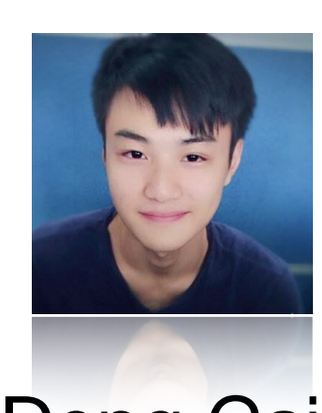
AMR Parsing via Graph⇆Sequence Iterative Inference



Deng Cai and Wai Lam

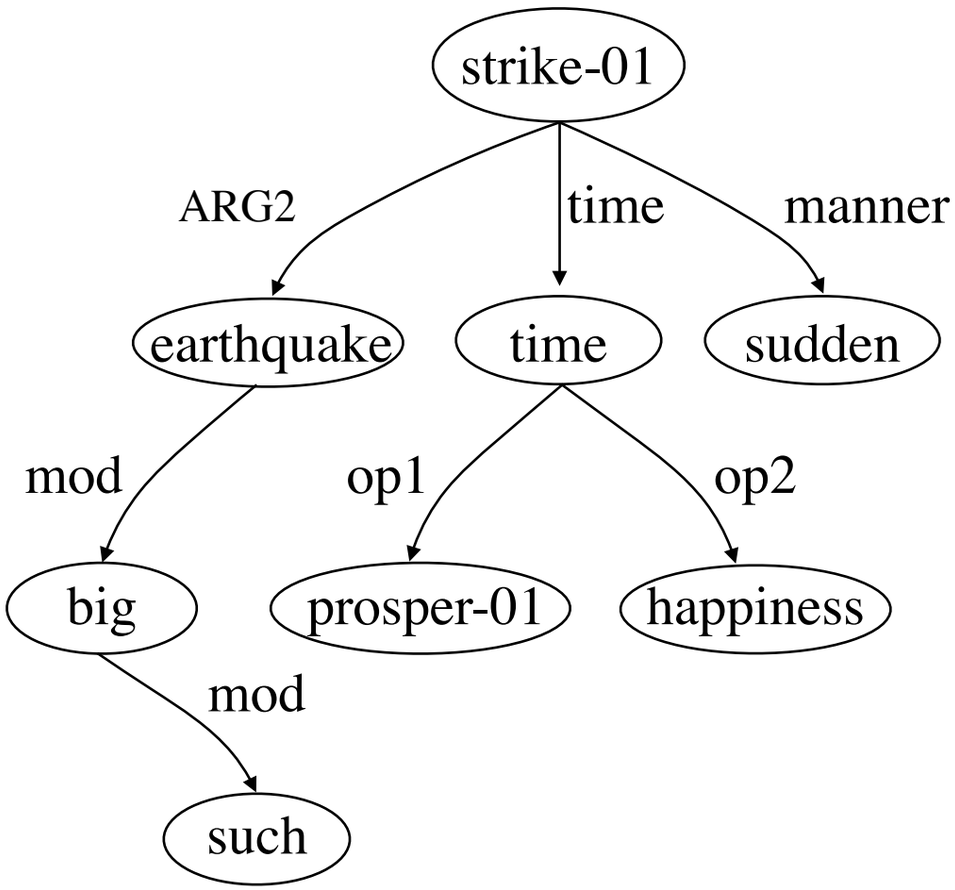
The Chinese University of Hong Kong

ACL2020

Background

Natural Language ⟷ Meaning Representation
(AMR)

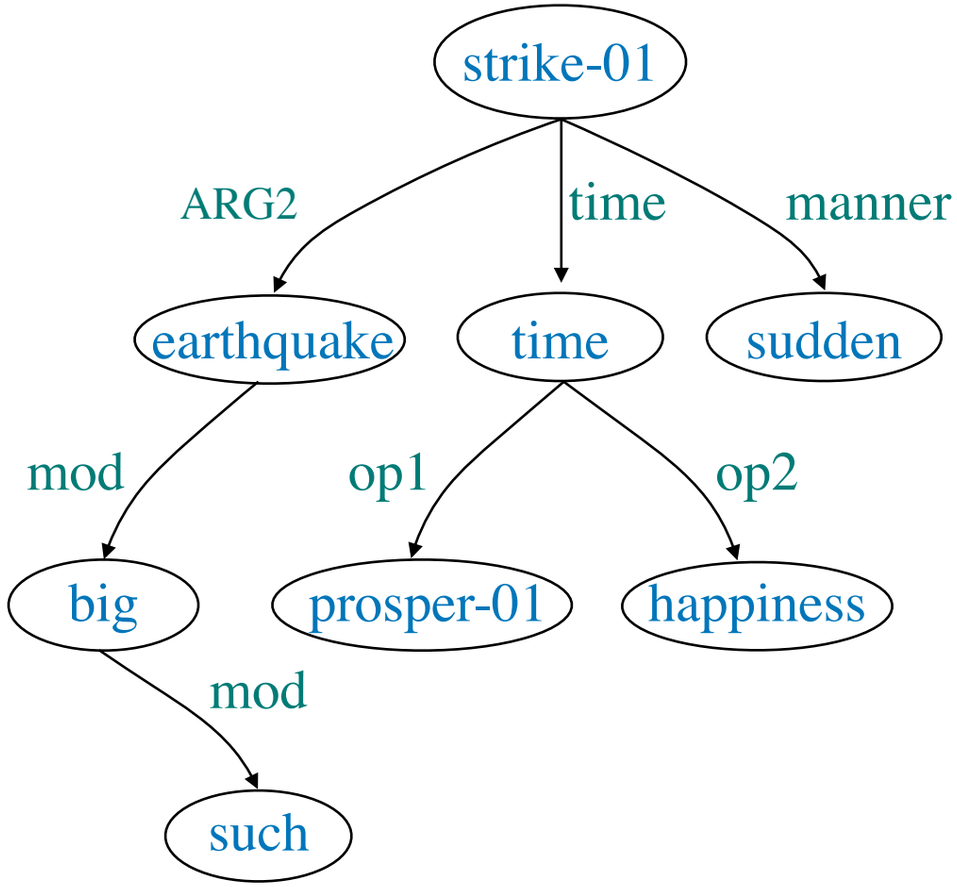
During a time of prosperity and happiness, such a big earthquake suddenly struck.



Background

- Abstract Meaning Representation (AMR)
  - rooted, labeled, and directed acyclic graph
  - nodes represent concepts
  - edges represent relations

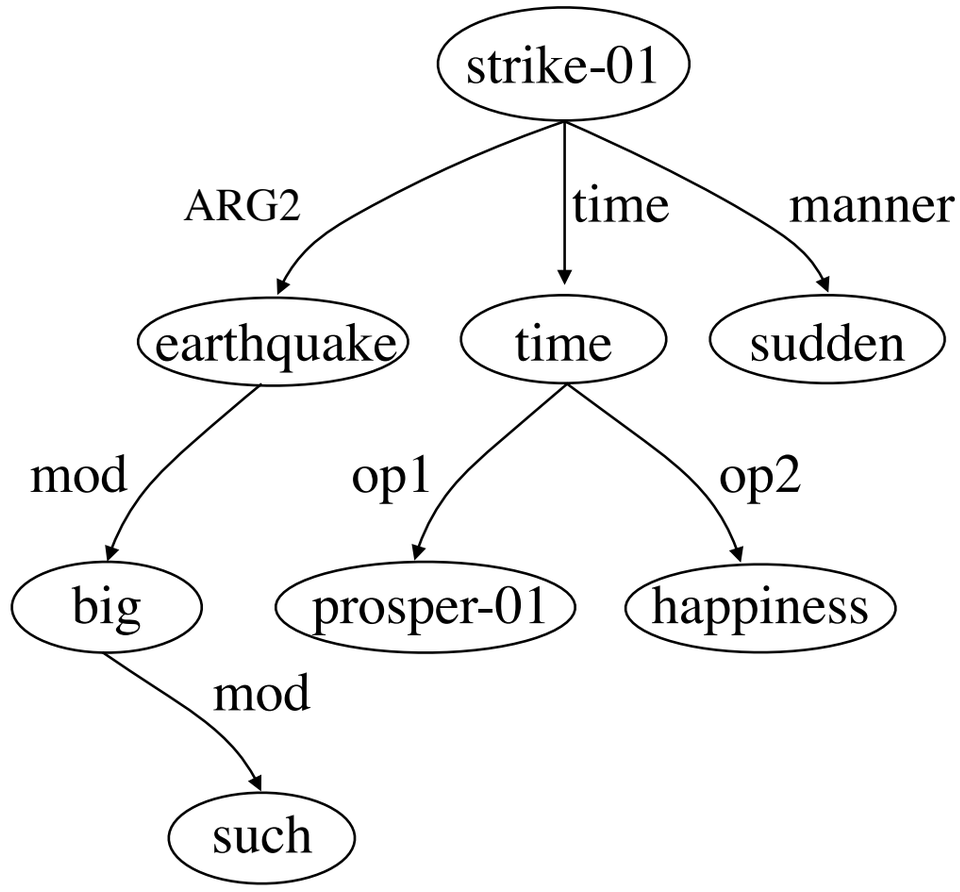
During a time of prosperity and happiness, such a big earthquake suddenly struck.



