Northeastern University
Khoury College of
Computer Sciences

Argument Generation with Retrieval, Planning, and Realization

Xinyu Hua, Zhe Hu, and Lu Wang

Project page: https://xinyuhua.github.io/Resources/acl19

July 30, 2019
ACL
Florence, Italy

Motivation

- One of the most useful function of language is persuasion.

Motivation

- One of the most useful function of language is persuasion.
- Argumentation, in its various forms, affects critical decisions we make about our lives.

Political debate

Decision making

Legal debate

Motivation

- Constructing high quality arguments is difficult.
  - Abundant world knowledge is necessary.
  - Argumentative structure requires careful planning.

Motivation

- Constructing high quality arguments is difficult.
  - Abundant world knowledge is necessary.
  - Argumentative structure requires careful planning.
- Recent deep neural generation models have shown impressive ability in creating long fluent text.

GPT-2's take on argumentative text

I believe death penalty should be abolished.

Credit: OpenAI GPT-2 https://talktotransformer.com/

GPT-2's take on argumentative text

I believe death penalty should be abolished. But the only place we can go is to the bar."

He said he supports the right to bear arms, and called on people to exercise their constitutionally guaranteed right to vote.

"Some of the people protesting in our city were actually protesting for their mother's rights, her rights to be able to walk the streets and not be threatened with a death sentence," Mr. O'Brien said.

Credit: OpenAI GPT-2 https://talktotransformer.com/

Motivation

- Human-level argumentation is still hard to generate with current conditional LM based approaches.

Motivation

- Human-level argumentation is still hard to generate with current conditional LM based approaches.
  - World knowledge has to be acquired only from training data.

Motivation

- Human-level argumentation is still hard to generate with current conditional LM based approaches.
  - World knowledge has to be acquired only from training data.
  - Token level sequence models do not produce explicit text plans.

Proposal

I believe death penalty should be abolished.

Proposal

I believe death penalty should be abolished.

Retrieval

THE WALL STREET JOURNAL.

The New York Times

WIKIPEDIA
The Free Encyclopedia

The problem of innocence in death penalty cases

The evidence in death penalty cases is not always very strong.

The Grim Facts About Lethal Injection

Our justice system is a joke and we are asking other people…

List of exonerated death row inmates.

There had been 156 exonerations of prisoners on death row

Proposal

I believe death penalty should be abolished.

Text planning

- ❑ imperfect justice system
- ❑ unreliable evidence
- ❑ wrongful conviction
- ❑ execution of innocent people
- ❑ little effect on crime rate
- ❑ <STOP>

Proposal

I believe death penalty should be abolished.

Sentence 1: [unreliable evidence; wrongful conviction];

Text planning

- ❑ imperfect justice system
- ✓ ❑ unreliable evidence
- ✓ ❑ wrongful conviction
- ❑ execution of innocent people
- ❑ little effect on crime rate
- ❑ <STOP>

Proposal

I believe death penalty should be abolished.

Text planning

- ☑ imperfect justice system
- ☐ unreliable evidence
- ☐ wrongful conviction
- ☑ execution of innocent people
- ☐ little effect on crime rate
- ☐ <STOP>

Sentence 1: [unreliable evidence; wrongful conviction];

Sentence 2: [imperfect justice system; execution of innocent people];

Proposal

I believe death penalty should be abolished.

Text planning

- ❑ imperfect justice system
- ❑ unreliable evidence
- ❑ wrongful conviction
- ❑ execution of innocent people
- ✓ ❑ little effect on crime rate
- ❑ <STOP>

Sentence 1: [unreliable evidence; wrongful conviction];

Sentence 2: [imperfect justice system; execution of innocent people];

Sentence 3: [little effect on crime rate]

Proposal

I believe death penalty should be abolished.

Text planning

- ❑ imperfect justice system
- ❑ unreliable evidence
- ❑ wrongful conviction
- ❑ execution of innocent people
- ❑ little effect on crime rate
- ☑ <STOP>

Sentence 1: [unreliable evidence; wrongful conviction];

Sentence 2: [imperfect justice system; execution of innocent people];

Sentence 3: [little effect on crime rate]

Proposal

I believe death penalty should be abolished.

Content realization

Sentence 1: [unreliable evidence; wrongful conviction];

Sentence 2: [imperfect justice system; execution of innocent people];

Sentence 3: [little effect on crime rate]

Proposal

I believe death penalty should be abolished.
Unreliable evidence might be used when there is no live witness, which results in **wrongful conviction**.

Content realization

Sentence 1: [**unreliable evidence; wrongful conviction**];

Sentence 2: [imperfect justice system; execution of innocent people];

Sentence 3: [little effect on crime rate]

Proposal

I believe death penalty should be abolished.
Unreliable evidence might be used when there is no live witness, which results in wrongful conviction. Our **justice system is not perfect**, court could order the **execution of innocent people**.

Content realization

Sentence 1: [unreliable evidence; wrongful conviction];

Sentence 2: [**imperfect justice system; execution of innocent people**];

Sentence 3: [little effect on crime rate]

Proposal

I believe death penalty should be abolished.
Unreliable evidence might be used when there is no live witness, which results in wrongful conviction. Our justice system is not perfect, court could order the execution of innocent people. Lastly, study has shown that death penalty **does not reduce crime rate**.

Content realization

Sentence 1: [unreliable evidence; wrongful conviction];

Sentence 2: [imperfect justice system; execution of innocent people];

Sentence 3: [**little effect on crime rate**]

Proposal

Content realization

I believe death penalty should be abolished.
Unreliable evidence might be used when there is no live witness, which results in wrongful conviction. Our justice system is not perfect, court could order the execution of innocent people. Lastly, study has shown that death penal

Sentence 1: [unreliable evidence; wrongful conviction];

Sentence 2: [imperfect justice system; execution of innocent people];

e

CANDELA: Counter-Arguments with two step Neural Decoders and ExternaL knowledge Augmentation

Roadmap

- Prior Work
- Argument Retrieval
- Argument Generation Model
- Experiments
- Conclusion

Prior Work

- Rule-based models [Reed, Long, and Fox, 1996; Carenini and Moore, 2000]
- Retrieval-based systems [Sato et al., 2015; Reisert et al., 2015; Yanase et al., 2015]
- Concept-to-text generation [Moryossef, Goldberg, and Dagan, 2019; Koncel-Kedziorski et al., 2019; Le et al., 2018; Wiseman, Shieber, and Rush, 2017]

Roadmap

- Prior Work
- Argument Retrieval
- Argument Generation Model
- Experiments
- Conclusion

Argument Retrieval

- Goal: to collect both subjective and factual external resources that can form the argument talking points

Argument Retrieval

- Goal: to collect both subjective and factual external resources that can form the argument talking points
- Indexed data:

Source	# documents
Wikipedia	5,743,901
Washington Post	1,109,672
The New York Times	1,952,446
Reuters	1,052,592
Wall Street Journal	2,059,128
Total	**11,917,739**

Argument Retrieval

- Goal: to collect both subjective and factual external resources that can form the argument talking points
- Indexed data:

Objective, fact-based

Source	# documents
Wikipedia	5,743,901
Washington Post	1,109,672
The New York Times	1,952,446
Reuters	1,052,592
Wall Street Journal	2,059,128
Total	**11,917,739**

Argument Retrieval

- Goal: to collect both subjective and factual external resources that can form the argument talking points
- Indexed data:

Objective, fact-based

Left

Right

By https: //www.adfontesmedia.com/

Source	# documents
Wikipedia	5,743,901
Washington Post	1,109,672
The New York Times	1,952,446
Reuters	1,052,592
Wall Street Journal	2,059,128
Total	**11,917,739**

Ranking and Filtering

- **Step 1**: Documents are breakdown into passages (of 3 sentences).

> The American death penalty has a big innocence problem, and it is not going away. The events of last week show why.
> On Wednesday, Missouri planned to execute Marcellus Williams. The problem was that he may be innocent. Governor Eric Greitens wisely put that execution on hold while a panel investigates further. On Thursday, Florida did execute Mark Asay. We may never fully know whether he actually deserved the death penalty…

Ranking and Filtering

- **Step 1**: Documents are breakdown into passages (of 3 sentences).

> The American death penalty has a big innocence problem, and it is not going away. The events of last week show why.
> On Wednesday, Missouri planned to execute Marcellus Williams. The problem was that he may be innocent. Governor Eric Greitens wisely put that execution on hold while a panel investigates further. On Thursday, Florida did execute Mark Asay. We may never fully know whether he actually deserved the death penalty…

Ranking and Filtering

- **Step 1**: Documents are breakdown into passages (of 3 sentences).

The American death penalty has a big innocence problem, and it is not going away. The events of last week show why.
On Wednesday, Missouri planned to execute Marcellus Williams. The problem was that he may be innocent. Governor Eric Greitens wisely put that execution on hold while a panel investigates further. On Thursday, Florida did execute Mark Asay. We may never fully know whether he actually deserved the death penalty…

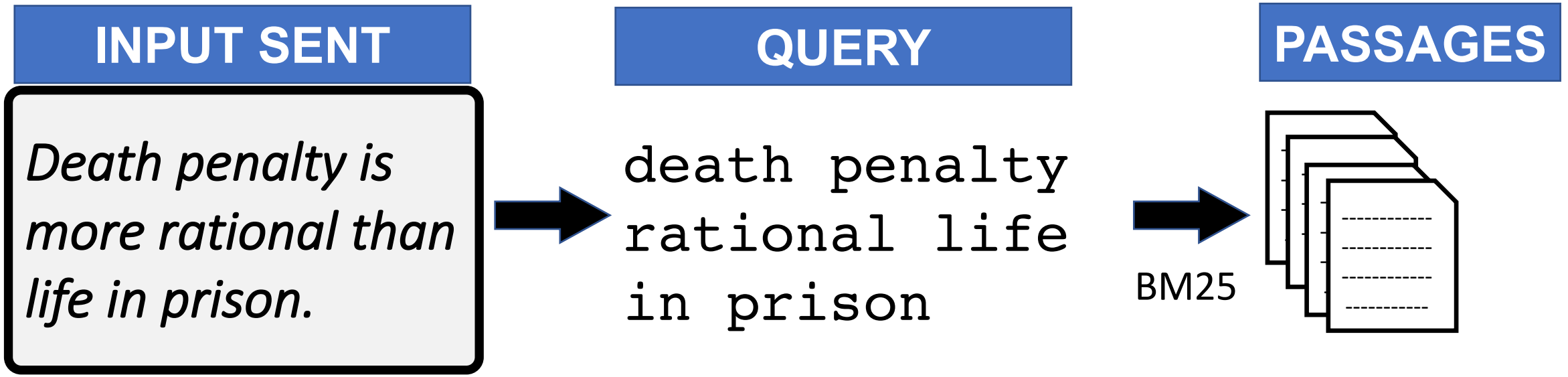
Ranking and Filtering

- **Step 1**: Documents are breakdown into passages (of 3 sentences).

