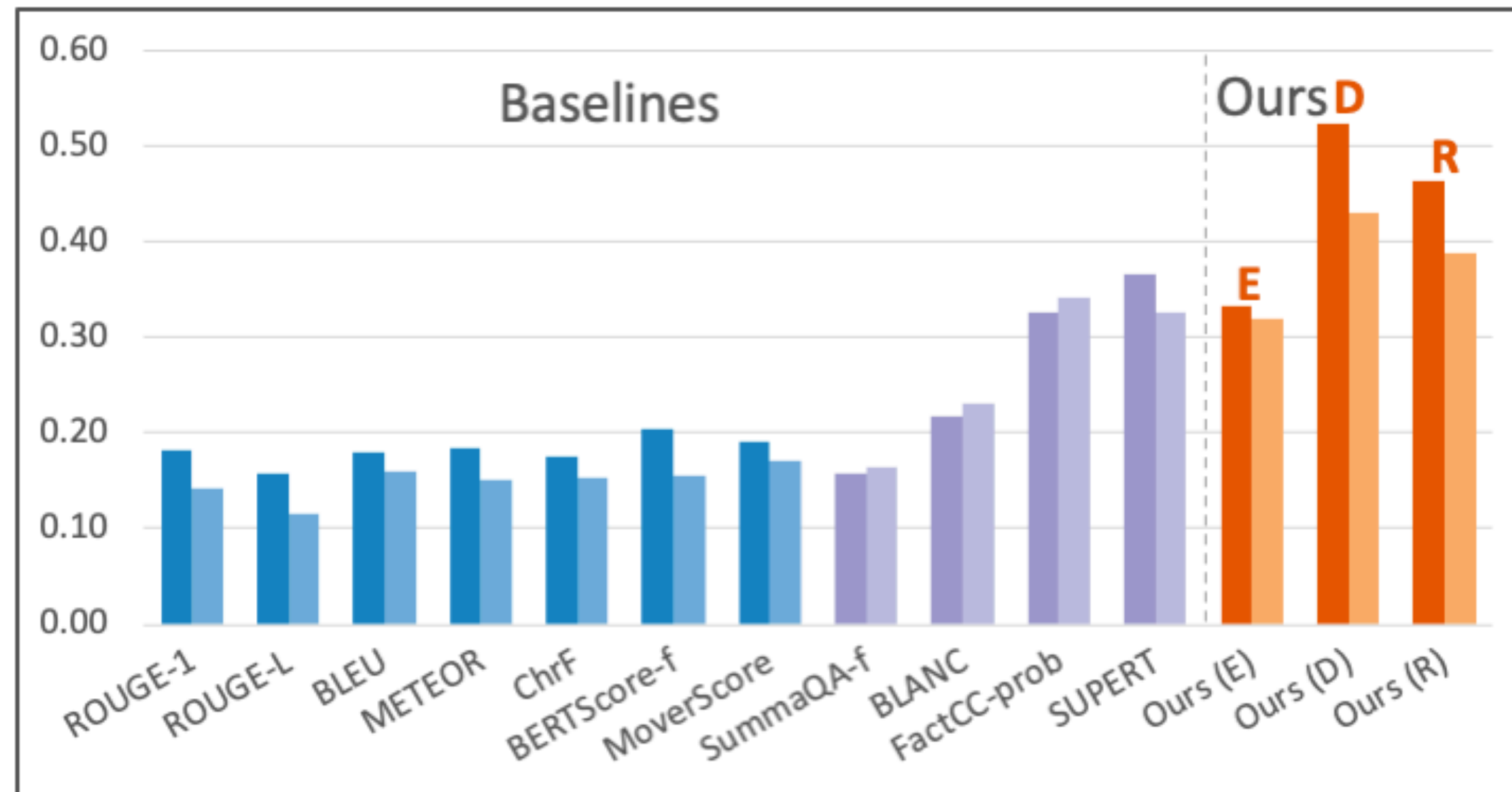


Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

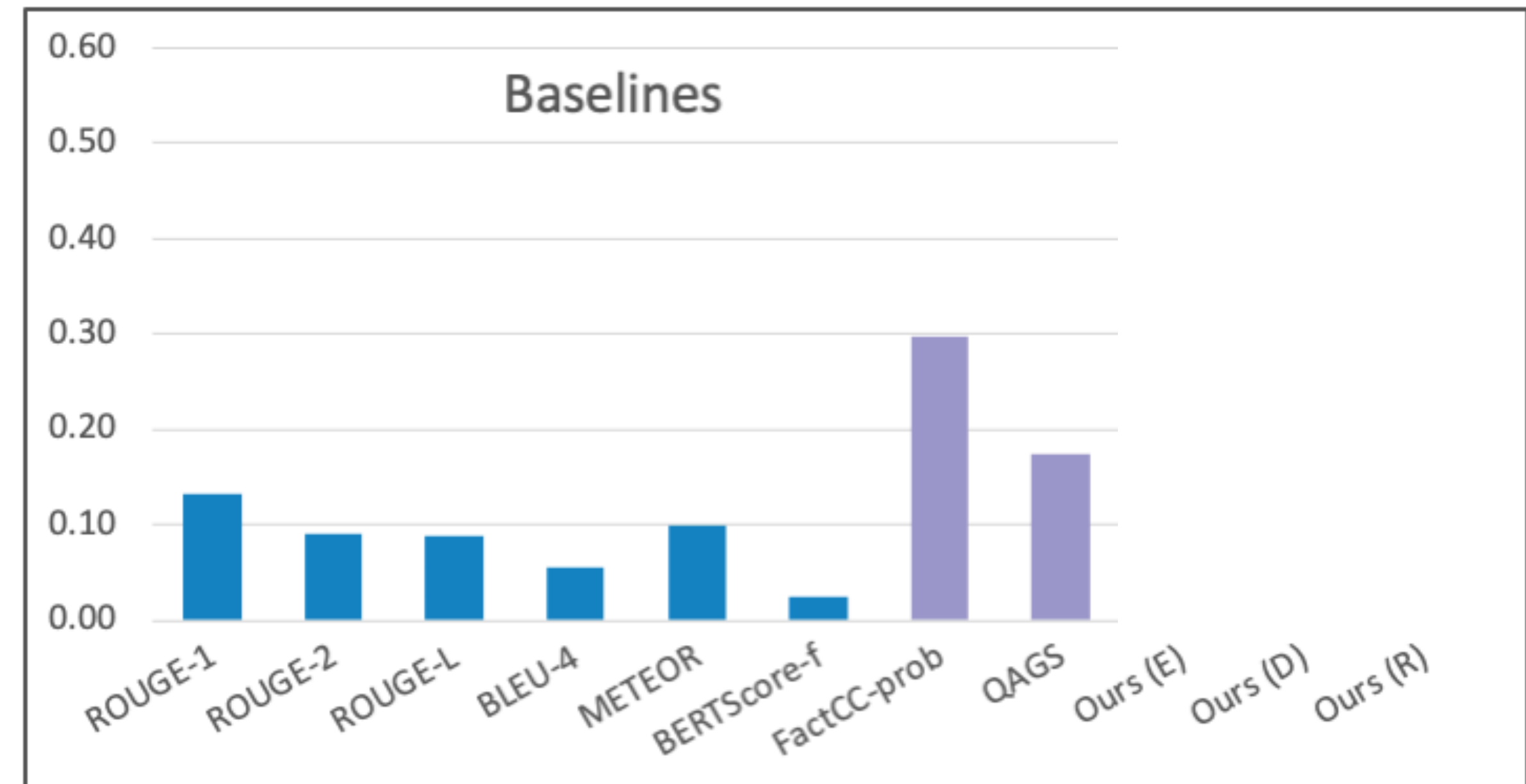
Compression metrics - consistency results

- **Dataset:** 1) SummEval on CNN/DM summarization dataset; 2) QAGS on XSUM
- **Results:** 1) On CNN/DM, our **D**- and **R**-based metrics clearly outperform baselines

Consistency (CNN/DM – SummEval)



Consistency (XSUM – QAGS)

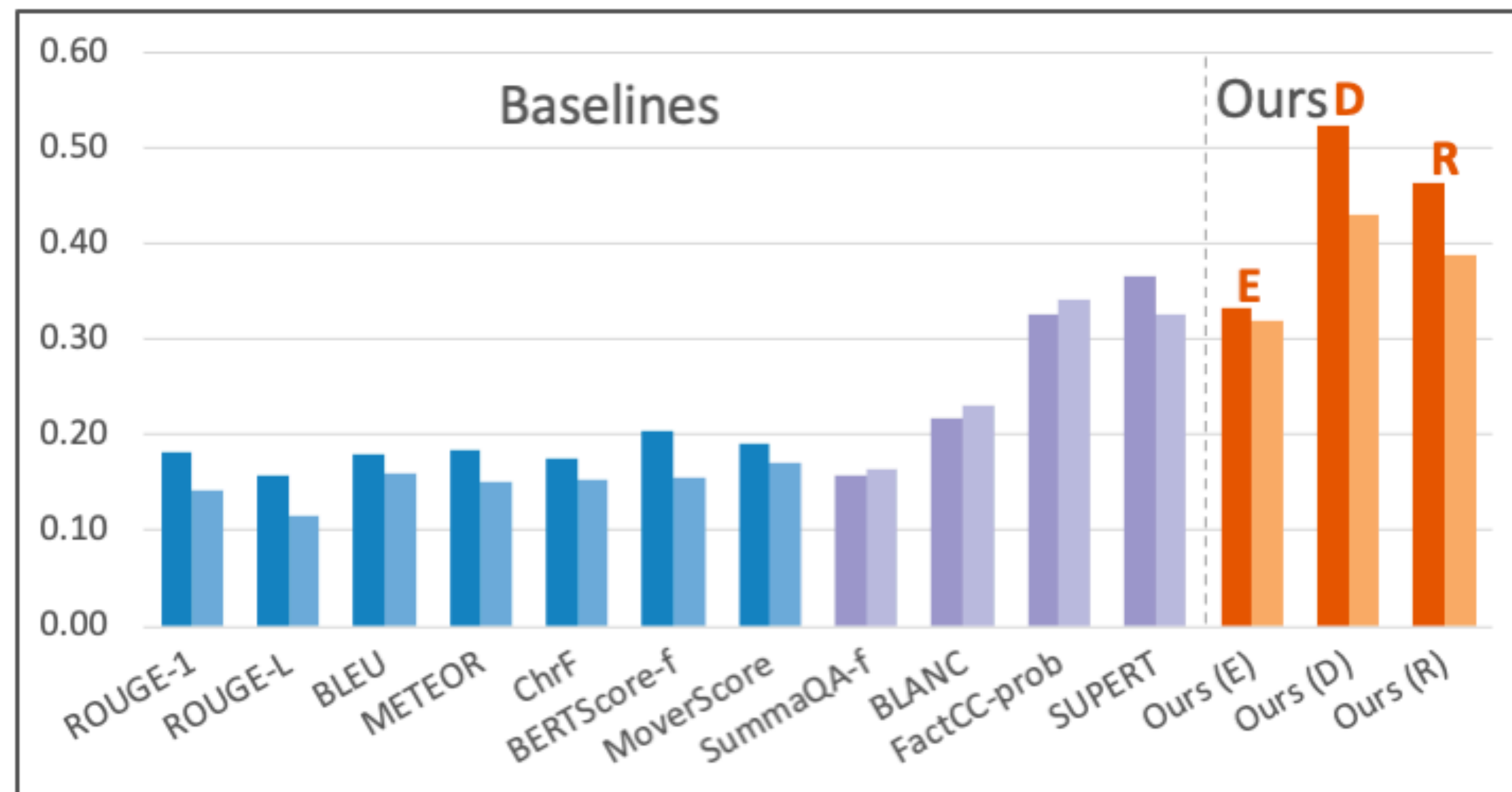


Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

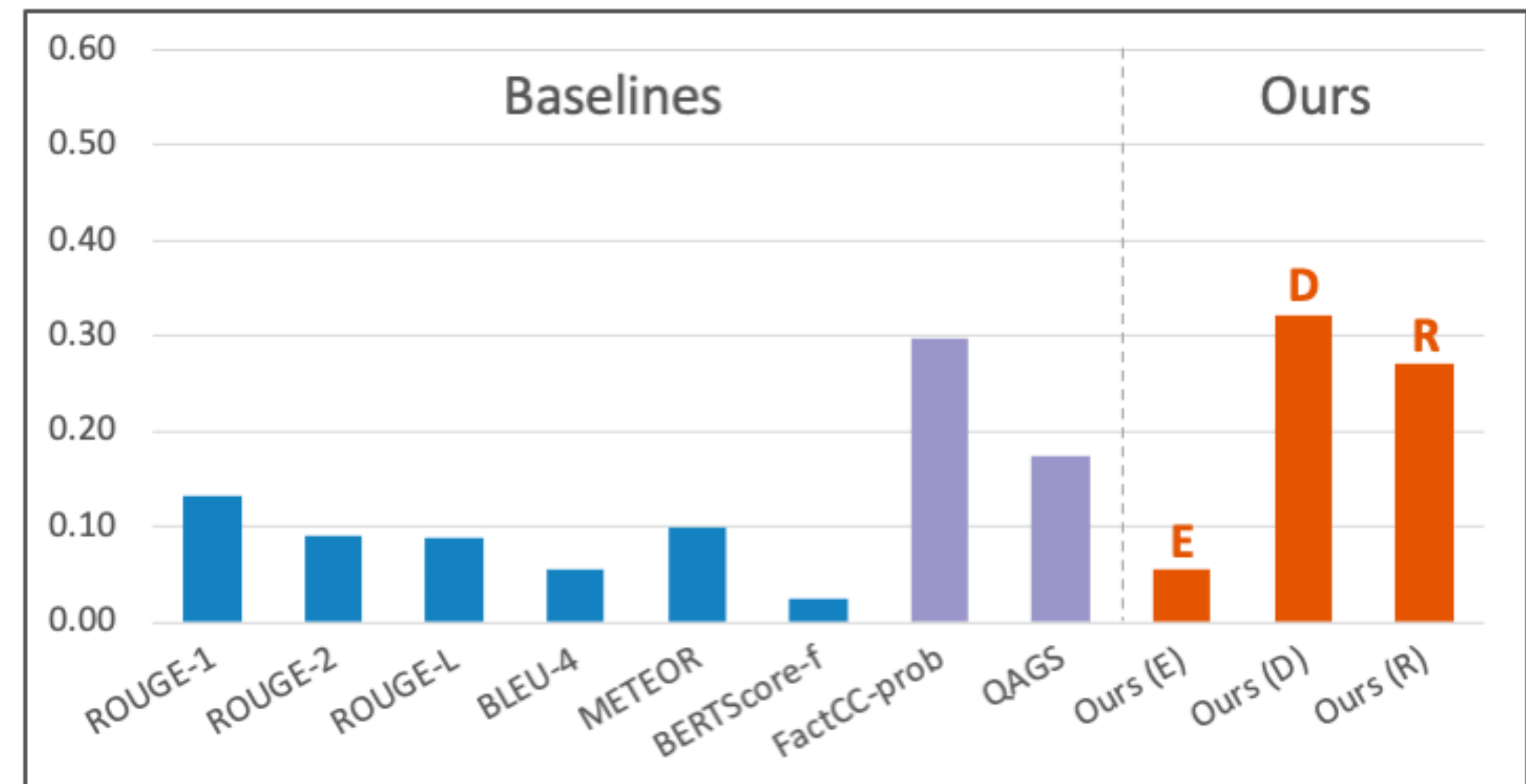
Compression metrics - consistency results

- **Dataset:** 1) SummEval on CNN/DM summarization dataset; 2) QAGS on XSUM
- **Results:** 1) On CNN/DM, our **D**- and **R**-based metrics clearly outperform baselines
2) On XSUM, our **D**-based metric also achieves the best performance

Consistency (CNN/DM – SummEval)



Consistency (XSUM – QAGS)



Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

Example: Word-level alignment prediction (1)



Article: **Darth Vader** and Imperial Stormtroopers have invaded a **Denbighshire seaside town** to welcome **the actor from Rhyl** who plays the infamous villain...

(Word)
(Score)

Summary: A **Welsh** actor who plays Darth Vader ... has been honored at **the London Film Festival**

0.94	0.79	0.98	1.00	0.99	0.98
0.99	0.97	0.91	0.89	0.83	0.56 0.47

0.56 **0.63** (Human Consistency Score: 0)

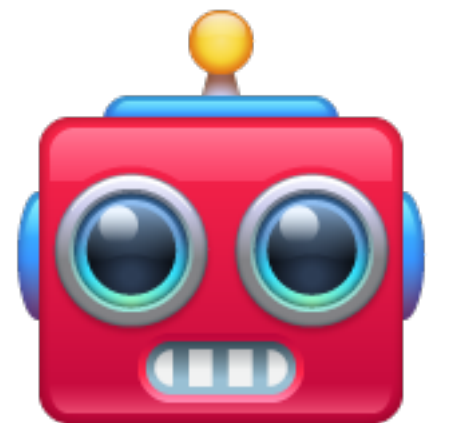


Example: *Word-level alignment prediction (2)*



Article: **Darth Vader** and Imperial Stormtroopers have invaded a **Denbighshire seaside town** to welcome **the actor from Rhyl** who plays the infamous villain..