Background

- Abstract Meaning Representation (AMR)
  - Named Entity Recognition
  - Word Sense Disambiguation
  - Semantic Role Labeling
  - Coreference Resolution
  - …

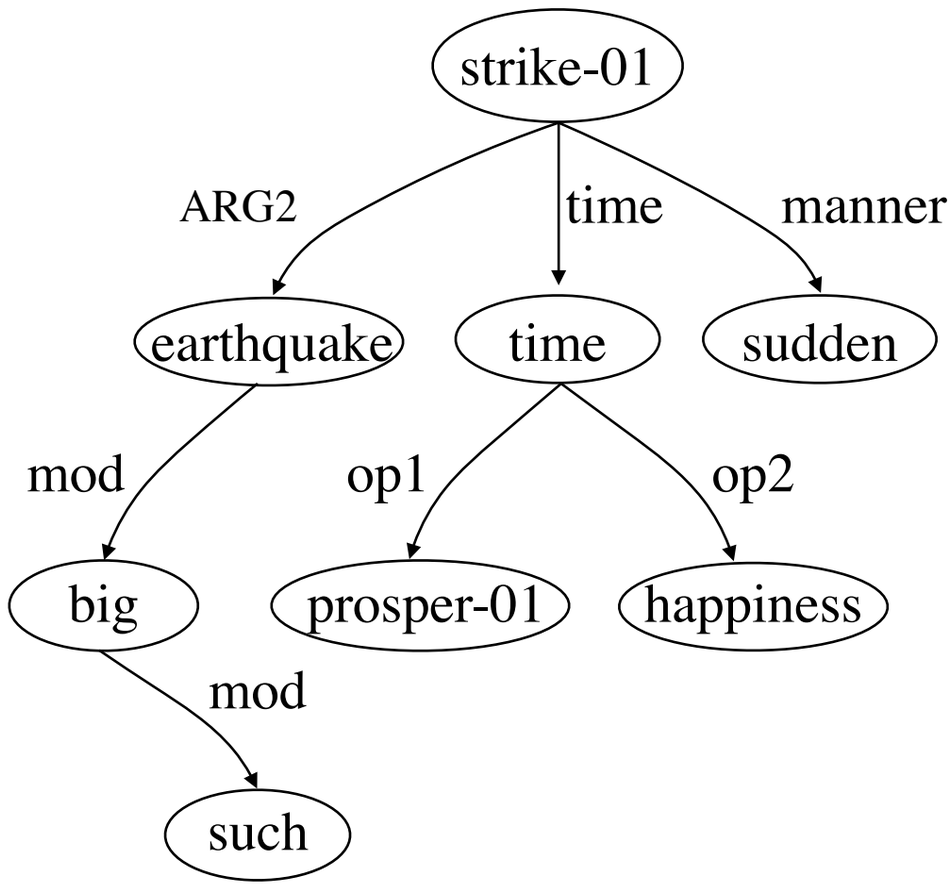
During a time of prosperity and happiness, such a big earthquake suddenly struck.



Challenges

During a time of prosperity and happiness, such a big earthquake suddenly struck.

AMR Parsing



- Concept Prediction:
  - No explicit alignment of graph nodes and sentence tokens
  - Large and sparse concept vocabulary vs. Limited training data
- Relation Prediction:
  - Frequent reentrancies and non-projective arcs

Existing Work

- Two-stage Parsing (Flanigan et al., 2014; Lyu and Titov, 2018; Zhang et al., 2019a)
  - first predict all concepts
  - then predict all relations

- One-stage Parsing (Wang et al., 2016; Damonte et al., 2017; Ballesteros and Al-Onaizan, 2017; Peng et al., 2017; Guo and Lu, 2018; Liu et al., 2018; Wang and Xue, 2017; Naseem et al., 2019; Barzdins and Gosko, 2016; Konstas et al., 2017; van Noord and Bos, 2017; Peng et al., 2018; Cai and Lam, 2019; Zhang et al., 2019b)
  - Construct a parse graph incrementally

- Grammar-based Parsing (Peng et al., 2015;Pust et al., 2015;Artzi et al., 2015; Groschwitz et al., 2018; Lindemann et al., 2019)

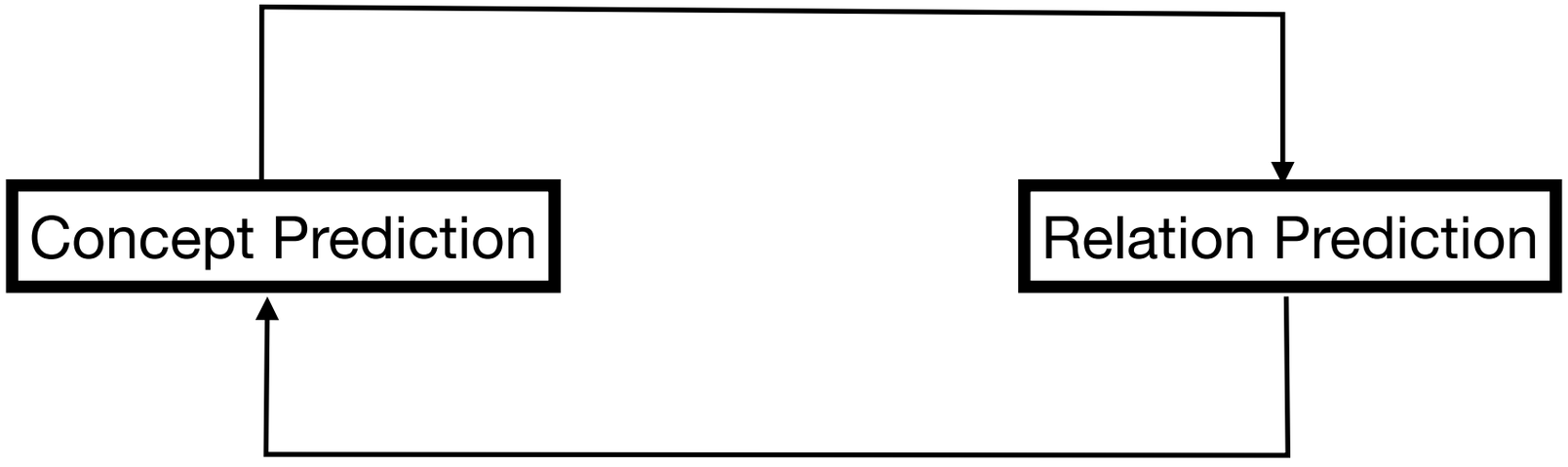
Existing Work

- Two-stage Parsing (Flanigan et al., 2014; Lyu and Titov, 2018; Zhang et al., 2019a)
  - Pipeline: concept prediction -> relation prediction
- One-stage Parsing
  - Transition-based (Wang et al., 2016; Damonte et al., 2017; Ballesteros and Al-Onaizan, 2017; Peng et al., 2017; Guo and Lu, 2018; Liu et al., 2018; Wang and Xue, 2017; Naseem et al., 2019)
    - Insert node and build edge sequentially
  - Seq2seq-based (Barzdins and Gosko, 2016; Konstas et al., 2017; van Noord and Bos, 2017; Peng et al., 2018)
    - Nodes and edges are mixed in the same output space
  - Graph-based (Cai and Lam, 2019; Zhang et al., 2019b)
    - A new node and its connections to existing nodes are jointly decoded in order or in parallel.

