Association for Computational Linguistics

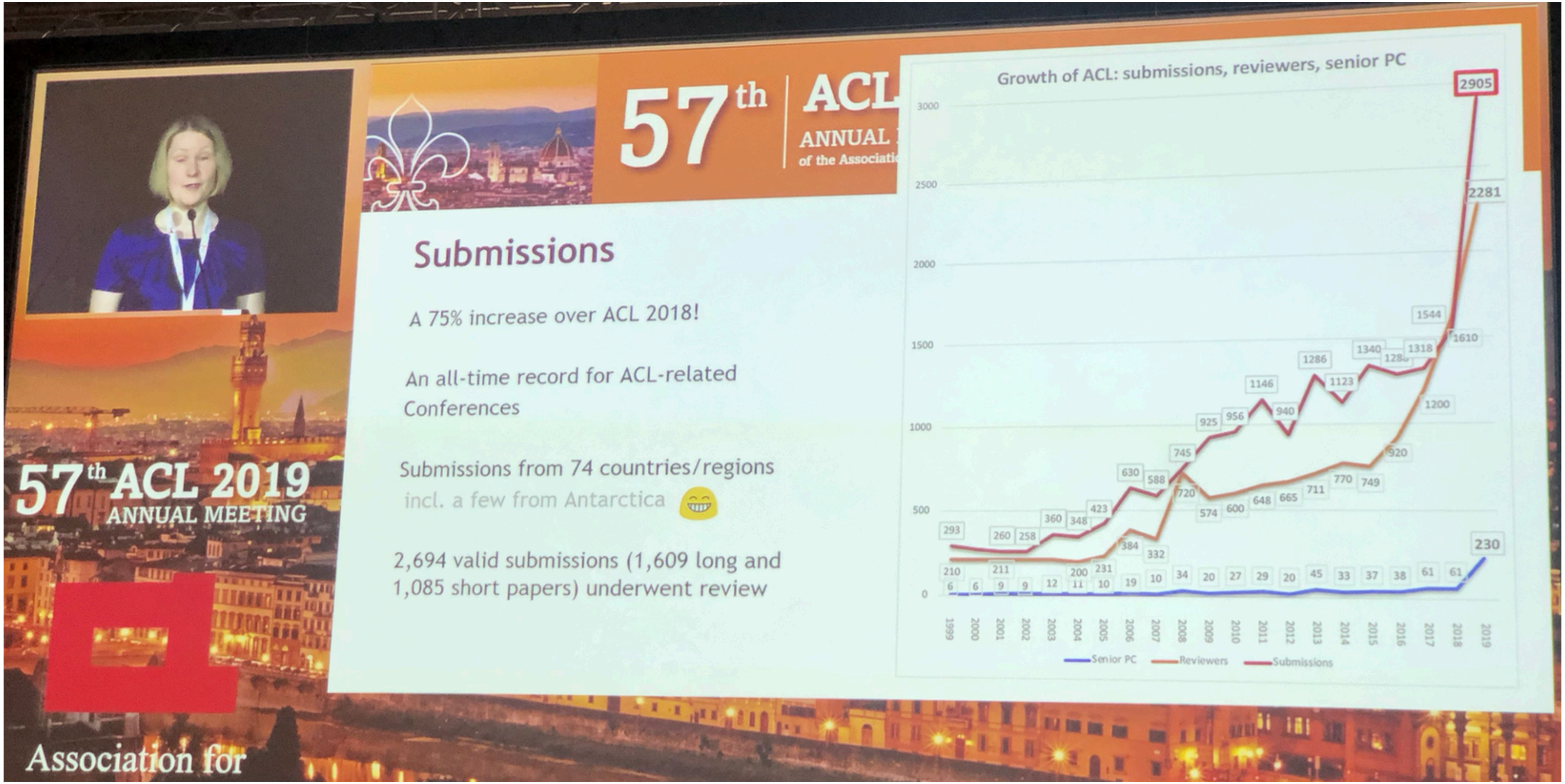
57th | ACL 2019 Florence
ANNUAL MEETING July 28th - August 2nd
of the Association for Computational Linguistics