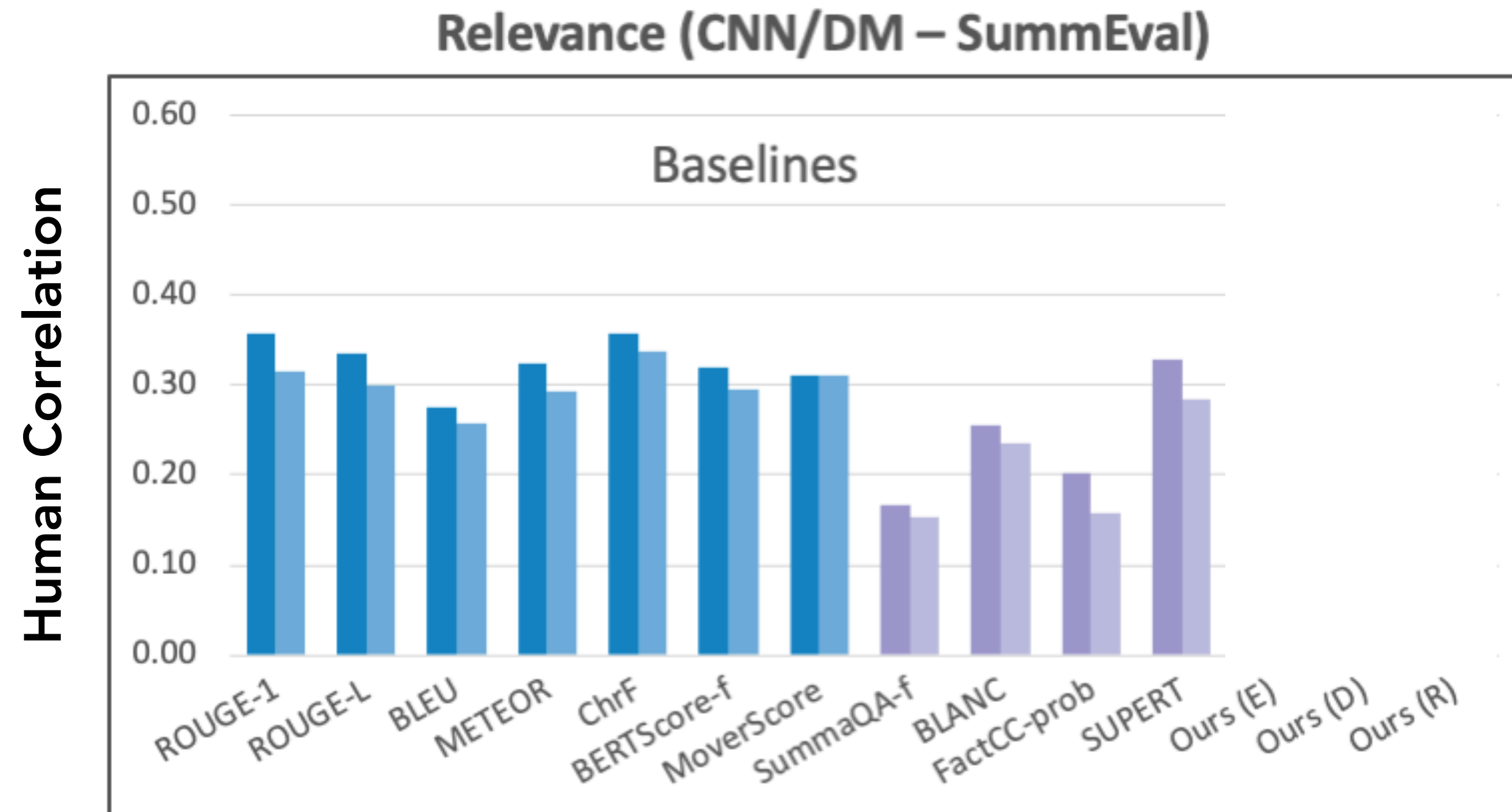
Gibberish: the the the the the the
0.83 0.61 0.56 0.53 0.49 0.48
the the the the the the the
0.50 0.53 0.57 0.58 0.57 0.56 0.57 0.55



Compression metrics - *relevance* results

- Dataset:** SummEval on CNN/DM summarization dataset

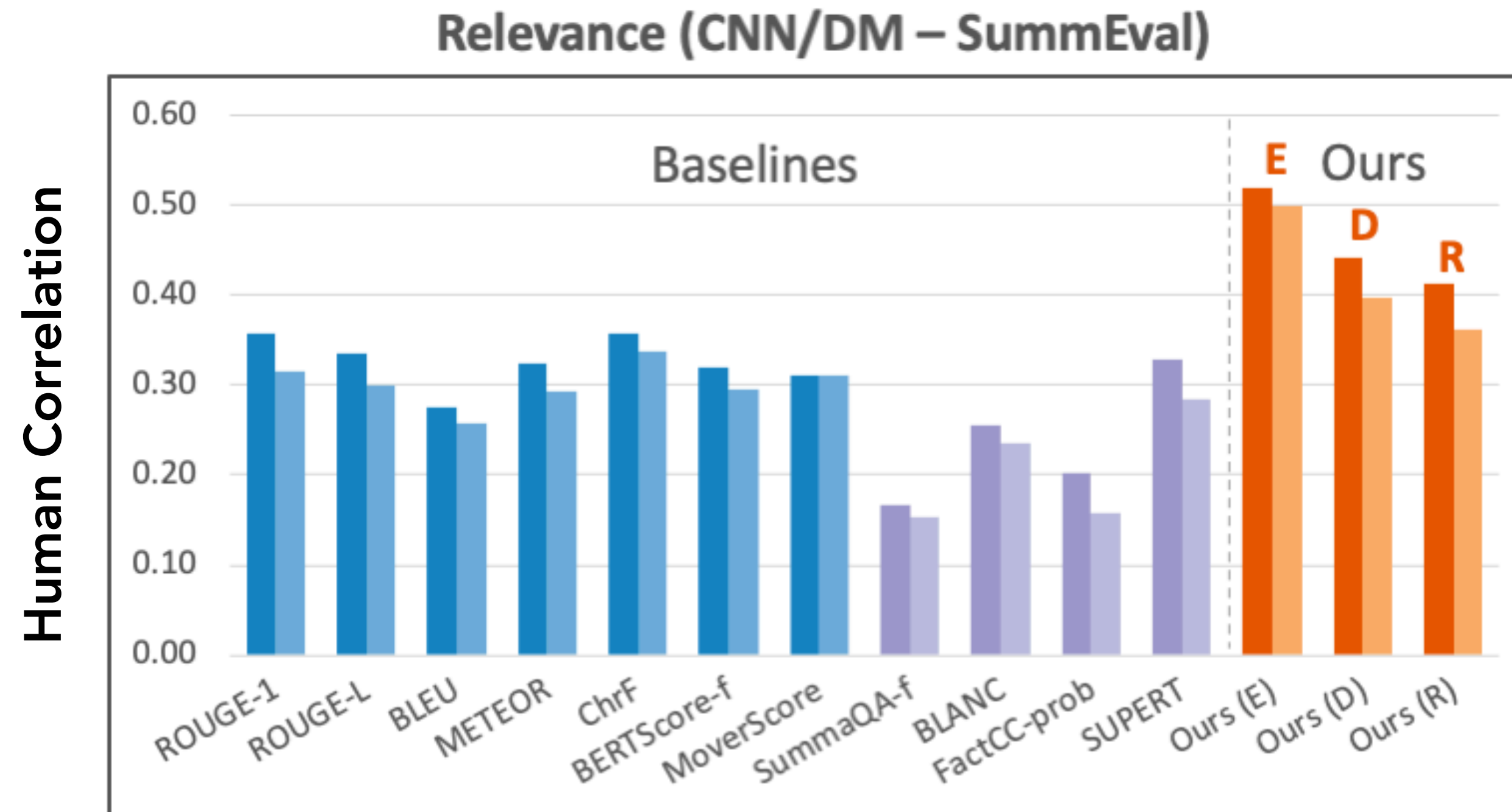
- Results:**



Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

Compression metrics - *relevance* results

- **Dataset:** SummEval on CNN/DM summarization dataset
- **Results:** 1) Our metrics strongly outperform all other baselines
2) **E**-based metric better than **D**- and **R**-based variants

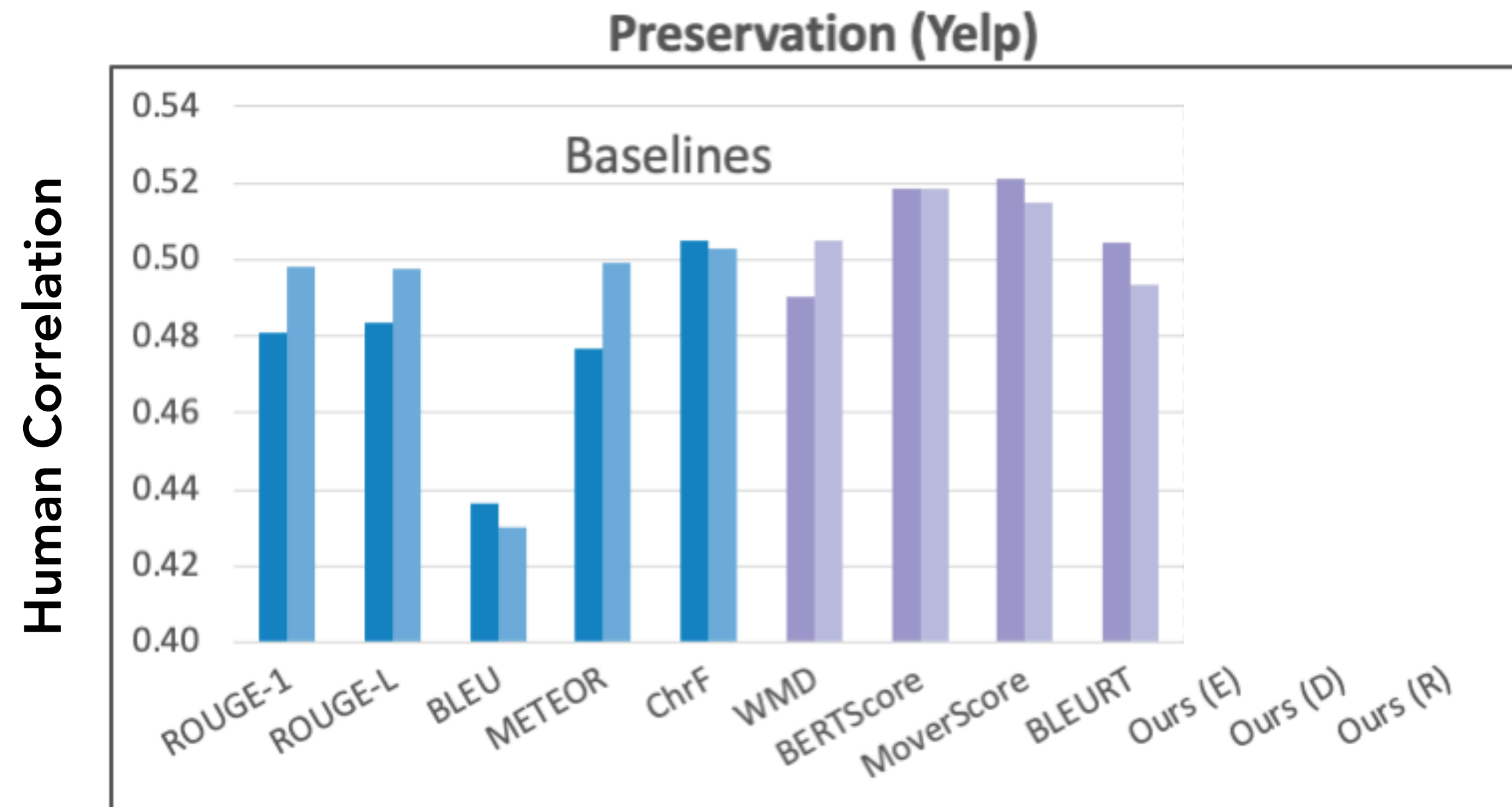


Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

Transduction metric - preservation results

- Dataset:** Mir et al. (2019) on Yelp style transfer dataset

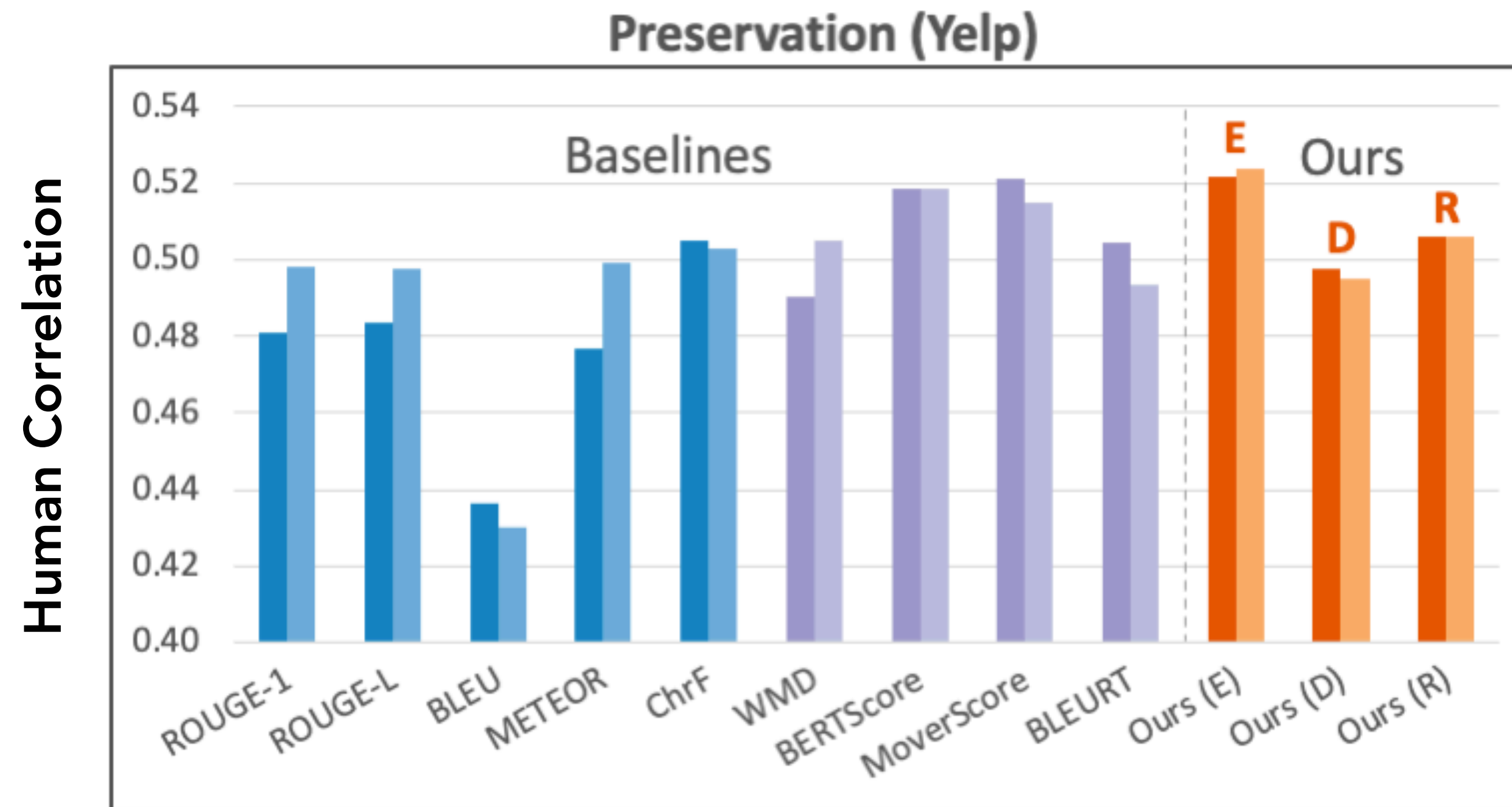
- Results:**



Lexical-matching metrics are in blue, embedding-/model-based metrics in purple and our metrics in red/orange

Transduction metric - preservation results

- **Dataset:** Mir et al. (2019) on Yelp style transfer dataset
- **Results:** Our **E**-based metric is competitive with or better than all previous metrics