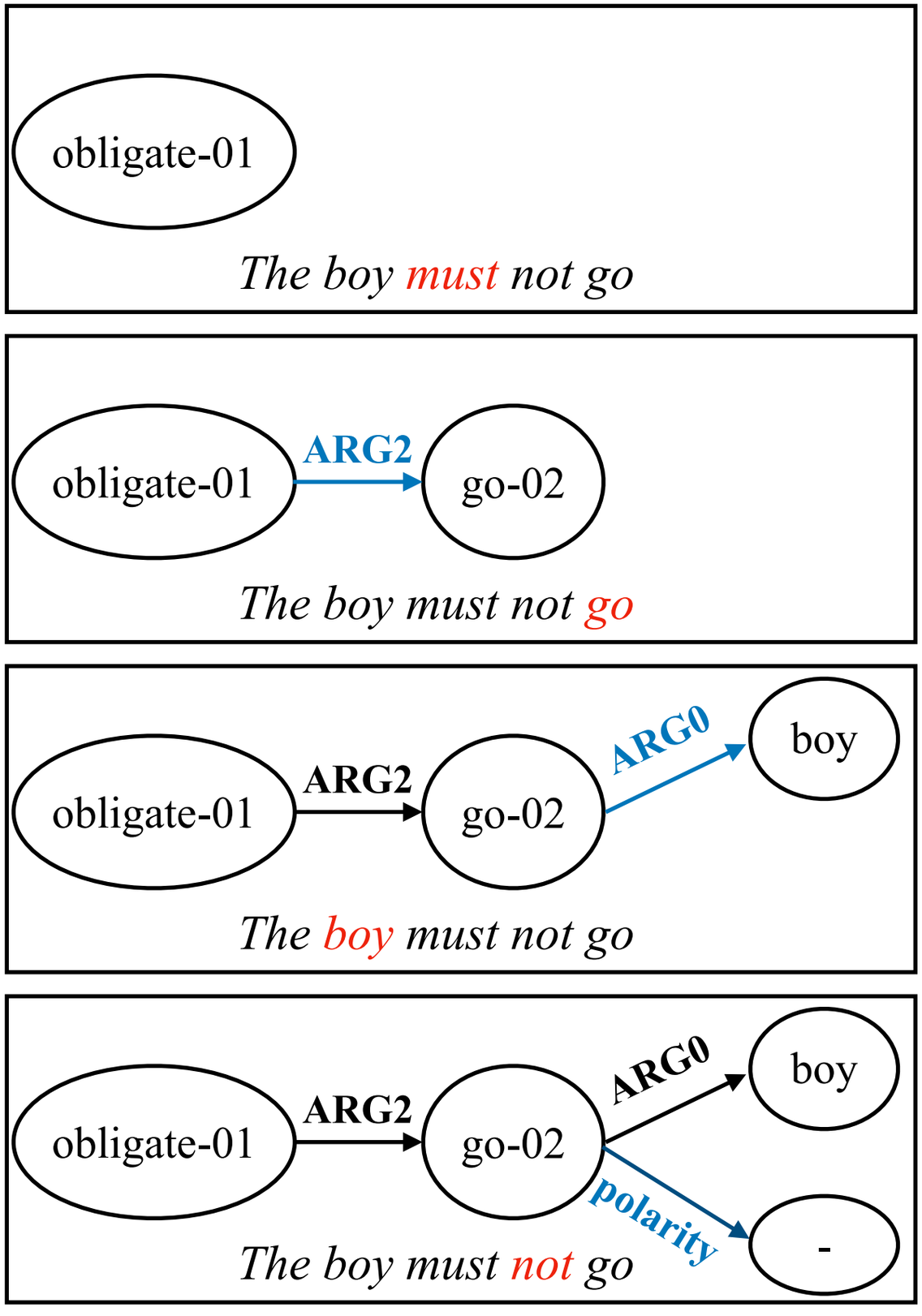
Motivation

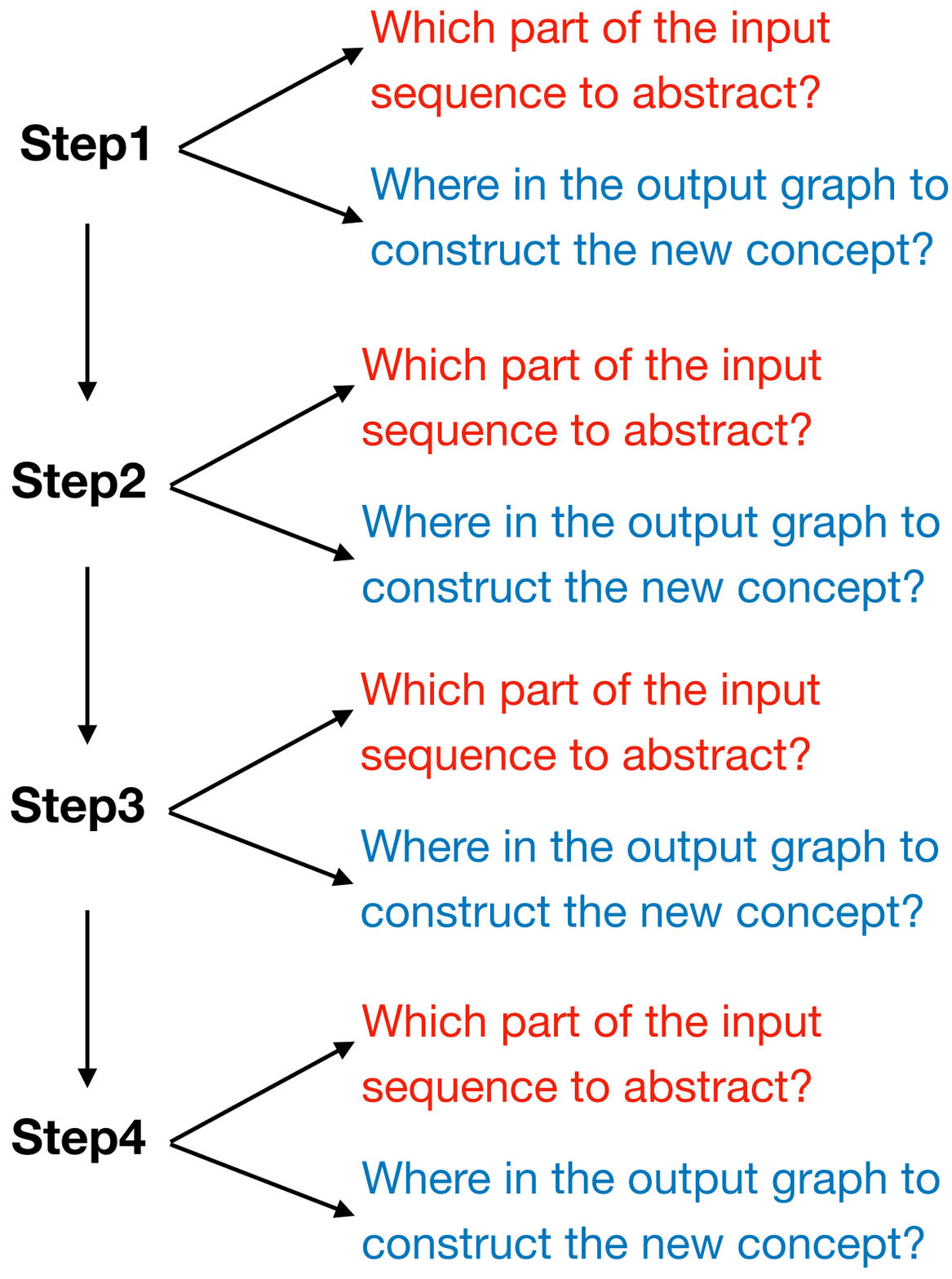
Our hypothesis for unsatisfactory parsing accuracy:

The lack of the modeling capability of the **interactions between concept prediction and relation prediction**

