

On the Transparency and Reliability of Automatic Summarization



Carnegie Mellon University
Language Technologies Institute



Information Overload



Summarized Information Consumption

The New York Times

For Biden's Climate Pledge to Succeed, America Would Need Big Changes

Hitting President Biden's emissions targets would require rapid and sweeping changes to virtually every corner of the nation's economy, experts say.

It would require a rapid shift to electric vehicles, the expansion of forests, new carbon-capture technology and many other changes, researchers said.



SEMANTIC SCHOLAR

DOI: 10.18653/v1/P18-1378 • Corpus ID: 223637

Simple and Effective Multi-Paragraph Reading Comprehension

Christopher Clark, Matt Gardner • Published 2018 • Computer Science • ArXiv

[Highlight Information](#)

[Methods](#)

[Results](#)

We consider the problem of adapting neural paragraph-level question answering models to the case where entire documents are given as input. Our proposed solution trains models to produce well-calibrated confidence scores for their results on individual paragraphs. We sample multiple paragraphs from the documents during training, and use a shared-normalization training objective that encourages the model to produce globally correct output. We combine this method with a state-of-the-art pipeline for training models on document QA data. Experiments demonstrate strong performance on several document QA datasets. Overall, we are able to achieve a score of 71.3 F1 on the web portion of TriviaQA, a large improvement from the 56.7 F1 of the previous best system. [Collapse](#)



India records most COVID-19 cases and deaths in last 24 hours

short by Anmol Sharma / 10:01 am on 23 Apr 2021, Friday

As many as 3,32,730 coronavirus cases were confirmed in India in the last 24 hours, marking the world's biggest one-day jump so far and taking the total number of cases to 1,62,63,695. Meanwhile, 2,263 deaths have been recorded in the last 24 hours, marking India's biggest one-day jump and taking the death toll to 1,86,920.

read more at MOHFW

 **inshorts**
stay informed

Neural Abstractive Summarization

Article : Electronic Devices allowed on Southwest

[1] Southwest Airlines has received Federal Aviation Administration approval to allow passengers to use many portable electronic devices in all phases of flight.
[2] Under the new rules, passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue Airways and Delta Air Lines moved quickly to get FAA approval to allow devices on board on November 1.
....
....
....
....
[12] The new expanded use of electronics does not apply to making or taking calls, which are still prohibited in flight.

Source Article



[1] Southwest is newest airline to allow use of portable electronic devices.
[2] Passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue, Delta, United and others have also moved to get FAA approval.

Neural
Model

Generated
Summary

Limitations of Neural Methods for Summarization

Article : Electronic Devices allowed on Southwest

[1] Southwest Airlines has received Federal Aviation Administration approval to allow passengers to use many portable electronic devices in all phases of flight.
[2] Under the new rules, passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue Airways and Delta Air Lines moved quickly to get FAA approval to allow devices on board on November 1.
....
....
....
....
[12] The new expanded use of electronics does not apply to making or taking calls, which are still prohibited in flight.



[1] Southwest is newest airline to allow use of portable electronic devices.
[2] Passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue, Delta, United and others have also moved to get FAA approval.

Limited Abstractiveness!

Limitations of Neural Methods for Summarization

Article : Electronic Devices allowed on Southwest

[1] Southwest Airlines has received Federal Aviation Administration approval to allow passengers to use many portable electronic devices in all phases of flight.
[2] Under the new rules, passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue Airways and Delta Air Lines moved quickly to get FAA approval to allow devices on board on November 1. ✗
....
....
....
....
[12] The new expanded use of electronics does not apply to making or taking calls, which are still prohibited in flight. ✓



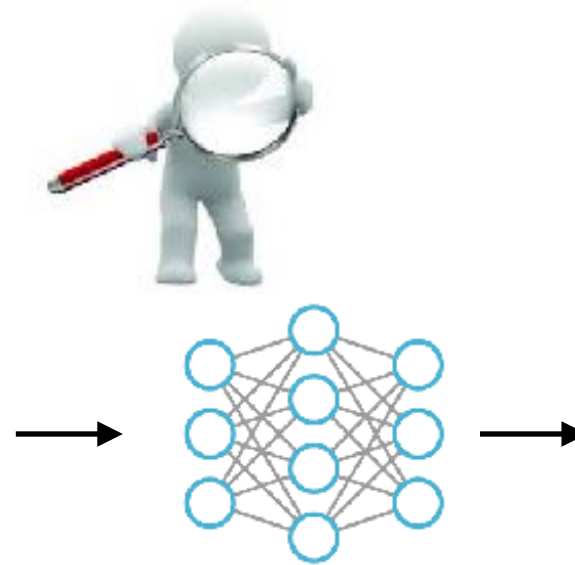
[1] Southwest is newest airline to allow use of portable electronic devices.
[2] Passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue, Delta, United and others have also moved to get FAA approval.

Overfit to Layout Biases!

Limitations of Neural Methods for Summarization

Article : Electronic Devices allowed on Southwest

[1] Southwest Airlines has received Federal Aviation Administration approval to allow passengers to use many portable electronic devices in all phases of flight.
[2] Under the new rules, passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue Airways and Delta Air Lines moved quickly to get FAA approval to allow devices on board on November 1.
....
....
....
....
[12] The new expanded use of electronics does not apply to making or taking calls, which are still prohibited in flight.



[1] Southwest is newest airline to allow use of portable electronic devices.
[2] Passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue, Delta, United and others have also moved to get FAA approval.

Lack of Transparency!

Limitations of Neural Methods for Summarization

Article : Electronic Devices allowed on Southwest

[1] **Southwest Airlines** has received Federal Aviation Administration approval to allow passengers to use many portable electronic devices in all phases of flight.
[2] Under the new rules, passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] **JetBlue Airways and Delta Air Lines** moved quickly to get FAA approval to allow devices on board on November 1.
....
....
....
....
[12] The new expanded use of electronics does not apply to making or taking calls, which are still prohibited in flight.



[1] **JetBlue Airways** is newest airline to allow use of portable electronic devices. ✗
[2] Passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing. ✓
[3] **Southwest**, Delta, **United** and others have also moved to get FAA approval. ✗

Generates Factually Inconsistent Content!

Today's Talk

- Framework for Incorporating Document Structure

StructSum: Summarization via Structured Representations Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, Yulia Tsvetkov. In *Proc. EACL'21*.

- Benchmark for Evaluating Factuality of Generated Summaries

Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics Artidoro Pagnoni, Vidhisha Balachandran, Yulia Tsvetkov *To Appear In NAACL'21*.

Today's Talk

- Framework for Incorporating Document Structure

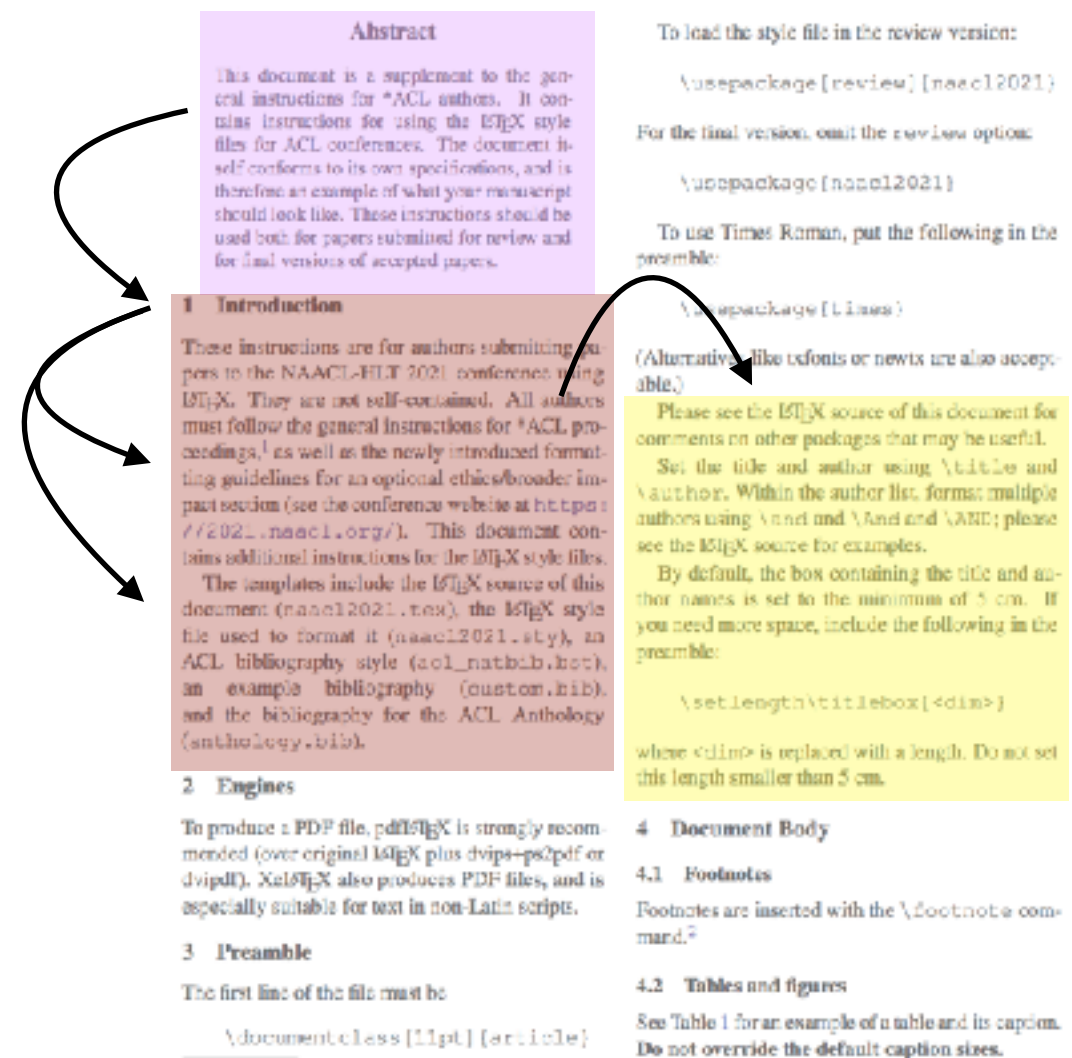
StructSum: Summarization via Structured Representations Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, Yulia Tsvetkov. In *Proc. EACL'21*.

- Benchmark for Evaluating Factuality of Generated Summaries

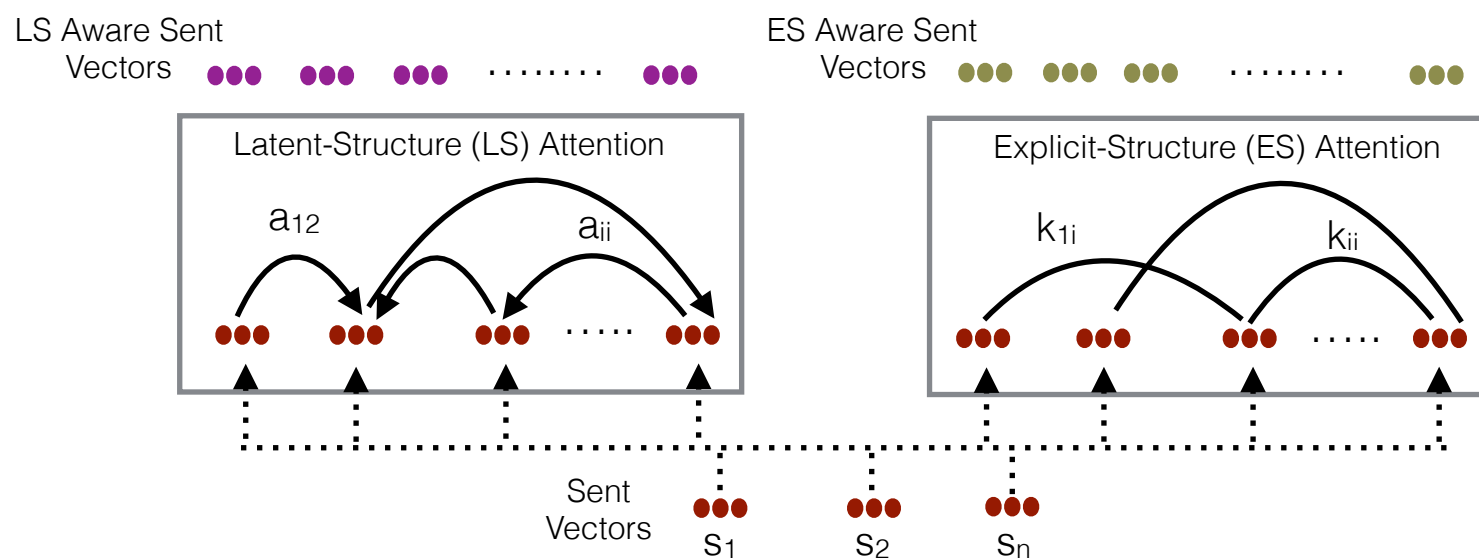
Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics Artidoro Pagnoni, Vidhisha Balachandran, Yulia Tsvetkov *To Appear In NAACL'21*.

Language Structure for Document Representations

- Limitations we aim to address:
 - Limited Abstractiveness
 - Overfit to Layout Bias
 - Lack of Transparency
- Incorporate knowledge on document structure - improve representation and understanding
- Encourage learning narrative between events and entities



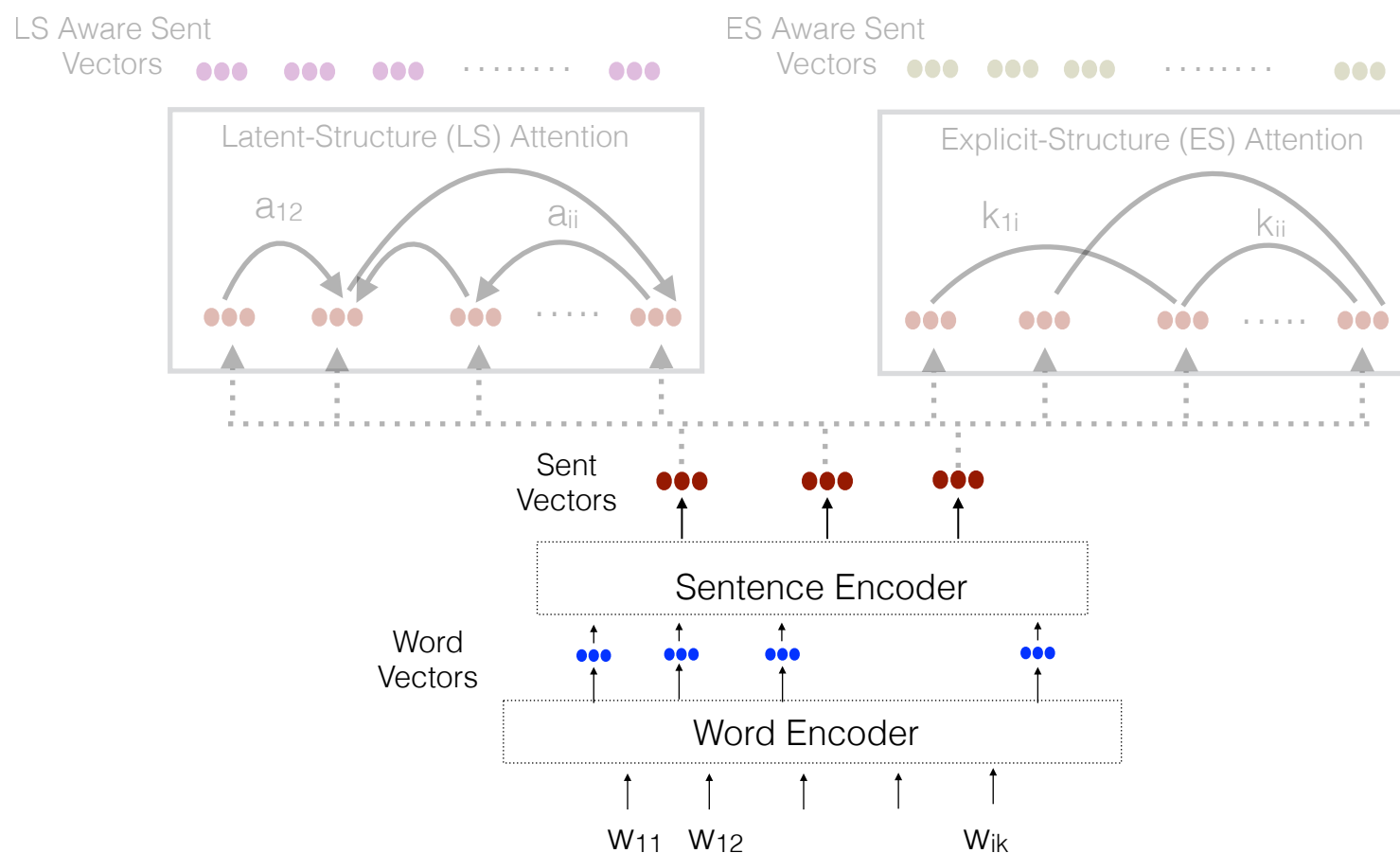
StructSum - Structured Document Representations for Summarization



- Latent Structure - induce latent structured representations (Liu and Lapata, 2017)
- Learn task-specific document structures
- Explicit Structure - incorporate external linguistic structure
- Incorporate domain-specific inductive biases