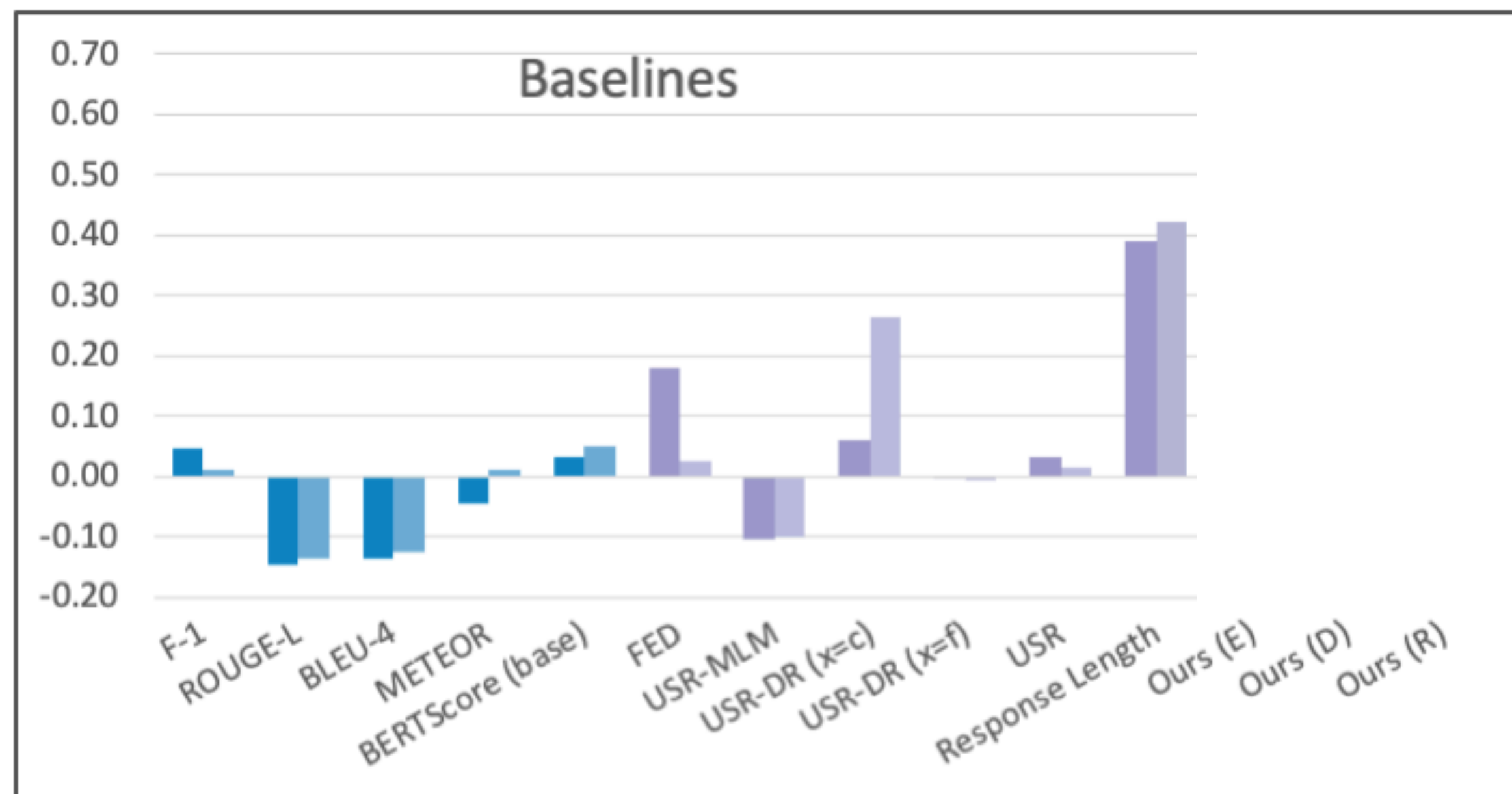


Lexical-matching metrics are in blue, embedding-/model-based metrics in purple and our metrics in red/orange

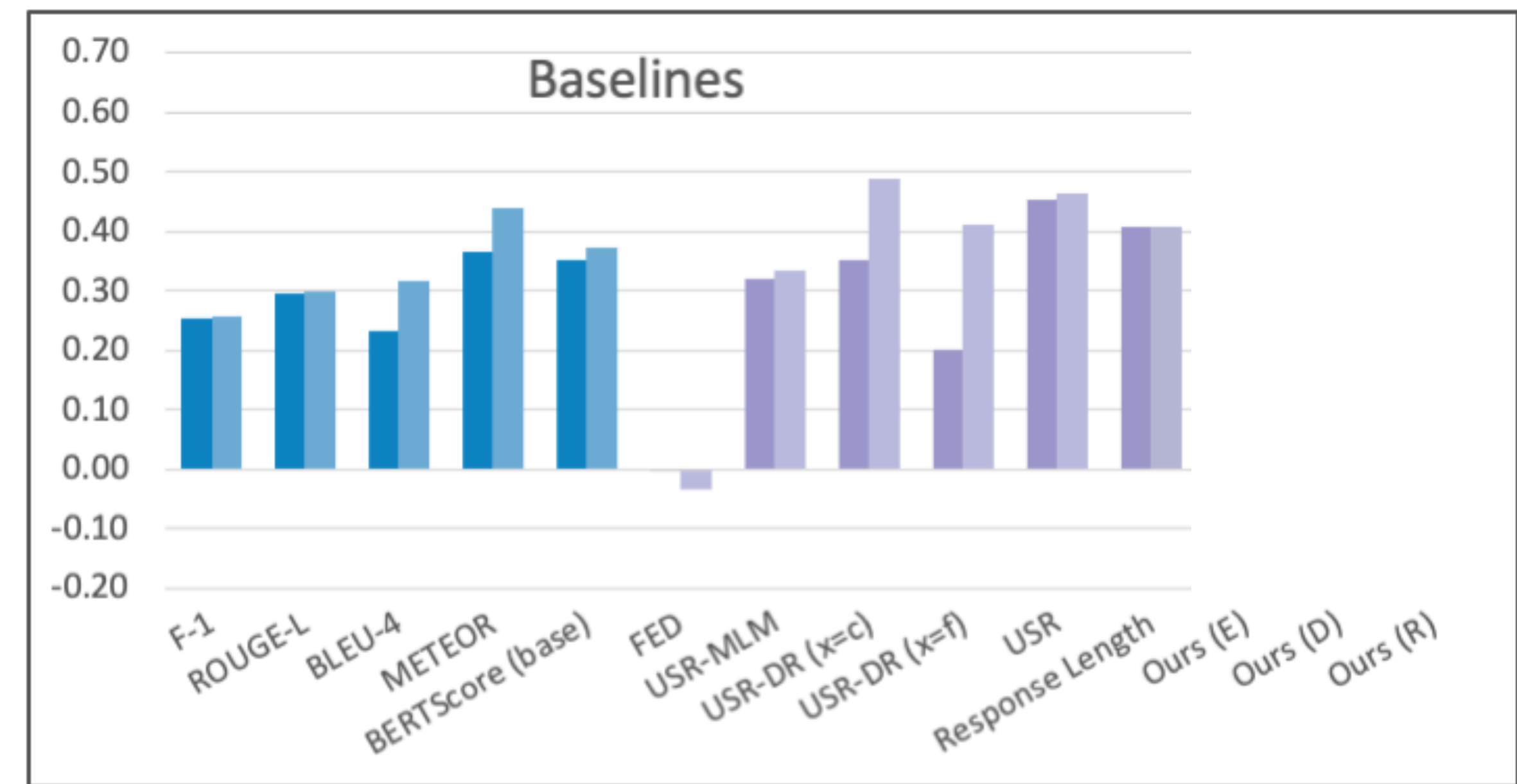
Creation metrics - engagingness results

- **Dataset:** Mehri and Eskenazi (2020) on 1) PersonaChat and 2) TopicalChat knowledge-based dialog datasets
- **Results:**

Engagingness (PersonaChat)



Engagingness (TopicalChat)

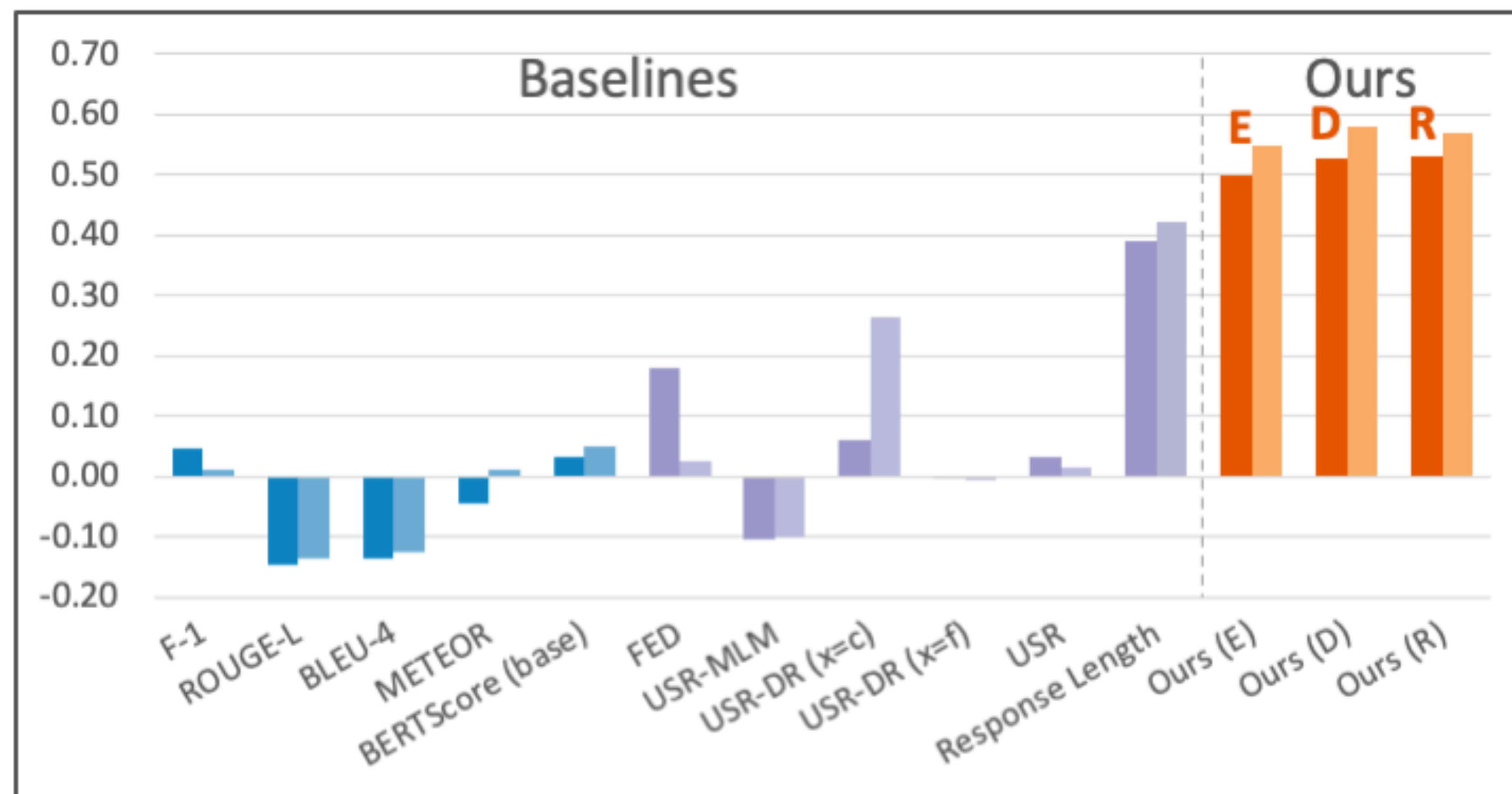


Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

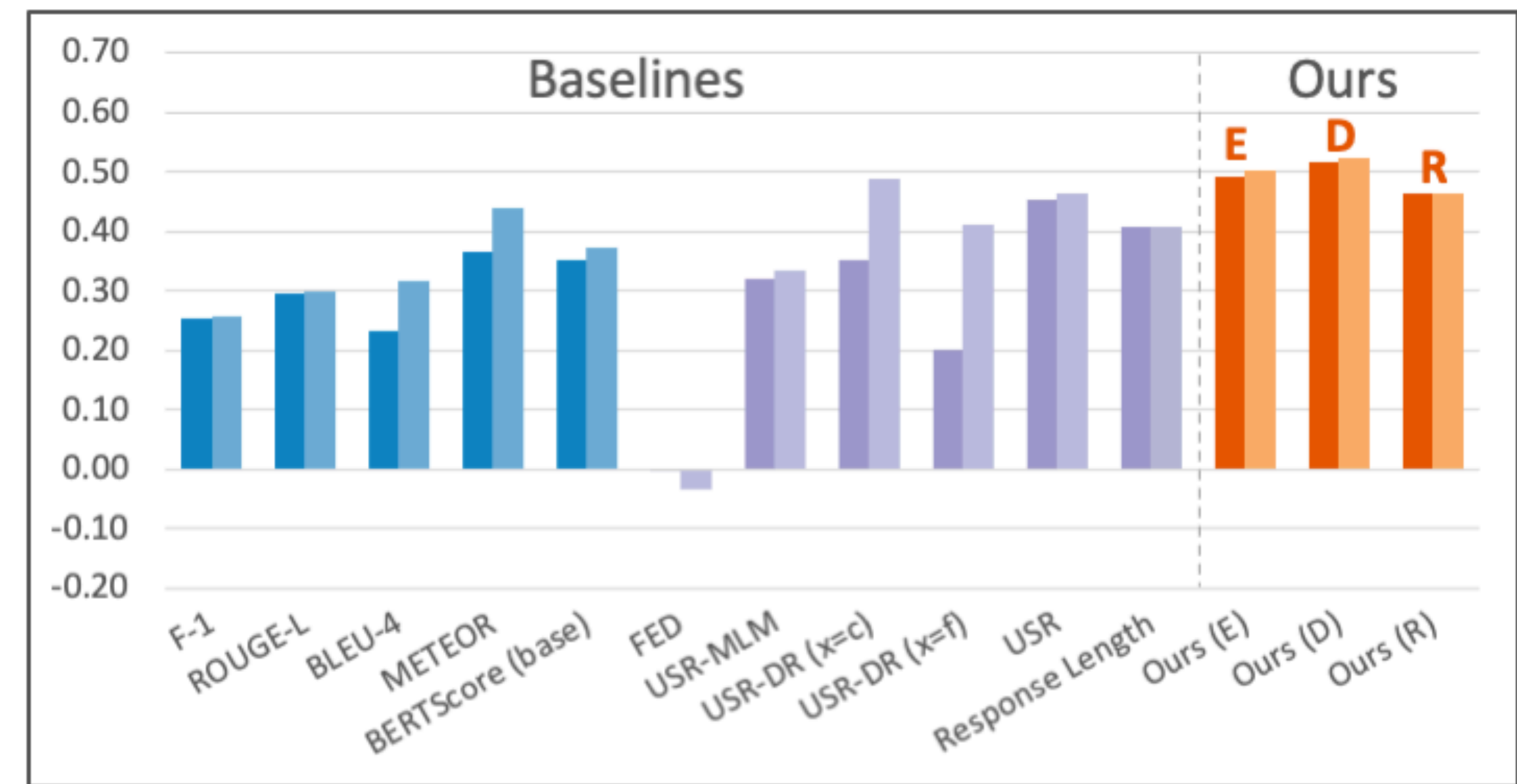
Creation metrics - engagingness results

- **Dataset:** Mehri and Eskenazi (2020) on 1) PersonaChat and 2) TopicalChat knowledge-based dialog datasets
- **Results:** Our metrics all improve over previous methods by large margins on the two datasets

Engagingness (PersonaChat)



Engagingness (TopicalChat)