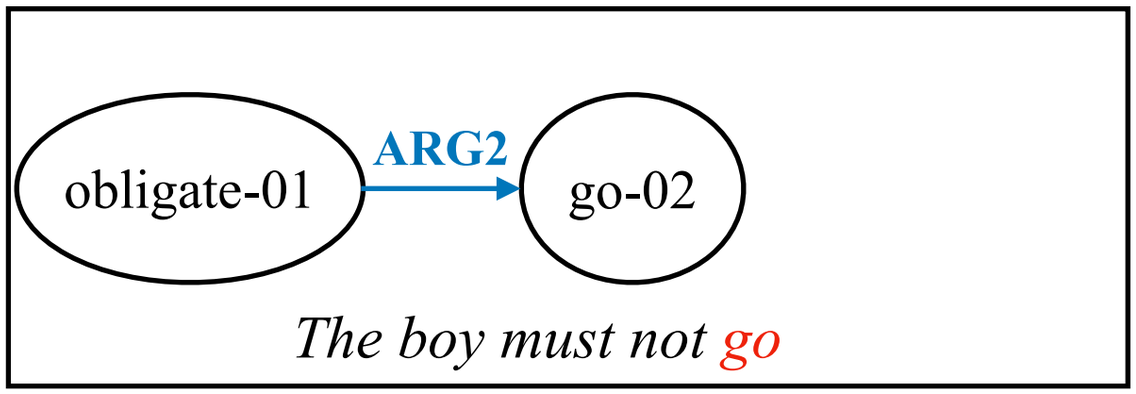


Model Overview



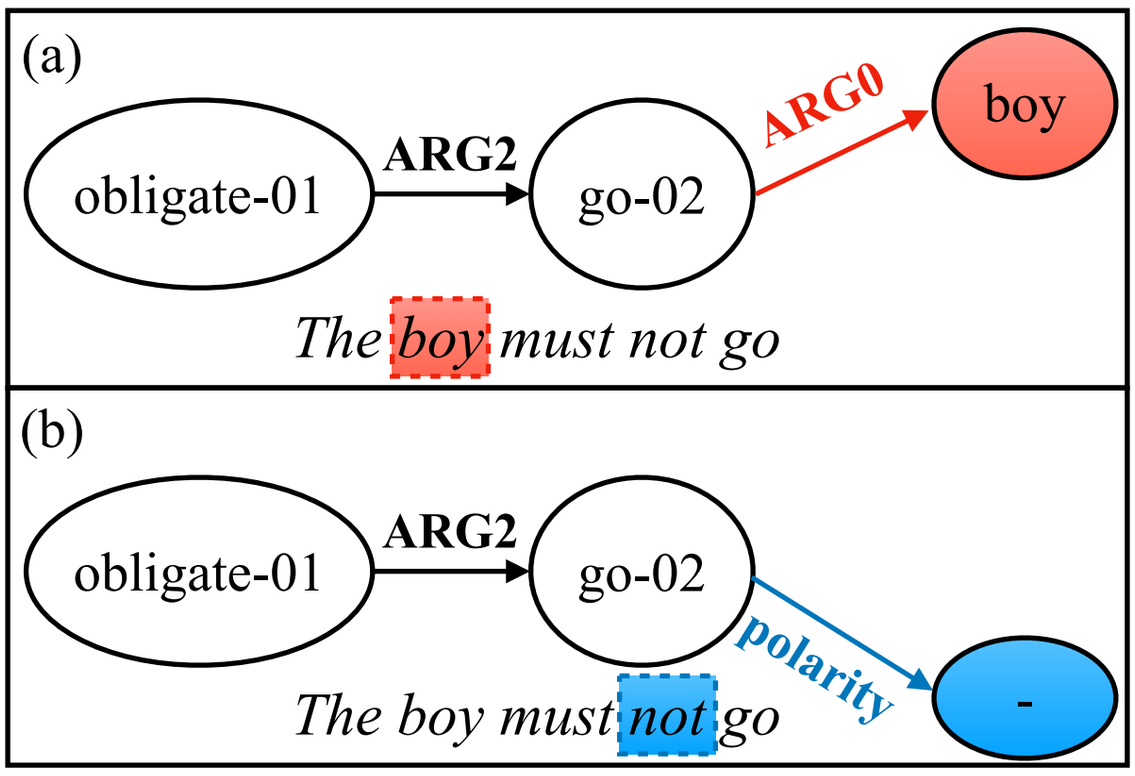


Model Overview



Step2

- Which part of the input sequence to abstract?
- Where in the output graph to construct the new concept?



Model Overview

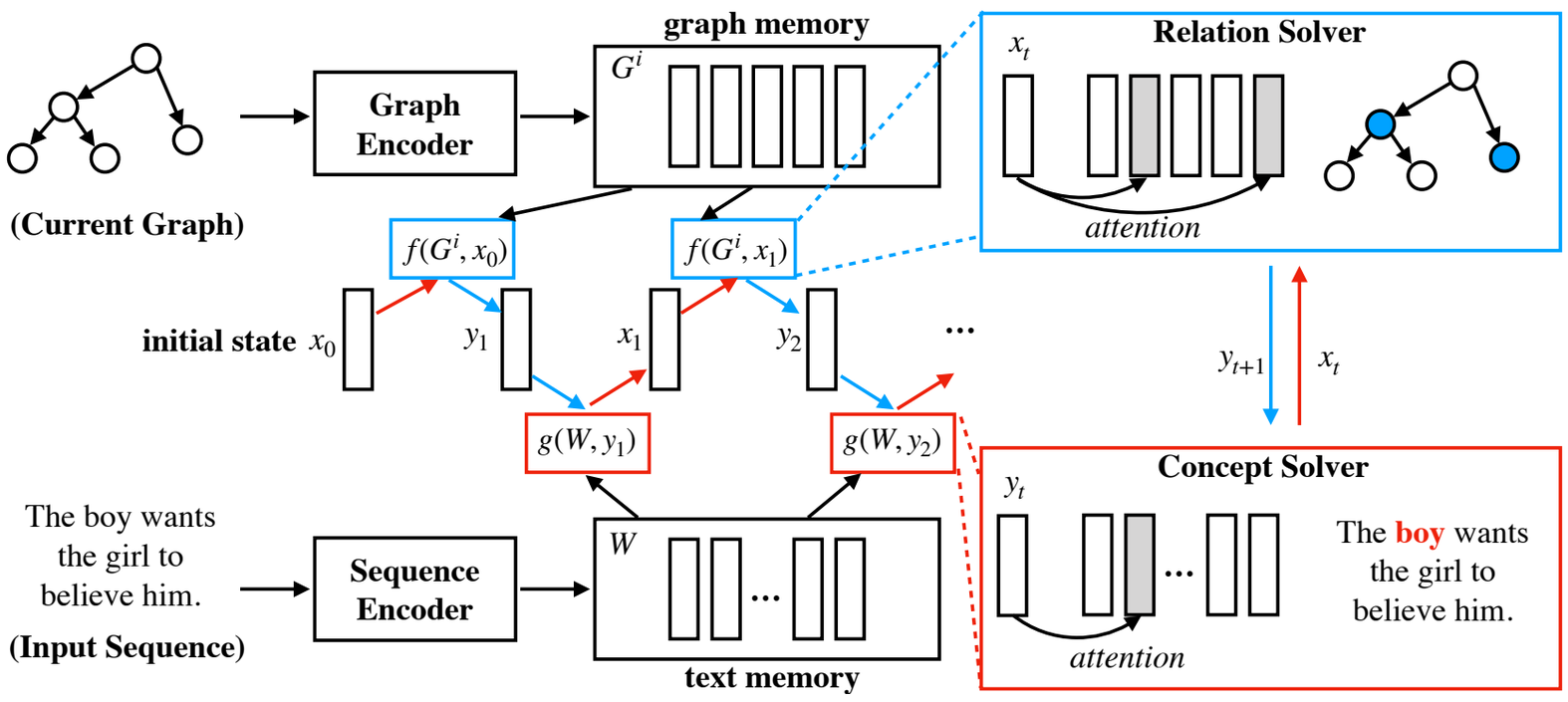
W: the input sentence

G^i: the current graph

x_t: the t-th hypothesis for where to construct (relation prediction)

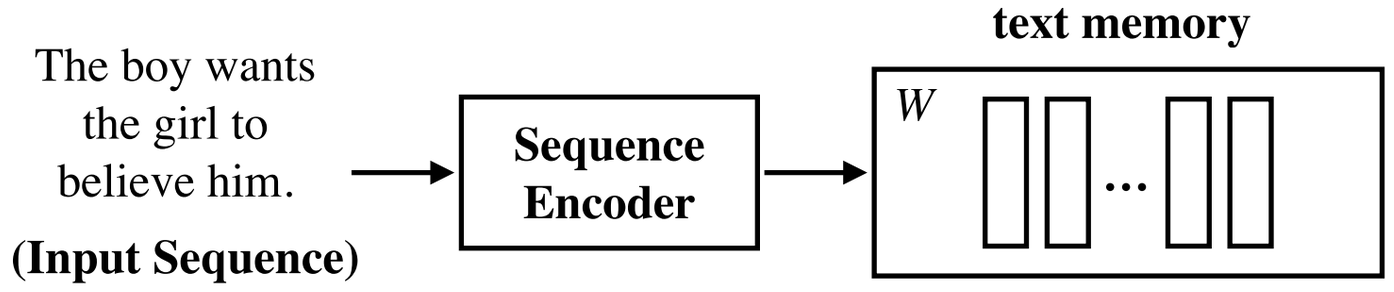
y_t: the t-th hypothesis for what to abstract (concept prediction)