Conference Report

AI Lab – NLP center

Jiangtong Li

Basic Statistics



Basic Statistics

Conference		Submissions	Accepts	Accept rate (%)
ACL 2019	**All**	**2905**	**660**	**22.7**
	Long	1737	447	25.7
	Short	1168	213	18.2
ACL 2018	**All**	**1544**	**384**	**24.9**
	Long	1018	258	25.3
	Short	526	126	24.0
ACL 2017	**All**	**1297**	**302**	**23.3**
	Long	737	195	26.5
	Short	560	107	19.1

Basic Statistics

	Area	All submissions	Accepts	Accept rate (%)
1.	Applications	136	32	23.5
2.	Dialogue and Interactive Systems	183	52	28.4
3.	Discourse and Pragmatics	55	15	27.3
4.	Document Analysis	81	15	18.5
5.	Generation	153	40	26.1
16.	Information Extraction and Text Mining	247	51	20.6
7.	Linguistic Theories, Cognitive Modeling and Psycholinguistics	60	14	23.3
8.	Machine Learning	223	56	25.1
8.	Machine Translation	205	46	22.4
10.	Multidisciplinary and Area Chair COI	112	35	31.3
11.	Multilinguality	75	21	28.0
12.	Phonology Morphology and Word Segmentation	43	9	20.9
13.	Question Answering	155	39	25.2
14.	Resources and Evaluation	128	36	28.1
15.	Sentence-level semantics	111	22	19.8
15.	Sentiment Analysis and Argument Mining	150	33	22.0
17.	Social Media	93	23	24.7
18.	Summarization	81	21	25.9
19.	Tagging Chunking Syntax and Parsing	99	27	27.3
20.	Textual Inference and Other Areas of Semantics	74	21	28.0
21.	Vision Robotics Multimodal Grounding and Speech	80	24	30.0
22.	Word-level Semantics	135	28	20.7
	Desk reject or withdrawn	225		
	Total	**2905**	**660**	**22.7**

	Area	Long submissions	Accepts	Accept rate (%)
1.	Applications	65	14	28.8
2.	Dialogue and Interactive Systems	126	38	30.2
3.	Discourse and Pragmatics	33	7	21.2
4.	Document Analysis	48	8	16.7
5.	Generation	96	32	33.3
6.	Information Extraction and Text Mining	155	37	23.9
7.	Linguistic Theories, Cognitive Modeling and Psycholinguistics	39	9	23.1
8.	Machine Learning	148	38	25.7
8.	Machine Translation	102	27	26.5
10.	Multidisciplinary and Area Chair COI	69	21	30.4
11.	Multilinguality	43	11	25.6
12.	Phonology Morphology and Word Segmentation	26	7	26.9
13.	Question Answering	99	32	32.3
14.	Resources and Evaluation	70	26	37.1
15.	Sentence-level semantics	69	14	20.3
15.	Sentiment Analysis and Argument Mining	91	24	26.4
17.	Social Media	51	14	27.5
18.	Summarization	48	11	22.9
19.	Tagging Chunking Syntax and Parsing	50	17	34.0
20.	Textual Inference and Other Areas of Semantics	44	16	36.4
21.	Vision Robotics Multimodal Grounding and Speech	56	20	35.7
22.	Word-level Semantics	78	20	25.6
	Desk reject or withdrawn	131		
	Total	**1737**	**447**	**25.7**

	Area	Short submissions	Accepts	Accept rate (%)
1.	Applications	71	43	19.7
2.	Dialogue and Interactive Systems	57	14	24.6
3.	Discourse and Pragmatics	22	8	36.4
4.	Document Analysis	33	7	21.2
5.	Generation	57	8	14.0
6.	Information Extraction and Text Mining	92	14	15.2
7.	Linguistic Theories, Cognitive Modeling and Psycholinguistics	21	5	23.8
8.	Machine Learning	75	18	24.0
8.	Machine Translation	103	19	18.4
10.	Multidisciplinary and Area Chair COI	43	14	32.6
11.	Multilinguality	32	10	31.3
12.	Phonology Morphology and Word Segmentation	17	2	11.8
13.	Question Answering	56	7	12.5
14.	Resources and Evaluation	58	10	17.2
15.	Sentence-level semantics	42	8	19.0
15.	Sentiment Analysis and Argument Mining	59	9	15.3
17.	Social Media	42	9	21.4
18.	Summarization	33	10	30.3
19.	Tagging Chunking Syntax and Parsing	49	10	20.4
20.	Textual Inference and Other Areas of Semantics	31	5	16.1
21.	Vision Robotics Multimodal Grounding and Speech	24	4	16.7
22.	Word-level Semantics	57	8	14.0
	Desk reject or withdrawn	94		
	Total	**1168**	**213**	**18.2**