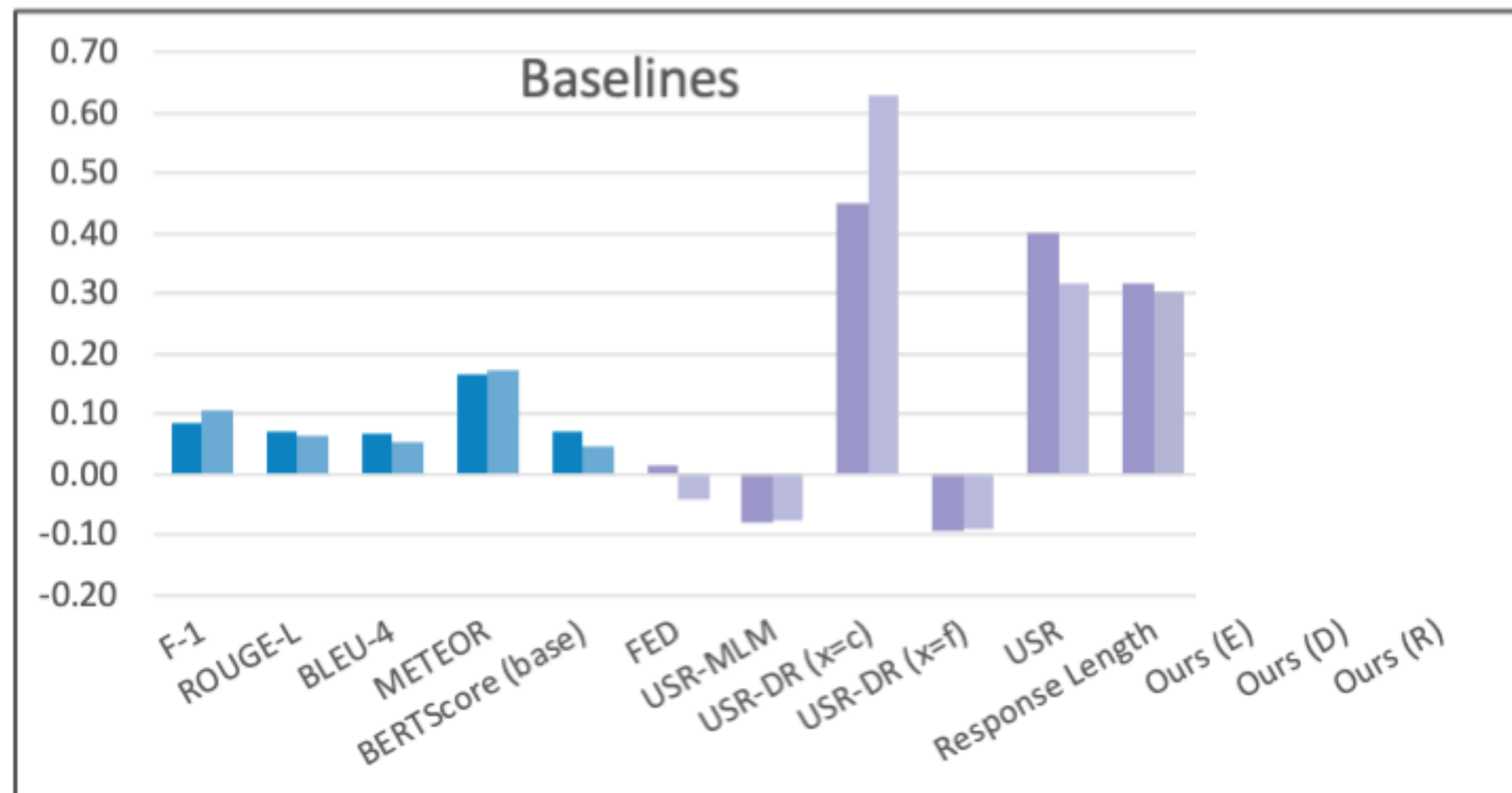
Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

Creation metrics - groundedness results

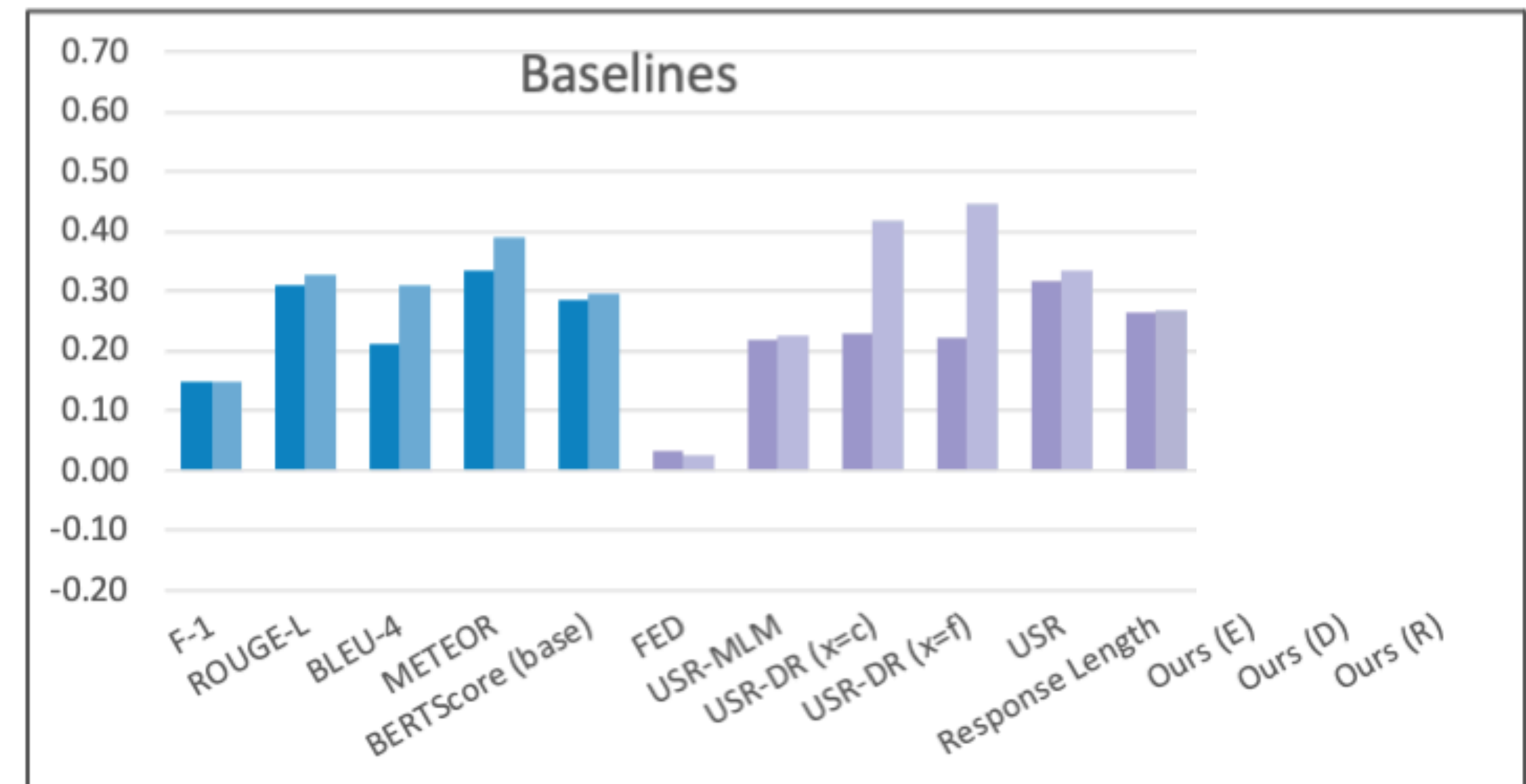
• **Dataset:** Mehri and Eskenazi (2020) on 1) PersonaChat and 2) TopicalChat knowledge-based dialog datasets

• **Results:**

Groundedness (PersonaChat)



Groundedness (TopicalChat)

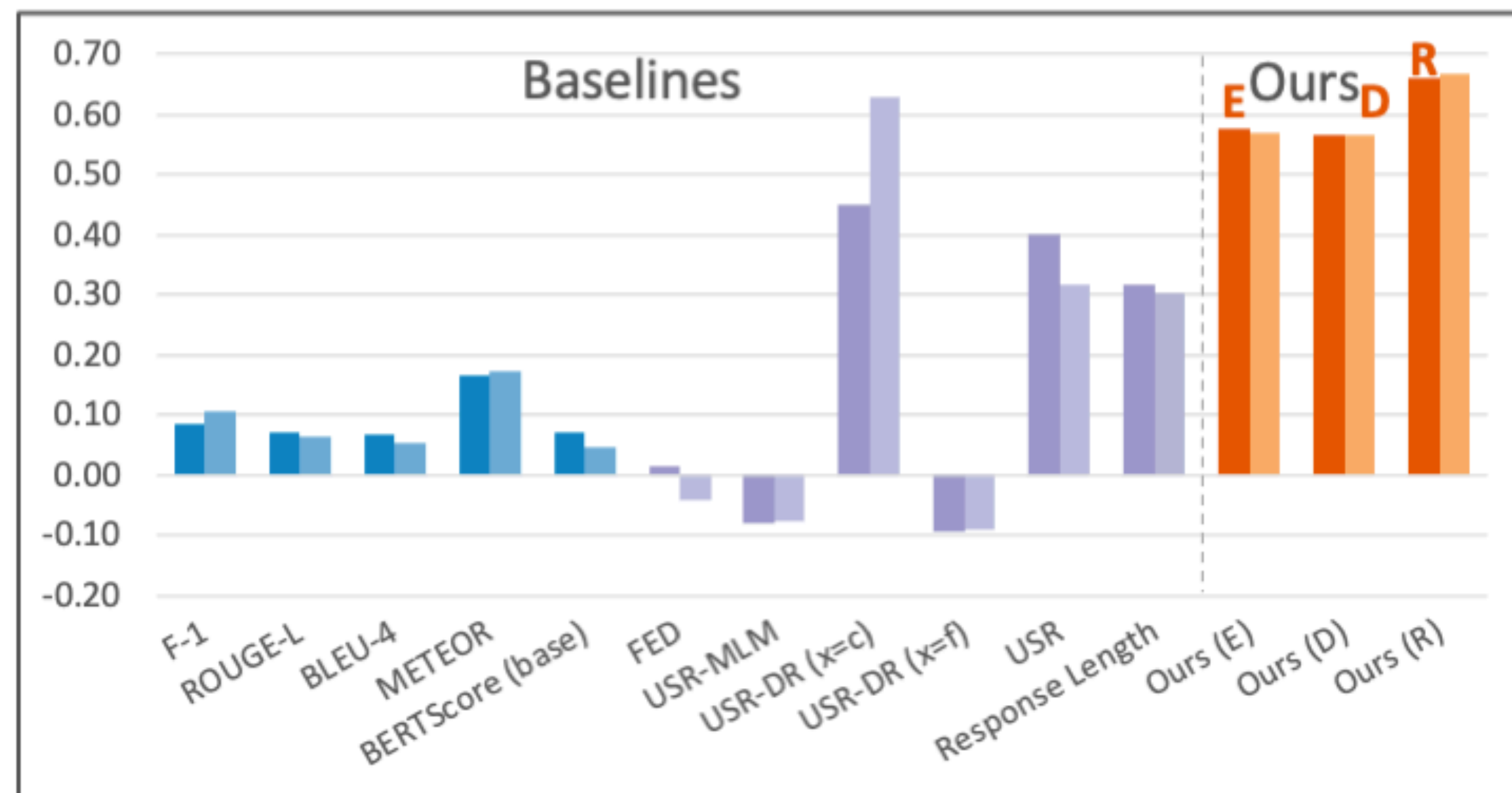


Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

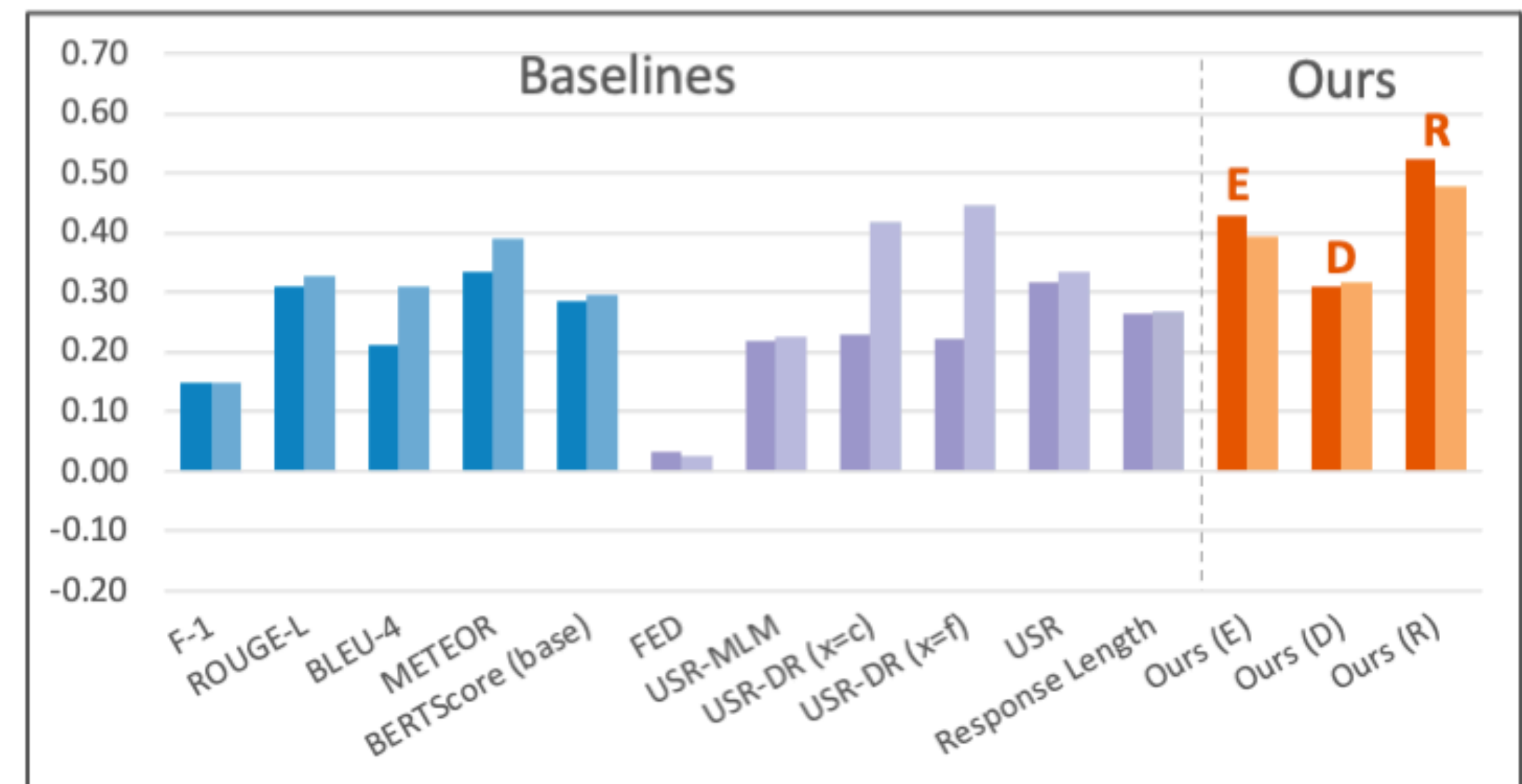
Creation metrics - groundedness results

- **Dataset:** Mehri and Eskenazi (2020) on 1) PersonaChat and 2) TopicalChat knowledge-based dialog datasets
- **Results:** 1) Our metrics again achieves strong correlations
2) Our **R**-based metric outperforms other implementations (**E** and **D**)

Groundedness (PersonaChat)