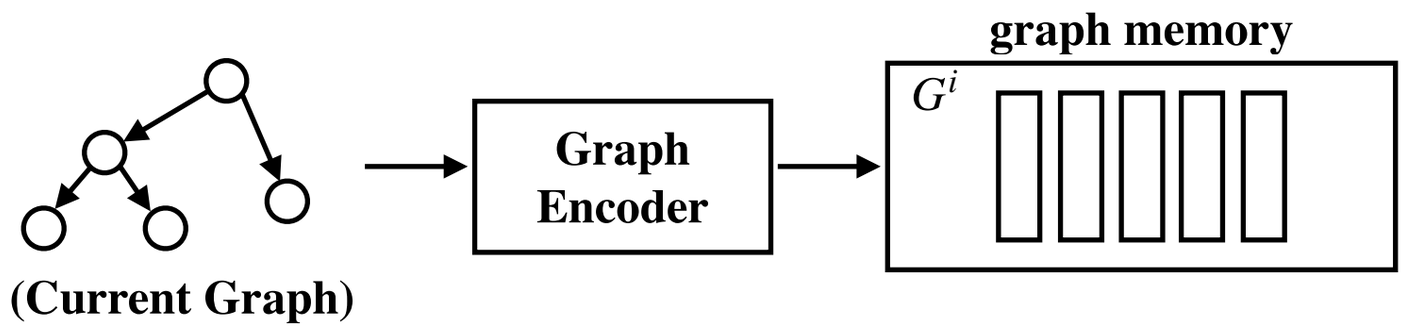
$$x_0 \rightarrow f(G^i, x_0) \rightarrow y_1 \rightarrow g(W, y_1) \rightarrow x_1 \rightarrow f(G^i, x_1) \rightarrow y_2 \rightarrow g(W, y_2) \rightarrow \cdots$$



Model Components

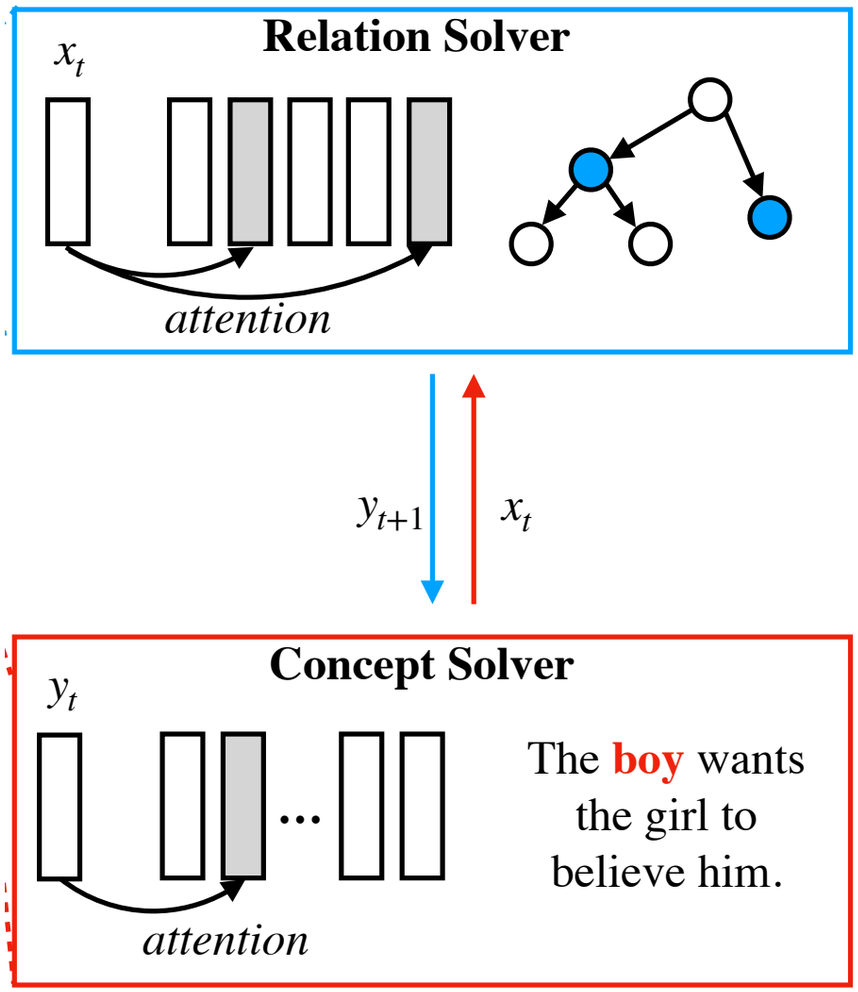
- Sequence Encoder
- Graph Encoder
- Relation Solver
- Concept Solver





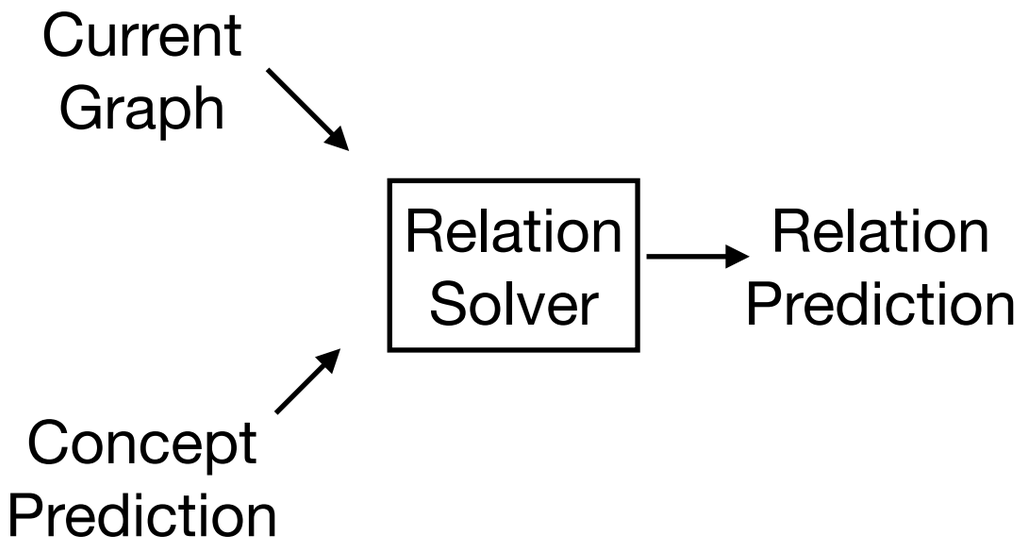
Model Components

- Sequence Encoder
- Graph Encoder
- Relation Solver $\rightarrow f(\cdot)$
- Concept Solver $\rightarrow g(\cdot)$



$$x_0 \rightarrow f(G^i, x_0) \rightarrow y_1 \rightarrow g(W, y_1) \rightarrow x_1 \rightarrow f(G^i, x_1) \rightarrow y_2 \rightarrow g(W, y_2) \rightarrow \cdots$$