The American death penalty has a big innocence problem, and it is not going away. The events of last week show why.
On Wednesday, Missouri planned to execute Marcellus Williams. The problem was that he may be innocent. Governor Eric Greitens wisely put that execution on hold while a panel investigates further. On Thursday, Florida did execute Mark Asay. We may never fully know whether he actually deserved the death penalty…

Ranking and Filtering

- **Step 1**: Documents are breakdown into passages (of 3 sentences).
- **Step 2**: Passages are retrieved and ranked based on input queries.



Ranking and Filtering

- **Step 1**: Documents are breakdown into passages (of 3 sentences).
- **Step 2**: Passages are retrieved and ranked based on input queries.
- **Step 3**: Passages with wrong stance are discarded.

INPUT SENT

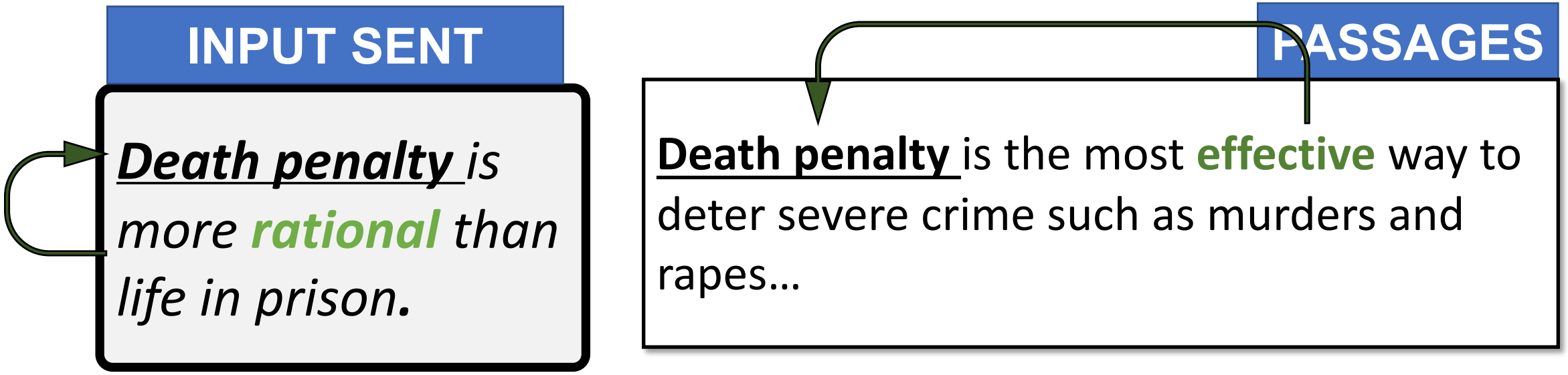
Death penalty is more rational than life in prison.

PASSAGES

Death penalty is the most effective way to deter severe crime such as murders and rapes…

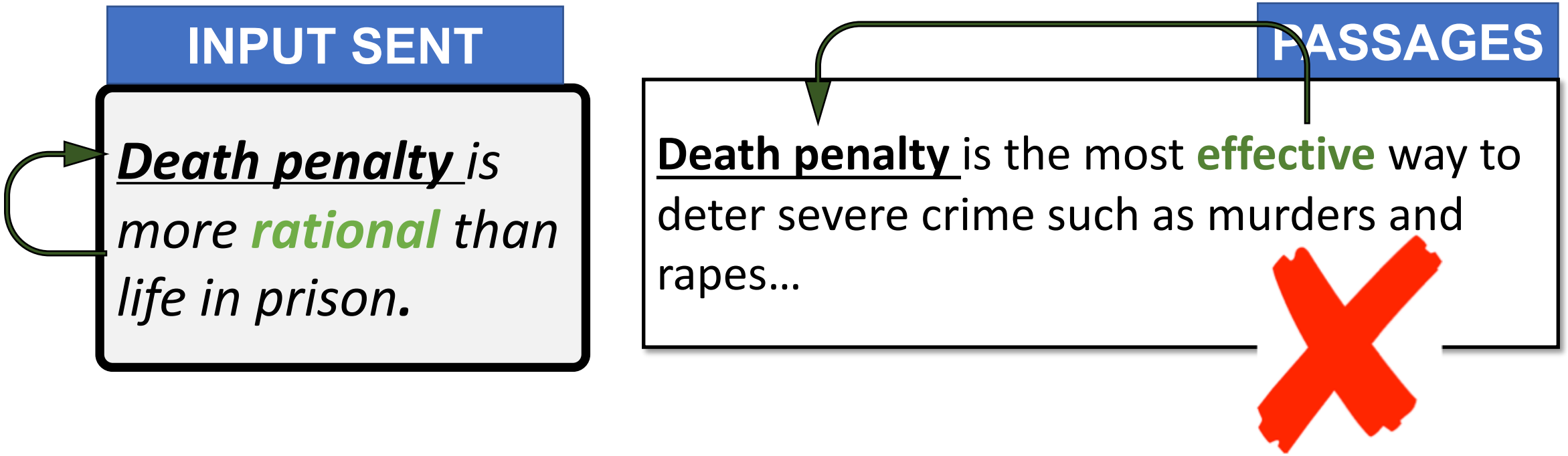
Ranking and Filtering

- **Step 1**: Documents are breakdown into passages (of 3 sentences).
- **Step 2**: Passages are retrieved and ranked based on input queries.
- **Step 3**: Passages with wrong stance are discarded.



Ranking and Filtering

- **Step 1**: Documents are breakdown into passages (of 3 sentences).
- **Step 2**: Passages are retrieved and ranked based on input queries.
- **Step 3**: Passages with wrong stance are discarded.



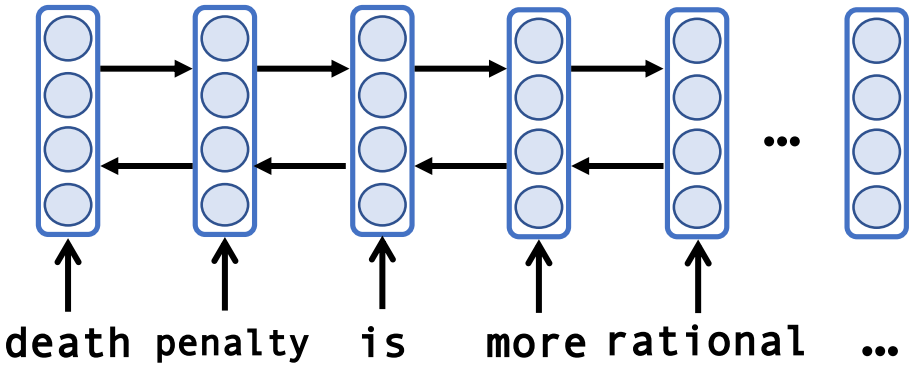
Roadmap

- Prior Work
- Argument Retrieval
- Argument Generation Model
- Experiments
- Conclusion

CANDELA Model

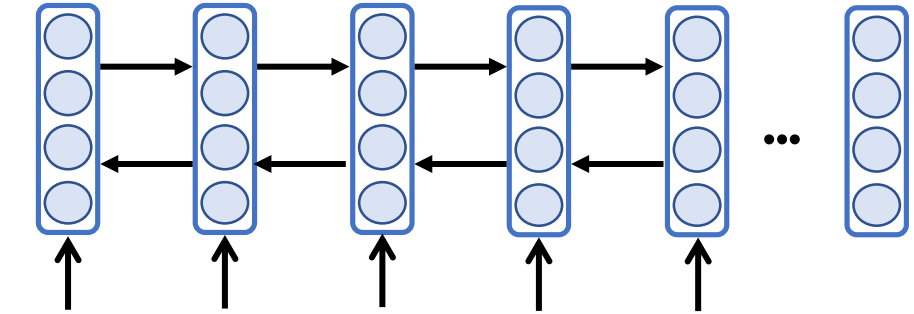
Encoding input with BiLSTM

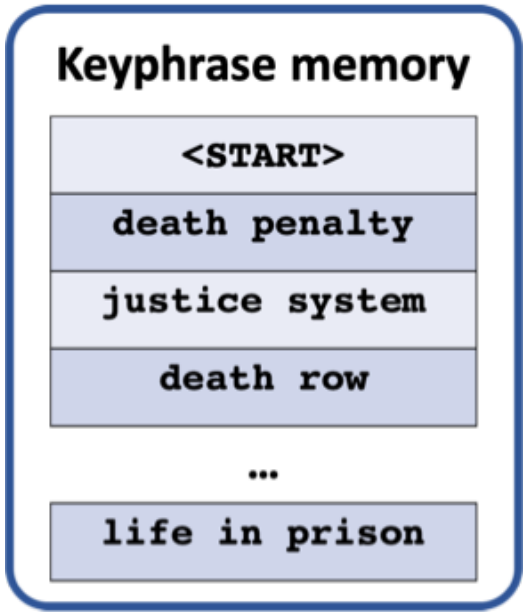
- Sequence-to-sequence framework:



CANDELA Model

- Sequence-to-sequence framework:

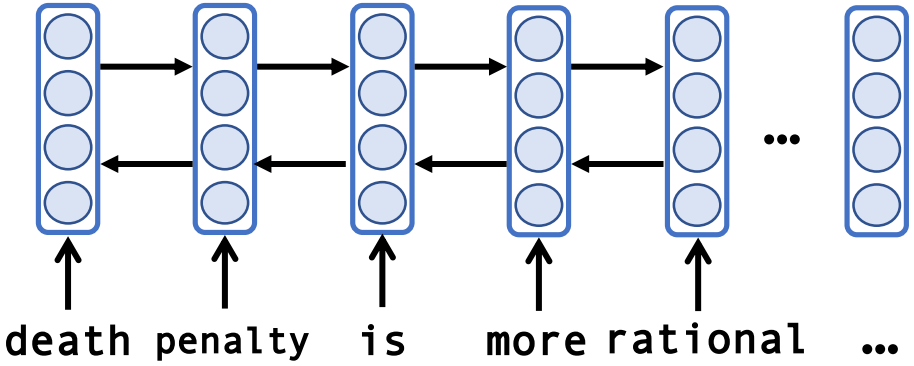


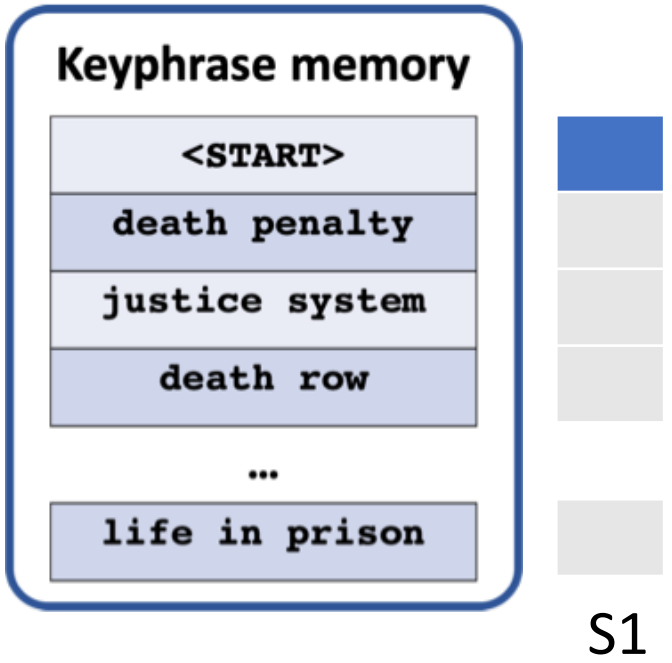


CANDELA Model

Text planning decoding

- Sequence-to-sequence framework:

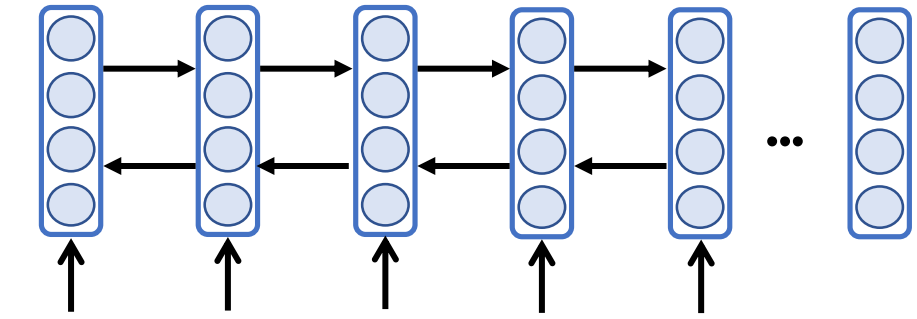


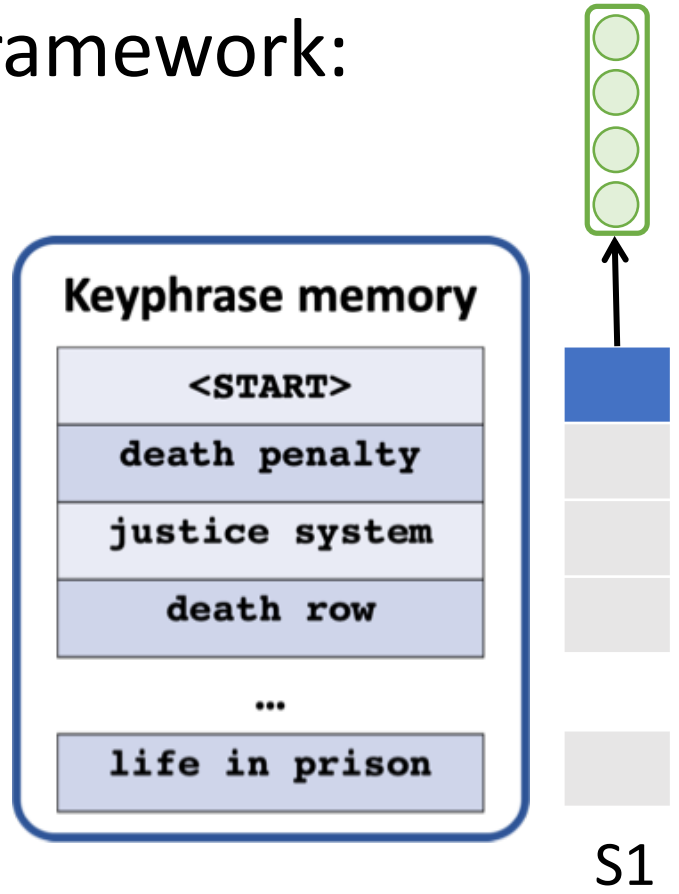


CANDELA Model

Text planning decoding

- Sequence-to-sequence framework:

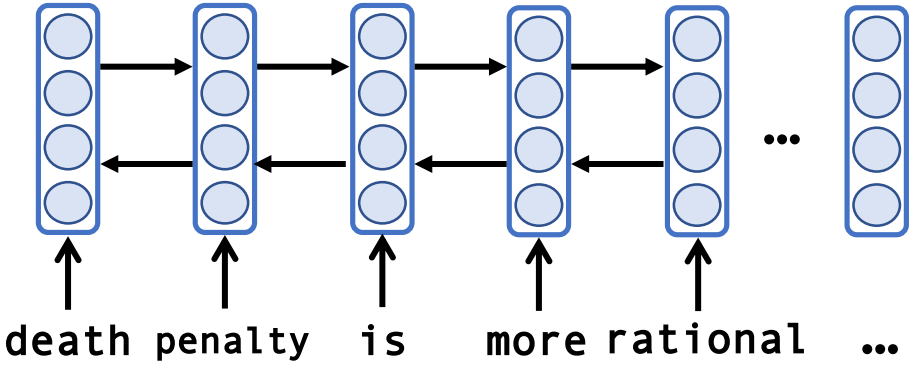


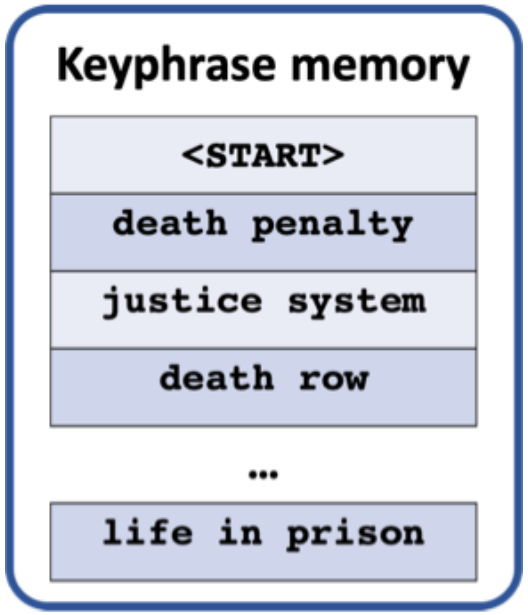


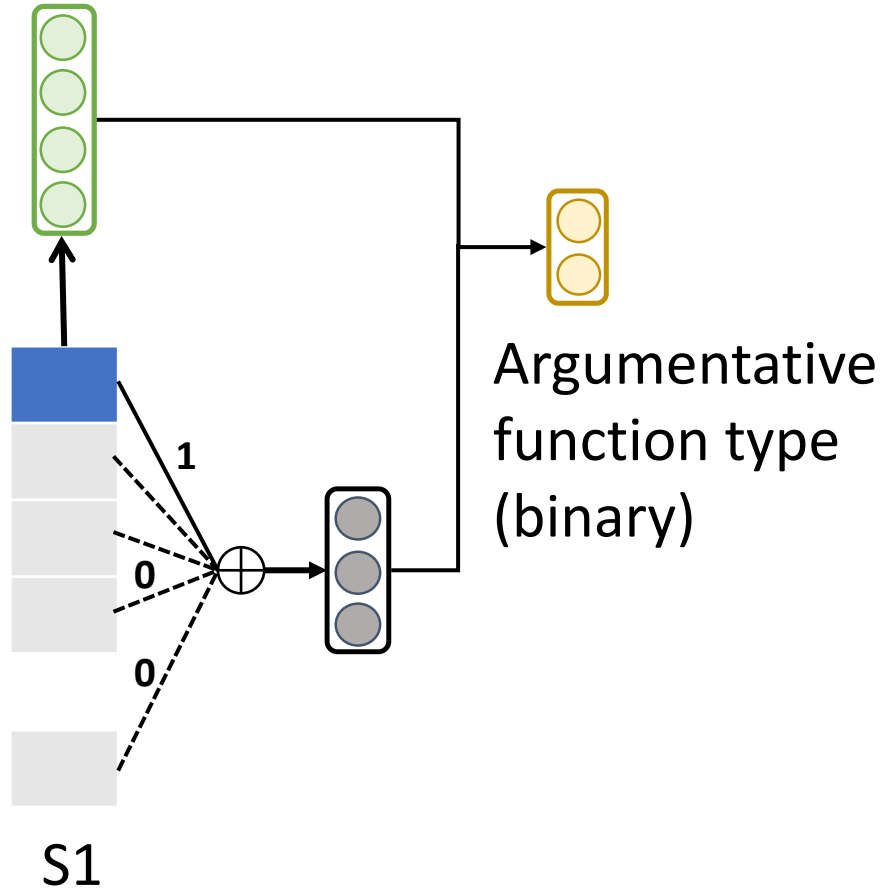
CANDELA Model

Text planning decoding

- Sequence-to-sequence framework:







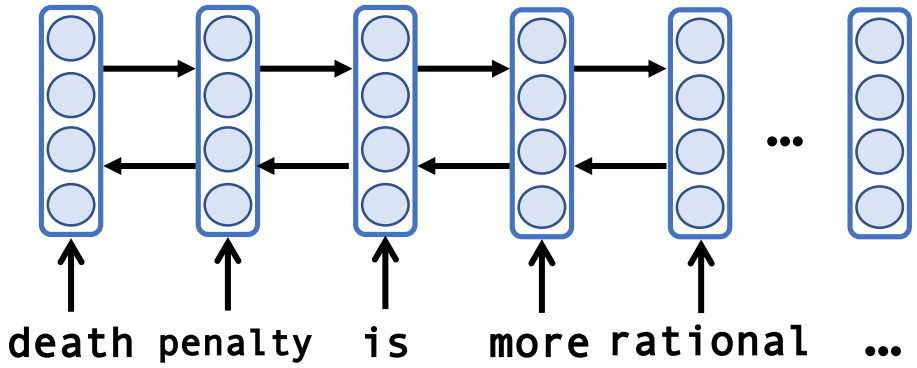
Argumentative Function Type

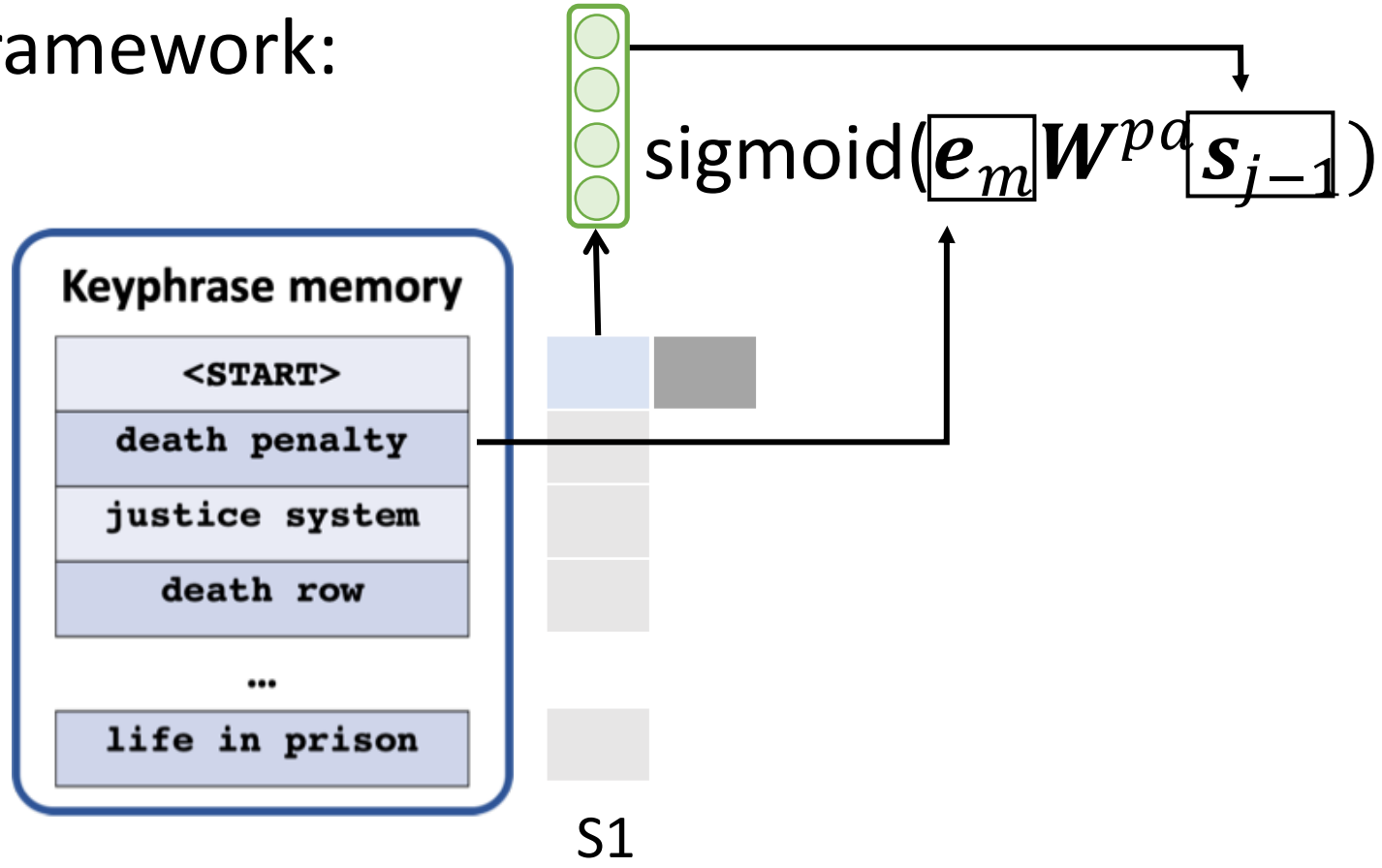
- Goal: to inform the realization decoder of the sentence style
- Argumentative content sentence
  - to deliver crucial ideas and supply evidence, e.g. "*unreliable evidence is used when there is no witness.*".
- Argumentative filler sentence
  - generic statement, e.g. "*in reality I agree, but in practice this is problematic*."

CANDELA Model

Content selection for next sentence

- Sequence-to-sequence framework:

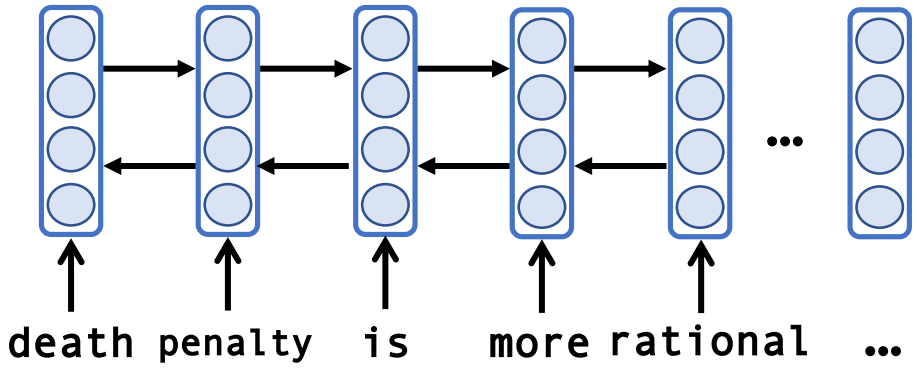


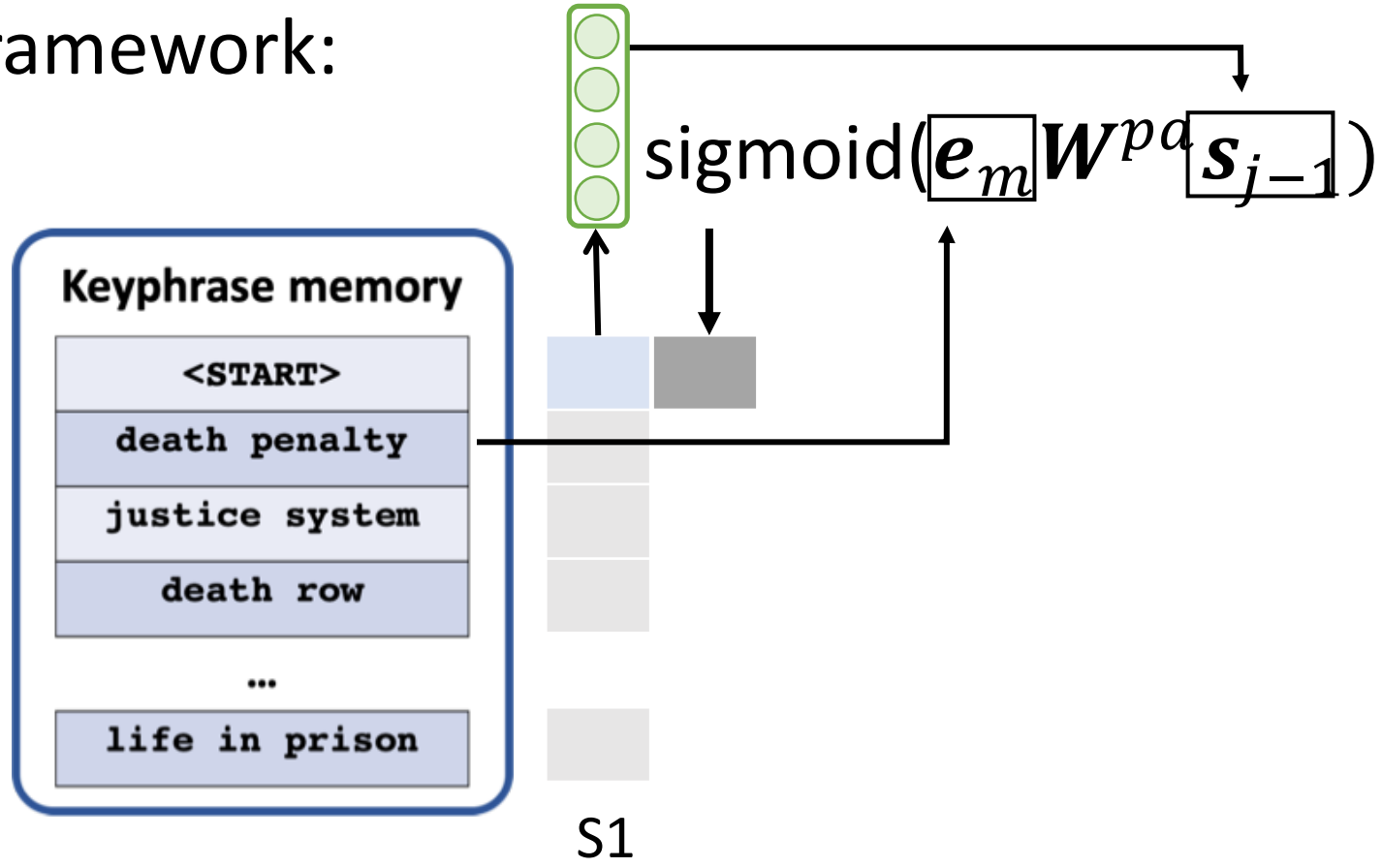


CANDELA Model

Content selection for next sentence

- Sequence-to-sequence framework:

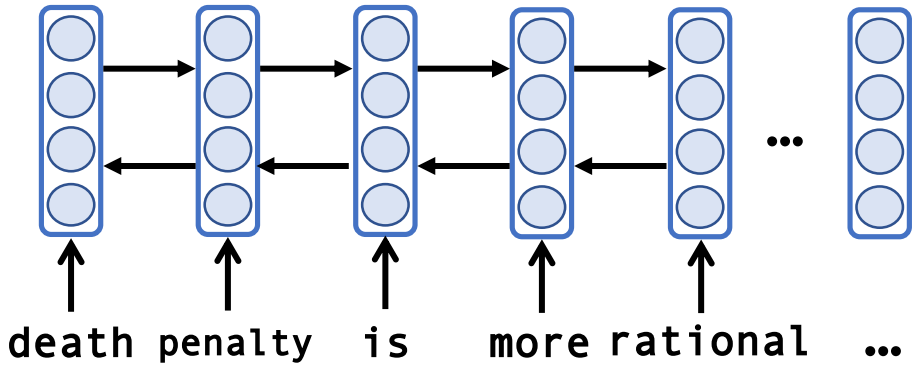


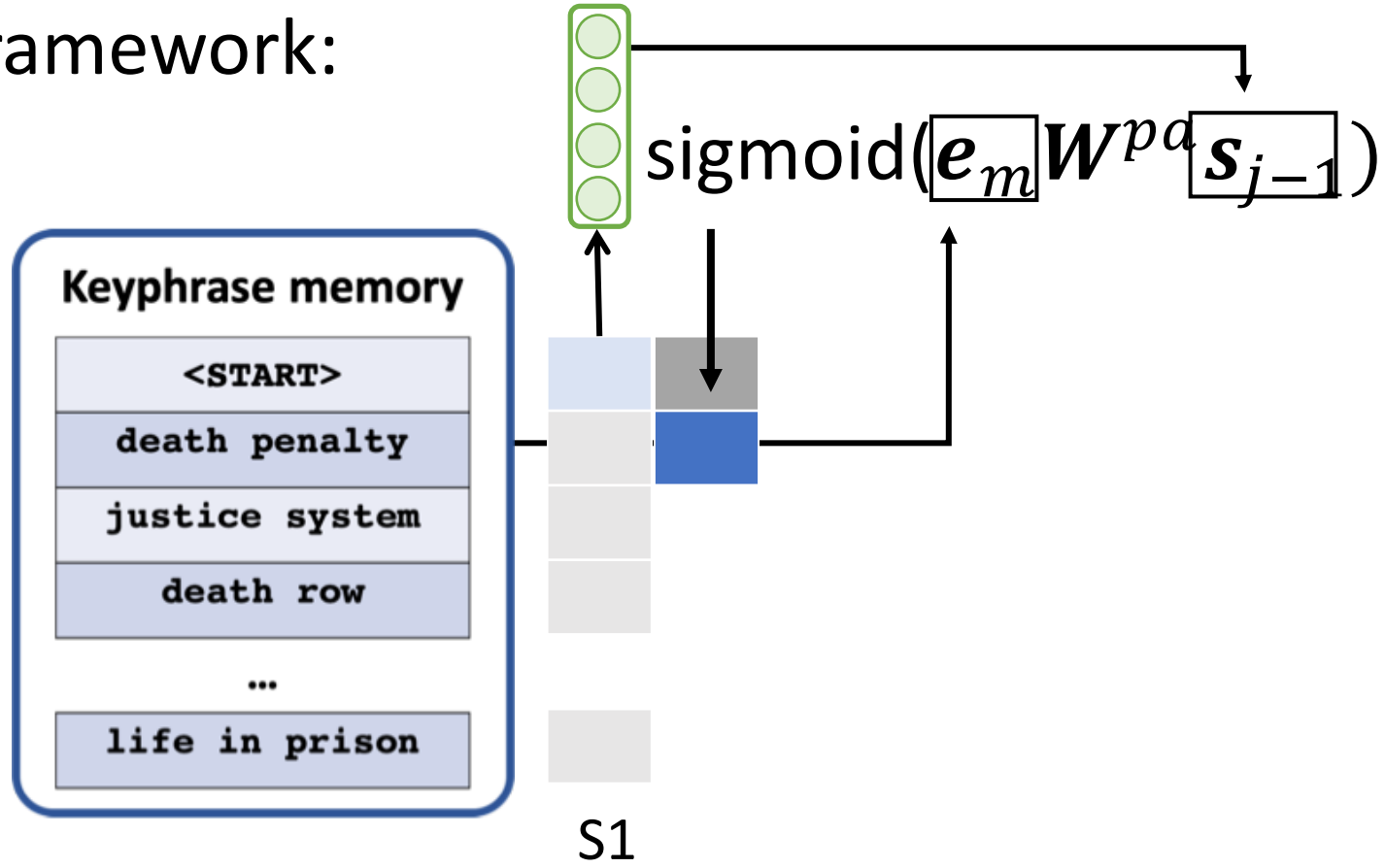


CANDELA Model

Content selection for next sentence

- Sequence-to-sequence framework:

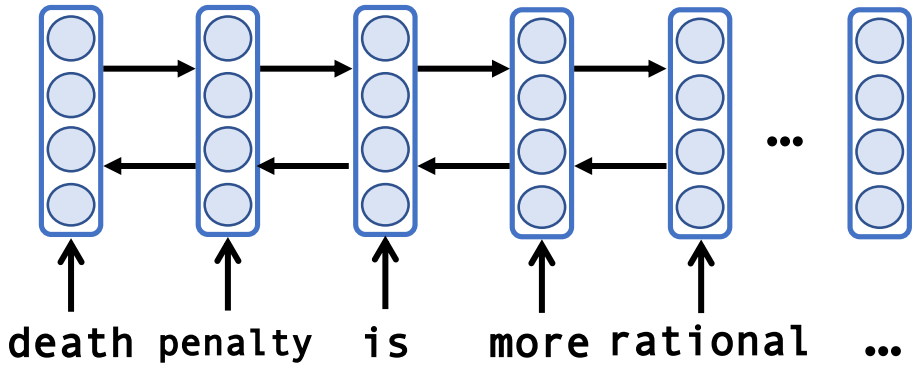


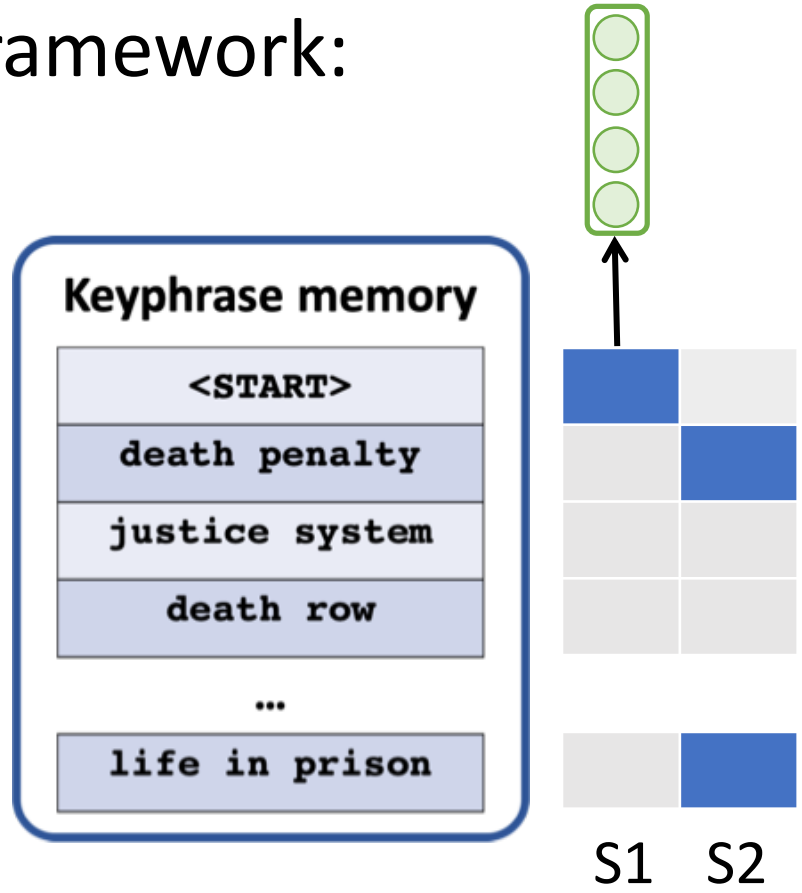


CANDELA Model

Content selection for next sentence

- Sequence-to-sequence framework:

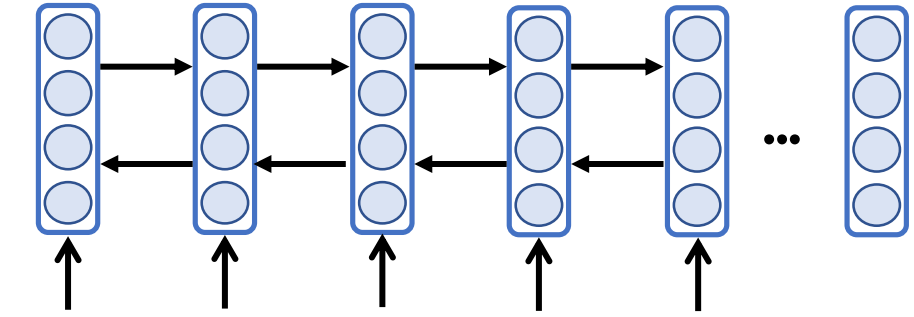


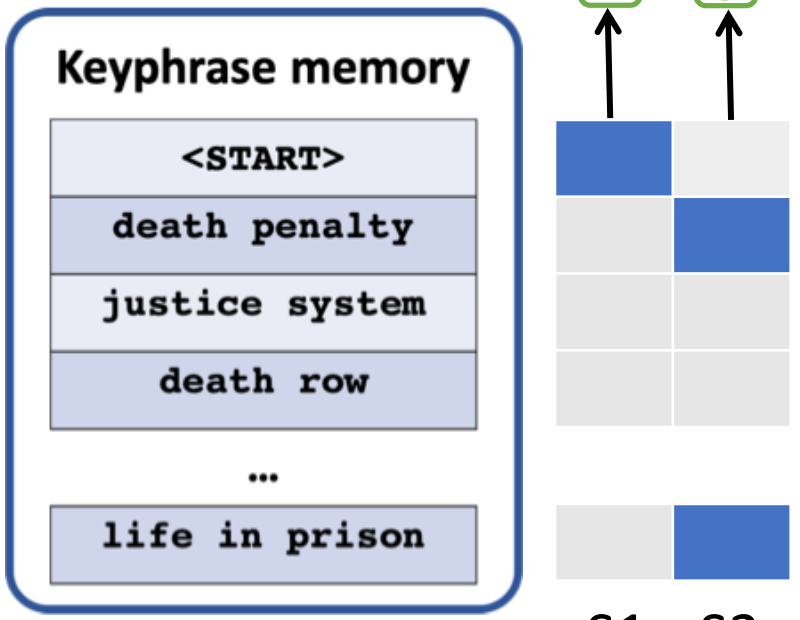


CANDELA Model

Recurrently learn sentence representation

- Sequence-to-sequence framework:

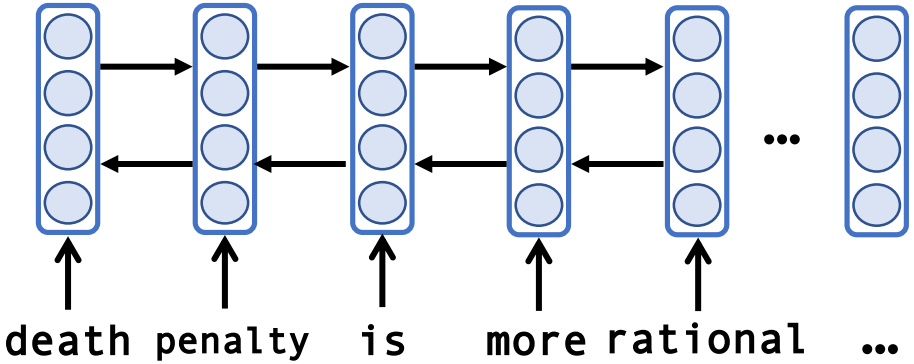


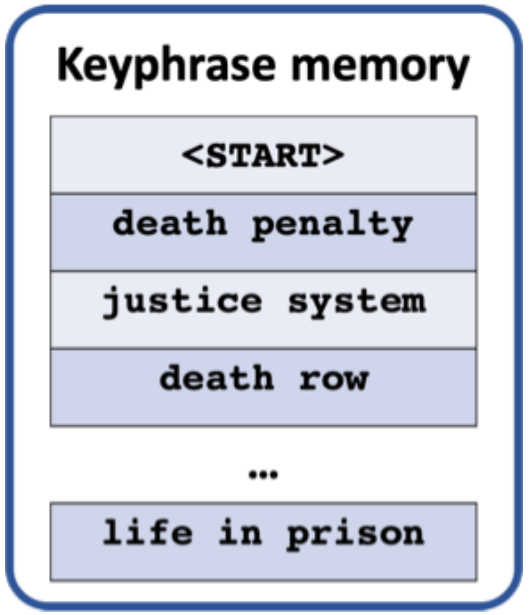


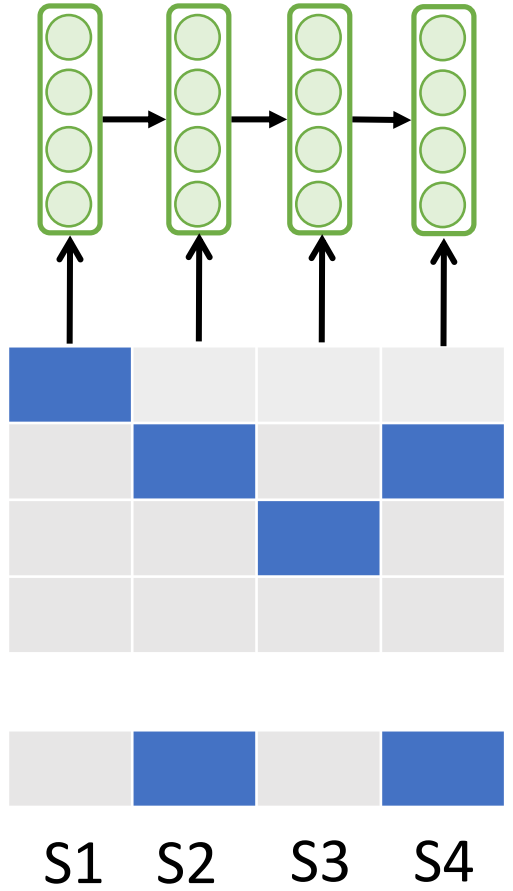
CANDELA Model

Recurrently learn sentence representation

- Sequence-to-sequence framework:

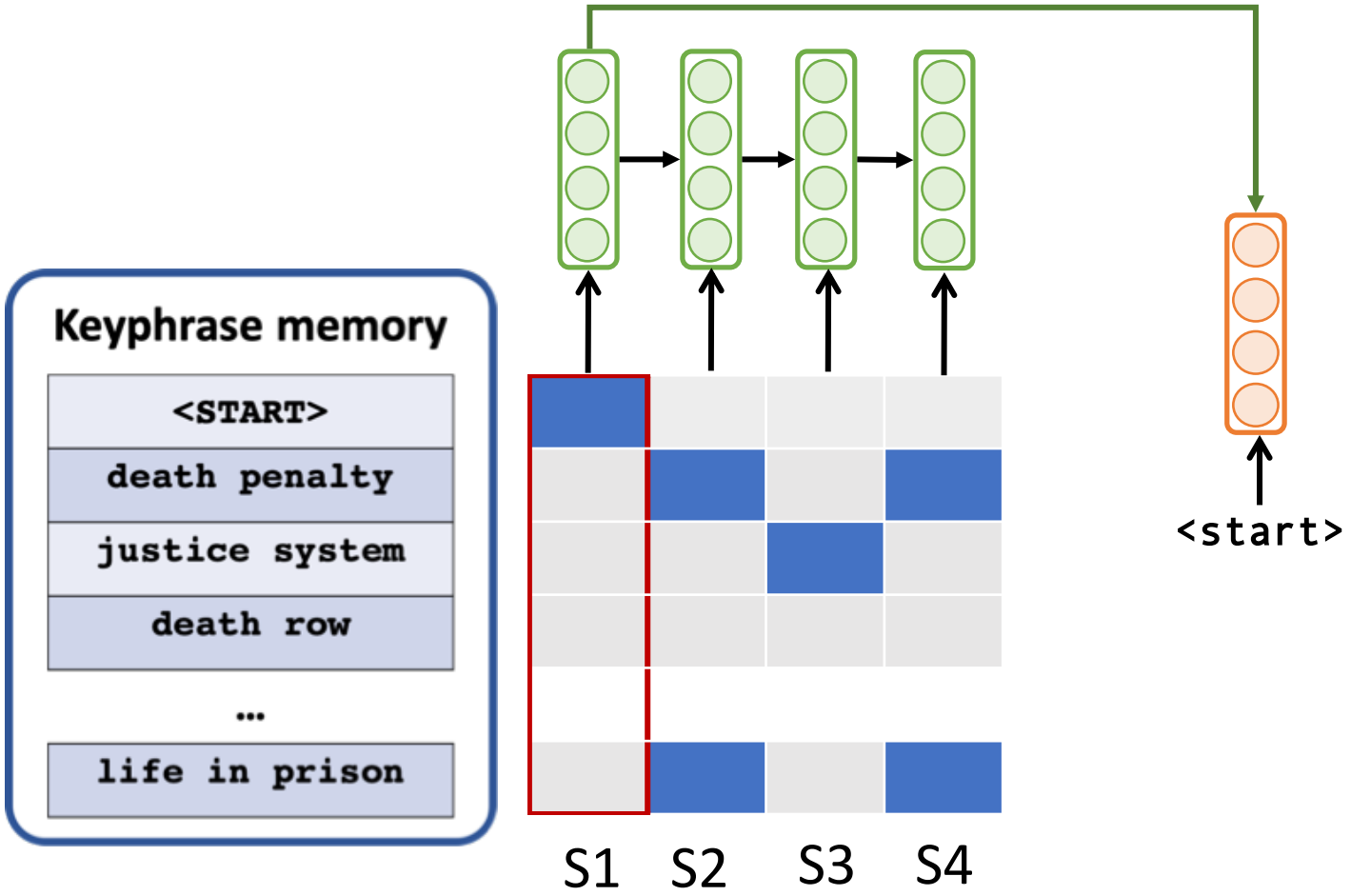






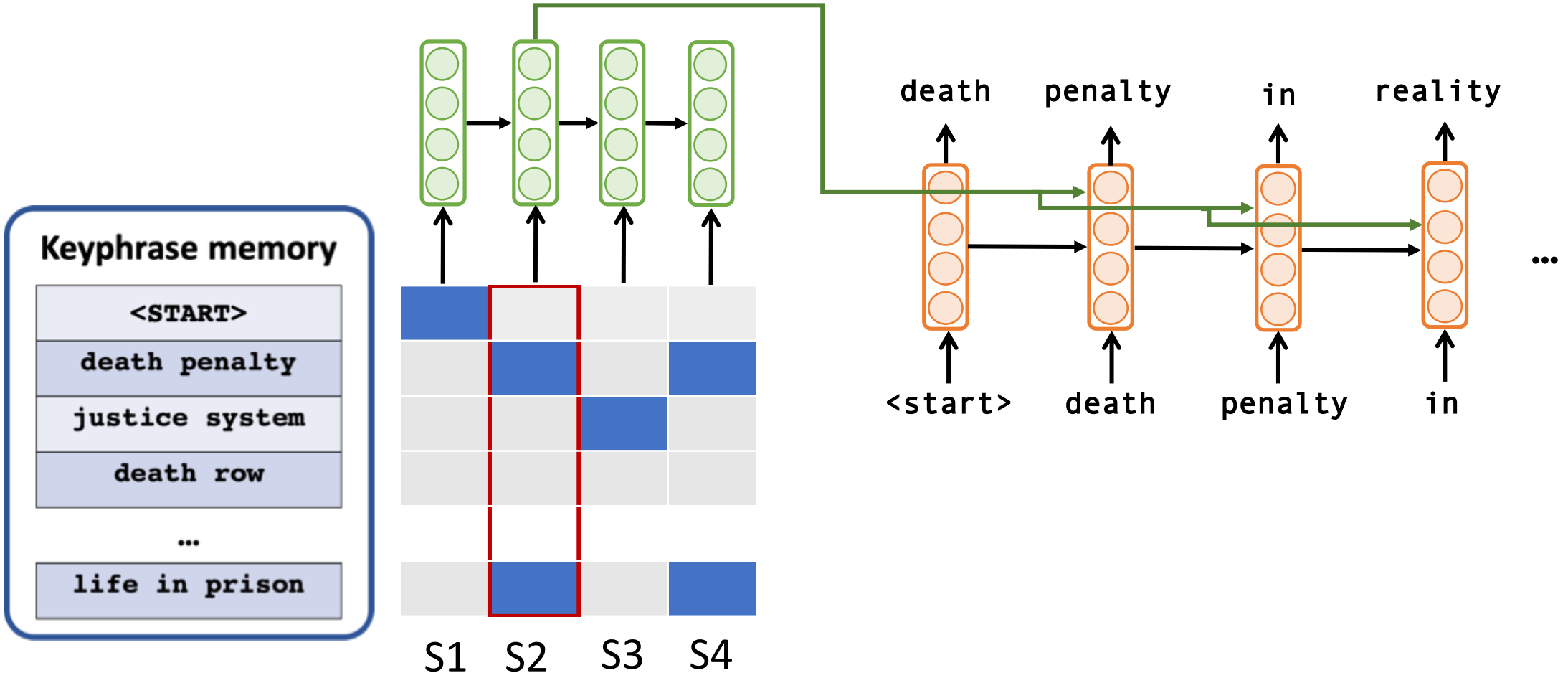
CANDELA Model

Content realization decoding



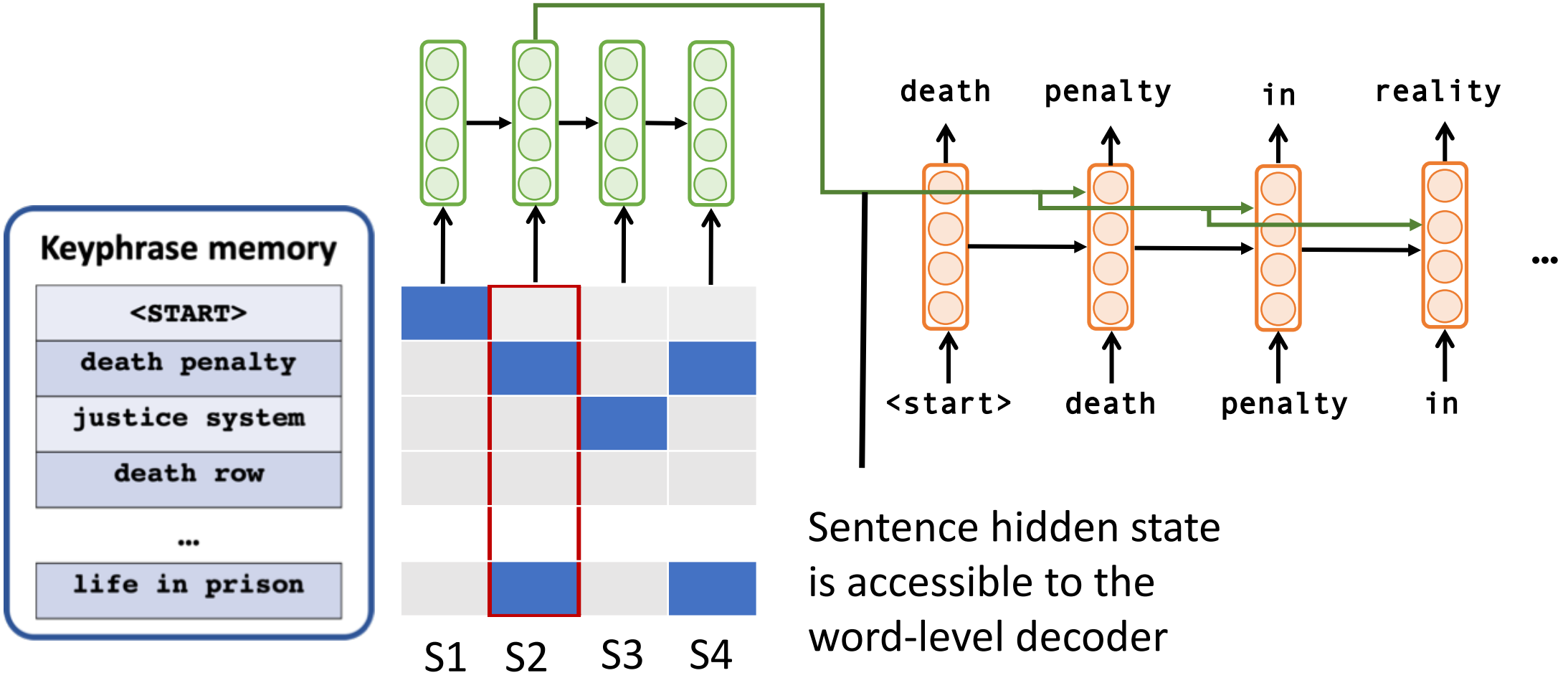
CANDELA Model

Content realization decoding



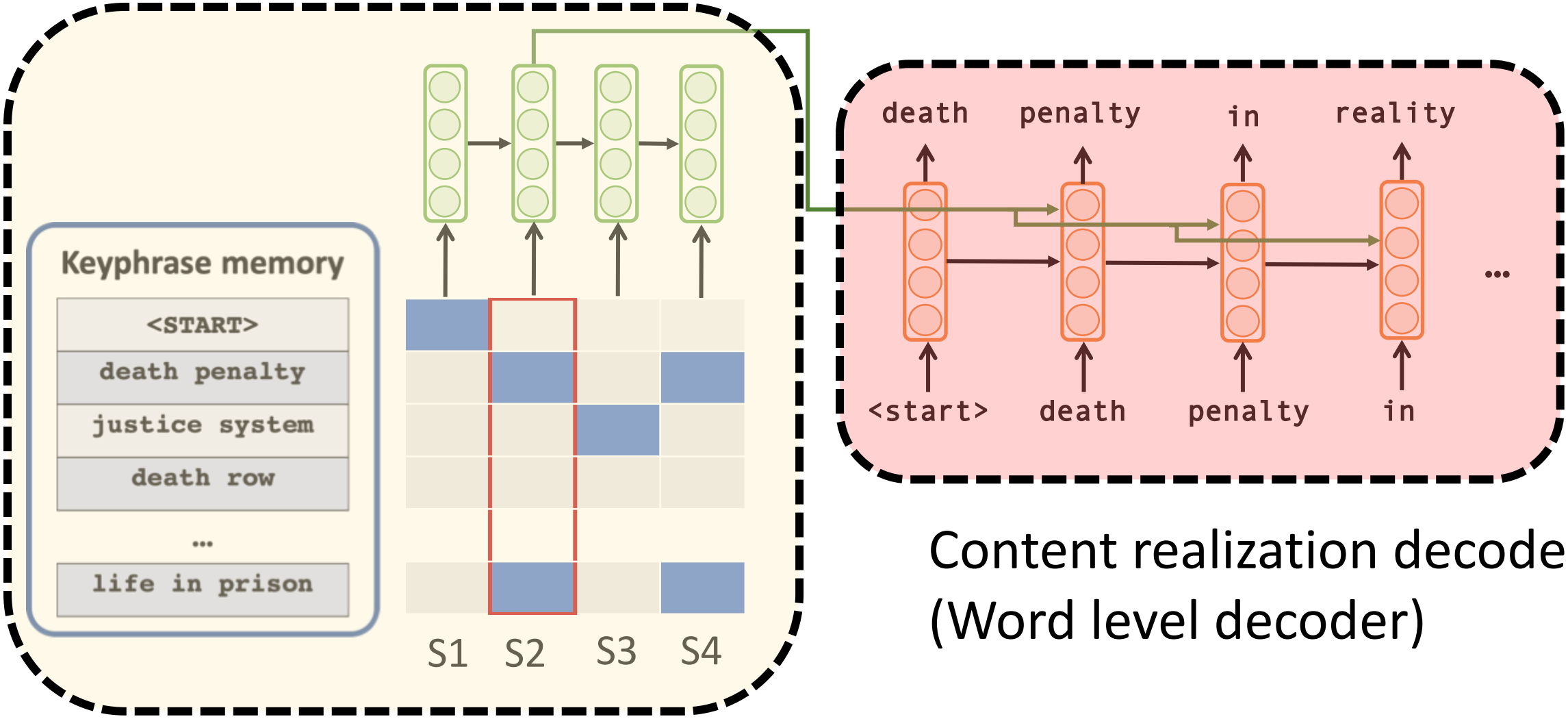
CANDELA Model

Content realization decoding



Sentence hidden state is accessible to the word-level decoder

CANDELA Model



Content realization decoder
(Word level decoder)

Learning Objective

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{arg}}(\theta) + \gamma \cdot \mathcal{L}_{\text{func}}(\theta) + \eta \cdot \mathcal{L}_{\text{sel}}(\theta)$$

Learning Objective

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{arg}}(\theta) + \gamma \cdot \mathcal{L}_{\text{func}}(\theta) + \eta \cdot \mathcal{L}_{\text{sel}}(\theta)$$

Cross-entropy on content realization decoding

Learning Objective

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{arg}}(\theta) + \gamma \cdot \mathcal{L}_{\text{func}}(\theta) + \eta \cdot \mathcal{L}_{\text{sel}}(\theta)$$

Cross-entropy on content realization decoding

Cross-entropy on argumentative function type

Learning Objective

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{arg}}(\theta) + \gamma \cdot \mathcal{L}_{\text{func}}(\theta) + \eta \cdot \mathcal{L}_{\text{sel}}(\theta)$$

Cross-entropy on content realization decoding

Cross-entropy on argumentative function type

Binary cross-entropy on content selection

Roadmap

- Prior Work
- Argument Retrieval
- Argument Generation Model
- Experiments
- Conclusion

Experiments - Data

- Dataset: statement-argument pairs from `/r/ChangeMyView` community
- 217K pairs for train, 33K and 36K for dev and test
- LM pre-training: an extended set of replies (353K)

Experiments - Data

- Dataset: statement-argument pairs from `/r/ChangeMyView` community
- 217K pairs for train, 33K and 36K for dev and test
- LM pre-training: an extended set of replies (353K)
- Topics: politics and policy making related
- Keyphrase memory: noun phrases/verb phrases that contains a Wikipedia title OR a topic signature word [Lin and Hovy, 2000]

Experiments - Data

Input	Average # words per statement	383.7
Output	Average # words per argument	66.0
	Average # passage	4.3
	Average # keyphrase	57.1

Experiments - Models

- Comparisons:
  - **Retrieval**: directly return the top ranked passage as output
  - **Seq2seq**: encode input and generate counter-argument
  - **Seq2seqAug**: encode input + retrieved passages + keyphrases
  - our prior work (**HW2018**): with only Wikipedia passages and an auxiliary keyphrase generation task

Experiments - Results

- Evaluation setups:
  - **System** (realistic): passages are retrieved by using the input statement only
  - **Oracle** (upper bound w.r.t retrieval): passages are retrieved by using the gold-standard arguments

Experiments - Results

- Automatic evaluation results (**System**)

System	BLEU-2	BLEU-4	ROUGE-2	METEOR	#Word

Experiments - Results

- Automatic evaluation results (**System**)

System	BLEU-2	BLEU-4	ROUGE-2	METEOR	#Word
Retrieval	7.55	1.11	8.64	14.38	123
Seq2seq	6.92	2.13	13.02	15.08	68

Experiments - Results

- Automatic evaluation results (**System**)

System	BLEU-2	BLEU-4	ROUGE-2	METEOR	#Word
Retrieval	7.55	1.11	8.64	14.38	123
Seq2seq	6.92	2.13	13.02	15.08	68
Seq2seqAug	8.26	2.24	13.79	15.75	78
HW2018	3.64	0.92	8.83	11.78	51

Experiments - Results

CANDELA model outperforms all comparisons.