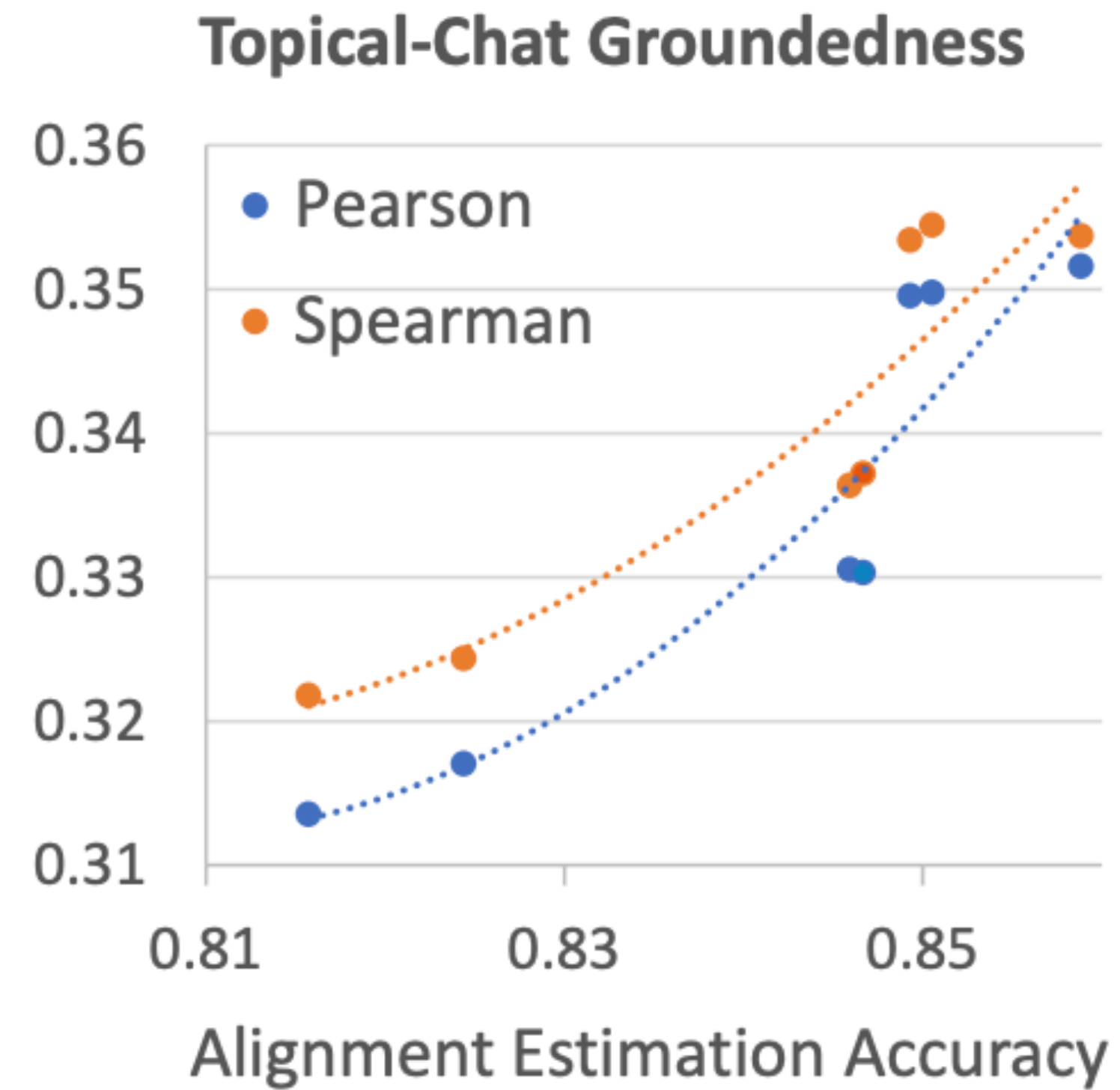
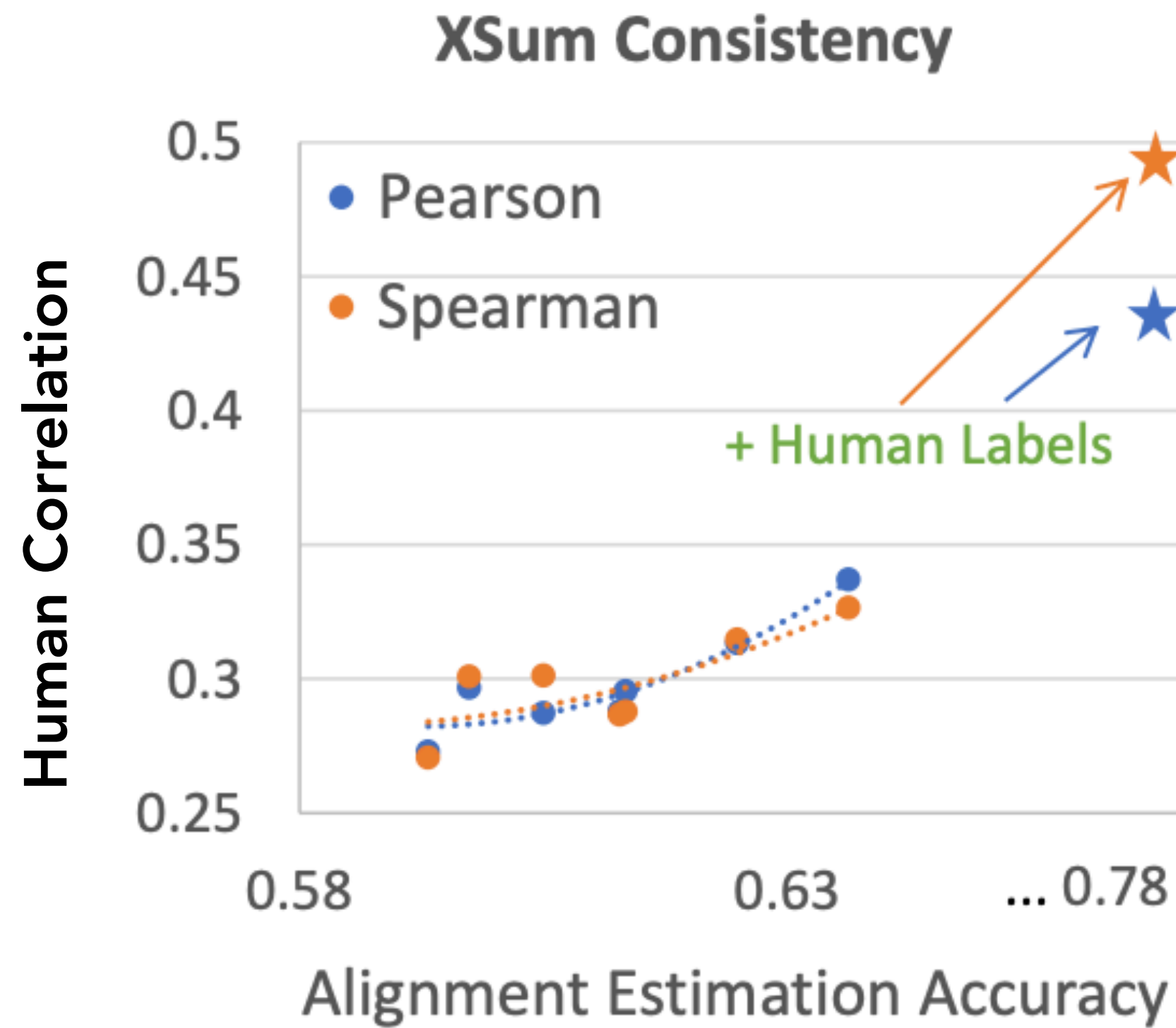


Groundedness (TopicalChat)

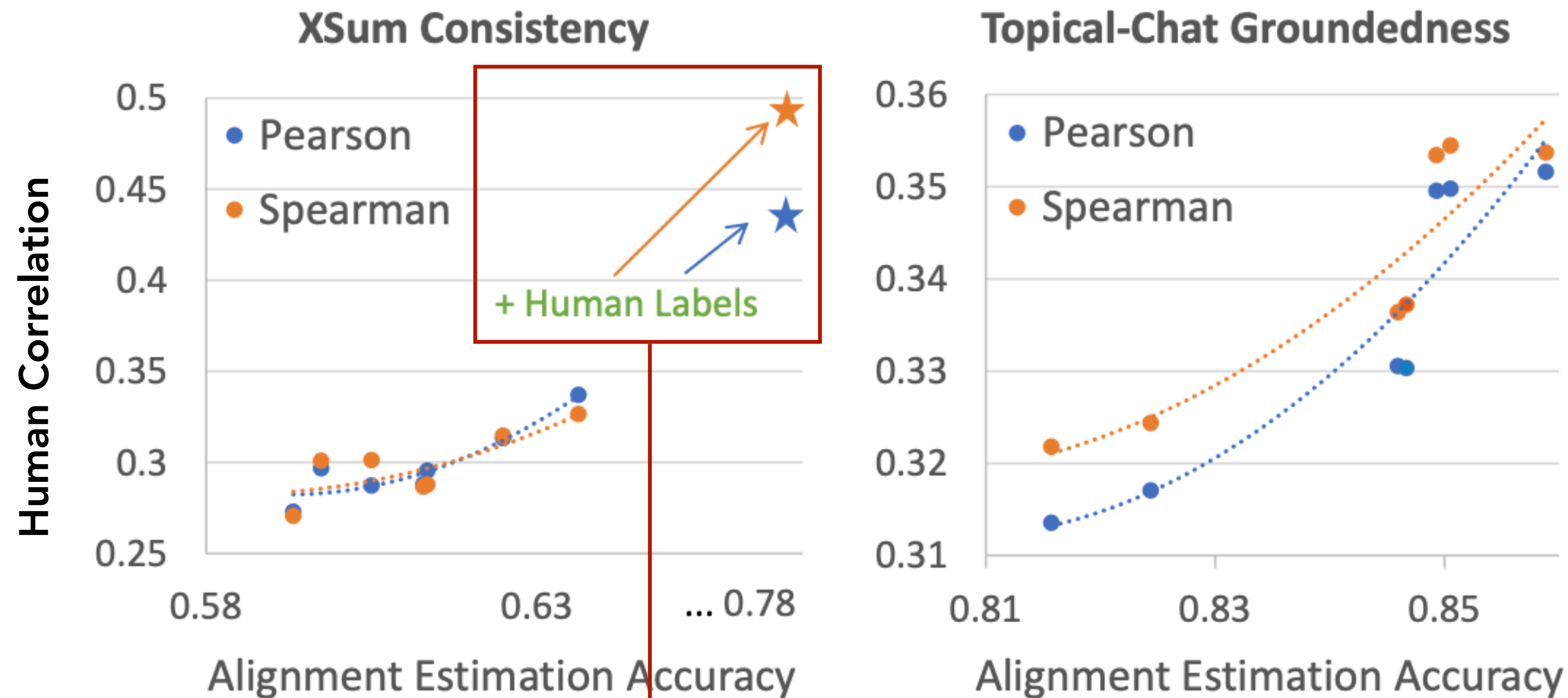


Reference-based metrics are in blue, reference-free metrics in purple and our metrics in red/orange

Higher alignment estimation accuracy, better correlation



Higher alignment estimation accuracy, better correlation



- Fine-tuning with token-level human labels further increases *both* alignment accuracy *and* human correlations

Higher alignment estimation accuracy, better correlation

XSum Consistency

Topical-Chat Groundedness

- Improvement in a single alignment model can immediately benefit a wide range of metrics
- Alignment modeling becomes a separate prediction task directly tied to the quality of evaluation metrics

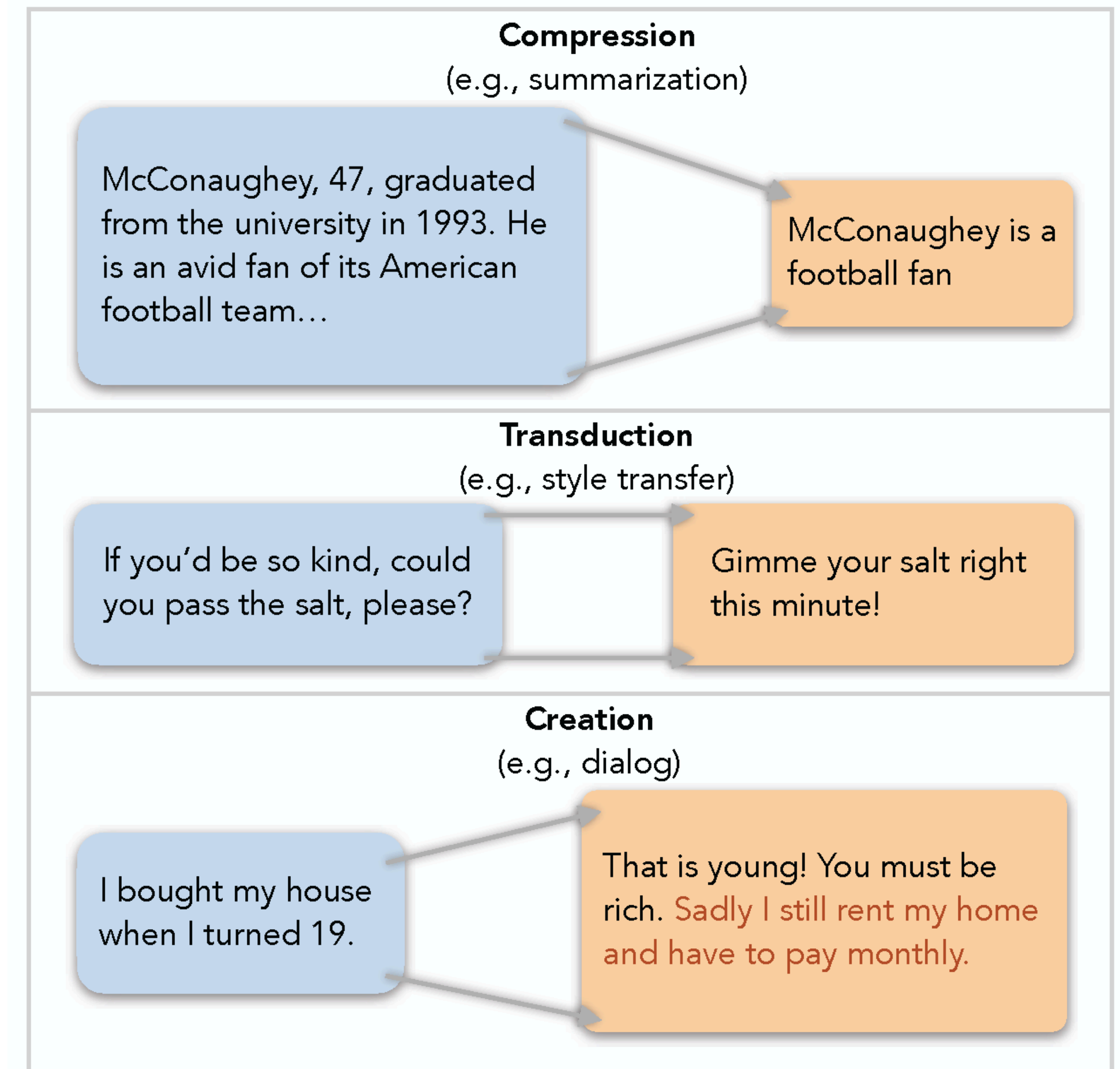
Alignment Estimation Accuracy

Alignment Estimation Accuracy

- Fine-tuning with token-level human labels further increases *both* alignment accuracy *and* human correlations