Relation Solver



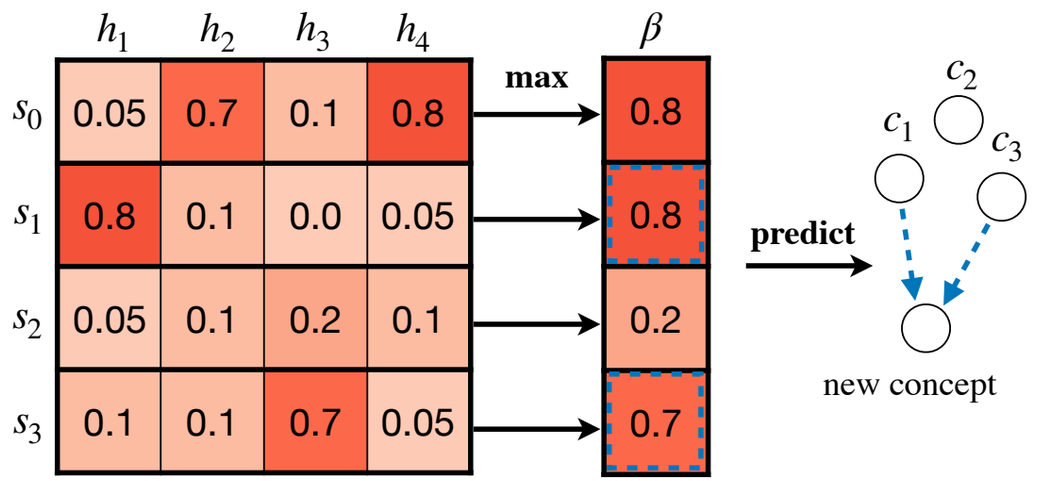
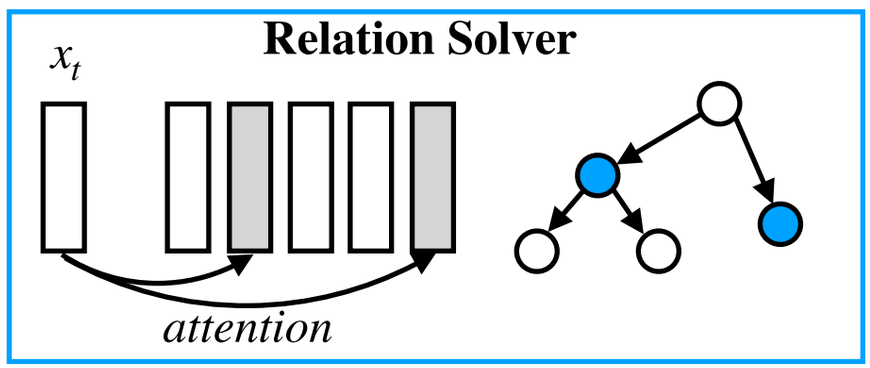
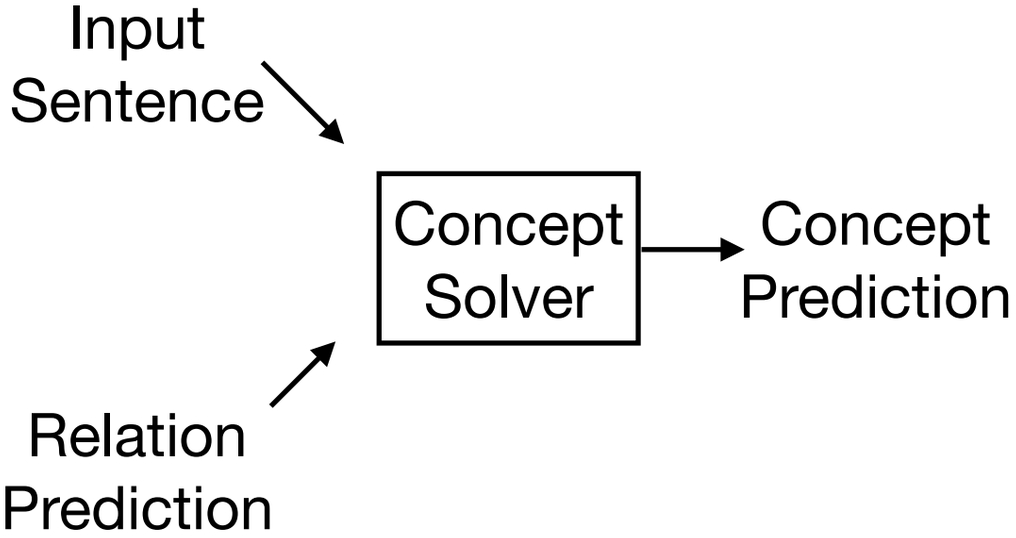
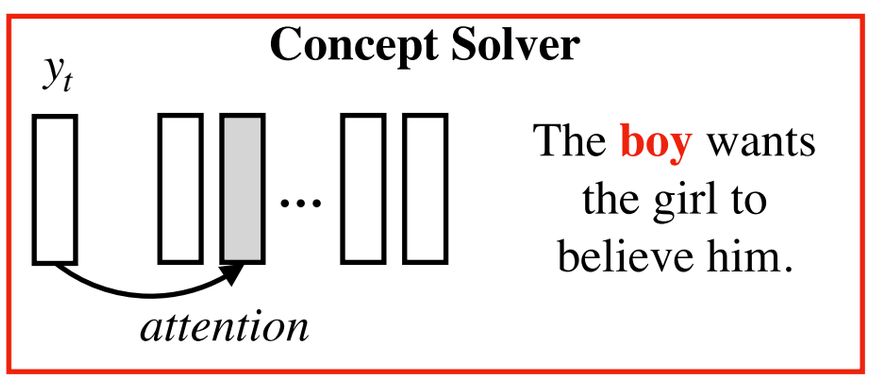


Figure 3: Multi-head attention for relation identification. At left is the attention matrix, where each column corresponds to a unique attention head, and each row corresponds to an existing node.



Concept Solver





Copy mechanisms

0: generate from the concept vocabulary
1: copy the lemma
2: copy the token string

$$P(c) = p_0 \cdot P^{(\text{vocab})}(c) + p_1 \cdot \Big(\sum_{i \in L(c)} \alpha_t[i]\Big) + p_2 \cdot \Big(\sum_{i \in T(c)} \alpha_t[i]\Big),$$

where $[i]$ indexes the i-th element and $L(c)$ and $T(c)$ are index sets of lemmas and tokens respectively that have the surface form as c.

Experiment Setup

- AMR2.0 (LDC2017T10)
  - The latest AMR sembank
  - ~37K, ~1K, and ~1K sentences in the training, development, and testing sets respectively
- AMR1.0 (LDC2014T12)
  - Same dev and test with AMR2.0, ~10K training sentences
  - good testbed to evaluate our model's sensitivity for data size

Evaluation Metrics

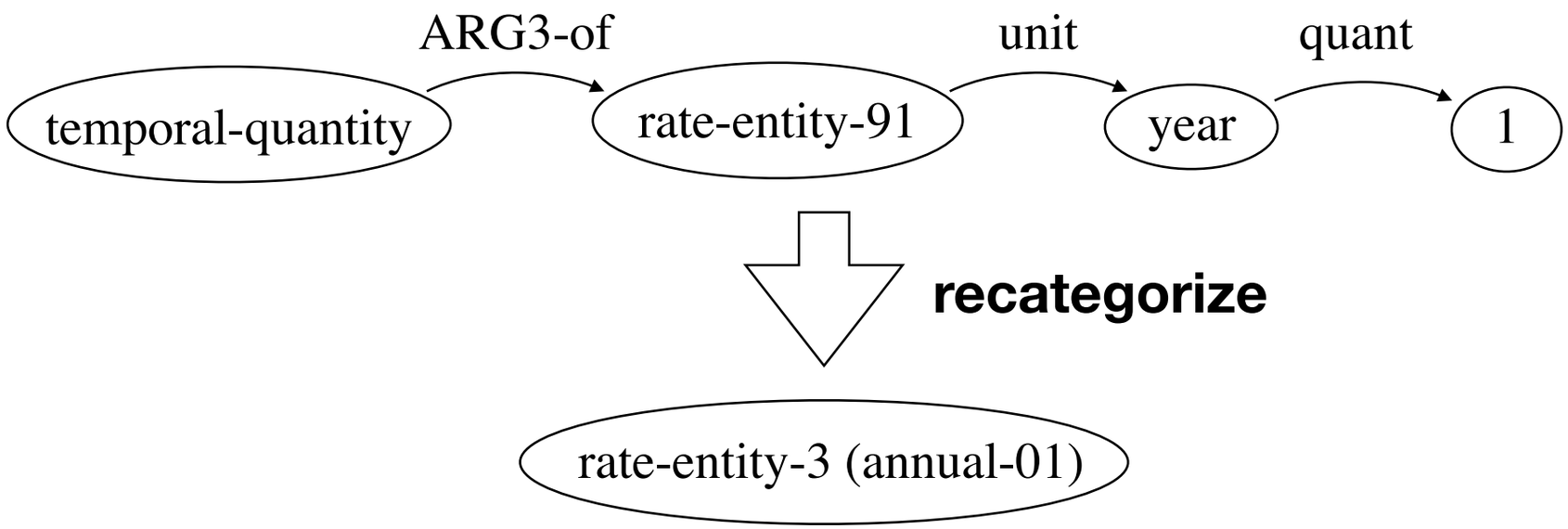
- Smatch (Cai and Knight, 2013) : seeks the maximum overlap after transforming graph into relation triples.
- Fine-grained metrics (Damonte et al, 2017) for individual sub-tasks.
  - NER, SRL, reentrancies, …

Ablation

(settings)

- Graph Re-categorization
- BERT

Graph Re-categorization



- Non-trivial. It requires exhaustive screening and expert-level manual efforts.