StructSum - Structured Document Representations for Summarization

- Word Encoder - produces word level contextual representations



$$w_{i1} \dots w_{ik} = BiLSTM(w_{i1} \dots w_{ik})$$

StructSum - Structured Document Representations for Summarization

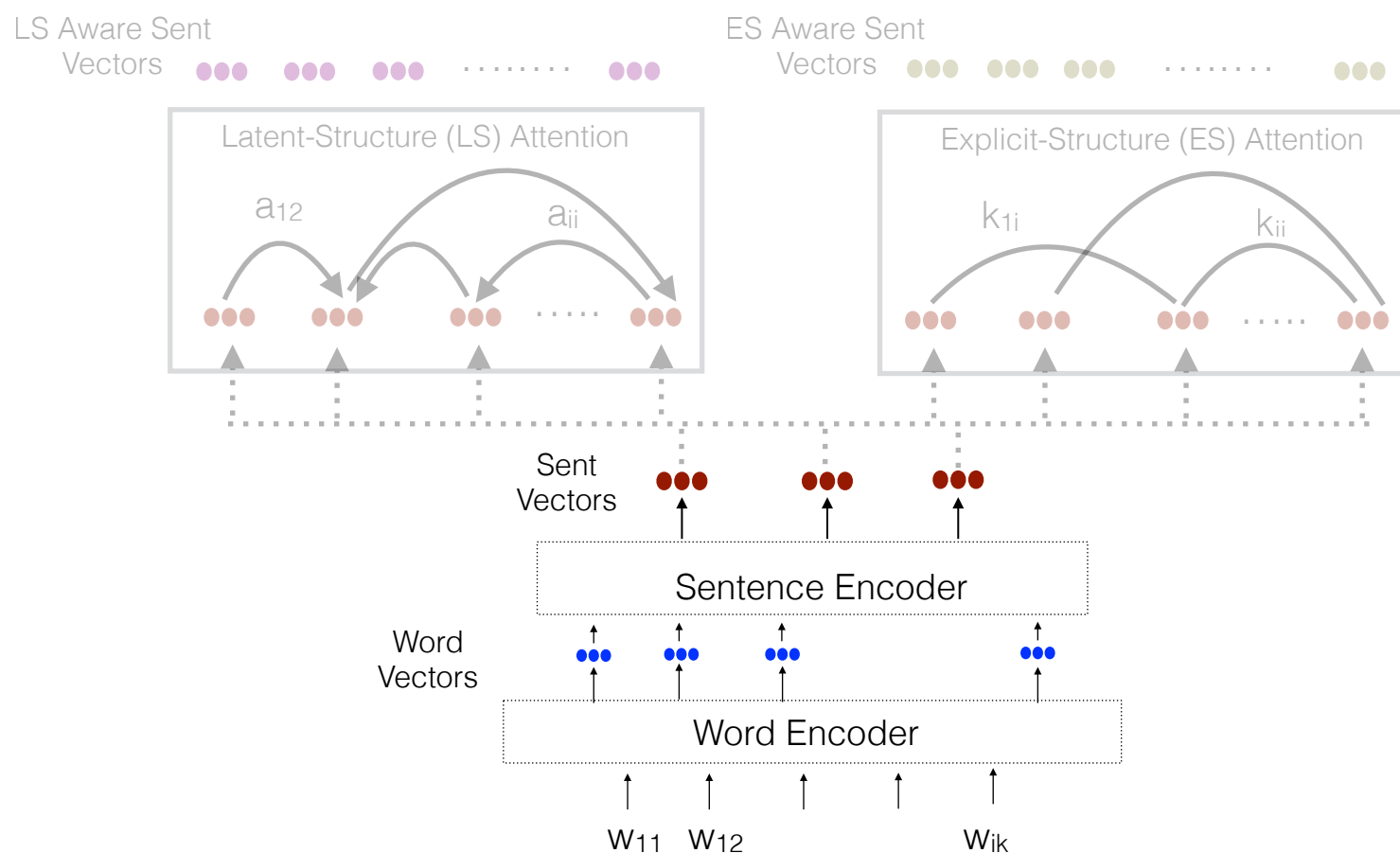
- Word Encoder - produces word level contextual representations

$$w_{i1}...w_{ik} = BiLSTM(w_{i1}..w_{ik})$$

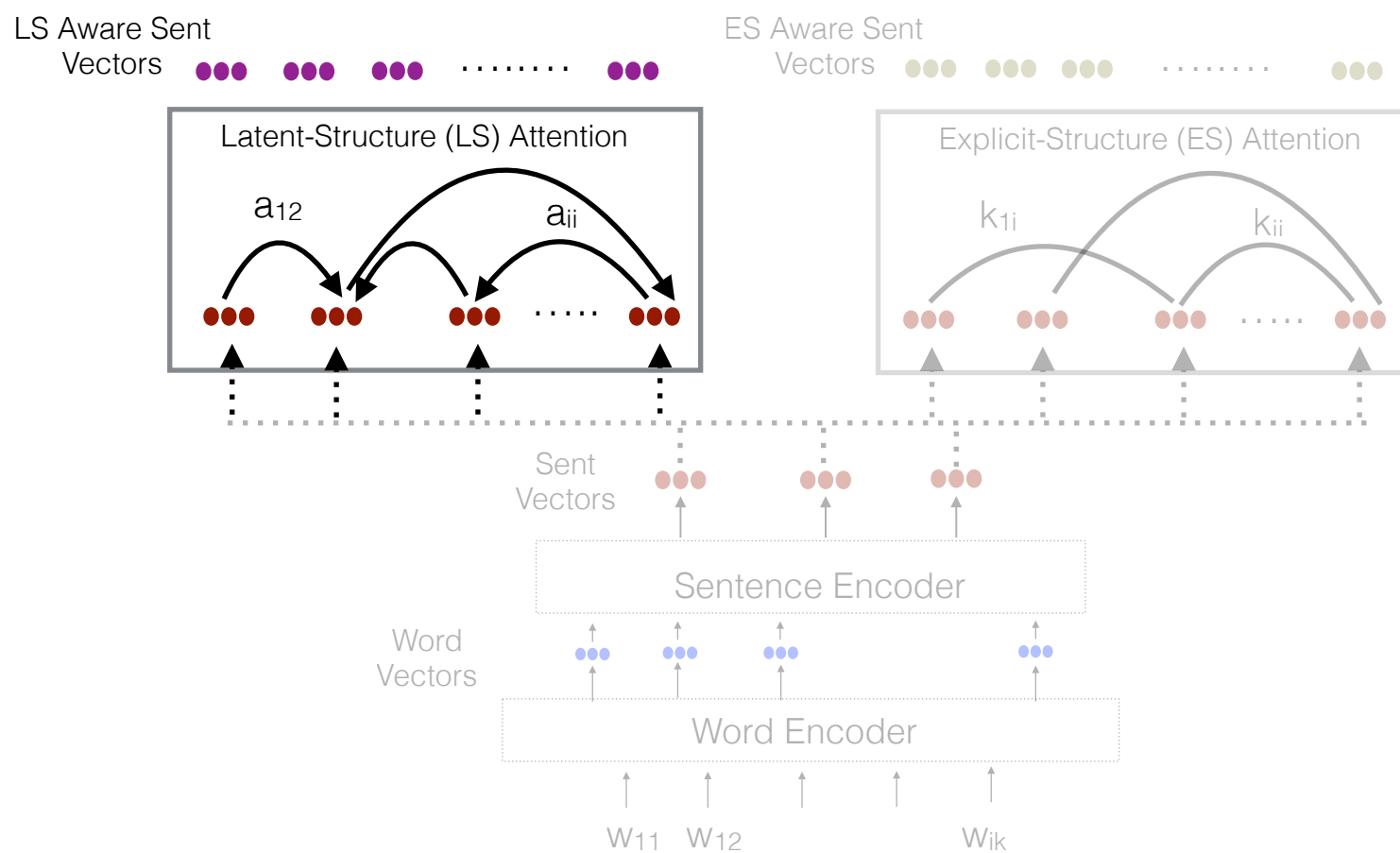
- Sentence Encoder - pool word representations and produce contextual sentence representations

$$s_i = f(w_{i1}...w_{ik}); \quad f = \max/\text{mean}$$

$$s_1...s_n = BiLSTM(s_1..s_n)$$



StructSum - Structured Document Representations for Summarization



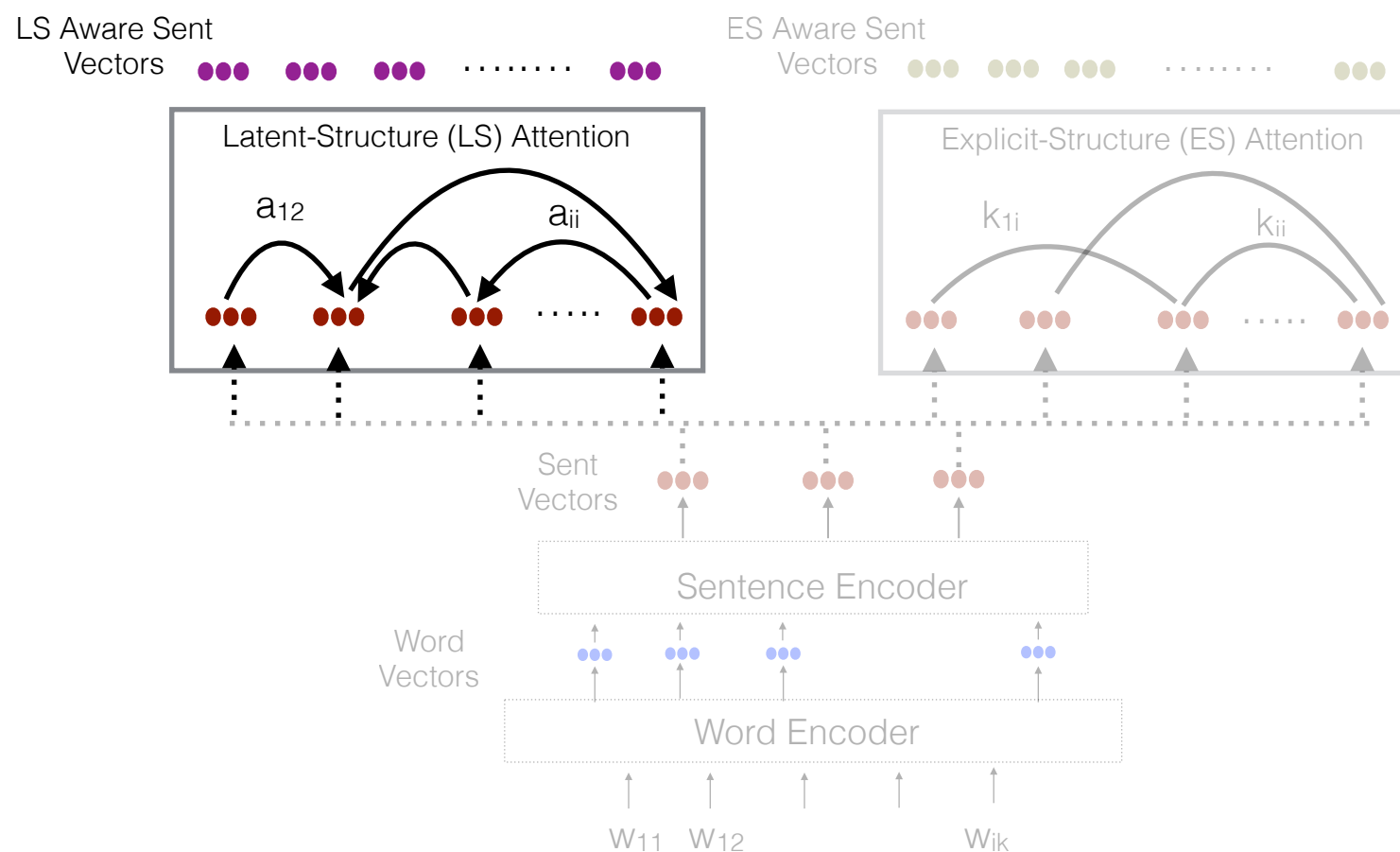
- Adapt Latent Structured Representations
 - induce dependency structure between sentences
 - form a non-projective dependency tree.

StructSum - Structured Document Representations for Summarization

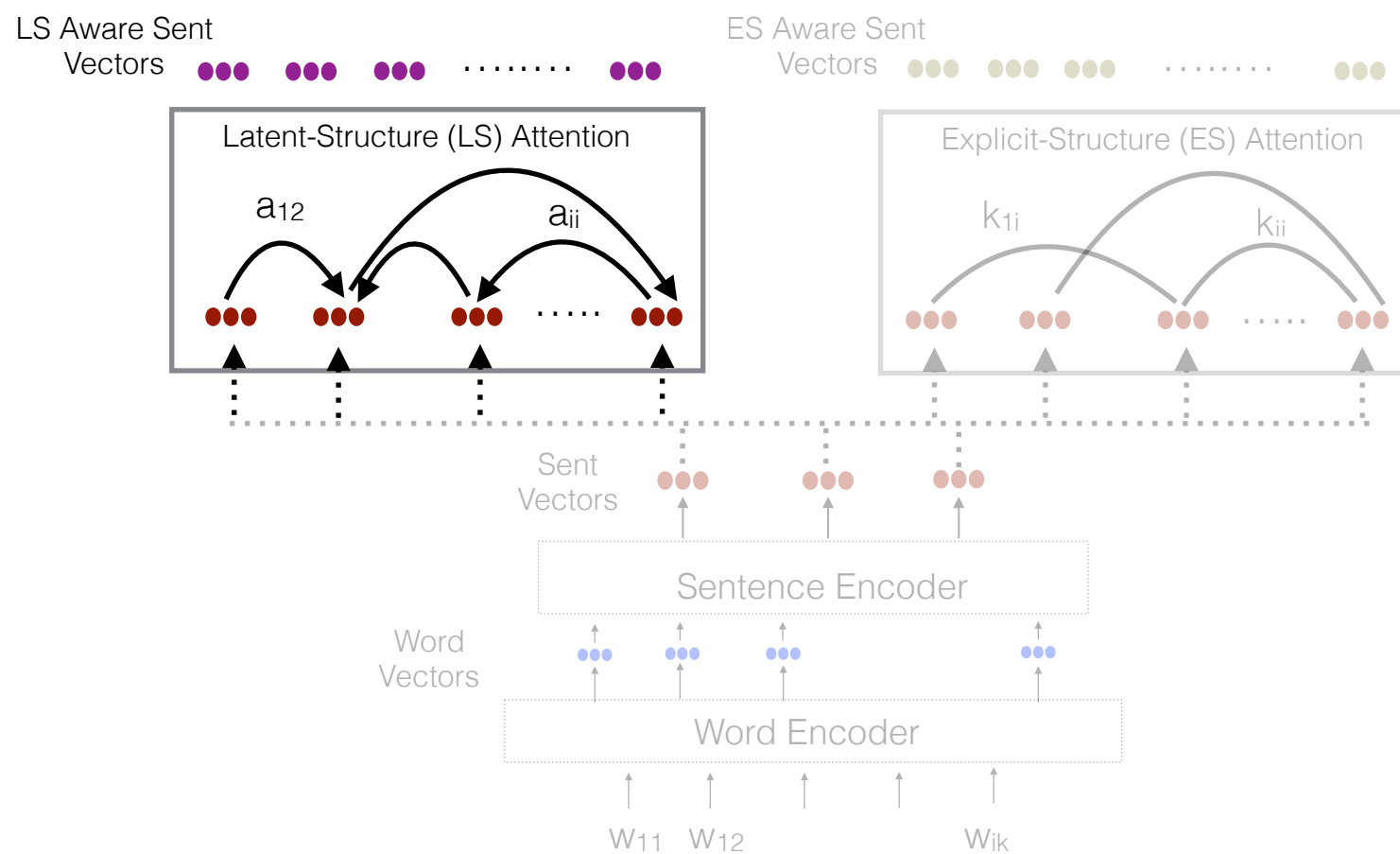
- Adapt Latent Structured Representations
 - induce dependency structure between sentences
 - form a non-projective dependency tree.