Outline

- 1. Bridging the Gap between Training and Inference for Neural Machine Translation
- 2. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs
- 3. Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study
- 4. Generating Fluent Adversarial Examples for Natural Languages
- 5. Dynamically Fused Graph Network for Multi-hop Reasoning
- 6. Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog

Bridging the Gap between Training and Inference for Neural Machine Translation

- Motivation
  - At training time, it predicts with the ground truth words as context while at inference it has to generate the entire sequence from scratch.
  - Word-level training requires strict matching between the generated sequence and the ground truth sequence which leads to overcorrection over different but reasonable translations.
- Solution
  - Use oracle/GT word as the prefix to predict the next word
  - Word-level oracle: Gumbel-Max sampling
  - Sentence-level oracle: Beam search sampling

Systems	**Architecture**	**MT03**	**MT04**	**MT05**	**MT06**	**Average**
Existing end-to-end NMT systems						
Tu et al. (2016)	Coverage	33.69	38.05	35.01	34.83	35.40
Shen et al. (2016)	MRT	37.41	39.87	37.45	36.80	37.88
Zhang et al. (2017)	Distortion	37.93	40.40	36.81	35.77	37.73
Our end-to-end NMT systems						
this work	RNNsearch	37.93	40.53	36.65	35.80	37.73
	+ SS-NMT	38.82	41.68	37.28	37.98	38.94
	+ MIXER	38.70	40.81	37.59	38.38	38.87
	+ OR-NMT	**40.40**‡†*	**42.63**‡†*	**38.87**‡†*	**38.44**‡	**40.09**
	Transformer	46.89	47.88	47.40	46.66	47.21
	+ word oracle	47.42	48.34	47.89	47.34	47.75
	+ sentence oracle	**48.31***	**49.40***	**48.72***	**48.45***	**48.72**

OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs

- Motivation & Tasks
    - While a large-scale knowledge graph (KG) includes vast knowledge, the core challenge is in the domain-agnostic and scalable prediction of a small subset from those reachable entities that follows natural conceptual threads that can keep conversations engaging and meaningful.
    - Given a set of KG entity mentions from current turn, and dialog history of all current and previous sentences, the goal is to build a robust model that can retrieve a set of natural entities to mention from a large-scale KG that resemble human responses.
- Solution