- The precise set of re-categorization rules differs among different models.

BERT

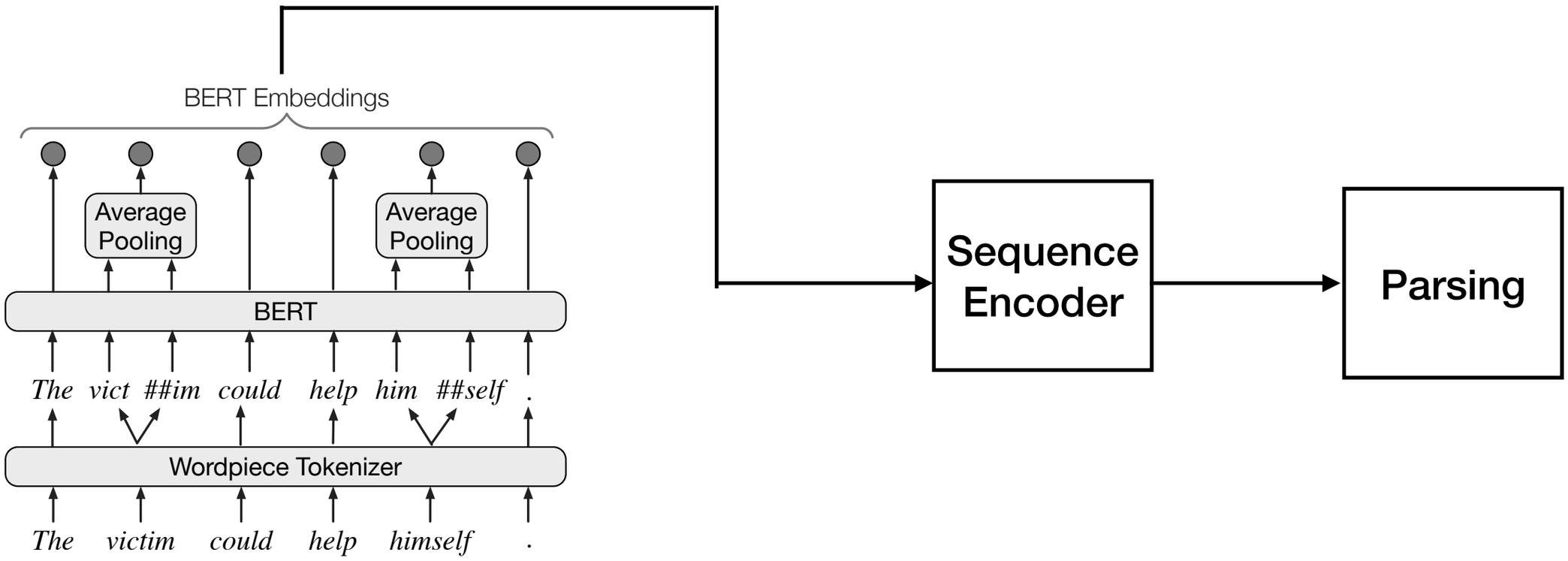


Figure 4: Word-level embeddings from BERT.

** left figure is from (Zhang et al., 2019a)*

Settings

Model	G. R.	BERT	SMATCH
van Noord and Bos (2017)	×	×	71.0
Groschwitz et al. (2018)	✓	×	71.0
Lyu and Titov (2018)	✓	×	74.4
Cai and Lam (2019)	×	×	73.2
Lindemann et al. (2019)	✓	✓	75.3
Naseem et al. (2019)	✓	✓	75.5
Zhang et al. (2019a)	✓	×	74.6
Zhang et al. (2019a)	✓	✓	76.3
Zhang et al. (2019b)	✓	✓	77.0

Table 1: SMATCH scores on the test set of AMR 2.0.

Model	G. R.	BERT	SMATCH
Flanigan et al. (2016)	×	×	66.0
Pust et al. (2015)	×	×	67.1
Wang and Xue (2017)	✓	×	68.1
Guo and Lu (2018)	✓	×	68.3
Zhang et al. (2019a)	✓	✓	70.2
Zhang et al. (2019b)	✓	✓	71.3

Table 2: SMATCH scores on the test set of AMR 1.0.

Settings

Model	G. R.	BERT	SMATCH
van Noord and Bos (2017)	×	×	71.0
Groschwitz et al. (2018)	✓	×	71.0
Lyu and Titov (2018)	✓	×	74.4
Cai and Lam (2019)	×	×	73.2
Lindemann et al. (2019)	✓	✓	75.3
Naseem et al. (2019)	✓	✓	75.5
Zhang et al. (2019a)	✓	×	74.6
Zhang et al. (2019a)	✓	✓	76.3
Zhang et al. (2019b)	✓	✓	77.0
Ours	×	×	
	✓	×	
	×	✓	
	✓	✓	

Table 1: SMATCH scores on the test set of AMR 2.0.

Model	G. R.	BERT	SMATCH
Flanigan et al. (2016)	×	×	66.0
Pust et al. (2015)	×	×	67.1
Wang and Xue (2017)	✓	×	68.1
Guo and Lu (2018)	✓	×	68.3
Zhang et al. (2019a)	✓	✓	70.2
Zhang et al. (2019b)	✓	✓	71.3
Ours	×	×	
	✓	×	
	×	✓	
	✓	✓	

Table 2: SMATCH scores on the test set of AMR 1.0.

Main Results

Model	G. R.	BERT	SMATCH
van Noord and Bos (2017)	×	×	71.0
Groschwitz et al. (2018)	✓	×	71.0
Lyu and Titov (2018)	✓	×	74.4
Cai and Lam (2019)	×	×	73.2
Lindemann et al. (2019)	✓	✓	75.3
Naseem et al. (2019)	✓	✓	75.5
Zhang et al. (2019a)	✓	×	74.6
Zhang et al. (2019a)	✓	✓	76.3
Zhang et al. (2019b)	✓	✓	77.0
Ours	×	×	74.5
	✓	×	77.3
	×	✓	78.7
	✓	✓	**80.2**

+3.2

Table 1: SMATCH scores on the test set of AMR 2.0.

Model	G. R.	BERT	SMATCH
Flanigan et al. (2016)	×	×	66.0
Pust et al. (2015)	×	×	67.1
Wang and Xue (2017)	✓	×	68.1
Guo and Lu (2018)	✓	×	68.3
Zhang et al. (2019a)	✓	✓	70.2
Zhang et al. (2019b)	✓	✓	71.3
Ours	×	×	68.8
	✓	×	71.2
	×	✓	74.0
	✓	✓	**75.4**

Table 2: SMATCH scores on the test set of AMR 1.0.

Main Results

Model	G. R.	BERT	SMATCH
van Noord and Bos (2017)	×	×	71.0
Groschwitz et al. (2018)	✓	×	71.0
Lyu and Titov (2018)	✓	×	74.4
Cai and Lam (2019)	×	×	73.2
Lindemann et al. (2019)	✓	✓	75.3
Naseem et al. (2019)	✓	✓	75.5
Zhang et al. (2019a)	✓	×	74.6
Zhang et al. (2019a)	✓	✓	76.3
Zhang et al. (2019b)	✓	✓	77.0
Ours	×	×	74.5
	✓	×	77.3
	×	✓	78.7
	✓	✓	**80.2**

Table 1: SMATCH scores on the test set of AMR 2.0.

+0.3 w/o BERT

Model	G. R.	BERT	SMATCH
Flanigan et al. (2016)	×	×	66.0
Pust et al. (2015)	×	×	67.1
Wang and Xue (2017)	✓	×	68.1
Guo and Lu (2018)	✓	×	68.3
Zhang et al. (2019a)	✓	✓	70.2
Zhang et al. (2019b)	✓	✓	71.3
Ours	×	×	68.8
	✓	×	71.2
	×	✓	74.0
	✓	✓	**75.4**

Table 2: SMATCH scores on the test set of AMR 1.0.