z_{ij} : dependency edge from $s_i \rightarrow s_j$

z_i^r : s_i is root



StructSum - Structured Document Representations for Summarization

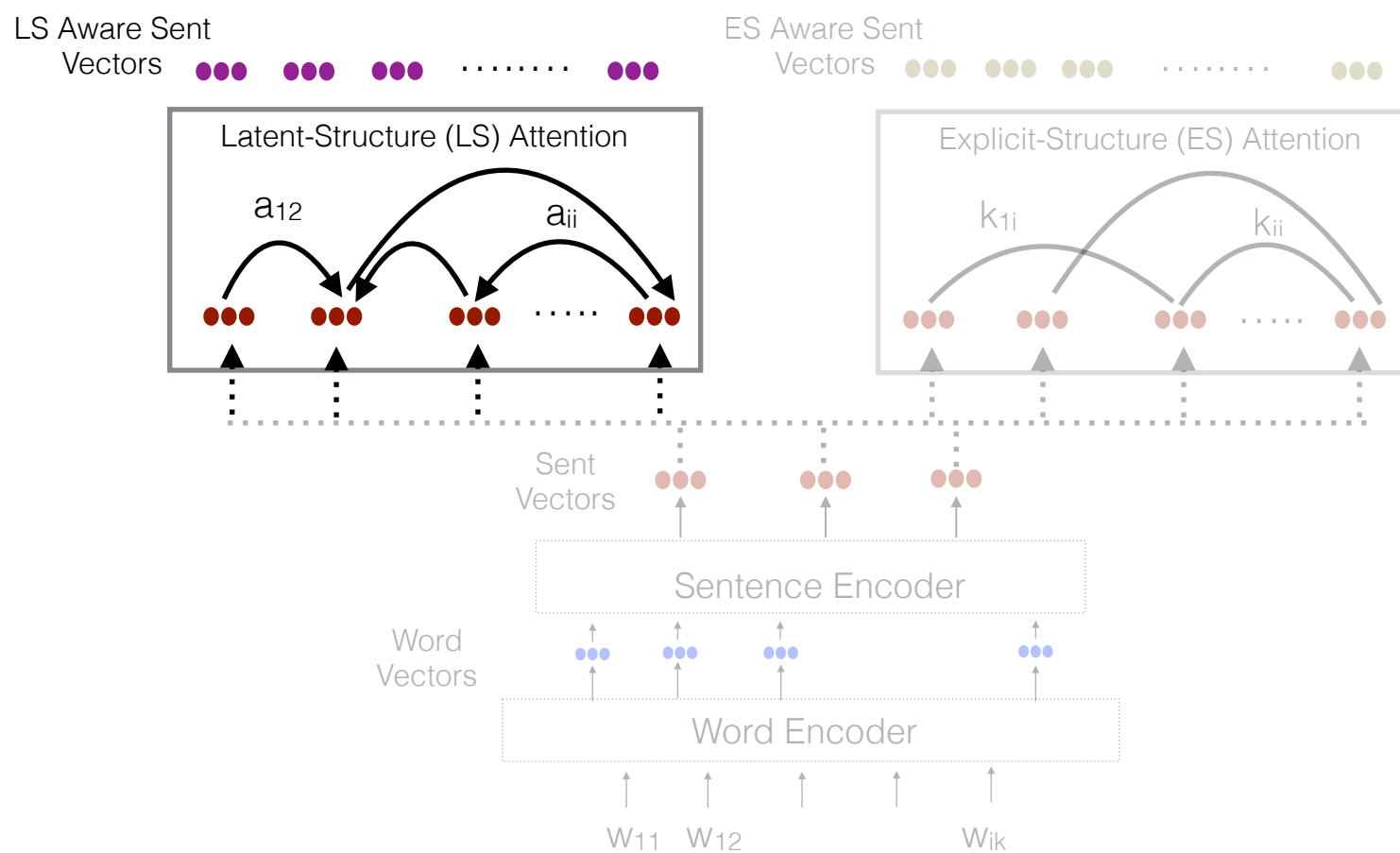


z_{ij} : dependency edge from $s_i \rightarrow s_j$

z_i^r : s_i is root

$a_{ij} = p(z_{ij} = 1); \quad a_i^r = p(z_i^r = 1)$

StructSum - Structured Document Representations for Summarization



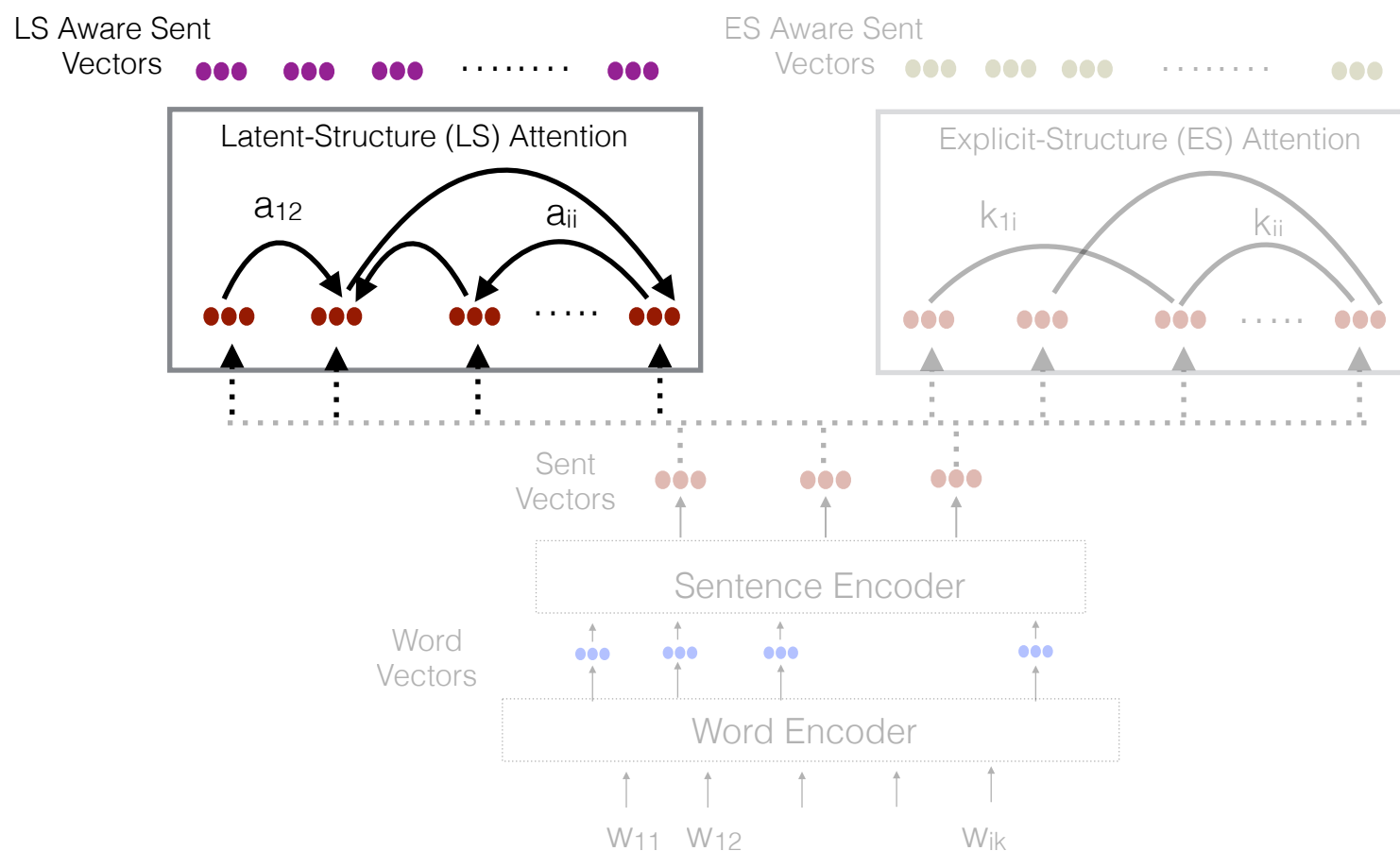
z_{ij} : dependency edge from $s_i \rightarrow s_j$

z_i^r : s_i is root

$$a_{ij} = p(z_{ij} = 1); \quad a_i^r = p(z_i^r = 1)$$

$$t_{ij} = F(s_i)^T W_a F(s_j); \quad r_i = F(s_i)$$

StructSum - Structured Document Representations for Summarization



z_{ij} : dependency edge from $s_i \rightarrow s_j$

z_i^r : s_i is root

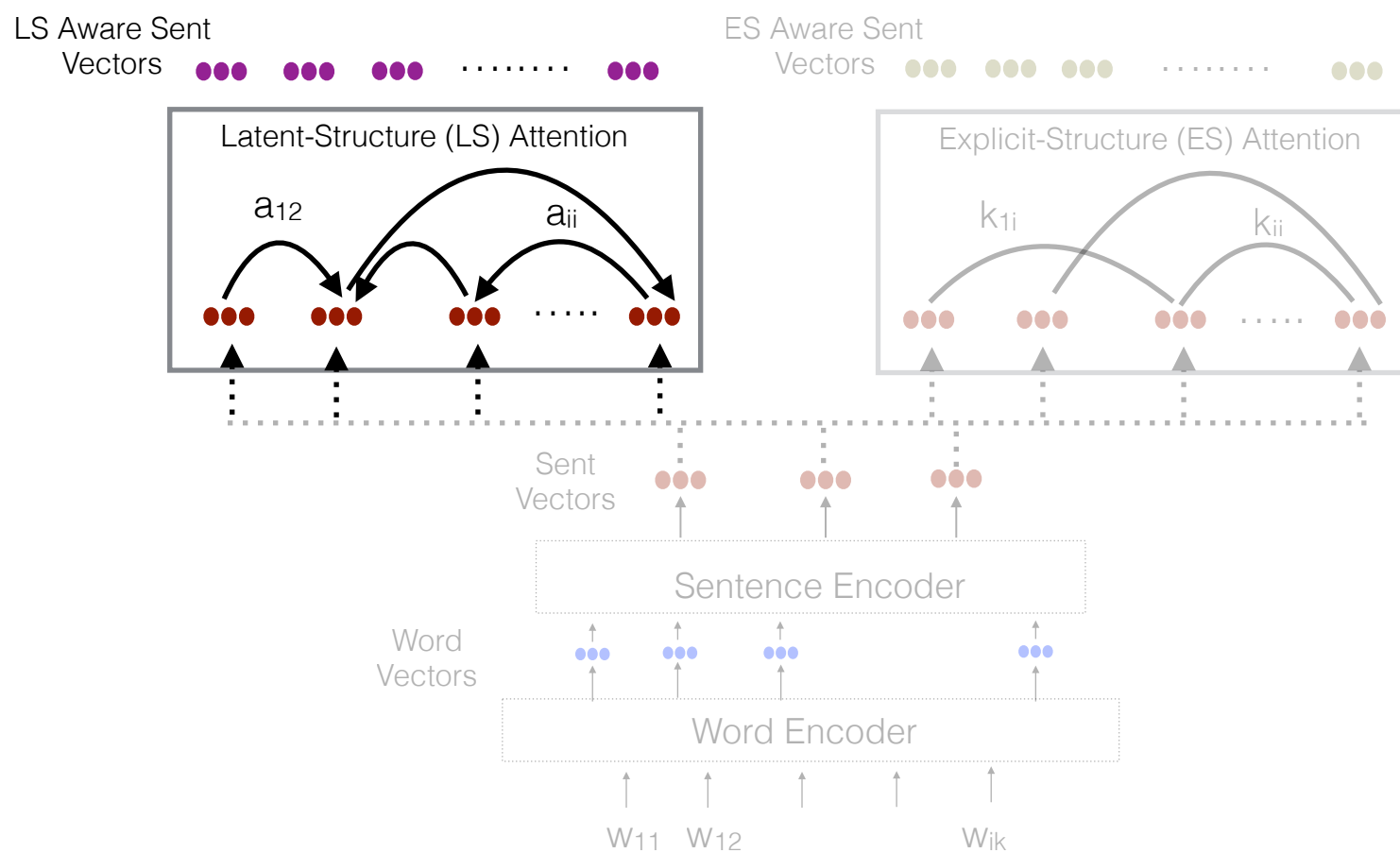
$$a_{ij} = p(z_{ij} = 1); \quad a_i^r = p(z_i^r = 1)$$

$$t_{ij} = F(s_i)^T W_a F(s_j); \quad r_i = F(s_i)$$

$$t_{ij} \rightarrow p(z_{ij} = 1); \quad r_i \rightarrow p(z_i^r = 1)$$

Using Kirchoff's Matrix Tree theorem

StructSum - Structured Document Representations for Summarization



z_{ij} : dependency edge from $s_i \rightarrow s_j$

z_i^r : s_i is root

$$a_{ij} = p(z_{ij} = 1); \quad a_i^r = p(z_i^r = 1)$$

$$t_{ij} = F(s_i)^T W_a F(s_j); \quad r_i = F(s_i)$$

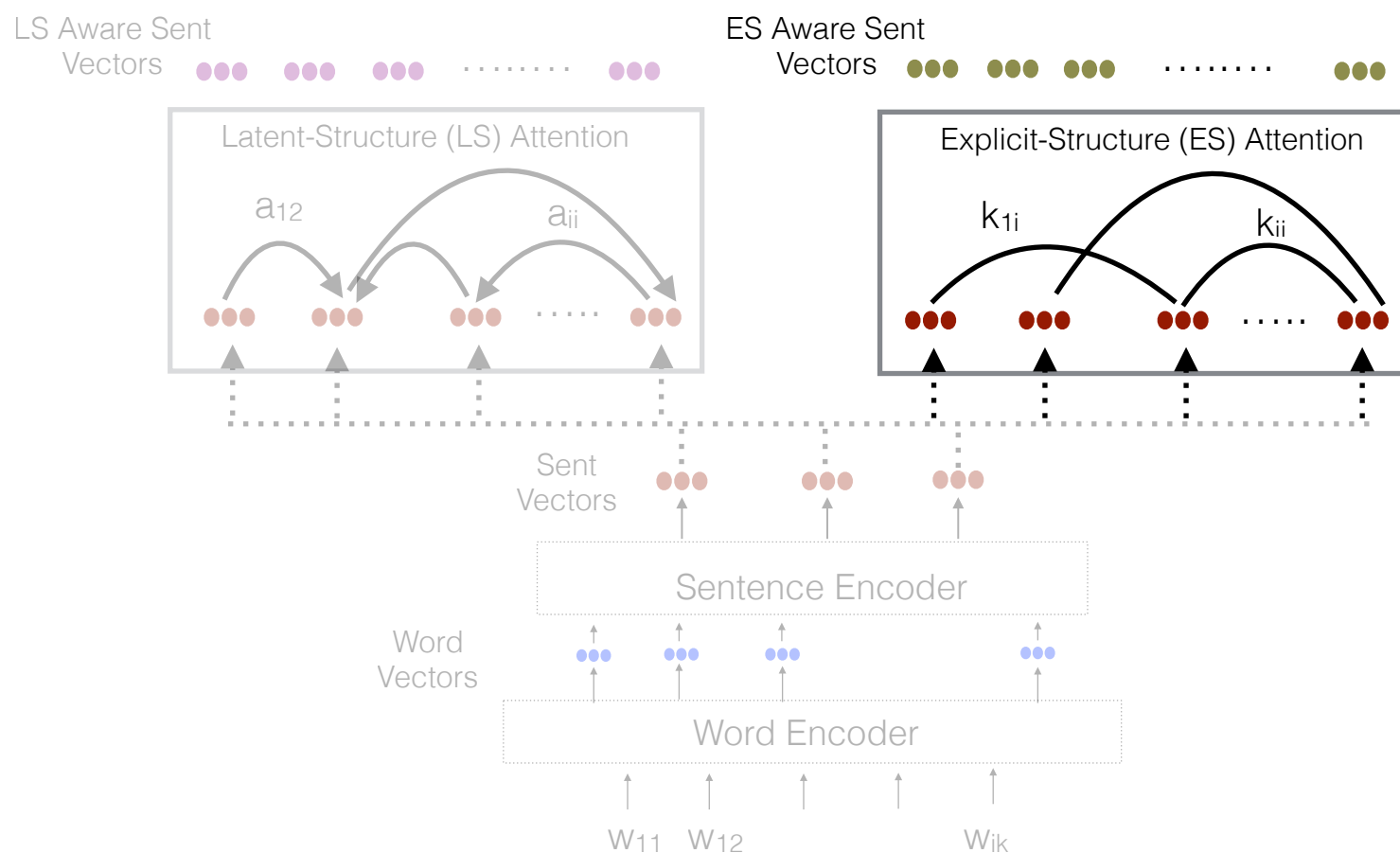
$$t_{ij} \rightarrow p(z_{ij} = 1); \quad r_i \rightarrow p(z_i^r = 1)$$