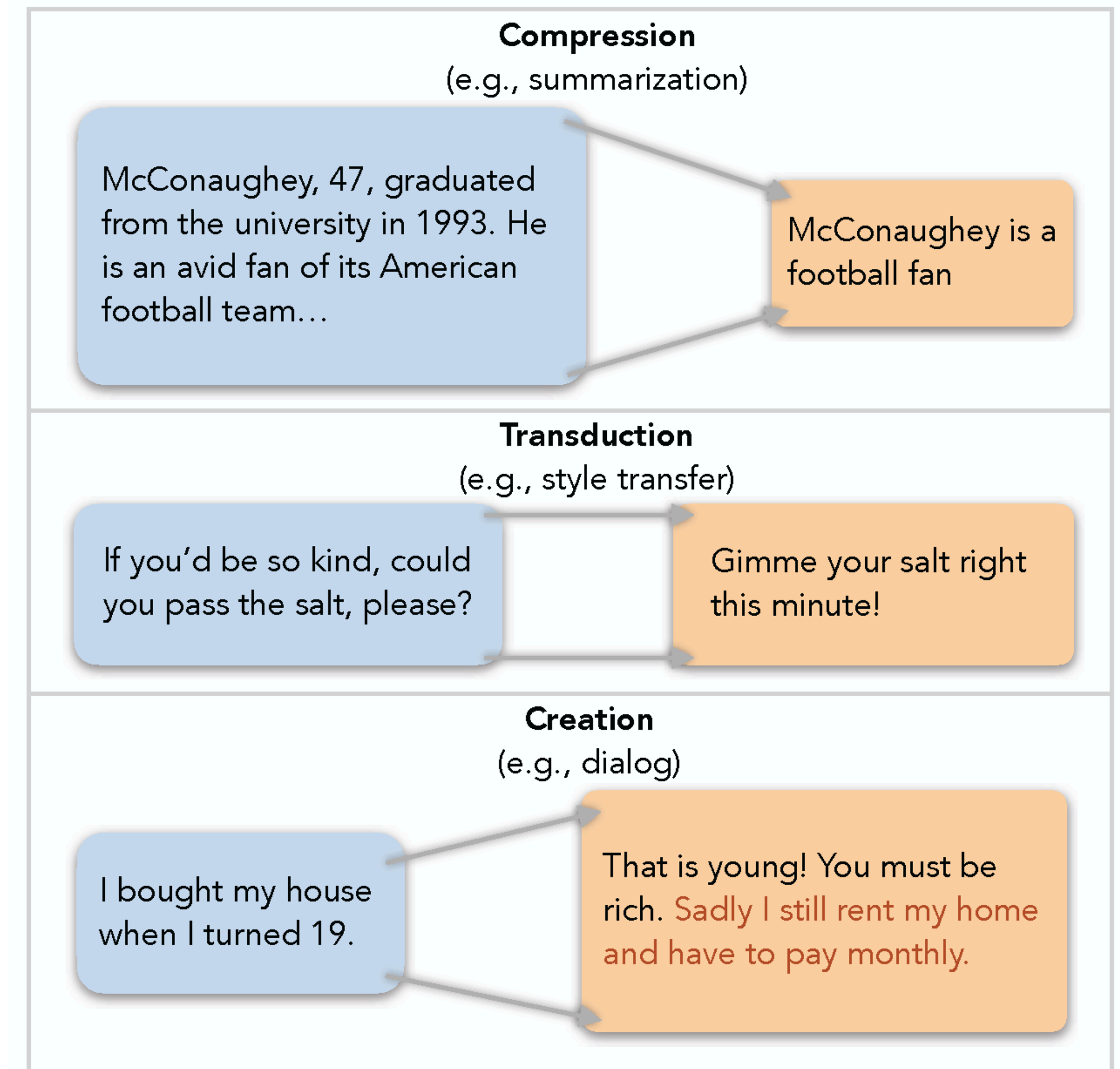
Summary so far

- A general evaluation framework for NLG tasks



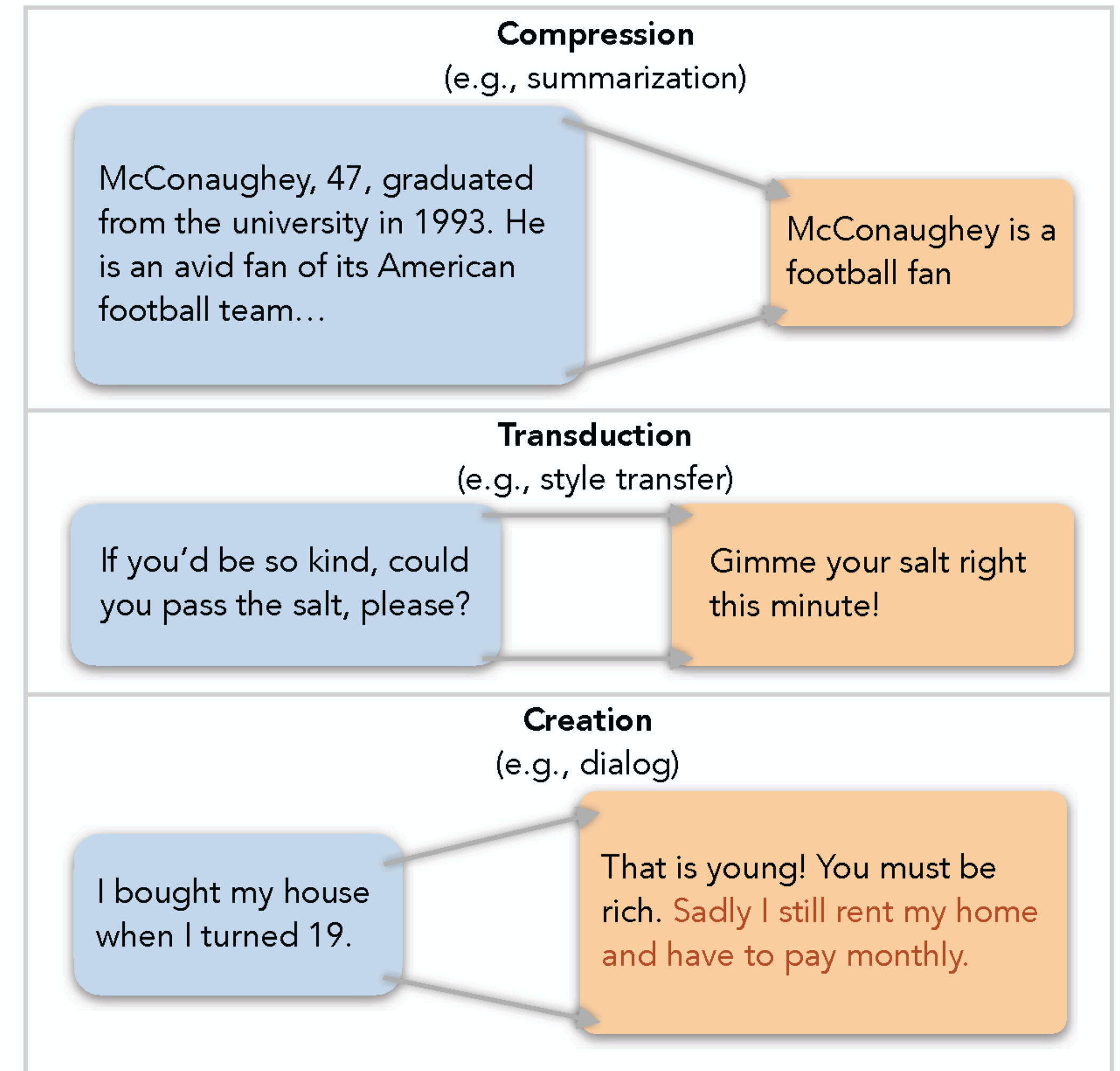
Summary so far

- A general evaluation framework for NLG tasks
- Unified evaluation of all types of tasks in terms of info. alignment



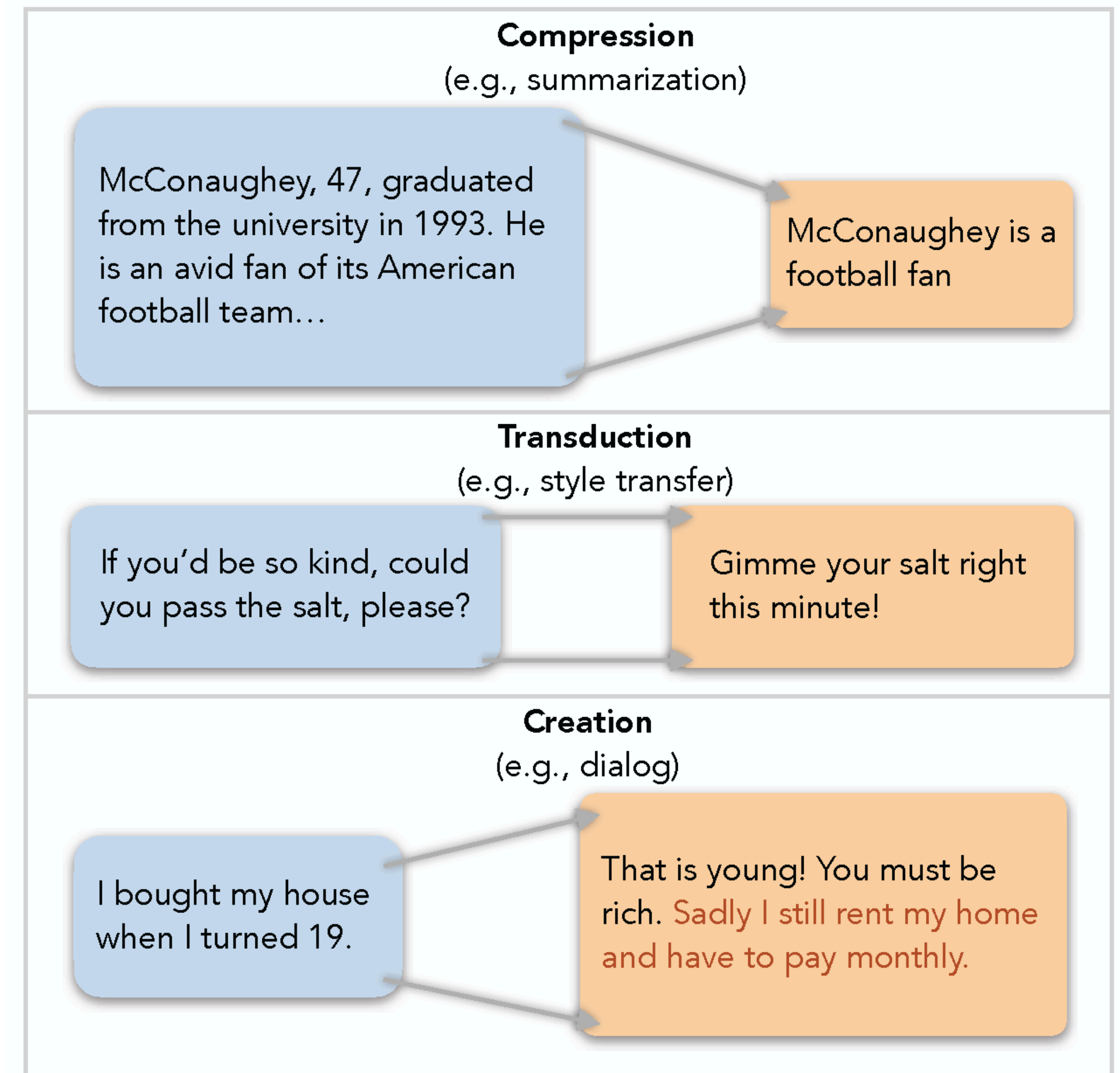
Summary so far

- A general evaluation framework for NLG tasks
- Unified evaluation of all types of tasks in terms of info. alignment
- Empirically, our uniformly-designed metrics outperform previous specially-designed metrics



Summary so far

- A general evaluation framework for NLG tasks
- Unified evaluation of all types of tasks in terms of info. alignment
- Empirically, our uniformly-designed metrics outperform previous specially-designed metrics
- Improving one alignment estimation model benefits a wide range of metrics in framework



Text Generation Basics

- Model
- Learning
- Inference (Decoding)
- Evaluation

Two Central Goals

- Generating human-like, grammatical, and readable text
 - I.e., generating **natural** language
- Generating text that contains desired information inferred from inputs
 - Machine translation
 - Source sentence --> target sentence w/ the same meaning
 - Data description
 - Table --> data report describing the table
 - Attribute control
 - Sentiment: positive --> "I like this restaurant"
 - Conversation control
 - Control conversation strategy and topic

Two Central Goals

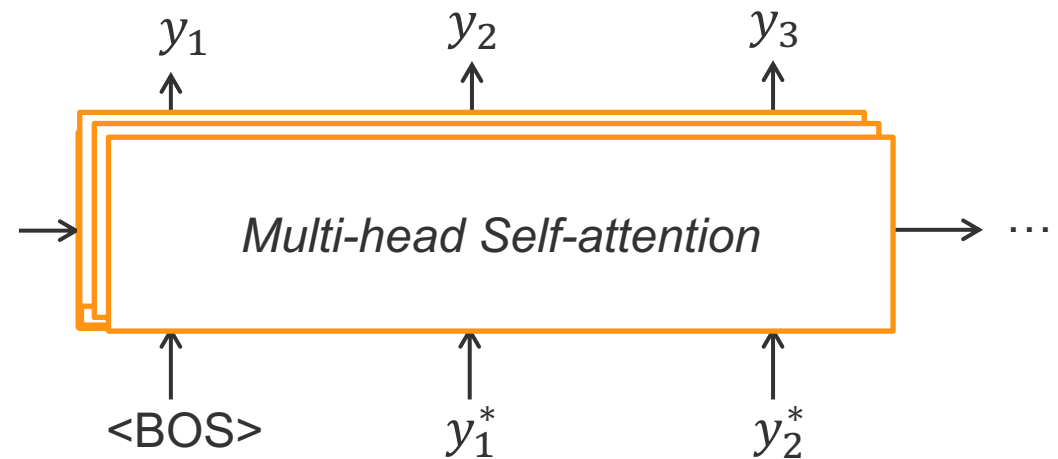
- Generating human-like, grammatical, and readable text
 - I.e., generating **natural** language
- Generating text that contains desired information inferred from inputs
 - Machine translation
 - Source sentence --> target sentence w/ the same meaning
 - Data description
 - Table --> data report describing the table
 - Attribute control
 - Sentiment: positive --> "I like this restaurant"
 - Conversation control
 - Control conversation strategy and topic

Common Learning Algorithm: Maximum Likelihood Estimation (MLE)

- Training
 - Maximize data log-likelihood
 - Given ground-truth data

$$\mathbf{y}^* = (y_1^*, y_2^* \dots, y_{T^*}^*)$$

$$\mathcal{L}_{\text{MLE}}(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{y}^* | \mathbf{x}) = \log \prod_t p_{\boldsymbol{\theta}}(y_t^* | \mathbf{y}_{1:t-1}^*, \mathbf{x})$$



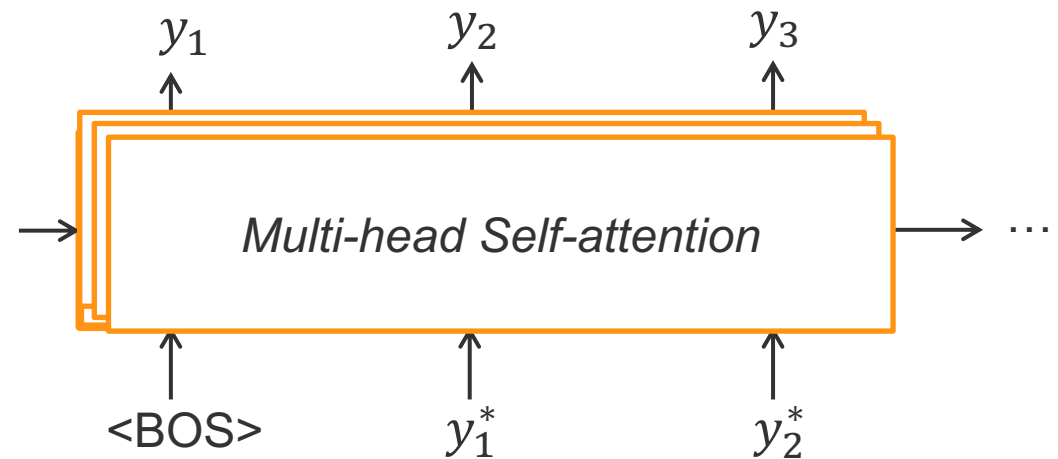
Common Learning Algorithm: Maximum Likelihood Estimation (MLE)

- Training
 - Maximize data log-likelihood
 - Given ground-truth data

$$\mathbf{y}^* = (y_1^*, y_2^*, \dots, y_{T^*}^*)$$

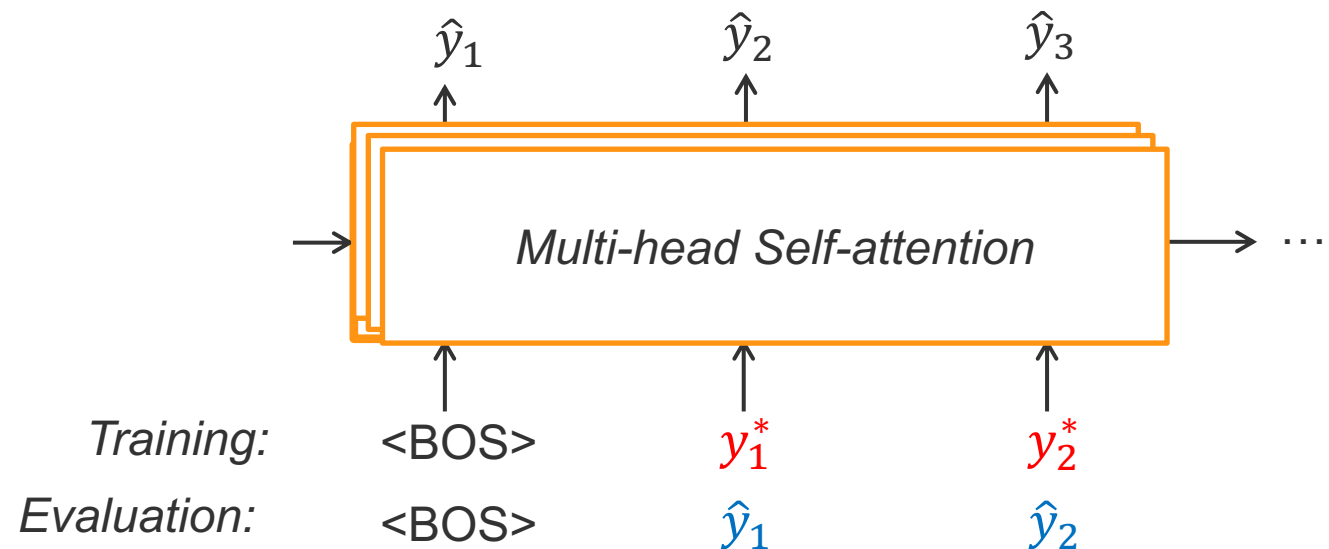
$$\mathcal{L}_{\text{MLE}}(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{y}^* | \mathbf{x}) = \log \prod_t p_{\boldsymbol{\theta}}(y_t^* | \mathbf{y}_{1:t-1}^*, \mathbf{x})$$

- Evaluation
 - Task-specific metrics
 - BLEU for machine translation
 - ROUGE for summarization
 -