Main Results

Model	G. R.	BERT	SMATCH
van Noord and Bos (2017)	×	×	71.0
Groschwitz et al. (2018)	✓	×	71.0
Lyu and Titov (2018)	✓	×	74.4
Cai and Lam (2019)	×	×	73.2
Lindemann et al. (2019)	✓	✓	75.3
Naseem et al. (2019)	✓	✓	75.5
Zhang et al. (2019a)	✓	×	74.6
Zhang et al. (2019a)	✓	✓	76.3
Zhang et al. (2019b)	✓	✓	77.0
Ours	×	×	74.5
	✓	×	77.3
	×	✓	78.7
	✓	✓	**80.2**

Table 1: SMATCH scores on the test set of AMR 2.0.

Model	G. R.	BERT	SMATCH
Flanigan et al. (2016)	×	×	66.0
Pust et al. (2015)	×	×	67.1
Wang and Xue (2017)	✓	×	68.1
Guo and Lu (2018)	✓	×	68.3
Zhang et al. (2019a)	✓	✓	70.2
Zhang et al. (2019b)	✓	✓	71.3
Ours	×	×	68.8
	✓	×	71.2
	×	✓	74.0
	✓	✓	**75.4**

+4.1

comparable with models on AMR2.0

Table 2: SMATCH scores on the test set of AMR 1.0.

Main Results

Model	G. R.	BERT	SMATCH
van Noord and Bos (2017)	×	×	71.0
Groschwitz et al. (2018)	✓	×	71.0
Lyu and Titov (2018)	✓	×	74.4
Cai and Lam (2019)	×	×	73.2
Lindemann et al. (2019)	✓	✓	75.3
Naseem et al. (2019)	✓	✓	75.5
Zhang et al. (2019a)	✓	×	74.6
Zhang et al. (2019a)	✓	✓	76.3
Zhang et al. (2019b)	✓	✓	77.0
Ours	×	×	74.5
	✓	×	77.3
	×	✓	78.7
	✓	✓	**80.2**

Table 1: SMATCH scores on the test set of AMR 2.0.

Model	G. R.	BERT	SMATCH
Flanigan et al. (2016)	×	×	66.0
Pust et al. (2015)	×	×	67.1
Wang and Xue (2017)	✓	×	68.1
Guo and Lu (2018)	✓	×	68.3
Zhang et al. (2019a)	✓	✓	70.2
Zhang et al. (2019b)	✓	✓	71.3
Ours	×	×	68.8
	✓	×	71.2
	×	✓	74.0
	✓	✓	**75.4**

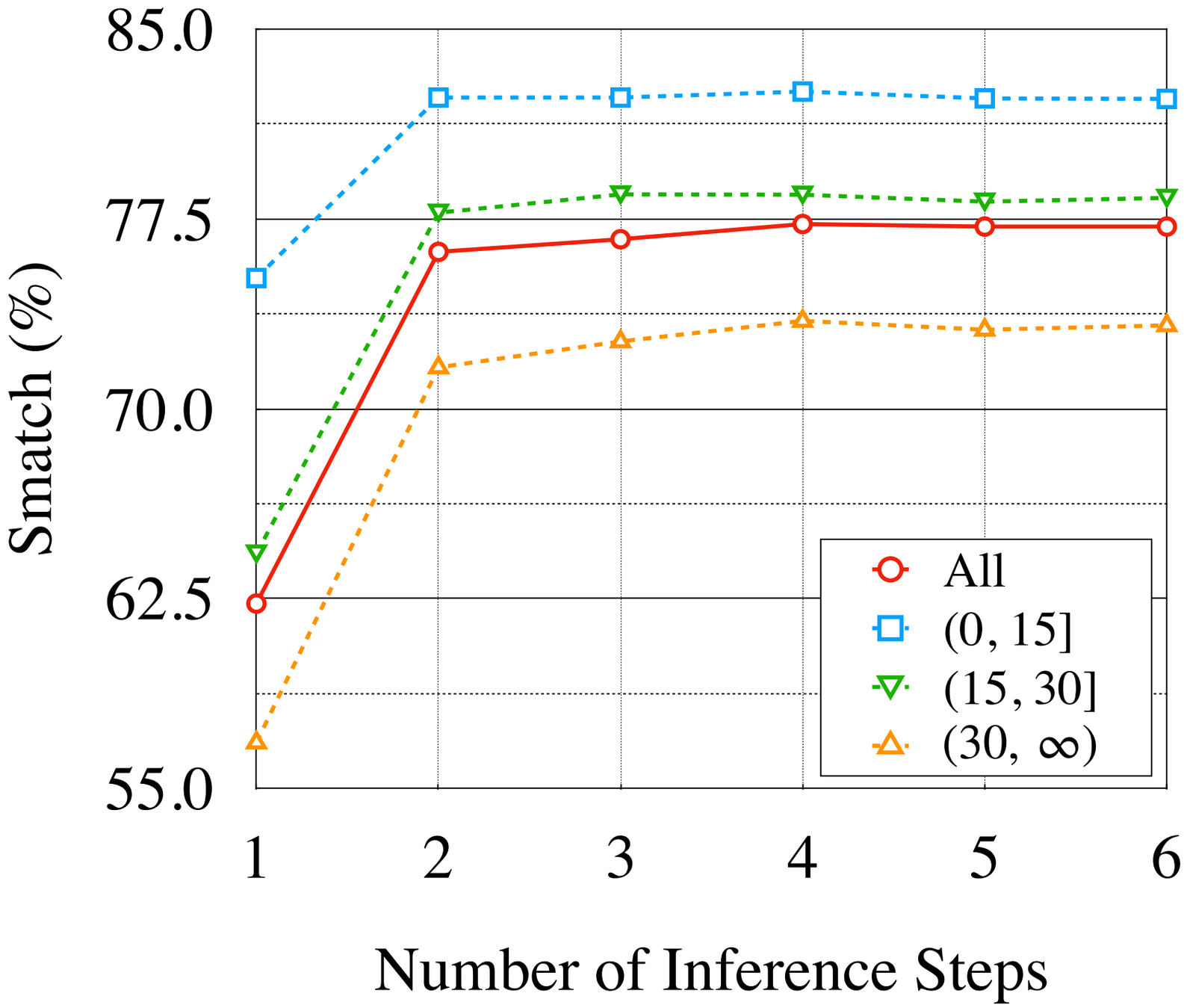
Table 2: SMATCH scores on the test set of AMR 1.0.

Fine-grained Results

Model	G. R.	BERT	SMATCH	fine-grained evaluation							
				Unlabeled	No WSD	Concept	SRL	Reent.	Neg.	NER	Wiki
van Noord and Bos (2017)	×	×	71.0	74	72	82	66	52	62	79	65
Groschwitz et al. (2018)	✓	×	71.0	74	72	84	64	49	57	78	71
Lyu and Titov (2018)	✓	×	74.4	77.1	75.5	85.9	69.8	52.3	58.4	86.0	75.7
Cai and Lam (2019)	×	×	73.2	77.0	74.2	84.4	66.7	55.3	62.9	82.0	73.2
Lindemann et al. (2019)	✓	✓	75.3	-	-	-	-	-	-	-	-
Naseem et al. (2019)	✓	✓	75.5	80	76	86	72	56	67	83	80
Zhang et al. (2019a)	✓	×	74.6	-	-	-	-	-	-	-	-
Zhang et al. (2019a)	✓	✓	76.3	79.0	76.8	84.8	69.7	60.0	75.2	77.9	85.8
Zhang et al. (2019b)	✓	✓	77.0	80	78	86	71	61	77	79	86
Ours	×	×	74.5	77.8	75.1	85.9	68.5	57.7	65.0	82.9	81.1
	✓	×	77.3	80.1	77.9	86.4	69.4	58.5	75.6	78.4	86.1
	×	✓	78.7	81.5	79.2	88.1	**74.5**	63.8	66.1	**87.1**	81.3
	✓	✓	**80.2**	**82.8**	**80.8**	**88.1**	74.2	**64.6**	**78.9**	81.1	**86.3**

Table 3 : Fine-grained results on the test set of AMR 2.0.

Effect of Iterative Inference



AMR Parsing via Graph⇆Sequence Iterative Inference

Thanks!

Deng Cai and Wai Lam
The Chinese University of Hong Kong

https://github.com/jcyk/AMR-gs
thisisjcykcd@gmail.com