Using Kirchoff's Matrix Tree theorem

$$LS_i = f(a_i^r, a_{ij}, a_{ki}, s_i, s_j, s_k) \quad \forall j, k : 1..n$$

StructSum - Structured Document Representations for Summarization

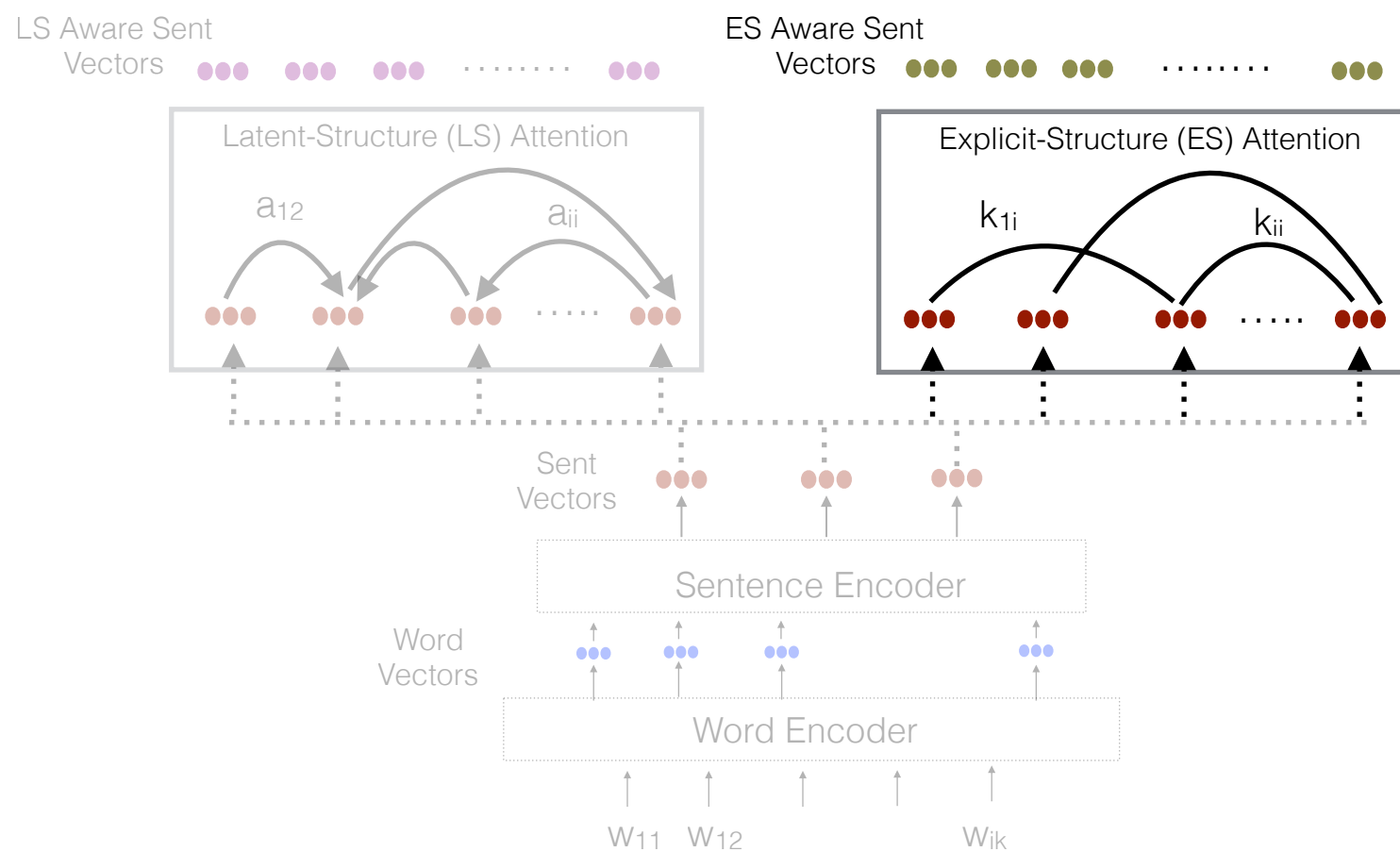
- Augment linguistic knowledge (inductive biases) from external parsers.
 - incorporate sentence level graph structures
 - in this work - coreference based graphs.
 - connect 2 sentences if they have coreferring mentions.



z_{ij} : co-referring mentions between S_i, S_j

StructSum - Structured Document Representations for Summarization

z_{ij} : co-referring mentions between S_i, S_j

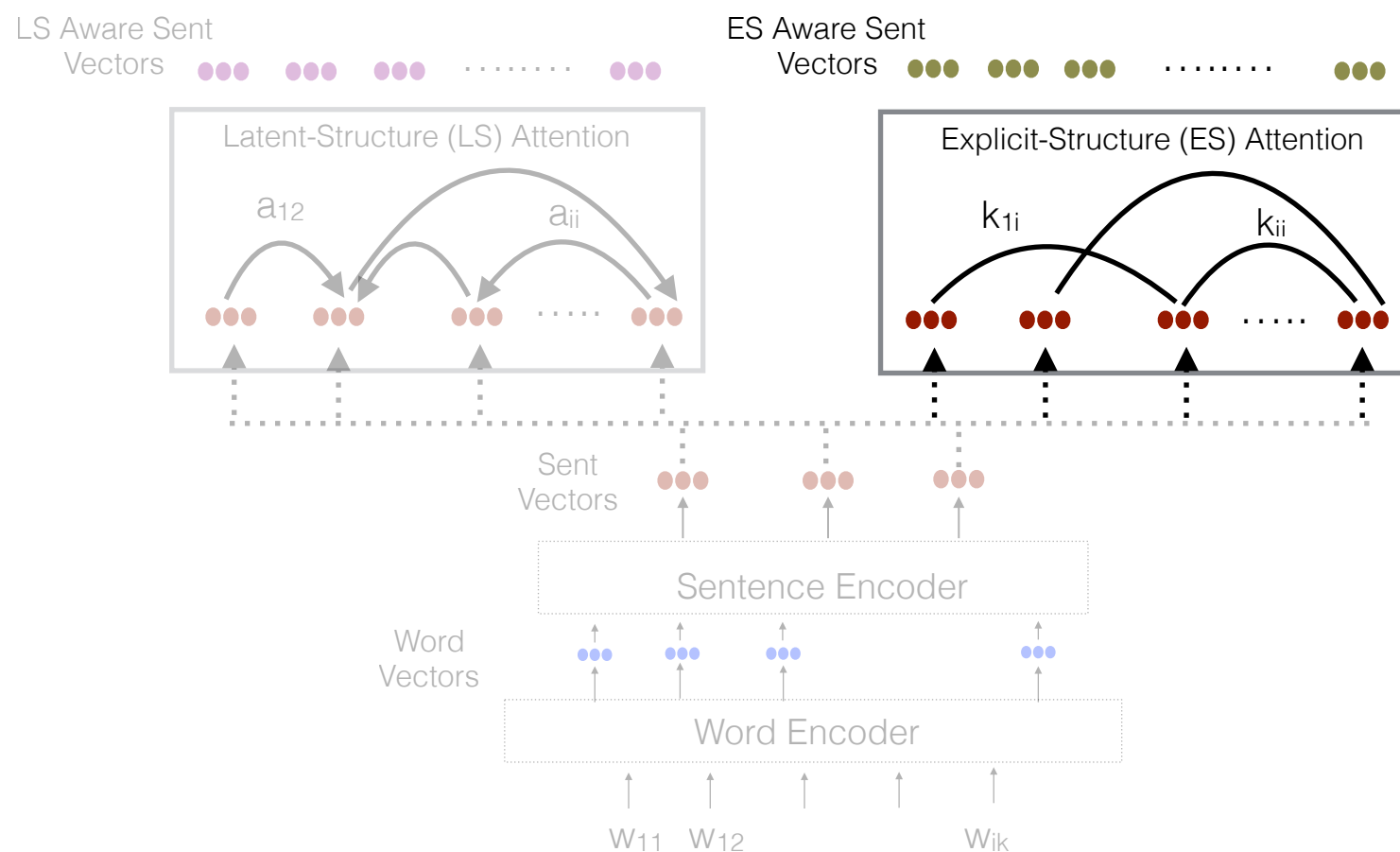


StructSum - Structured Document Representations for Summarization

z_{ij} : co-referring mentions between S_i, S_j

$$k_{ij} = p(z_{ji} = 1)$$

$$k_{ij} \propto \text{no: co-referring mentions}$$



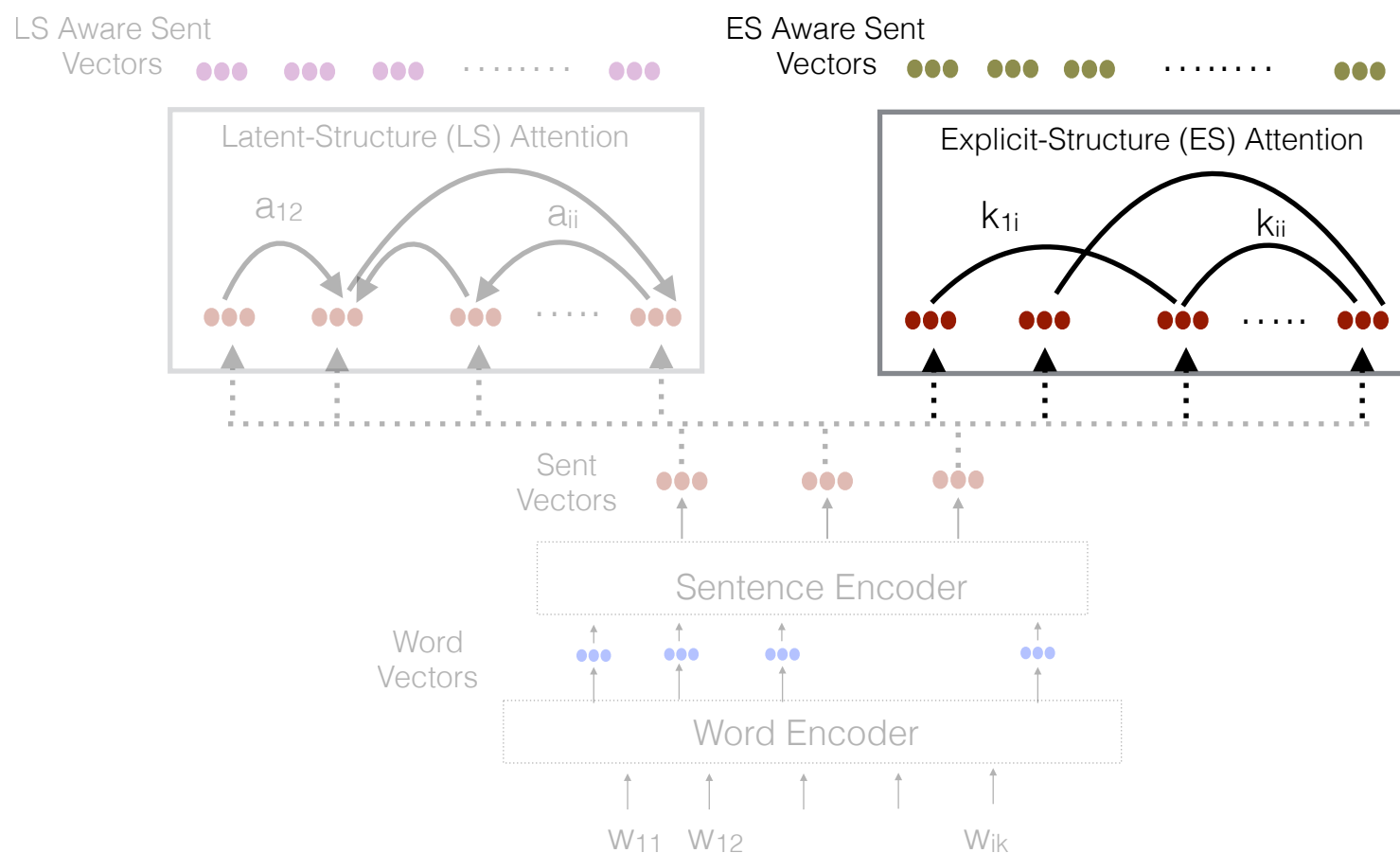
StructSum - Structured Document Representations for Summarization

z_{ij} : co-referring mentions between S_i, S_j

$$k_{ij} = p(z_{ji} = 1)$$

$$k_{ij} \propto \text{no: co-referring mentions}$$

$$k_{ij} = \frac{\text{count}(m_i \cap m_j) + \epsilon}{\sum_{v=1..n} \text{count}(m_i \cap m_v)}$$



StructSum - Structured Document Representations for Summarization

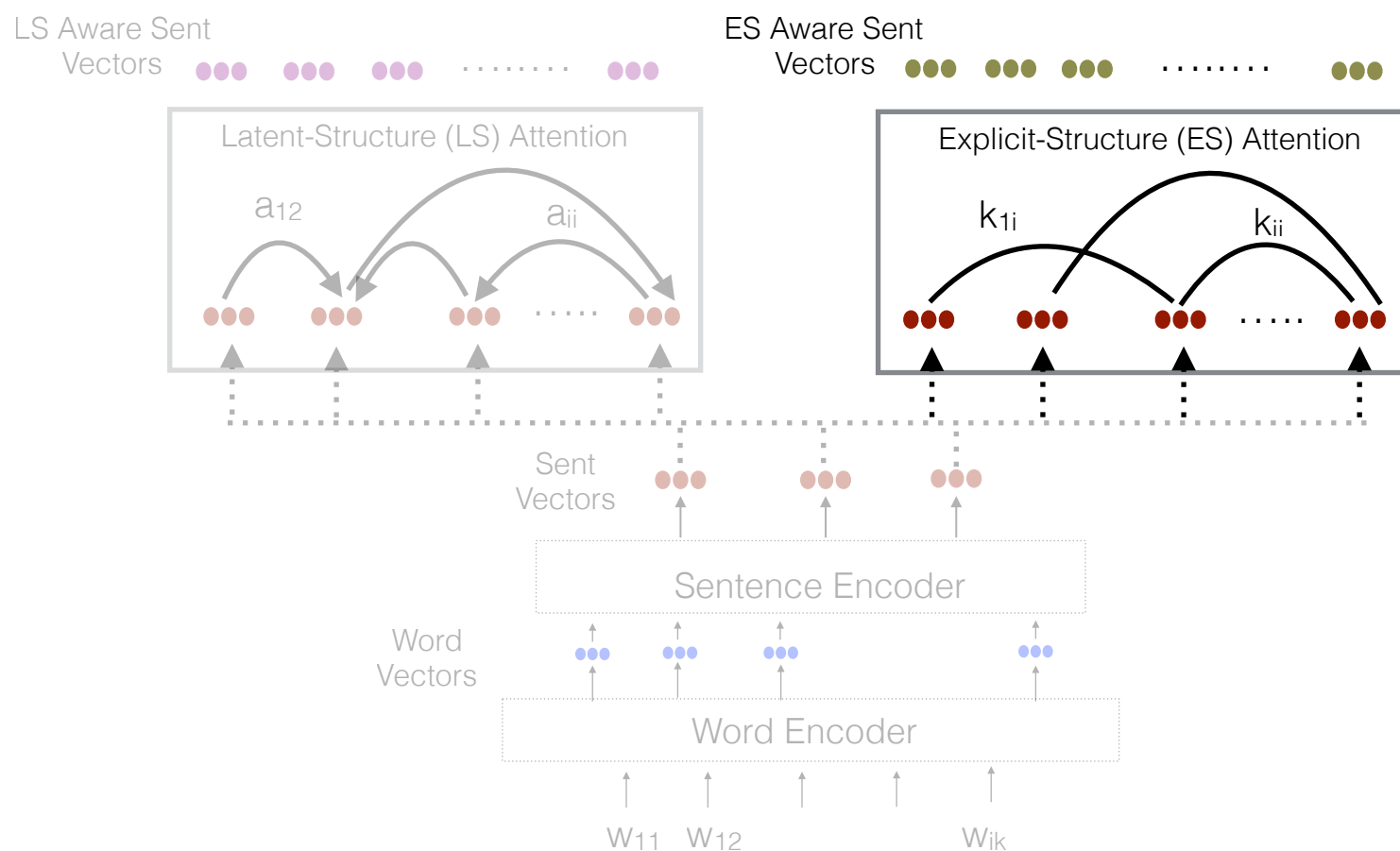
z_{ij} : co-referring mentions between S_i, S_j

$$k_{ij} = p(z_{ji} = 1)$$

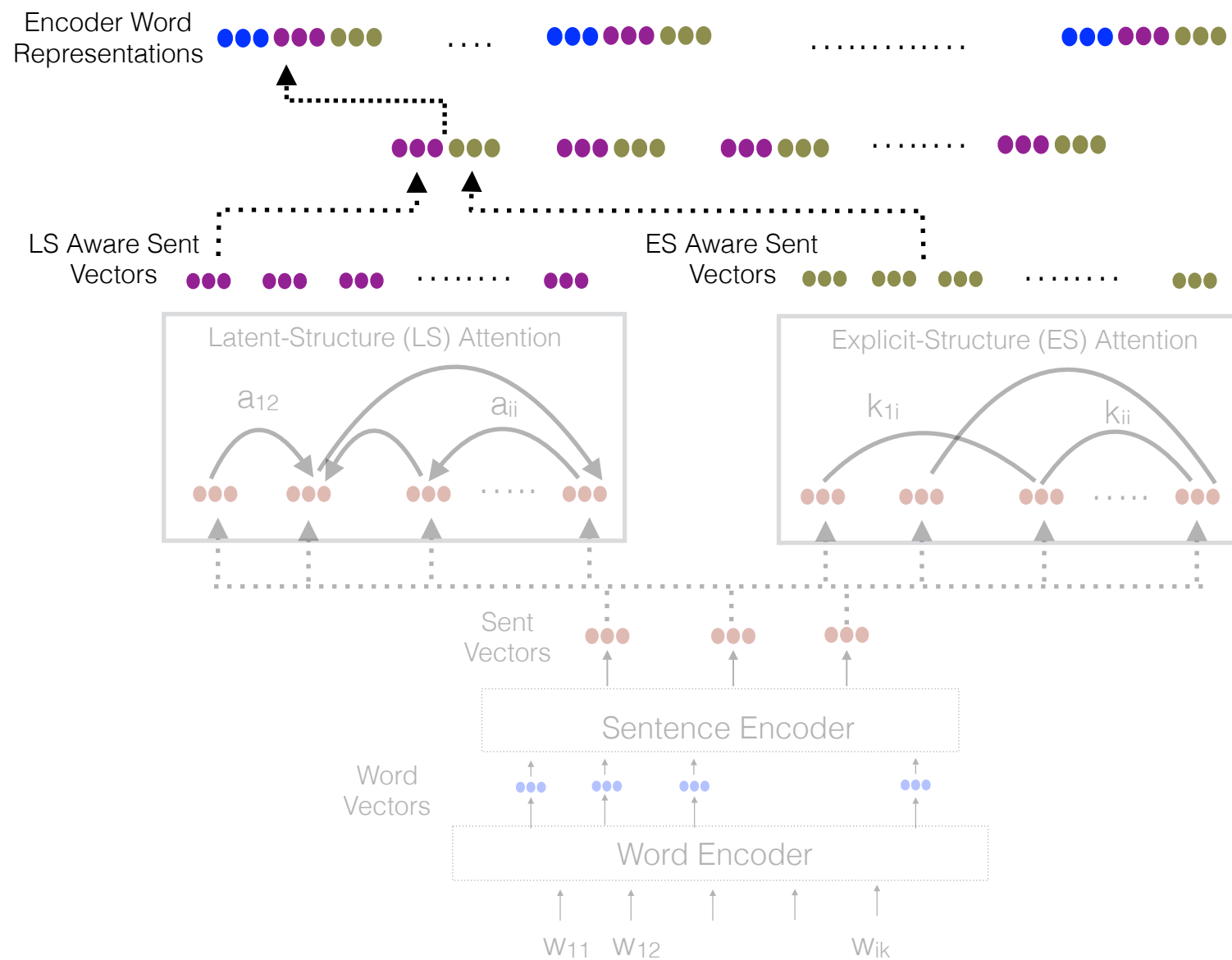
$$k_{ij} \propto \text{no: co-referring mentions}$$