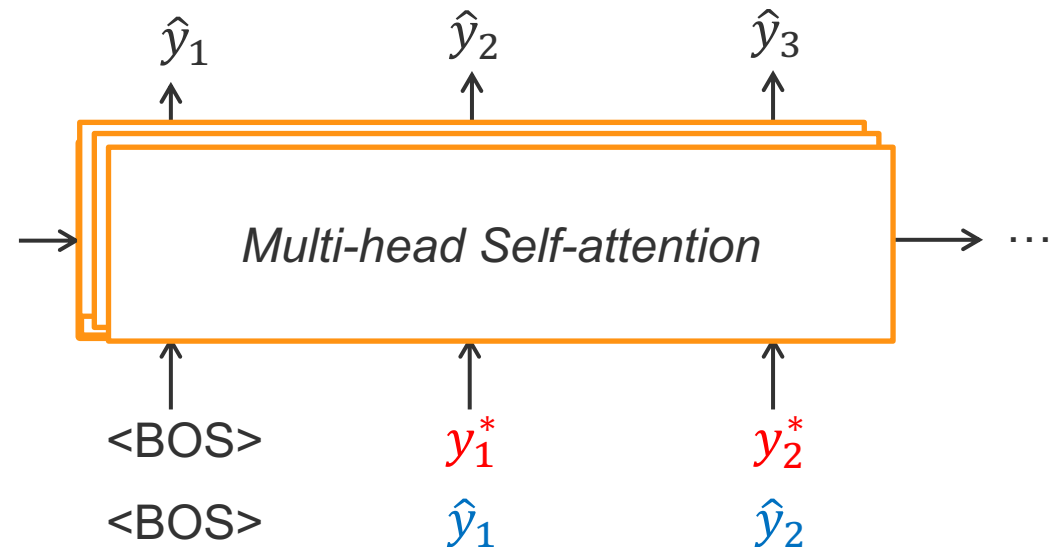
Two Issues of MLE

- Exposure bias [Ranzato et al., 2015]
 - **Training:** predict next token given the previous **ground-truth sequence**
 - **Evaluation:** predict next token given the previous **sequence that are generated by the model itself**



Two Issues of MLE

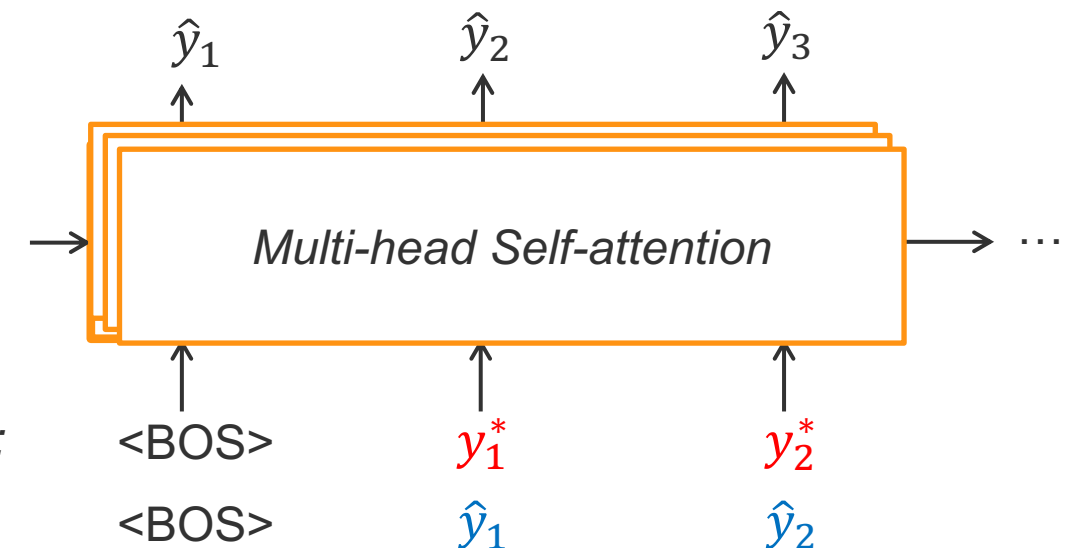
- Exposure bias [Ranzato et al., 2015]
 - **Training:** predict next token given the previous **ground-truth sequence**
 - **Evaluation:** predict next token given the previous **sequence that are generated by the model itself**
- Mismatch between training & evaluation criteria
 - Train to maximize **data log-likelihood**
 - Evaluate with, e.g., **BLEU**



Two Issues of MLE

Solution: Reinforcement learning for text generation (next lecture)

- Exposure bias [Ranzato et al., 2015]
 - **Training:** predict next token given the previous **ground-truth sequence**
 - **Evaluation:** predict next token given the previous **sequence that are generated by the model itself**
- Mismatch between training & evaluation criteria
 - Train to maximize **data log-likelihood**
 - Evaluate with, e.g., **BLEU**



Two Central Goals

- Generating human-like, grammatical, and readable text
 - Progressive generation
 - Exposure bias, criteria mismatch: reinforcement learning (next lecture)
- Generating text that contains desired information inferred from inputs
 - Machine translation
 - Source sentence --> target sentence w/ the same meaning
 - Data description
 - Table --> data report describing the table
 - Attribute control
 - Sentiment: positive --> "I like this restaurant"
 - Modify sentiment from positive to negative
 - Conversation control
 - Control conversation strategy and topic

Two Central Goals

- Generating human-like, grammatical, and readable text
 - Progressive generation
 - Exposure bias, criteria mismatch: reinforcement learning (next lecture)
- Generating text that contains desired information inferred from inputs
 - Machine translation
 - Source sentence --> target sentence w/ the same meaning
 - Data description
 - Table --> data report describing the table
 - Attribute control
 - Sentiment: positive --> "I like this restaurant"
 - Modify sentiment from positive to negative
 - Conversation control
 - Control conversation strategy and topic

Two Central Goals

- Generating human-like, grammatical, and readable text
 - Progressive generation
 - Exposure bias, criteria mismatch: reinforcement learning (next lecture)
- Generating text that contains desired information inferred from inputs #supervision data
 - Machine translation
 - Source sentence --> target sentence w/ the same meaning -----> 10s of millions
 - Data description
 - Table --> data report describing the table -----> 10s of 1000s
 - Attribute control
 - Sentiment: positive --> "I like this restaurant" -----> 10s of 1000s
 - Modify sentiment from positive to negative -----> 0
 - Conversation control
 - Control conversation strategy and topic -----> 0

Two Central Goals

Controlled generation in unsupervised settings

- Generating human-like, grammatical, and readable text
 - Progressive generation
 - Exposure bias, criteria mismatch: reinforcement learning (next lecture)
- Generating text that contains desired information inferred from inputs #supervision data
 - Machine translation
 - Source sentence --> target sentence w/ the same meaning -----> 10s of millions
 - Data description
 - Table --> data report describing the table -----> 10s of 1000s
 - Attribute control
 - Sentiment: positive --> "I like this restaurant" -----> 10s of 1000s
 - Modify sentiment from positive to negative -----> 0
 - Conversation control
 - Control conversation strategy and topic -----> 0

Unsupervised Controlled Generation of Text

- Sentence-level control
 - Text attribute transfer (style transfer) [Hu et al., 2017; Yang et al., 2018]
 - Text content manipulation [Wang, Hu et al., 2019]
- Conversation-level control
 - Target-guided Open-domain Conversation

Unsupervised Controlled Generation of Text

- Sentence-level control
 - **Text attribute transfer** (style transfer) [Hu et al., 2017; Yang et al., 2018]
 - Text content manipulation [Wang, Hu et al., 2019]
- Conversation-level control
 - Target-guided Open-domain Conversation

Text Attribute Transfer

- Modify a given sentence to
 - Have desired attribute values
 - While keeping all other aspects unchanged
- Attribute: sentiment, tense, voice, gender, ...

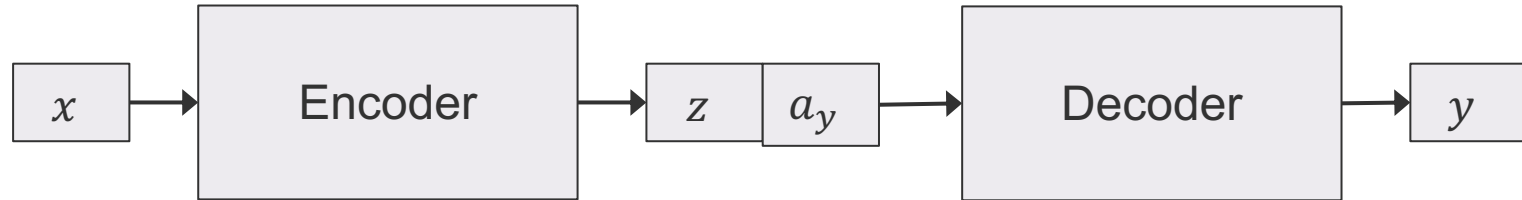
- E.g., transfer sentiment from **negative** to **positive**:
 - “It was super **dry** and had a **weird** taste to the entire slice .”
 - “It was super **fresh** and had a **delicious** taste to the entire slice .”
- Applications:
 - Personalized article writing, emotional conversation systems, ...

Text Attribute Transfer

- Original sentence \mathbf{x} , original attribute \mathbf{a}_x
- Target sentence \mathbf{y} , target attribute \mathbf{a}_y
- Task: $(\mathbf{x}, \mathbf{a}_y) \rightarrow \mathbf{y}$
 - \mathbf{y} has the desired attribute \mathbf{a}_y
 - \mathbf{y} keeps all attribute-independent properties of \mathbf{x}
- Usually, only have pairs of $(\mathbf{x}, \mathbf{a}_x)$, but no $((\mathbf{x}, \mathbf{a}_x), (\mathbf{y}, \mathbf{a}_y))$ for training
 - E.g., two sets of sentences: one with positive sentiment, the other with negative

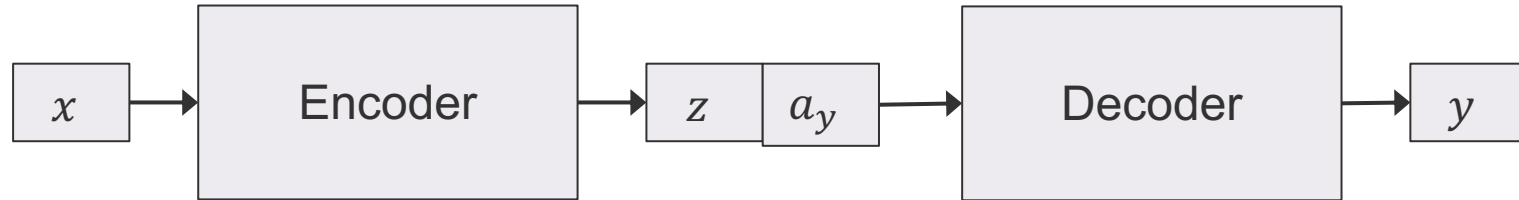
Text Attribute Transfer: Solution

- Task: $(x, a_y) \rightarrow y$
 - y has the desired attribute a_y
 - y keeps all attribute-independent properties of x
- Model $p_\theta(y|x, a_y)$

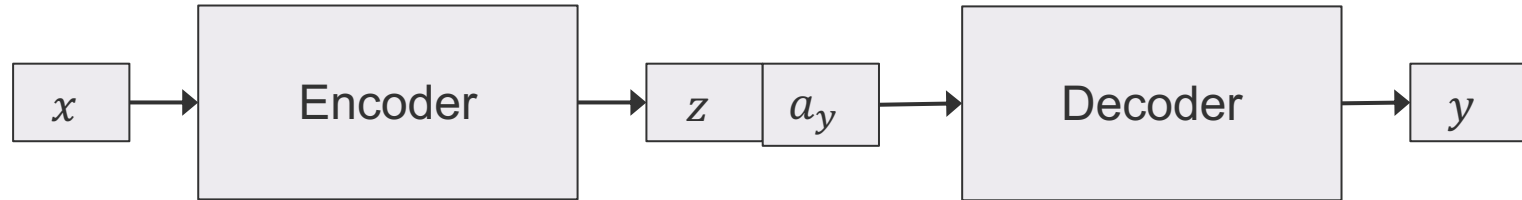


Text Attribute Transfer: Solution

- Task: $(x, a_y) \rightarrow y$
 - y has the desired attribute a_y
 - y keeps all attribute-independent properties of x
- Model $p_\theta(y|x, a_y)$
- Key intuition for learning:
 - Decompose the task into competitive sub-objectives
 - Use direct supervision for each of the sub-objectives

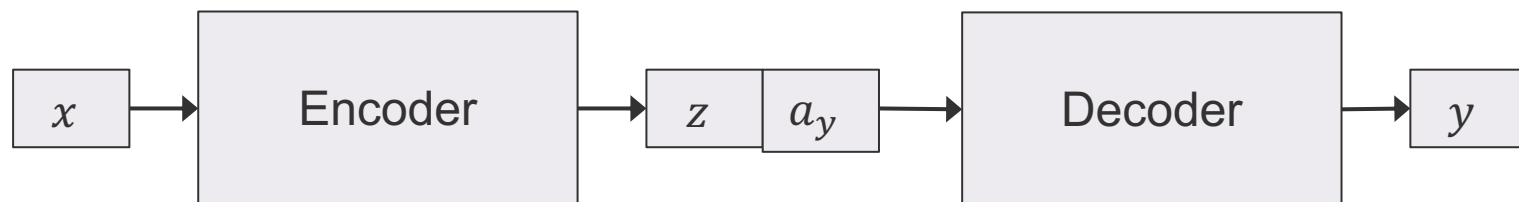


Text Attribute Transfer: Solution



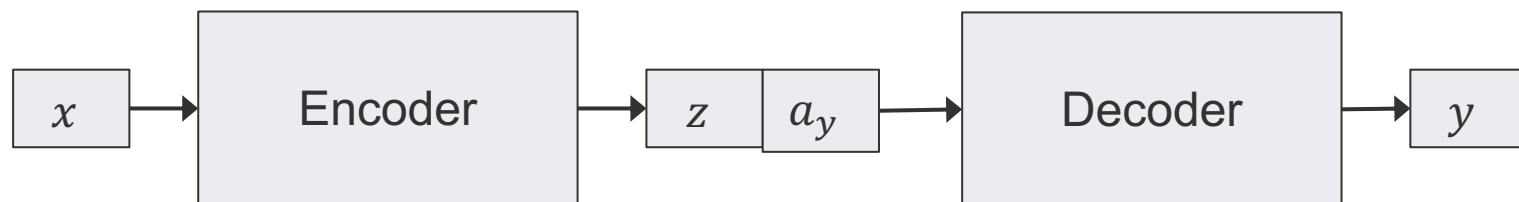
- Task: $(x, a_y) \rightarrow y$
 - y has the desired attribute a_y
 - y keeps all attribute-independent properties of x
- Model $p_\theta(y|x, a_y)$
- Key intuition for learning:
 - Decompose the task into competitive sub-objectives
 - Use direct supervision for each of the sub-objectives
- Auto-encoding loss: $(x, a_x) \rightarrow x$

Text Attribute Transfer: Solution

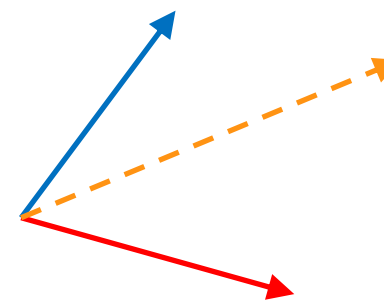


- Task: $(x, a_y) \rightarrow y$
 - y has the desired attribute a_y
 - y keeps all attribute-independent properties of x
- Model $p_\theta(y|x, a_y)$
- Key intuition for learning:
 - Decompose the task into competitive sub-objectives
 - Use direct supervision for each of the sub-objectives
- Auto-encoding loss: $(x, a_x) \rightarrow x$
- Classification loss: $\hat{y} \sim p_\theta(y|x, a_y), f(\hat{y}) \rightarrow a_y$
 - where f is a pre-trained attribute classifier

Text Attribute Transfer: Solution



- Task: $(x, a_y) \rightarrow y$
 - y has the desired attribute a_y
 - y keeps all attribute-independent properties of x
- Model $p_\theta(y|x, a_y)$
- Key intuition for learning:
 - Decompose the task into competitive sub-objectives
 - Use direct supervision for each of the sub-objectives
- Auto-encoding loss: $(x, a_x) \rightarrow x$
- Classification loss: $\hat{y} \sim p_\theta(y|x, a_y), f(\hat{y}) \rightarrow a_y$
 - where f is a pre-trained attribute classifier
- The above two losses are competitive; minimize jointly to avoid collapse



Text Attribute Transfer: Results & Improvement

- Performance on sentiment:
 - Accuracy: 92%
 - BLEU against input sentence: 54

Text Attribute Transfer: Results & Improvement

- Performance on sentiment:
 - Accuracy: 92%
 - BLEU against input sentence: 54
- Problem:
 - Language quality is often not good
 - LM perplexity: 239.8

Original: if i could give them a zero star review i would !

Output: if i **lite** give them a **sweetheart** star review i would !

Original: uncle george is very friendly to each guest

Output: uncle george is very **lackluster** to each guest

Original: the food is fresh and the environment is good

Output: the food is **atrocious** and the environment is **atrocious**

Text Attribute Transfer: Results & Improvement

- Performance on sentiment:
 - Accuracy: 92%
 - BLEU against input sentence: 54
- Problem:
 - Language quality is often not good
 - LM perplexity: 239.8
- Improvement:
 - Use an LM as a direct supervision!
 - $\hat{y} \sim p_{\theta}(y|x, a_y), \max_{\theta} \text{LM}(\hat{y})$
 - Accuracy: 91%
 - BLEU against input sentence: 57
 - LM perplexity: 60.9

Original: if i could give them a zero star review i would !