- Automatic evaluation results (**System**)

System	BLEU-2	BLEU-4	ROUGE-2	METEOR	#Word
Retrieval	7.55	1.11	8.64	14.38	123
Seq2seq	6.92	2.13	13.02	15.08	68
Seq2seqAug	8.26	2.24	13.79	15.75	78
HW2018	3.64	0.92	8.83	11.78	51
CANDELA	12.02	**2.99**	**14.93**	**16.92**	119
CANDELA w/o psg	**12.33**	2.86	14.53	16.60	123

Experiments - Results

3-4 BLEU/METEOR points to be expected from Oracle retrieval

- Automatic evaluation results (**Oracle**)

System	BLEU-2	BLEU-4	ROUGE-2	METEOR	#Word
Retrieval	10.97	3.05	23.49	20.08	140
Seq2seq	6.92	2.13	13.02	15.08	68
Seq2seqAug	10.98	4.41	22.97	19.62	71
HW2018	8.51	2.86	18.89	17.18	58
CANDELA	15.80	**5.00**	**23.75**	**20.18**	116
CANDELA w/o psg	**16.33**	4.98	23.65	19.94	123

Human Evaluation

Grammaticality	Appropriateness	Content richness

Human Evaluation

Grammaticality	Appropriateness	Content richness
1 (**ungrammatical**): *"to sought sentencing is is numerous at to to"* 5 (**fluent**): *"in theory i agree, but in reality the justice system is not perfect."*		

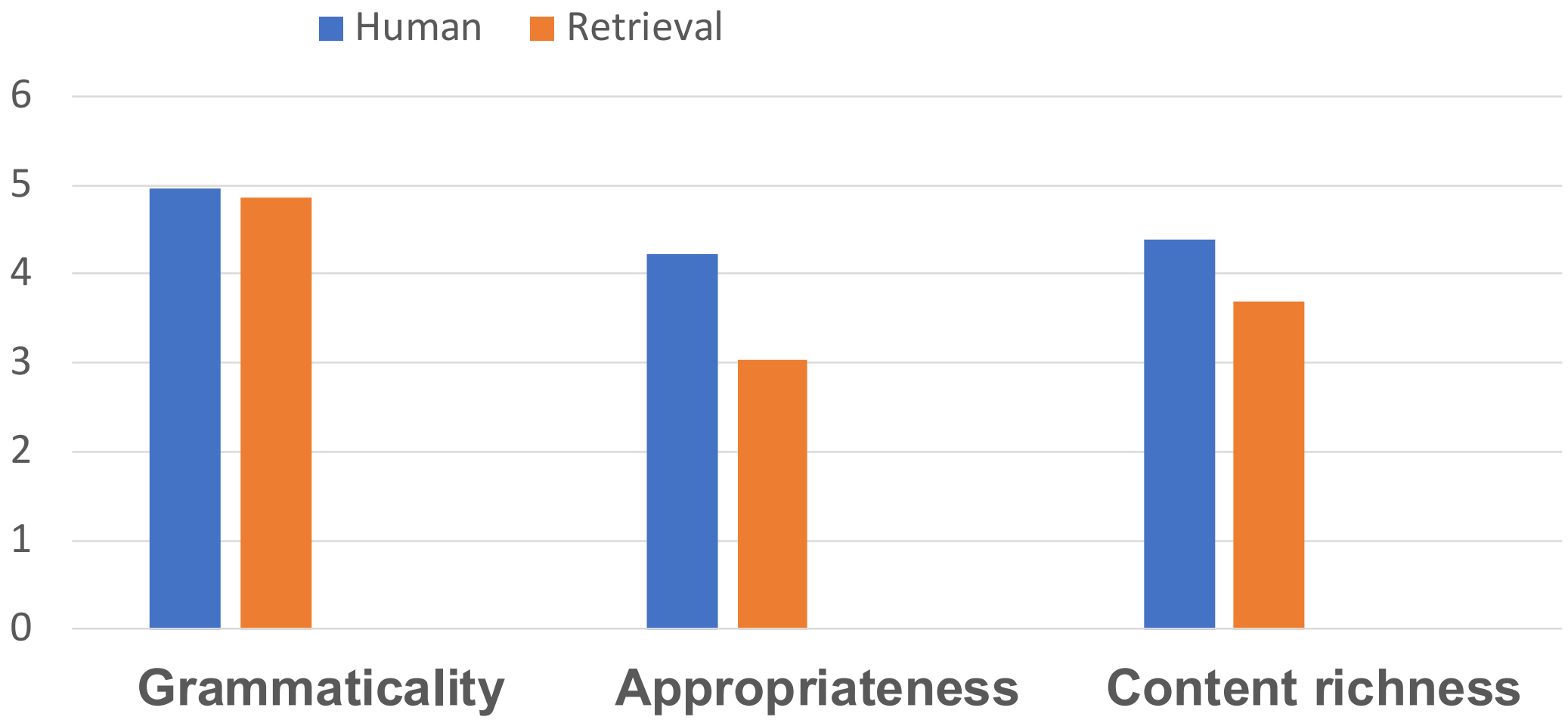
Human Evaluation

Grammaticality	Appropriateness	Content richness
1 (**ungrammatical**): *"to sought sentencing is is numerous at to to"* 5 (**fluent**): *"in theory i agree, but in reality the justice system is not perfect. "*	1 (**off-topic**): *"the gap between rich and poor is the major issue."* 5 (**on-topic, correct stance**): "the problem with death penalty is that wrongful conviction exists, and it is irreversible in such cases."	

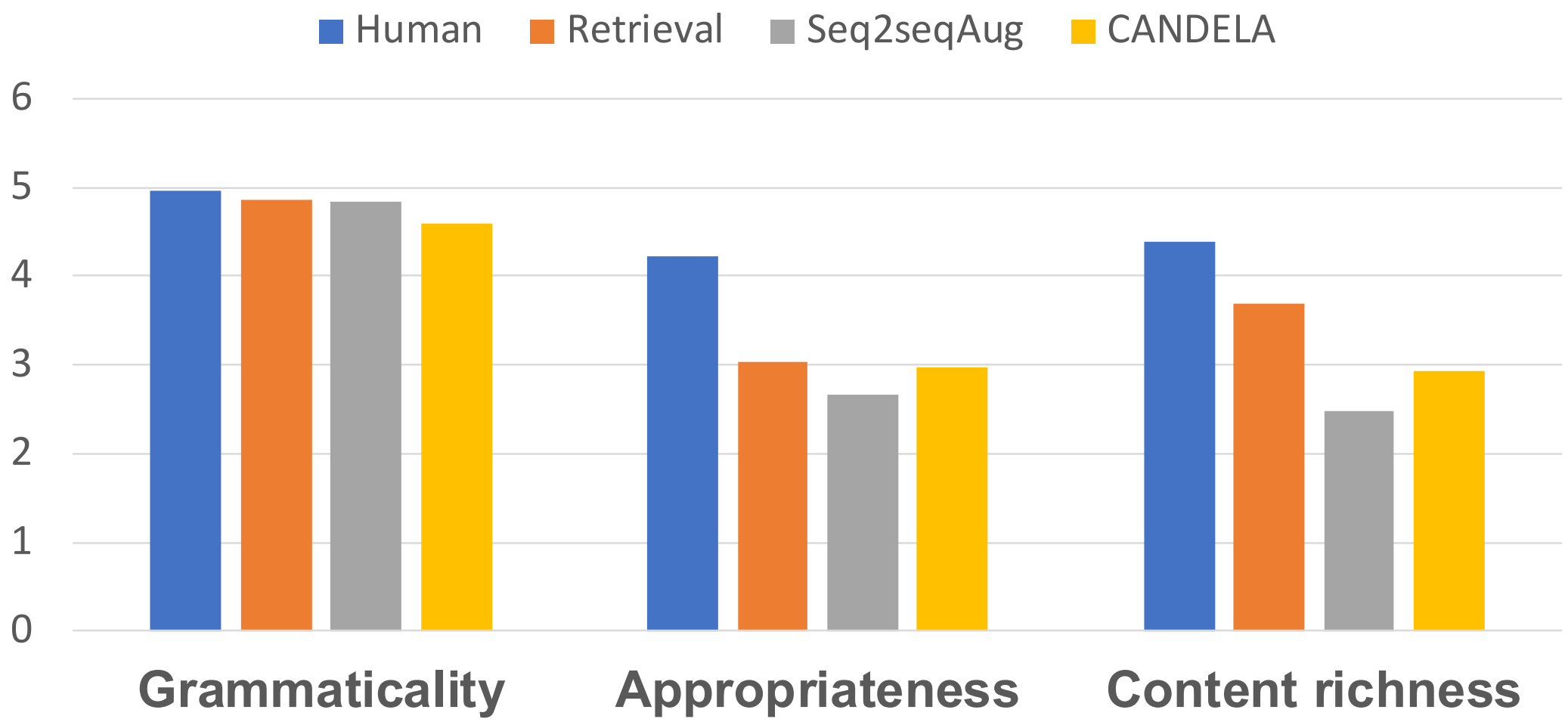
Human Evaluation

Grammaticality	Appropriateness	Content richness
1 (**ungrammatical**): *"to sought sentencing is is numerous at to to"*	1 (**off-topic**): *"the gap between rich and poor is the major issue."*	1 (**generic response**): *"i don't care."*
5 (**fluent**): *"in theory i agree, but in reality the justice system is not perfect. "*	5 (**on-topic, correct stance**): "the problem with death penalty is that wrongful conviction exists, and it is irreversible in such cases."	5 (**sufficient supports**): "a 2015 study showed that death penalty cases cost an average of one million more to prosecute, it has cost california more than four billion since 1978"

Human Evaluation - Results



Human Evaluation - Results



Roadmap

- Prior Work
- Argument Retrieval
- Argument Generation Model
- Experiments
- Conclusion

Conclusion

- We study the challenging argument generation task with diverse external knowledge.
- Our proposed CANDELA system is able to conduct text planning and content realization, unified through an end-to-end trained model.
- This task still remains far from being solved. Future directions include adding common-sense knowledge, and multi-turn argumentation.

Thank you!

- Online demo: https://xinyuhua.github.io/candela
- Project page: https://xinyuhua.github.io/Resources/acl19/
- Contact: hua.x@husky.neu.edu