$$k_{ij} = \frac{\text{count}(m_i \cap m_j) + \epsilon}{\sum_{v=1..n} \text{count}(m_i \cap m_v)}$$

$$ES_i = f(k_{ij}, s_i, s_j) \quad \forall j : 1..n$$

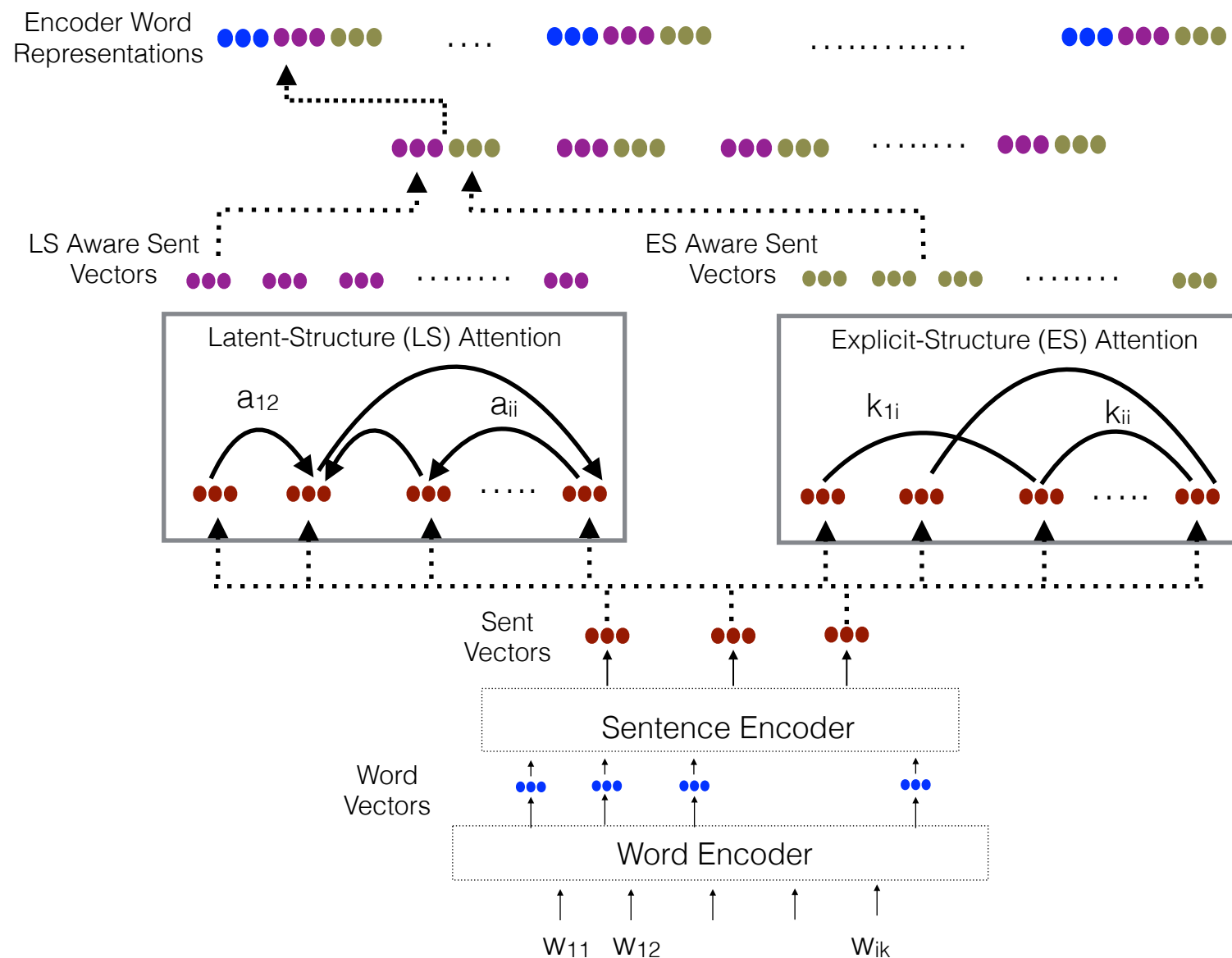


StructSum - Structured Document Representations for Summarization



- Concatenate LS_i and ES_i to corresponding word representations

StructSum - Structured Document Representations for Summarization

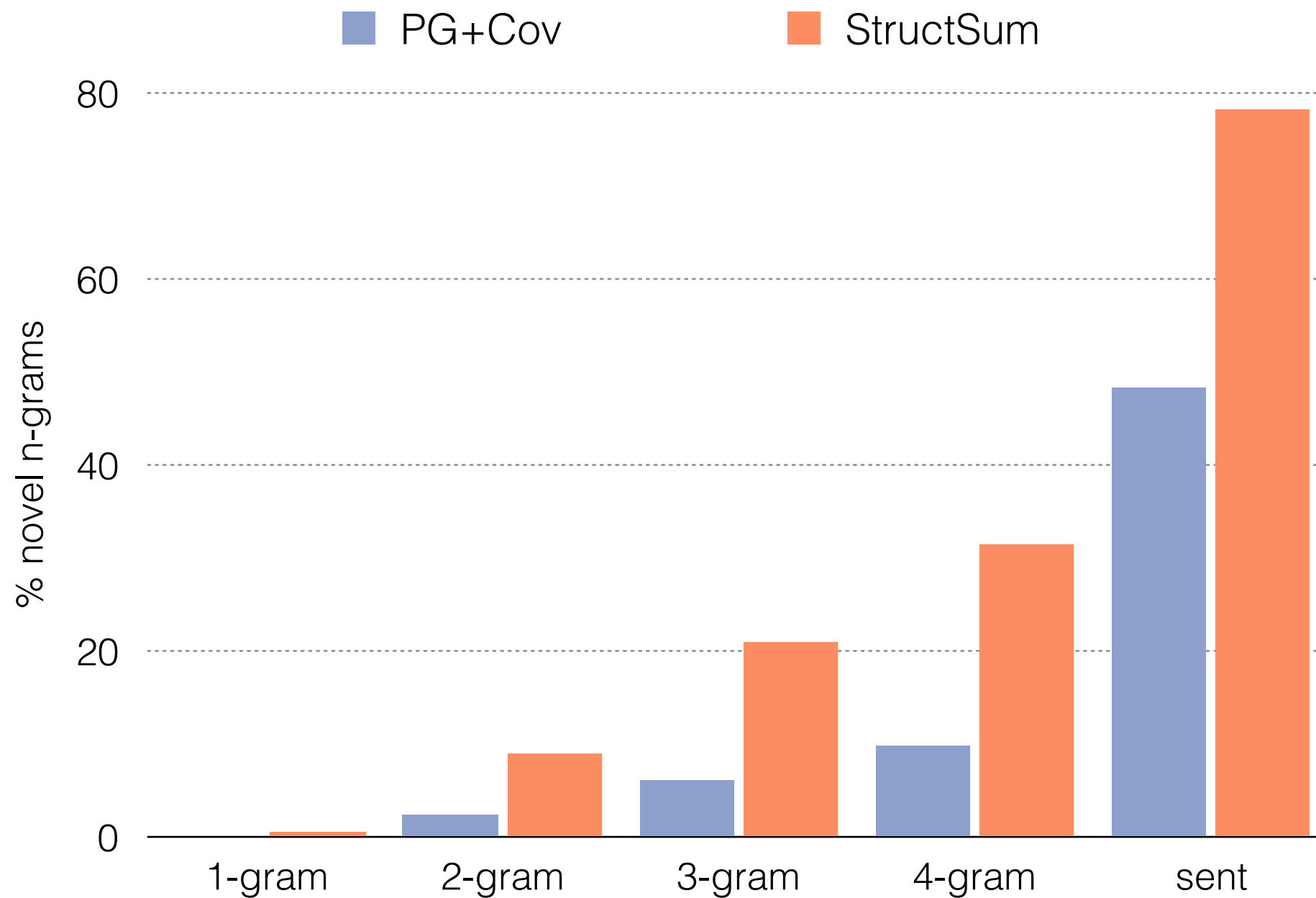


- Concatenate LS_i and ES_i to corresponding word representations
- Add the new structure-aware representations to a standard pointer-generator framework (See, et al 2017)

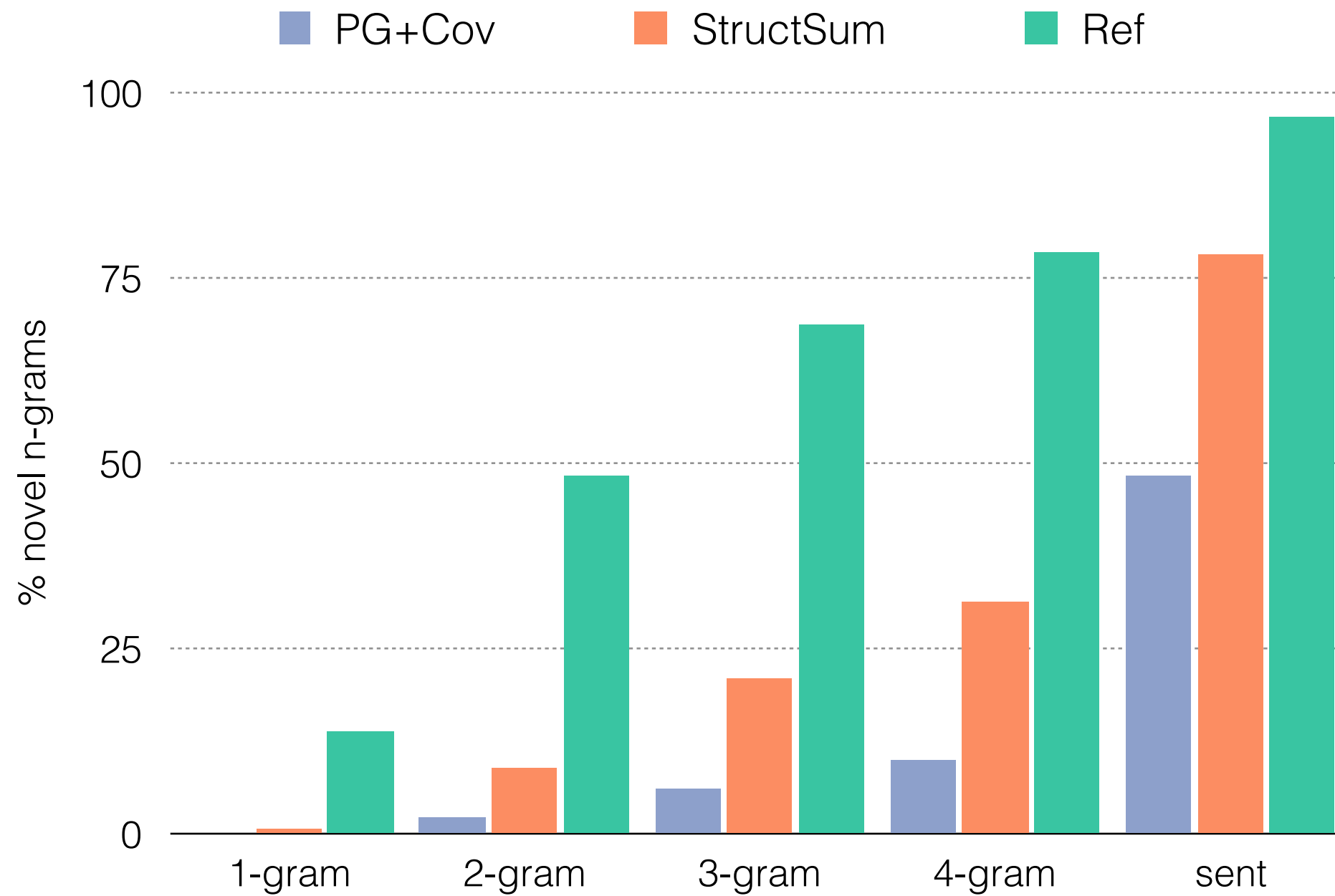
Rouge Results - CNN/DM

Model	ROUGE 1	ROUGE 2	ROUGE L
Pointer-Generator (See et al., 2017)	36.44	15.66	33.42
Pointer-Generator + Coverage (See et al., 2017)	39.53	17.28	36.38
Graph Attention (Tan et al., 2017)	38.10	13.90	34.00
Pointer-Generator + DiffMask (Gehrmann et al., 2018)	38.45	16.88	35.81
Pointer-Generator (Re-Implementation)	35.55	15.29	32.05
Pointer-Generator + Coverage (Re-Implementation)	39.07	16.97	35.87
Latent-Structure (LS) Attention	39.52	16.94	36.71
Explicit-Structure (ES) Attention	39.63	16.98	36.72
LS + ES Attention	39.62	17.00	36.95