Output: if i **lite** give them a **sweetheart** star review i would !

Original: uncle george is very friendly to each guest

Output: uncle george is very **lackluster** to each guest

Original: the food is fresh and the environment is good

Output: the food is **atrocious** and the environment is **atrocious**

Text Attribute Transfer: Results & Improvement

- Performance on sentiment:
 - Accuracy: 92%
 - BLEU against input sentence: 54
- Problem:
 - Language quality is often not good
 - LM perplexity: 239.8
- Improvement:
 - Use an LM as a direct supervision!
 - $\hat{y} \sim p_{\theta}(y|x, a_y), \max_{\theta} \text{LM}(\hat{y})$
 - Accuracy: 91%
 - BLEU against input sentence: 57
 - LM perplexity: 60.9

Original: if i could give them a zero star review i would !

Output: if i **like** give them a **sweetheart** star review i would !

+ LM: if i can give them a great star review i would !

Original: uncle george is very friendly to each guest

Output: uncle george is very **lackluster** to each guest

+ LM: uncle george is very rude to each guest

Original: the food is fresh and the environment is good

Output: the food is **atrocious** and the environment is **atrocious**

+ LM: the food is bland and the environment is bad .

Unsupervised Controlled Generation of Text

- Sentence-level control
 - Text attribute transfer (style transfer) [Hu et al., 2017; Yang et al., 2018]
 - Text content manipulation [Wang, Hu et al., 2019]
- Conversation-level control
 - Target-guided Open-domain Conversation

Key idea:

- Decompose the task into **competitive** sub-objectives
- Use **direct supervision** for each of the sub-objectives

Unsupervised Controlled Generation of Text

- Sentence-level control
 - Text attribute transfer (style transfer) [Hu et al., 2017; Yang et al., 2018]
 - **Text content manipulation** [Wang, Hu et al., 2019]
- Conversation-level control
 - Target-guided Open-domain Conversation

Key idea:

- Decompose the task into **competitive** sub-objectives
- Use **direct supervision** for each of the sub-objectives

Text Content Manipulation

- Generate a sentence to describe content in a given data record
- But language is rich with variation -- there are diverse possible ways of saying the same content (writing style):
 - word choice, expressions, transitions, tones, ...

Content Record	PLAYER	PT	RB	AS	PLAYER	PT
	LeBron_James	32	4	7	Kyrie_Irving	20
Reference Sentence	Jrue_Holiday led the way with 26 points and 6 assists , while Goran_Dragic scored 23 points and pulled down 8 rebounds .					
Output	LeBron_James led the way with 32 points , 7 assists and 4 rebounds , while Kyrie_Irving scored 20 points .					

Text Content Manipulation

- Generate a sentence to describe content in a given data record
- But language is rich with variation -- there are diverse possible ways of saying the same content (writing style):
 - word choice, expressions, transitions, tones, ...
- We want to control the writing style: use the writing style of a reference sentence

Content Record	PLAYER LeBron_James	PT 32	RB 4	AS 7	PLAYER Kyrie_Irving	PT 20
Reference Sentence	Jrue_Holiday led the way with 26 points and 6 assists , while Goran_Dragic scored 23 points and pulled down 8 rebounds .					
Output	LeBron_James led the way with 32 points , 7 assists and 4 rebounds , while Kyrie_Irving scored 20 points .					

Text Content Manipulation - Results

Content x	PLAYER	PTS	FGM	FGA	FG3M	FG3A	FTM	FTA	AST
	Gerald_Henderson	17	6	13	1	2	4	4	5
Reference y'	Kawhi_Leonard also had a solid offensive game , scoring 16 points (7 - 13 FG , 0 - 1 3Pt , 2 - 5 FT) and adding 5 assists and 5 rebounds .								
Rule-based	Gerald_Henderson also had a solid offensive game , scoring 17 points (6 - 13 FG , 1 - 2 3Pt , 4 - 4 FT) and adding 5 assists and 5 rebounds .								
AdvST	Gerald_Henderson also had a solid offensive game , scoring 13 points (13 - 13 FG , 2 - 2 3Pt , 4 - 4 FT) and adding 5 assists and 5 rebounds .								
Ours w/o Cover.	Gerald_Henderson also had a solid offensive game , scoring 17 points (6 - 13 FG , 1 - 2 3Pt , 4 - 4 FT) and adding 5 assists and 5 rebounds .								
Ours	Gerald_Henderson also had a solid offensive game , scoring 17 points (6 - 13 FG , 1 - 2 3Pt , 4 - 4 FT) and adding 5 assists .								

Text Content Manipulation - Results

	Model	Precision %	Recall %	Style BLEU
1	AttnCopy-S2S	88.71±2.45	60.64±1.31	39.15±5.48
	Rule-based	62.63	63.64	100
2	MAST	33.15±0.78	31.09±0.63	95.29±2.53
	AdvST	66.51±1.08	56.03±0.56	72.22±1.47
3	Ours w/o Cover.	75.61±1.03	62.93±0.53	75.09±2.17
	Ours	78.31±0.94	65.64±0.47	80.83±1.89

Unsupervised Controlled Generation of Text

- Sentence-level control
 - Text attribute transfer (style transfer) [Hu et al., 2017; Yang et al., 2018]
 - Text content manipulation [Wang, Hu et al., 2019]
- Conversation-level control
 - Target-guided Open-domain Conversation

Key idea:

- Decompose the task into **competitive** sub-objectives
- Use **direct supervision** for each of the sub-objectives

Unsupervised Controlled Generation of Text

- Sentence-level control
 - Text attribute transfer (style transfer) [Hu et al., 2017; Yang et al., 2018]
 - Text content manipulation [Wang, Hu et al., 2019]
- Conversation-level control
 - Target-guided Open-domain Conversation

Key idea:

- Decompose the task into **competitive** sub-objectives
- Use **direct supervision** for each of the sub-objectives

Target-guided Open-domain Conversation

- Task-oriented dialog:
 - Address a specific task, e.g., booking a flight
 - Close domain
- Open-domain chit-chat:
 - Improve user engagement
 - Random conversation, hard to control
- Target-guided conversation:
 - Open-domain conversation
 - Controlled conversation strategy to reach a *desired topic* in the end of conversation
 - Applications:
 - Bridges task-oriented dialog and open-domain chit-chat
 - Conversational recommender system, education, psychotherapy

Target-guided Open-domain Conversation

- Two goals:
 - Starting from any topic, reach a desired topic in the end of conversation
 - Natural conversation: smooth transition

Target: **e-books**

Agent : hi how are you doing

Human : not so good . i am really tired

Agent : oh i'm sorry to hear

Human : i have a lot of work to do

Agent : what kind was it ?

Human : i am writing a chatbot program

Agent : interesting . i read about coding from a book

Human: really

Agent : i prefer **e-books** over paperback book

Target-guided Open-domain Conversation

- Two goals:
 - Starting from any topic, reach a desired topic in the end of conversation
 - Natural conversation: smooth transition

Target: **e-books**

Agent : hi how are you doing

tired Human : not so good . i am really tired

sorry Agent : oh i'm sorry to hear

work Human : i have a lot of work to do

work Agent : what kind was it ?

program Human : i am writing a chatbot program

coding, book Agent : interesting . i read about coding from a book

Human: really

e-books Agent : i prefer **e-books** over paperback book

Target-guided Open-domain Conversation

- Two goals:
 - Starting from any topic, reach a desired topic in the end of conversation
 - Natural conversation: smooth transition

Target: **e-books**

Agent : hi how are you doing

tired Human : not so good . i am really tired

sorry Agent : oh i'm sorry to hear

work Human : i have a lot of work to do

work Agent : what kind was it ?

program Human : i am writing a chatbot program

coding, book Agent : interesting . i read about coding from a book

Human: really

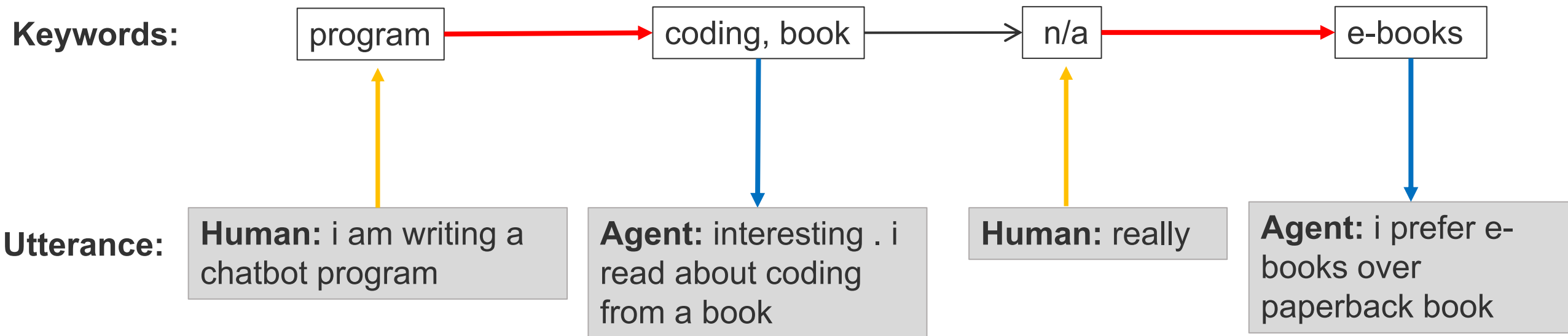
e-books Agent : i prefer **e-books** over paperback book

Challenge: No supervised data for the task

Solution: Use competitive sub-objectives and partial supervision

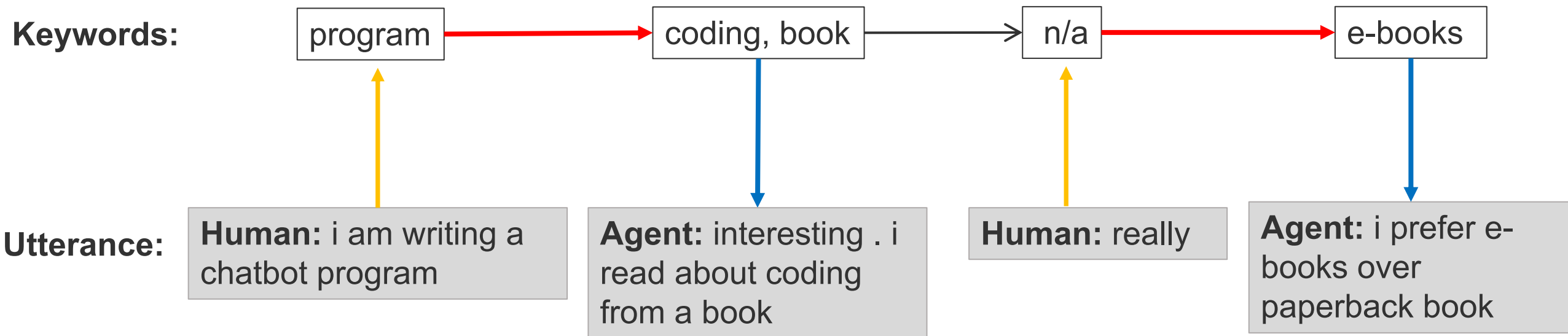
- **Natural conversation:** rich chit-chat data to learn smooth **single-turn** transition
- **Reaching desired target:** rule-based **multi-turn** planning

Target-guided Open-domain Conversation





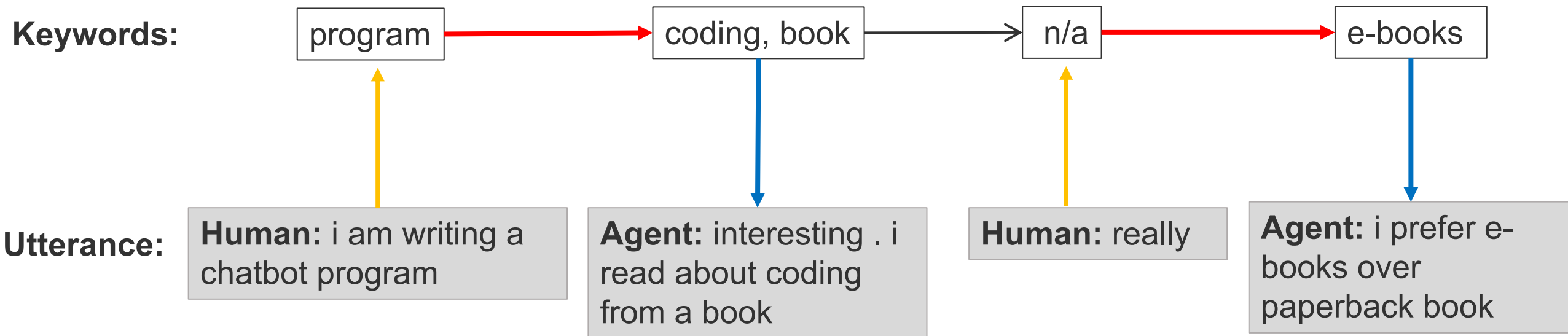
Target-guided Open-domain Conversation

- → keyword extraction






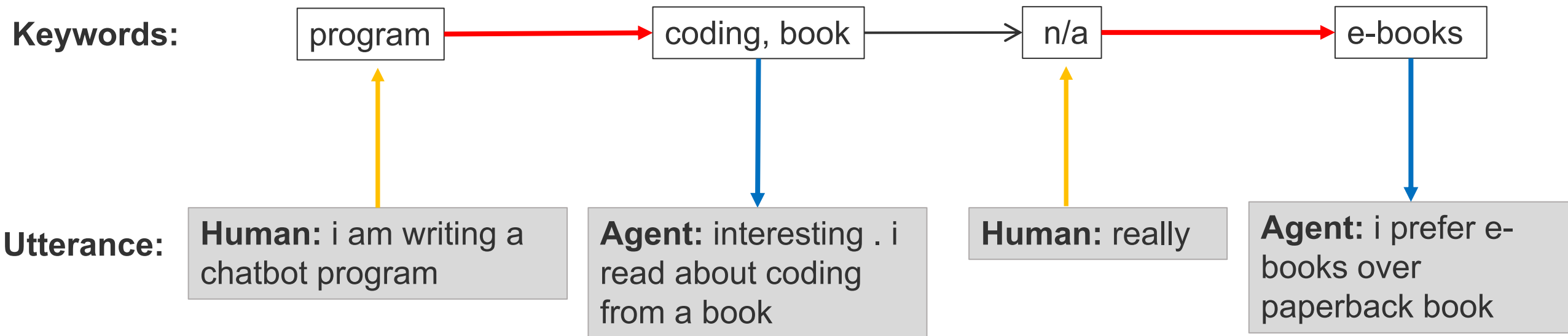
Target-guided Open-domain Conversation

-  keyword extraction
-  keyword conditional response retrieval



Target-guided Open-domain Conversation

-  keyword extraction
-  keyword conditional response retrieval
-  learned kernel-based topic transition + target-guided rule:
 - (1) transit to keywords that are close in the word embedding space – smooth
 - (2) next keywords must get closer to the target keyword – target-guided
 - Tune relative weight of (1) and (2) to control aggressiveness



Example Results

Target: dance

Agent : hi how is it going with you

Human : fine . I'm playing sports

Agent : i play basketball . u play ?

Human : yeah i also like basketball .

Agent : you like rap music ? i listen to a lot of rap music

Human : no i like classic music

Agent : its a taylor swift song

Human : taylor swift is a great singer

Agent : sing to me , what are you going to sing

Human : i know nothing about singing

Agent : oh i think i can dance lol

Example Results – Failure Case

Target: listen

Agent : hi ! how are you ? i am very exciting about my new job !

Human : fine . what is your new job ?

Agent : that sounds exciting .

Human : i like read books

Agent : what kind of books do you read ?

Human : i read comics

Agent : i do not have tv

Human : watch anime is fun

Agent : what position d you play ?

Human : i play a lot of sports

Agent : oh really ? what kind of music . i listen to classical

Unsupervised Controlled Generation of Text

- Sentence-level control
 - Text attribute transfer (style transfer) [Hu et al., 2017; Yang et al., 2018]
 - Text content manipulation [Wang, Hu et al., 2019]
- Conversation-level control
 - Target-guided Open-domain Conversation

Key idea:

- Decompose the task into **competitive** sub-objectives
- Use **direct supervision** for each of the sub-objectives

Key Takeaways: Two Central Goals

- Generating human-like, grammatical, and readable text
 - Progressive generation
 - Exposure bias, criteria mismatch: reinforcement learning (next lecture)
- Generating text that contains desired information inferred from inputs
 - Machine translation
 - Source sentence --> target sentence w/ the same meaning
 - Data description
 - Table --> data report describing the table
 - Attribute control
 - Sentiment: positive --> "I like this restaurant"
 - Conversation control
 - Control conversation strategy and topic

Questions?