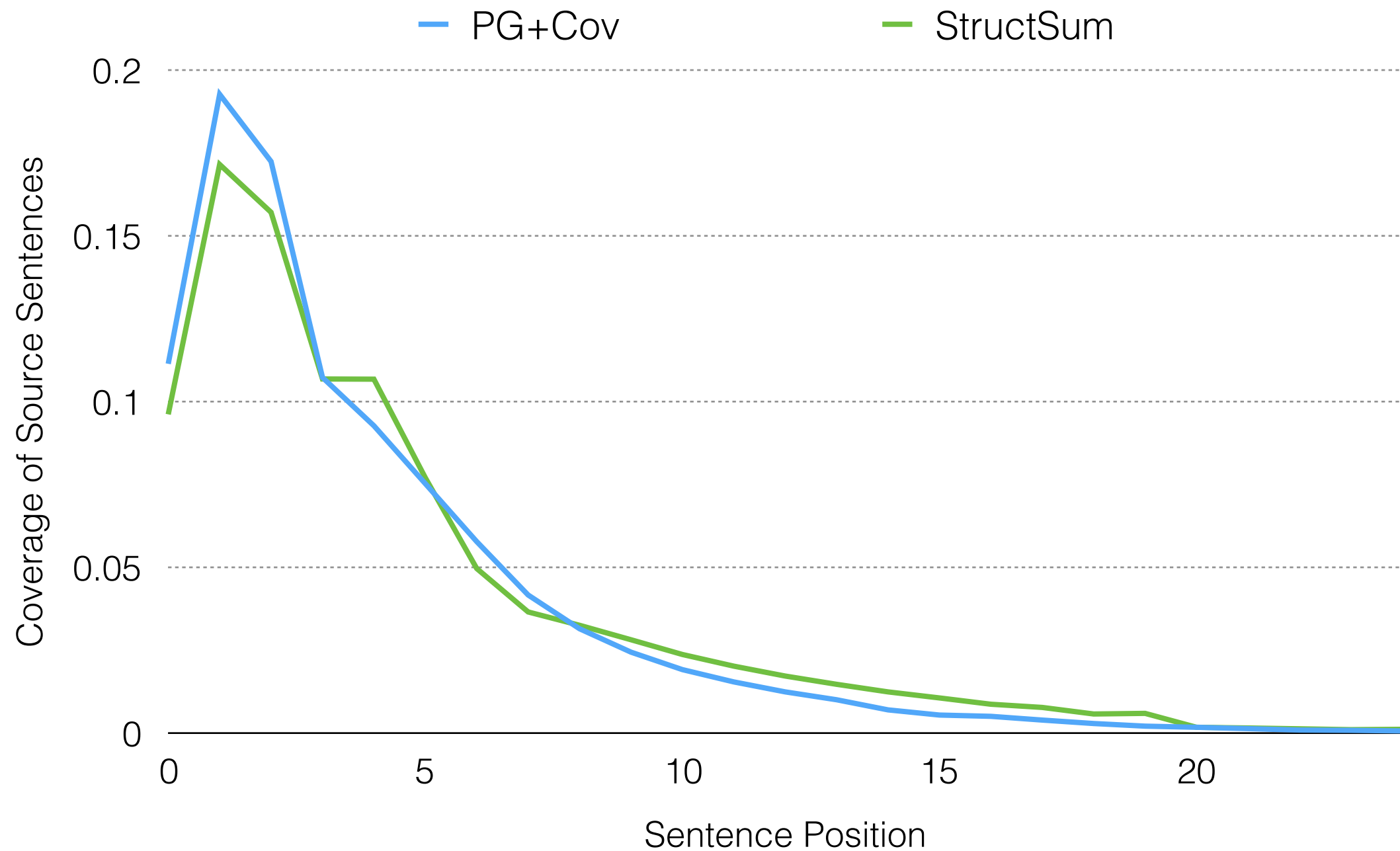
Abtractiveness - % Novel n-grams



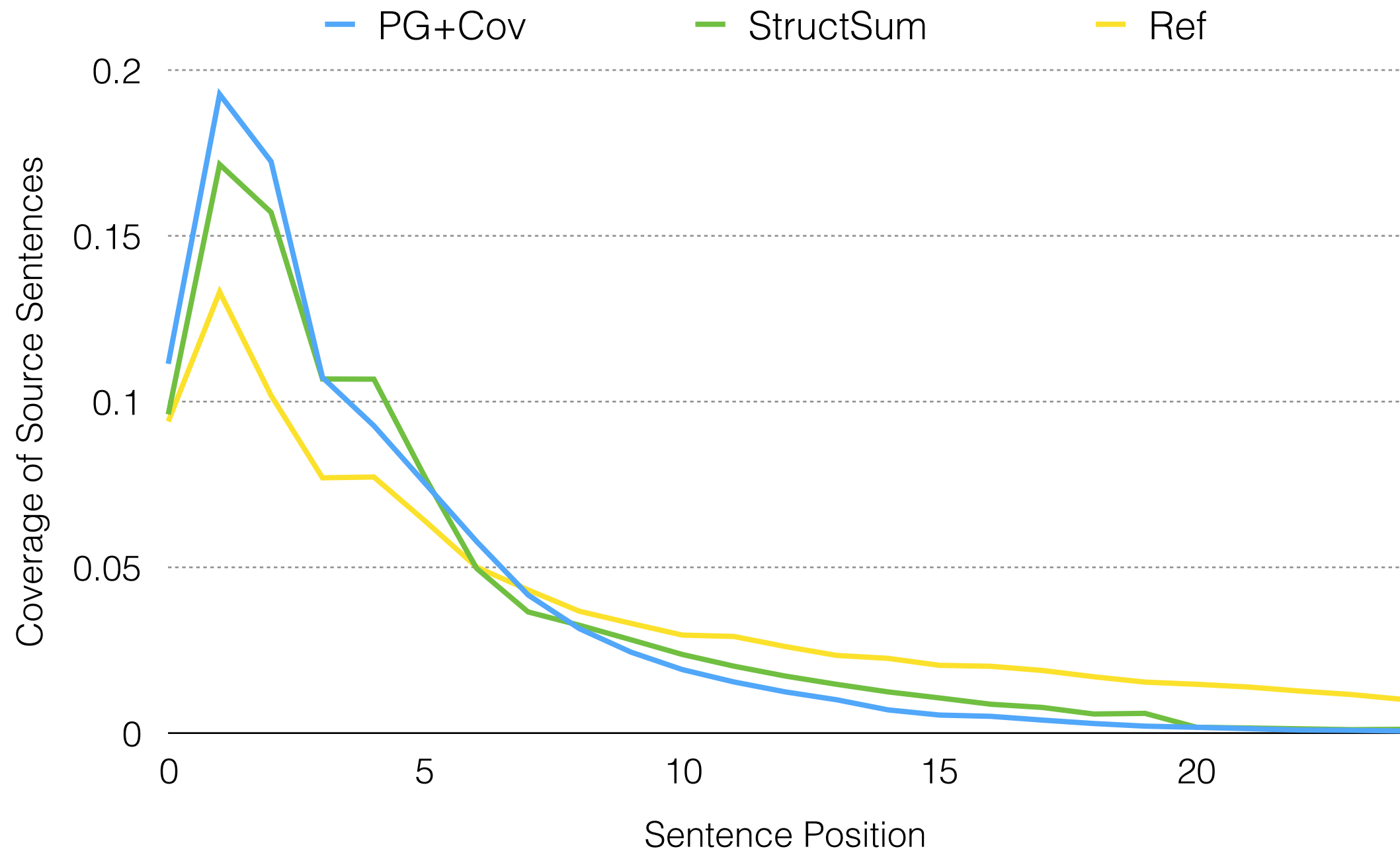
Abtractiveness - % Novel n-grams



Layout Bias - Coverage of Source Sentences



Layout Bias - Coverage of Source Sentences

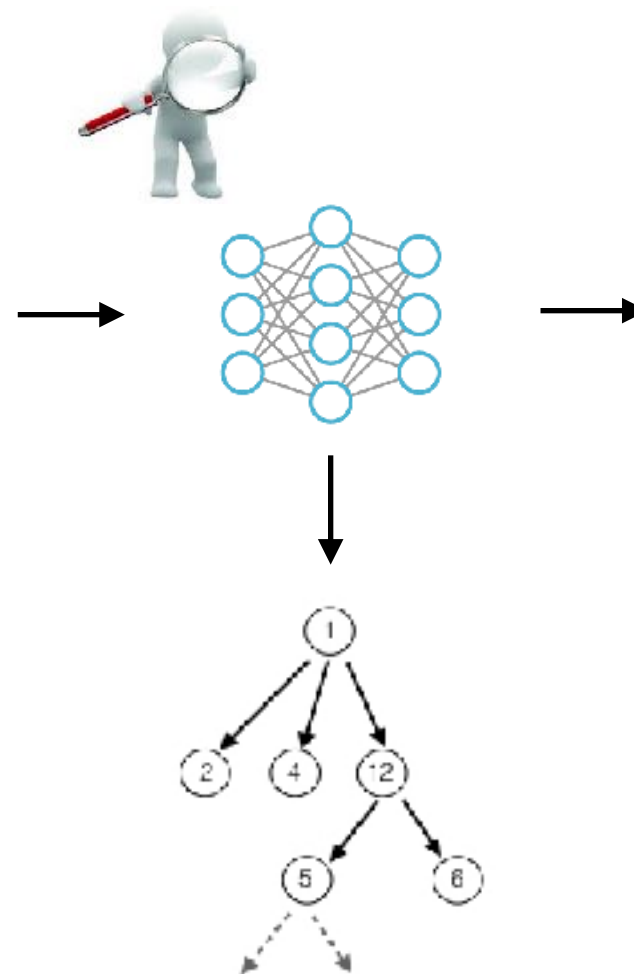


Interpretability - Analyse Latent Structures

Article : Electronic Devices allowed on Southwest

[1] Southwest Airlines has received Federal Aviation Administration approval to allow passengers to use many portable electronic devices in all phases of flight.
[2] Under the new rules, passengers may use certain electronic devices in "airplane mode" during taxiing, takeoff and landing.
[3] JetBlue Airways and Delta Air Lines moved quickly to get FAA approval to allow devices on board on November 1.
....
....
....
....
[12] The new expanded use of electronics does not apply to making or taking calls, which are still prohibited in flight.

Source Article



Neural Model +
Latent Structures

[1] Southwest is newest airline to allow use of portable electronic devices.
[2]
[3] Cell phone calls are still not permitted after the aircraft door is closed

Generated
Summary

Takeaways!

- Framework to encode latent and explicit forms of structure
- Encoding document structure useful for summarization
 - Improves abstractiveness
 - Reduces reliance of layout bias
 - Provides a means to visualize and interpret model

Today's Talk

- Framework for Incorporating Document Structure

StructSum: Summarization via Structured Representations Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, Yulia Tsvetkov. In *Proc. EACL'21*.

- Benchmark for Evaluating Factuality of Generated Summaries

Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics Artidoro Pagnoni, Vidhisha Balachandran, Yulia Tsvetkov *To Appear In NAACL'21*.

Factuality in Generated Summaries

Original: a recent poll finds that most americans feel that businesses like restaurants and event centers should not discriminate against same-sex weddings. public opinion has shifted on the issue since last fall after Indiana changed its ...

Factually Incorrect: Most americans say businesses should discriminate against same-sex weddings.



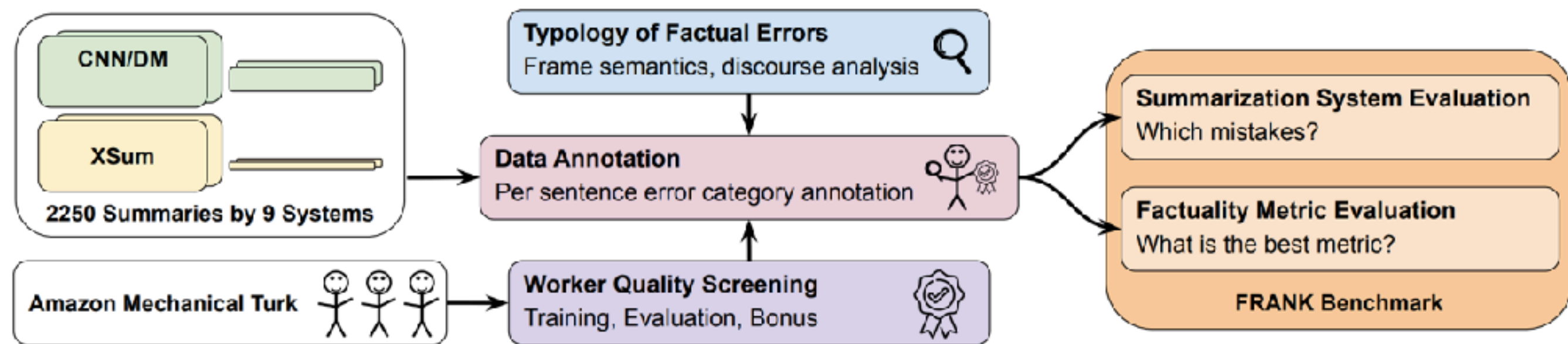
Factually correct: Most americans say businesses should not discriminate against same-sex weddings.




Research Questions

- How do we define factual errors?
- What kind of errors do different models make?
- What type of errors do various metrics capture?

Analyzing fine-grained factuality



Research Questions

- How do we define factual errors? 
- What kind of errors do different models make?
- What type of errors do various metrics capture?

Typology of Factual Errors

Semantic Frame
Errors

Discourse
Errors

Content Verifiability
Errors

Relation Error

Coreference Error

Out Of Article Error

Entity Error

Discourse Link Error

Grammatical Error

Circumstance Error

Relation Error (PredE)

Original Fact

The first vaccine for Ebola was approved by the FDA in 2019 in the US

Incorrect Fact

The Ebola vaccine **was rejected** by the FDA in 2019

Entity Error (EntE)

Original Fact

The first vaccine for Ebola was approved by the FDA in 2019 in the US

Incorrect Fact

The COVID-19 vaccine was approved by the FDA in 2019

Circumstance Error (CircE)

Original Fact

The first vaccine for Ebola was approved by the FDA in 2019 in the US

Incorrect Fact

The first vaccine for Ebola was approved by the FDA in 2014.

Coreference Error (CorefE)

Original Fact

The first vaccine for Ebola was approved by the FDA in 2019. Scientists say a vaccine for COVID-19 is unlikely to be ready this year.

Incorrect Fact

The first vaccine for Ebola was approved in 2019. **They** say a vaccine for COVID-19 is unlikely this year

Discourse Link Error (LinkE)

Original Fact

To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials

Incorrect Fact

To produce the vaccine, scientists have to show successful human trials, **then** sequence the DNA of the virus.

Out of Article Error (OutE)

Original Fact

Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.

Incorrect Fact

China has already started clinical trials of the COVID-19 vaccine.

Grammatical Error (GramE)

Original Fact

The first vaccine for Ebola was approved by the FDA in 2019 in the US.

Incorrect Fact

The Ebola vaccine **accepted have already started.**

Typology of Factual Errors

Semantic Frame
Errors

Discourse
Errors

Content Verifiability
Errors

Relation Error

Coreference Error

Out Of Article Error

Entity Error

Discourse Link Error

Grammatical Error

Circumstance Error

Typology of Factual Errors

Semantic Frame
Errors

Discourse
Errors

Content Verifiability
Errors

Relation Error

Coreference Error

Out Of Article Error

Entity Error

Discourse Link Error

Grammatical Error

Circumstance Error

Other Errors

Annotating Factual Errors

Article Text

Gaioz Nigalidze, the current Georgian champion, was expelled from the Dubai Open Chess tournament when he was found using his smartphone in the middle of a match. A chess grandmaster has been thrown out of an international tournament and faces a 15-year ban after he was caught sneaking to the toilet to check moves on his mobile phone. Gaioz Nigalidze, the current Georgian champion, was expelled from the Dubai Open Chess tournament when he was found using his phone in the middle of a match. The two-time national champion was exposed when his opponent lodged a complaint when he grew suspicious about his frequent trips to the lavatory. Tournament organisers found Nigalidze had stored a mobile phone in a cubicle, covered in toilet paper. They announced their decision to expel Nigalidze on Sunday morning on their Facebook page. The complaint was made by Nigalidze's opponent in their sixth-round match in the tournament, Armenia's Tigran Petrosian. He said: 'Nigalidze would promptly reply to my moves and then literally run to the toilet.' I noticed that he would always visit the same toilet partition, which was strange, since two other partitions weren't occupied. I informed the chief arbiter about my

Summary

UNK UNK was expelled from the dubai open chess tournament . **the two-time national champion was expelled from the dubai open chess tournament when he was found using his phone in the middle of a match .** the two-time national champion was expelled from the dubai open chess tournament when he was found using his phone in the middle of a match .

Annotating Factual Errors

Article Text

Gairoz Nigalidze, the current Georgian champion, was expelled from the Dubai Open Chess tournament when he was found using his smartphone in the middle of a match. A chess grandmaster has been thrown out of an international tournament and faces a 15-year ban after he was caught sneaking to the toilet to check moves on his mobile phone. Gairoz Nigalidze, the current Georgian champion, was expelled from the Dubai Open Chess tournament when he was found using his phone in the middle of a match. The two-time national champion was exposed when his opponent lodged a complaint when he grew suspicious about his frequent trips to the lavatory. Tournament organisers found Nigalidze had stored a mobile phone in a cubicle, covered in toilet paper. They announced their decision to expel Nigalidze on Sunday morning on their Facebook page. The complaint was made by Nigalidze's opponent in their sixth-round match in the tournament, Armenia's Tigran Petrosian. He said: 'Nigalidze would promptly reply to my moves and then literally run to the toilet.' 'I noticed that he would always visit the same toilet partition, which was strange, since two other partitions weren't occupied.' I informed the chief arbiter about my

Summary

UNK UNK was expelled from the dubai open chess tournament. **the two-time national champion was expelled from the dubai open chess tournament when he was found using his phone in the middle of a match.** the two-time national champion was expelled from the dubai open chess tournament when he was found using his phone in the middle of a match.

Question

Are the facts in the **highlighted** sentence in the summary correct?

- ☐ Yes
☒ No

Tip: Unsure about which category?

What kind of mistakes are present in the **highlighted** sentence? Select all that apply.

- ☐ Information not in article: entity or relation were not mentioned in the text.
- ☐ Grammatically meaningless: very wrong grammar cannot be understood.
- ☐ Misuse of pronoun: wrong pronoun ("he", "she", etc.) or referring expression ("the former", etc.).
- ☐ Wrong relationship between entities: what happened is wrong (typically described by the verb).
- ☐ Wrong entities in the relation: the "who", "what", "to whom", etc. is wrong. Relationship appears in the text but with different entities.
- ☐ Wrong circumstance: wrong location, time, date, goal, manner, adverbs etc.
- ☐ Wrong relationship between facts: logical or temporal link of facts is wrong.
- ☐ Other

Back

Next

Annotating Factual Errors

- Datasets
 - CNN/DM - 3 sentence summaries, less abstractive
 - XSum - 1 sentence summary, highly abstractive

Annotating Factual Errors

- Models
 - CNN/DM
 - LSTM Seq-to-Seq model (S2S) (Rush et al., 2015)
 - Pointer-Generator Network (PGN) model (See et al., 2017)
 - Bottom-Up Summarization (BUS) model (Gehrmann et al., 2018)
 - Bert Extractive-Abstractive model (BertSum) (Liu and Lapata, 2019)
 - Transformer encoder-decoder model BART (Lewis et al., 2019)
 - XSum

Annotating Factual Errors

- Models
 - CNN/DM
 - LSTM Seq-to-Seq model (S2S) (Rush et al., 2015)
 - Pointer-Generator Network (PGN) model (See et al., 2017)
 - Bottom-Up Summarization (BUS) model (Gehrmann et al., 2018)
 - Bert Extractive-Abstractive model (BertSum) (Liu and Lapata, 2019)
 - Transformer encoder-decoder model BART (Lewis et al., 2019)
 - XSum
 - Topic-Aware CNN Model (Narayan et al., 2018b)
 - Pointer-Generator Network (PGN)
 - Randomly initialized Transformer Seq2Seq (TransS2S) (Vaswani et al., 2017)
 - Initialized with Bert-Base (BertS2S) (Devlin et al., 2019)

Benchmark Statistics

- 250 instances per dataset
 - 1250 model outputs on CNN/DM
 - 1000 model outputs on XSum
 - Sentence level annotation
- 3 annotators per article, 1
- 4942 annotated sentences.

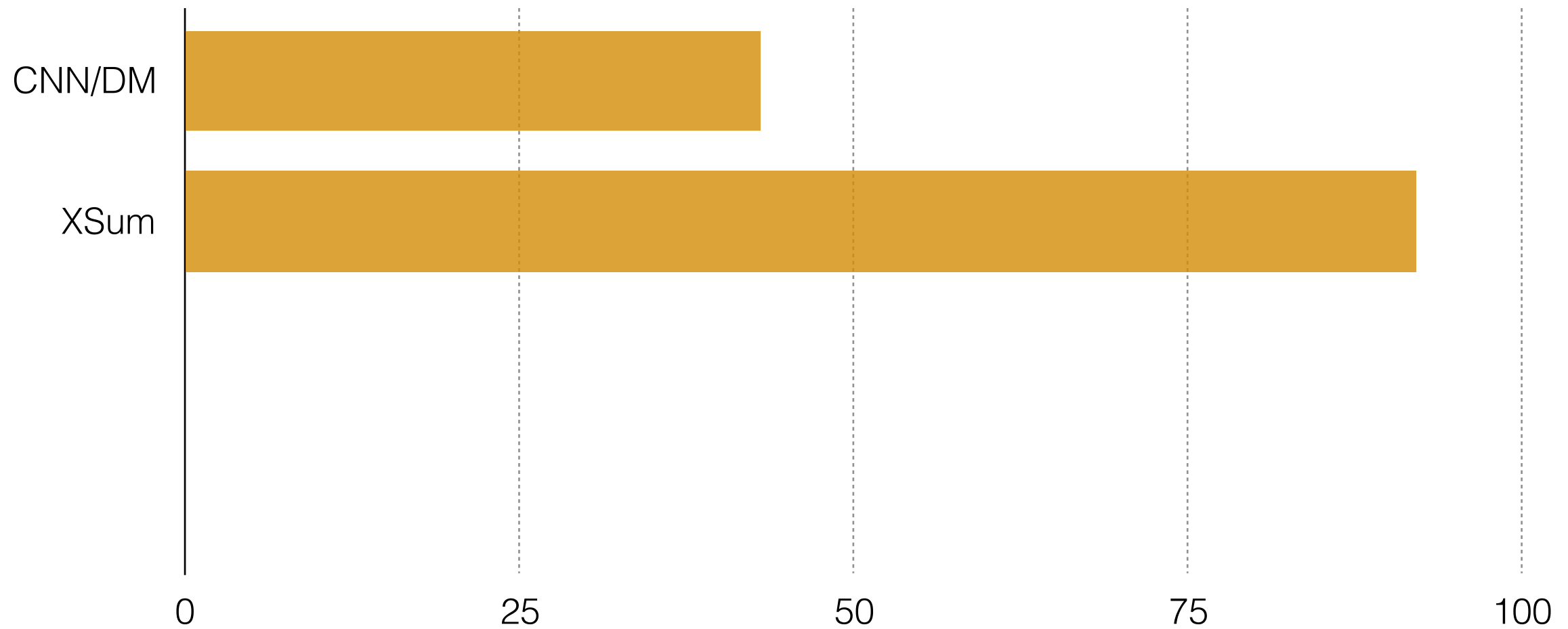
Annotation Quality

- Inter-Annotator Agreement Fleiss Kappa κ (Fleiss, 1971)
- Percentage p of annotators that agree with the majority class
- Sentence is factual or not - $\kappa = 0.58$, $p = 91\%$
- Category of error - $\kappa = 0.39$, $p = 73.9\%$
- Agreement with domain expert - Cohen Kappa $\kappa = 0.39$

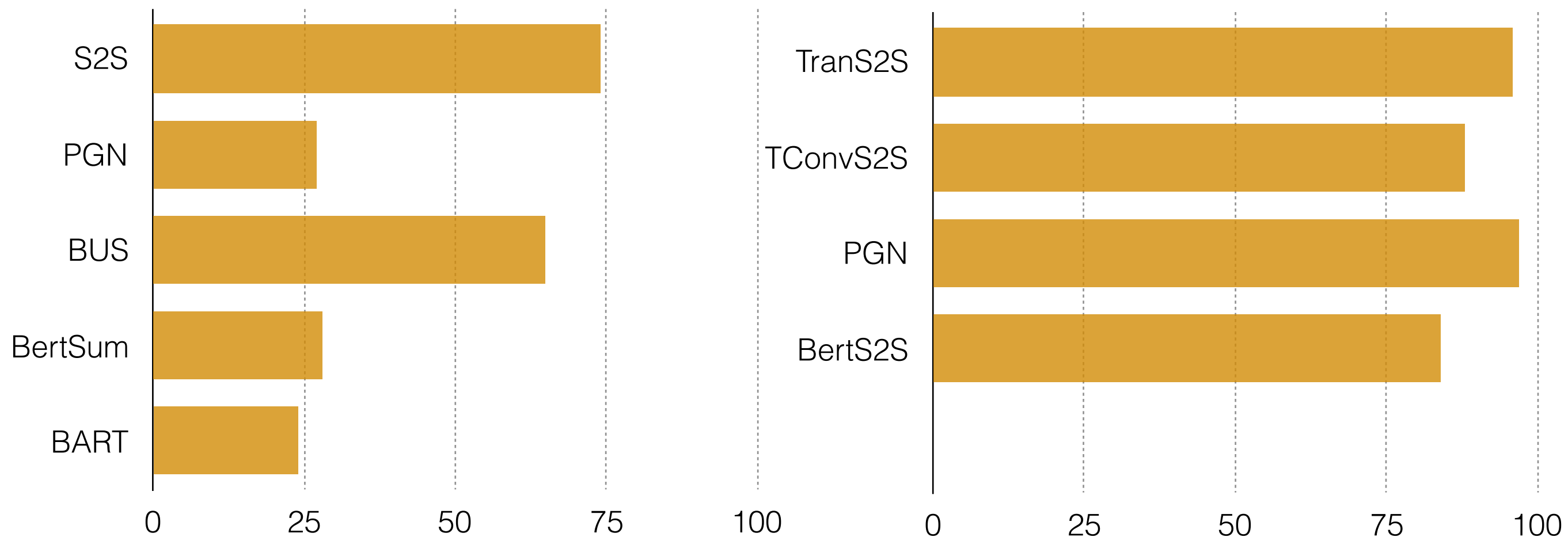
Research Questions

- How do we define factual errors?
- What kind of errors do different models make? ✓
- What type of errors do various metrics capture?

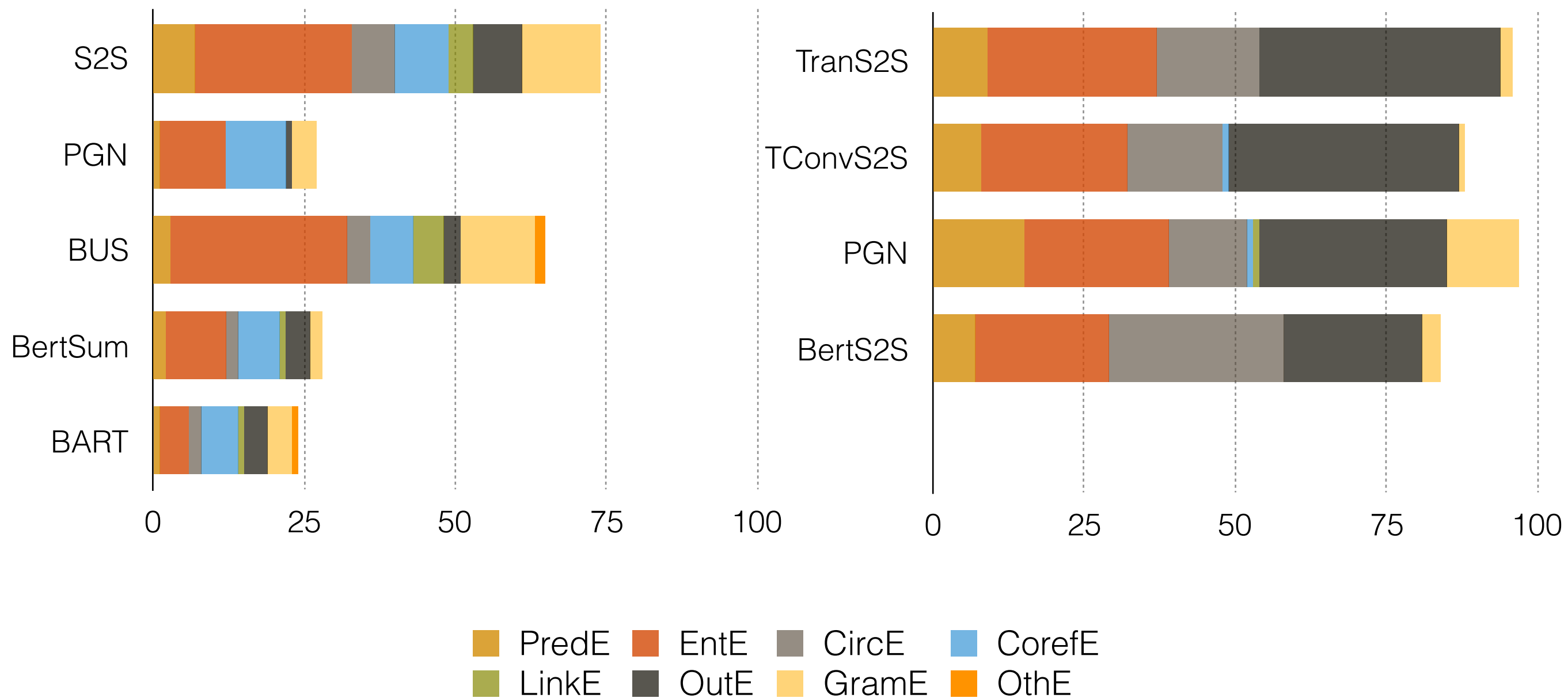
Factual Errors at Dataset Level



Factual Errors at Model Level



Factual Errors at Category Level



Findings

- Highly abstractive summaries (XSum) have high factual error rate.
- Pretrained LM based models have reduced factual error rate - but large gaps still exist.
- Entity and Circumstance errors are relatively high in all settings
- Coreference errors are high in long summaries.
- Out of Article errors are most in abstractive summaries (XSum)

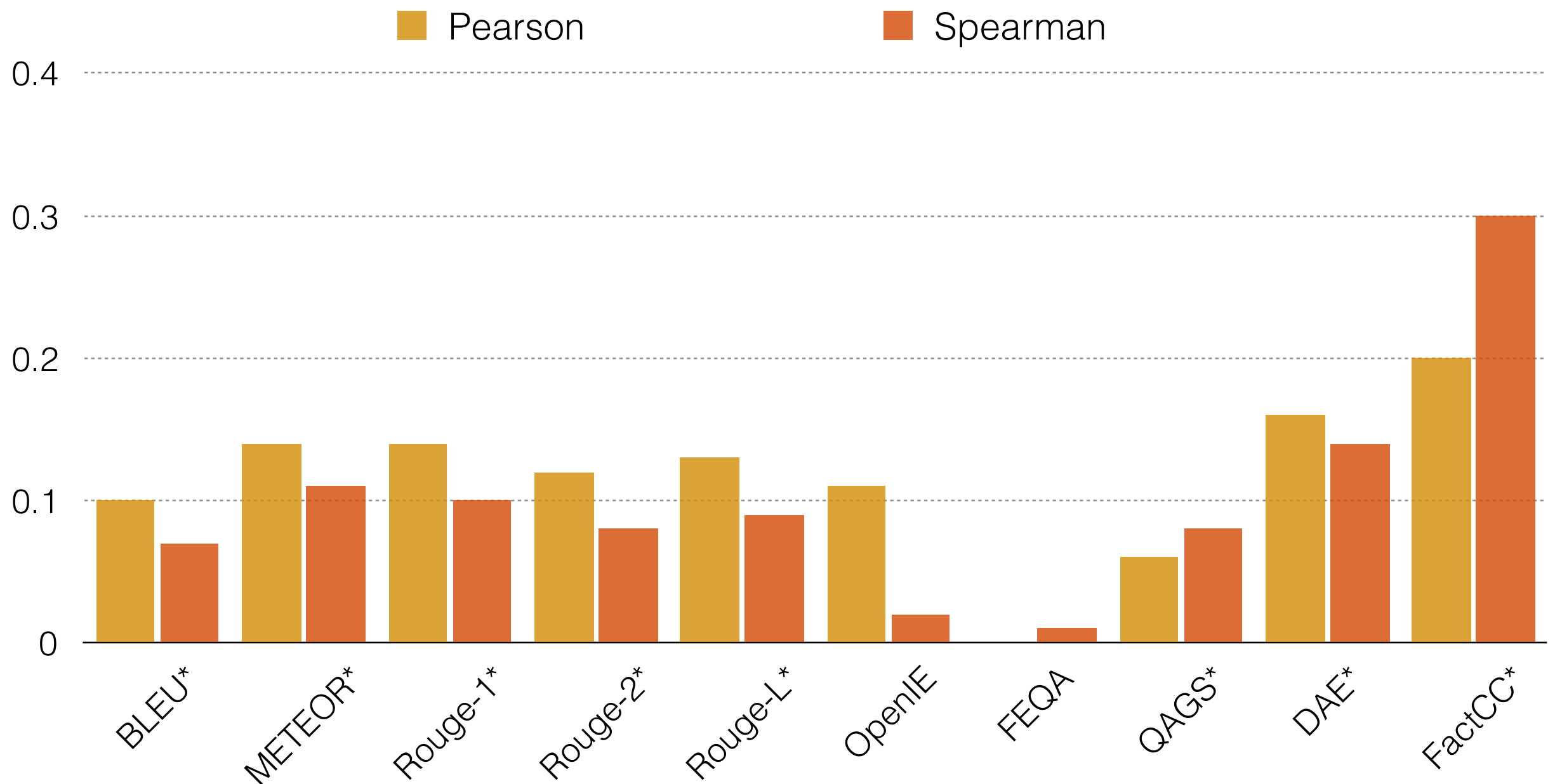
Research Questions

- How do we define factual errors?
- What kind of errors do different models make?
- What type of errors do various metrics capture? ✓

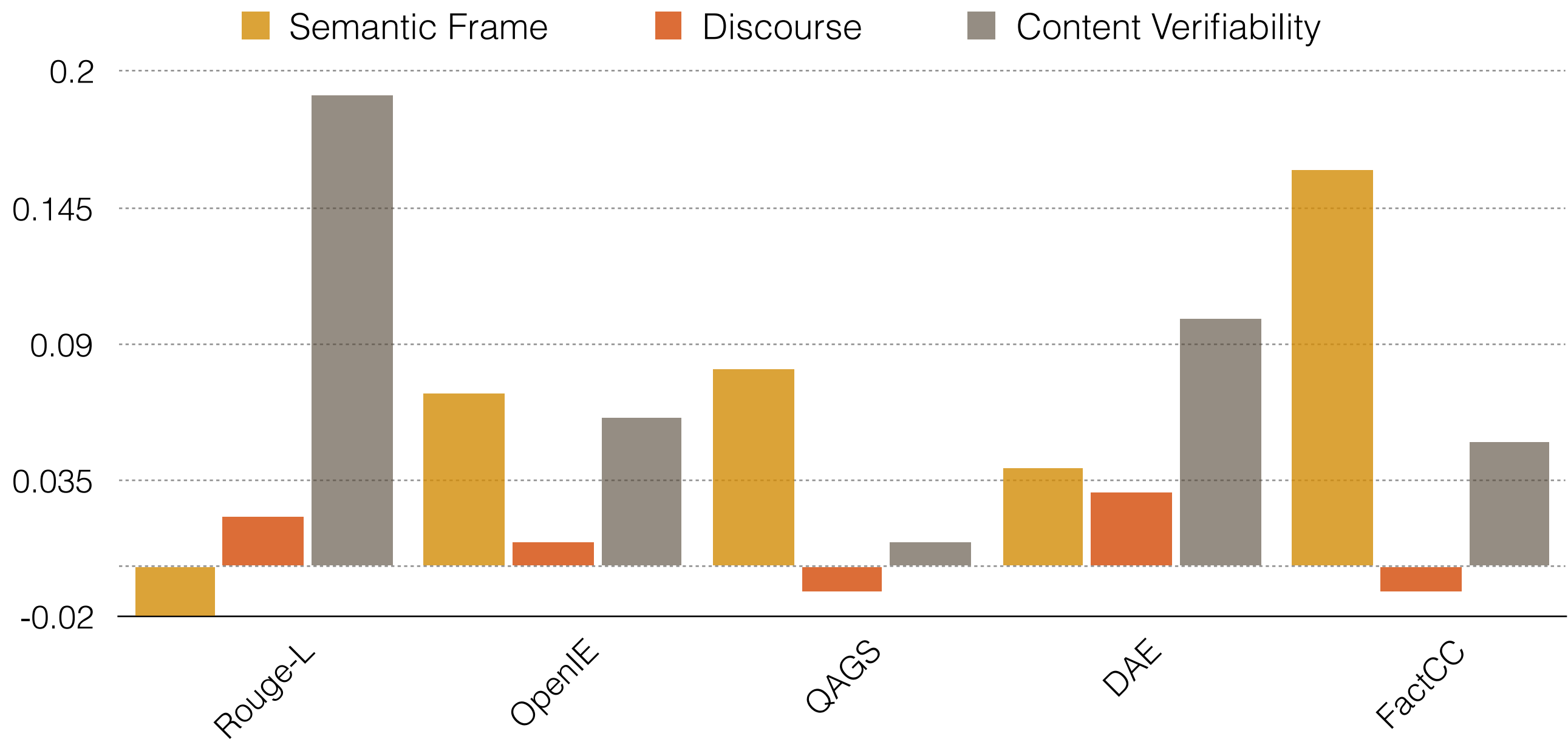
Evaluating Factuality

- Multiple recent methods for evaluating factuality
 - FactCC - Sentence level entailment based (Falke et al., 2019)
 - QAGS, FEQA - QA based (Wang et al., 2020; Durmus et al., 2020)
 - DAE - Dependency arc based (Goyal et al., 2020)

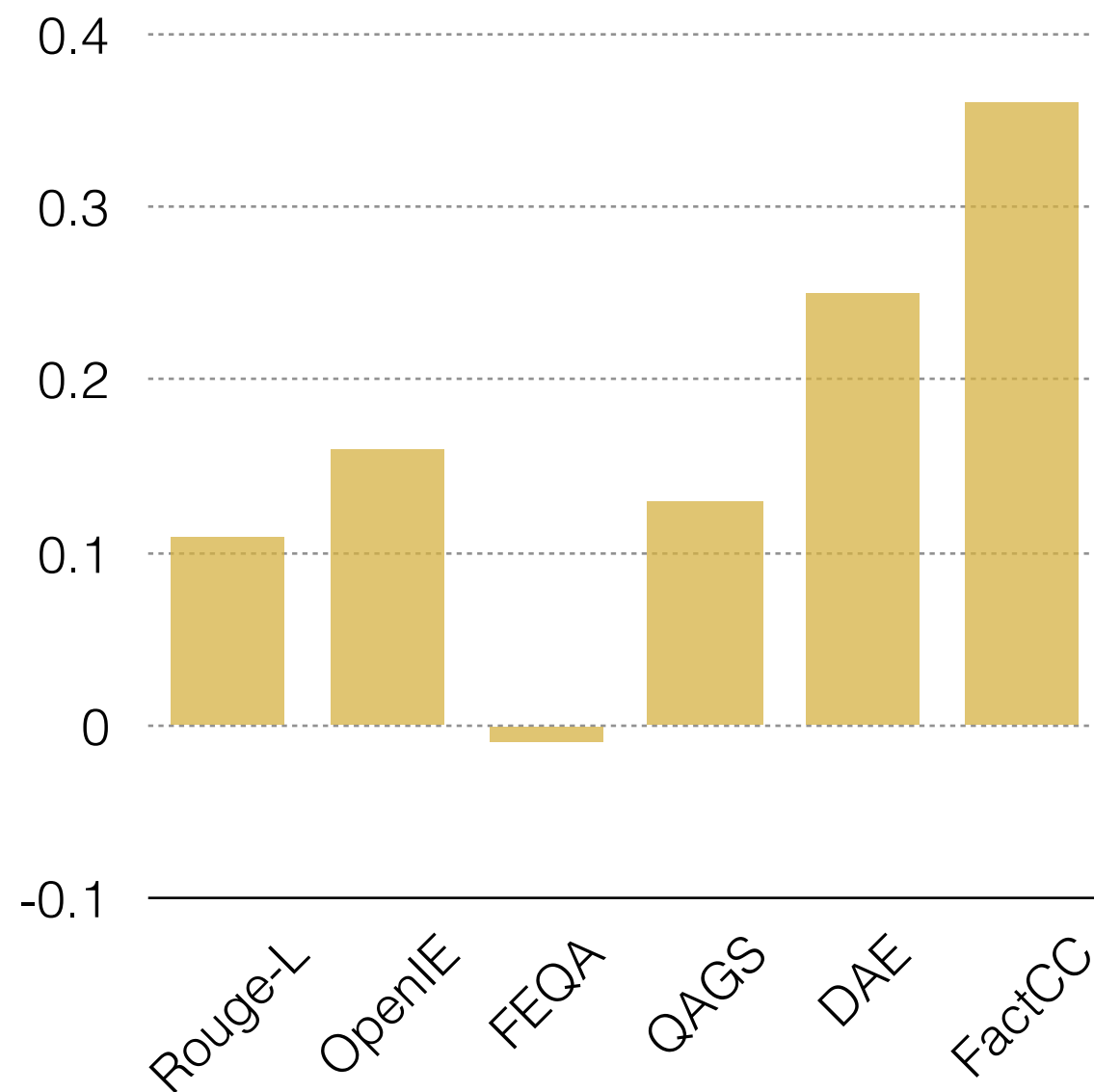
Correlation of Metrics



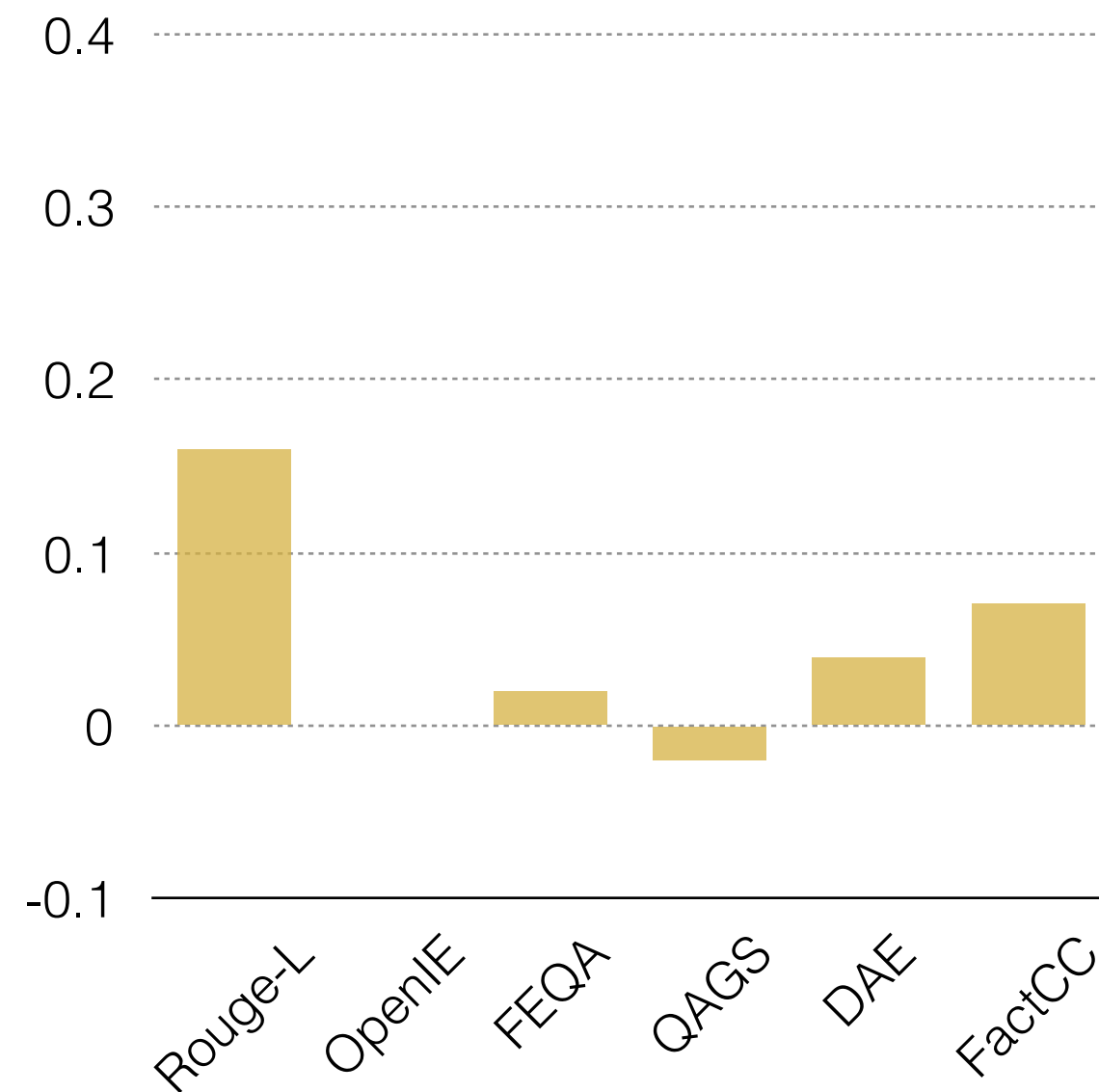
Correlation of Metrics at Category Level



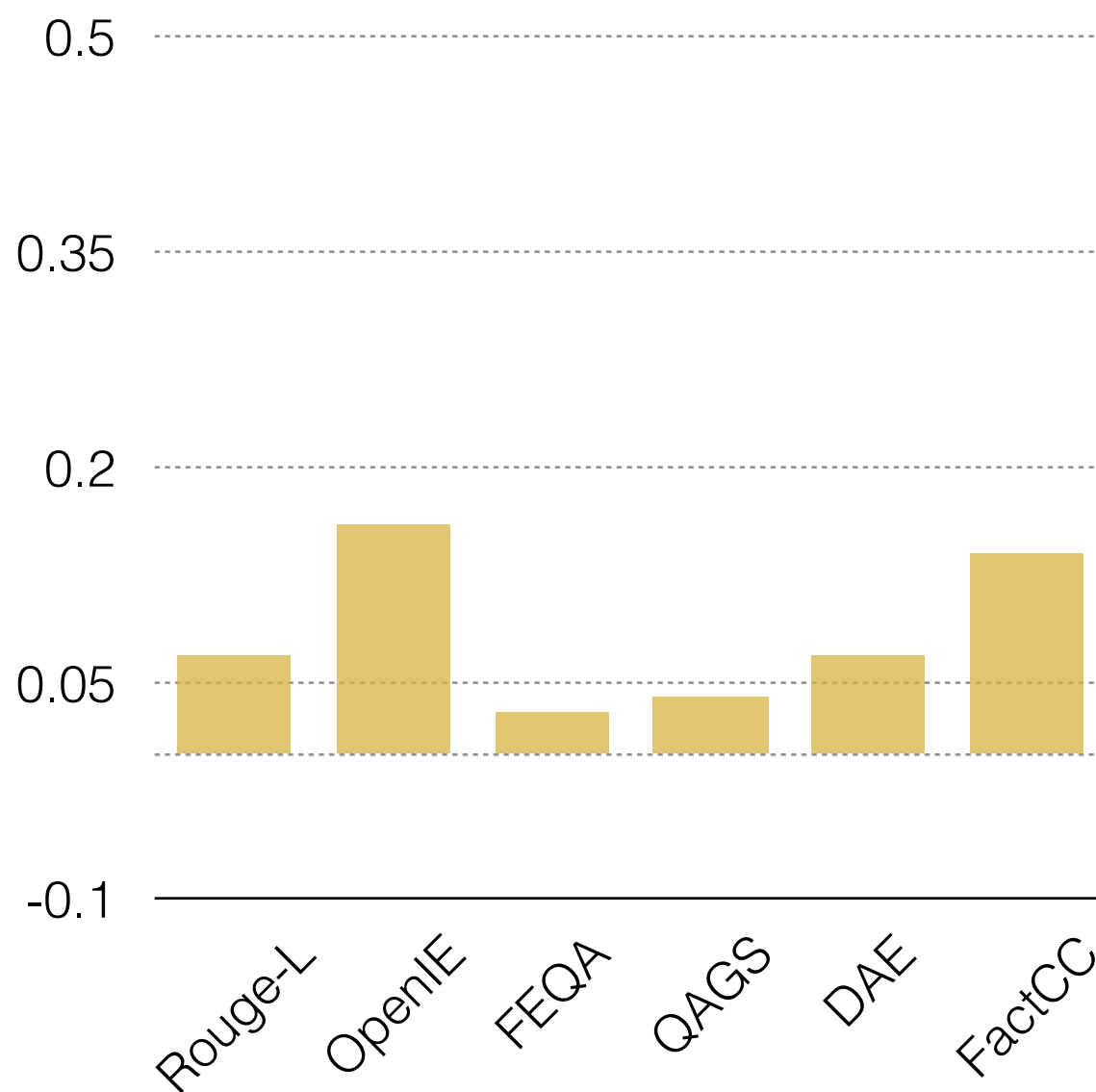
CNN/DM



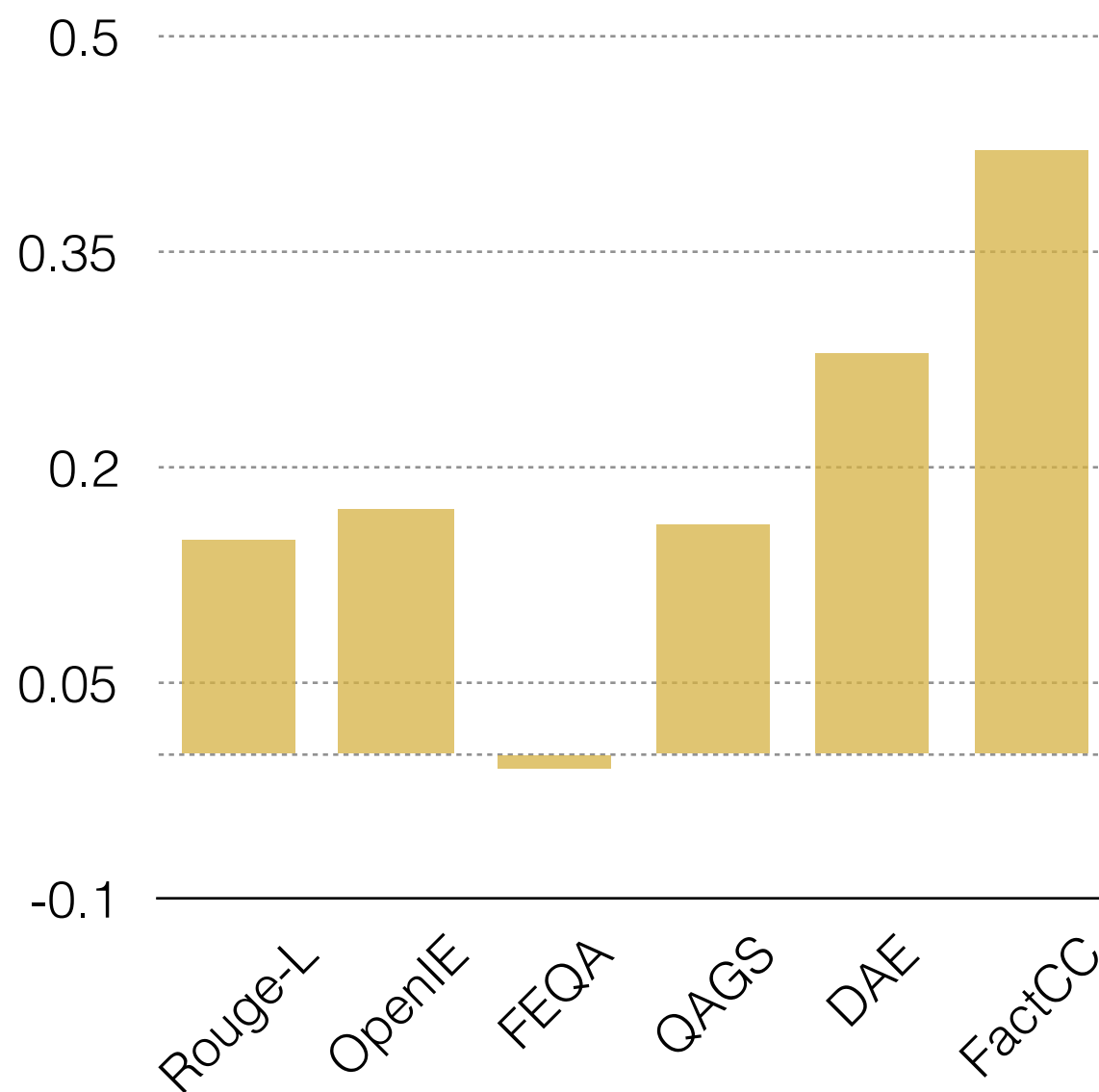
XSum



CNN/DM - pretrained



CNN/DM - non-pretrained



Findings

- Most metrics are weakly correlated with human judgements, especially for XSum.
- Most metrics capture surface level errors - weakly correlated for pretrained model errors.
- FactCC - well correlated for semantic frame errors.
- DAE - well correlated for discourse errors.
- QA methods - capture semantic errors - but weakly correlated.

Takeaways!

- Fine-grained typology of factual errors.
- Abstractive models and datasets - high error rate.
- Different datasets lead to different distribution of factual errors.
- Factuality metrics capture only specific set of errors - long way to go!

Future Directions

- Fine-grained metrics for detecting specific error types
- Improving factual correctness
- Structure Aware Representations for multilingual models

Thank You



Carnegie Mellon University
Language Technologies Institute



Thank You

Questions?

Contact: vbalacha@cs.cmu.edu



Carnegie Mellon University
Language Technologies Institute