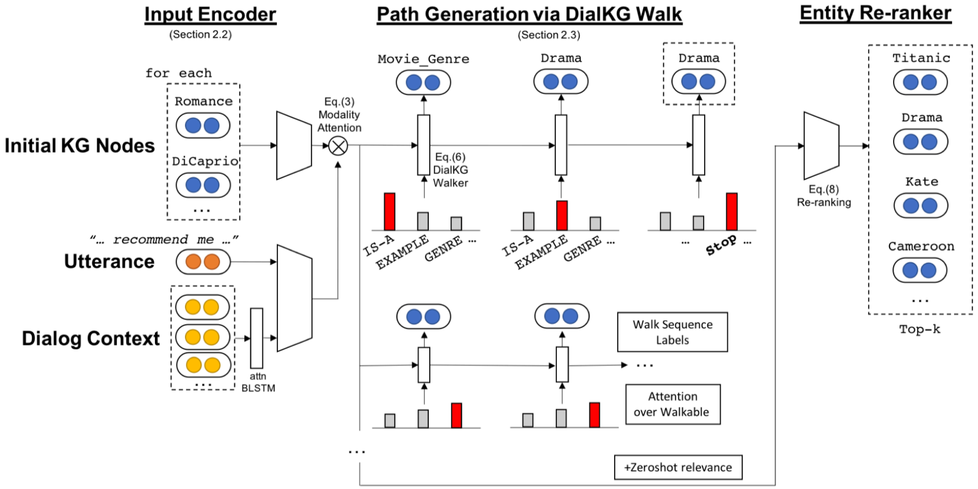


Figure 2: **Overall architecture**. $\mathbf{x} = \{\mathbf{x}_e; \mathbf{x}_s; \mathbf{x}_d\}$ is encoded with the input encoder (left), aggregated via multiple attention mechanism. The decoder (right) predicts both the optimal paths and the final entities $\mathbf{y} = \{\mathbf{y}_e; \mathbf{y}_r\}$ based on their zeroshot relevance scores as well as soft-attention based walk paths, which prunes unlikely entities.

Input	Model	All Domains → All					Movie → Movie				
		$r@1$	3	5	10	25	$r@1$	3	5	10	25
E + S + D	`seq2seq` (Sutskever et al., 2014)	3.1	18.3	29.7	44.1	60.2	3.0	13.4	23.4	38.5	55.5
E + S	`Tri-LSTM` (Young et al., 2018)	3.2	14.2	22.6	36.3	56.2	1.5	10.3	17.4	30.7	51.1
E + S	`Ext-ED` (Parthasarathi and Pineau, 2018)	1.9	5.8	9.0	13.3	19.0	1.3	5.4	7.8	11.8	15.8
E	`DialKG Walker` (ablation)	10.7	22.9	32.0	44.9	57.4	5.3	13.5	18.5	25.2	39.1
E + S	`DialKG Walker` (ablation)	11.3	23.3	31.0	44.0	60.5	7.2	19.2	**27.9**	**40.7**	**58.7**
E + S + D	`DialKG Walker` (**proposed**)	**13.2**	**26.1**	**35.3**	**47.9**	**62.2**	**7.8**	**20.0**	**27.9**	40.4	58.6

Table 2: In-domain (train/test on the same domain) response generation performance on the *OpenDialKG* dataset (metric: recall@k). Our proposed model is compared against state-of-the-art models as well as several ablation variations of the proposed model. All of the 100K+ KG entities are considered initial candidates for generation (before masking). E: entities, S: sentence, D: dialog contexts.

Input	Model	Movie → Book					Movie → Music				
		$r@1$	3	5	10	25	$r@1$	3	5	10	25
E + S + D	`seq2seq` (Sutskever et al., 2014)	2.9	21.3	35.1	50.6	64.2	1.5	12.1	19.7	34.9	49.4
E + S	`Tri-LSTM` (Young et al., 2018)	2.3	17.9	29.7	44.9	61.0	1.9	8.7	12.9	25.8	44.4
E + S	`Ext-ED` (Parthasarathi and Pineau, 2018)	2.0	7.9	11.2	16.4	22.4	1.3	2.6	3.8	4.1	8.3
E	`DialKG Walker` (ablation)	8.2	15.7	22.8	31.8	48.9	4.5	**16.7**	21.6	25.8	33.0
E + S	`DialKG Walker` (ablation)	12.6	28.6	38.6	**54.1**	**65.6**	**6.0**	15.9	**22.8**	**33.0**	**47.5**
E + S + D	`DialKG Walker` (**proposed**)	**13.5**	**28.8**	**39.5**	52.6	64.8	5.3	13.3	19.7	28.8	38.0

Table 3: Cross-domain (train/test on the different domain) response generation performance on the *OpenDialKG* dataset (metric: recall@k). E: entities, S: sentence, D: dialog contexts.

Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study

- Motivation & Tasks
  - A common criticism of current dialogue systems is that they understand or use the available dialog history effectively.
  - This paper take an empirical approach to understanding how these models use the available dialog history by studying the sensitivity of the models to artificially introduced unnatural changes or perturbations to their context at test time.
- Solution
  - Type of Perturbations
    - Utterance-level: (1) *Shuf* (2)*Rev* (3)*Drop* (4)*Truncate*
    - Word-level: (1)*Word-shuf* (2)*Rev* (3)*Word-drop* (4)*Noun-drop* (5)*Verb-drop*

Models	Test PPL	Only Last	Shuf	Rev	Drop First	Drop Last	Word Drop	Verb Drop	Noun Drop	Word Shuf	Word Rev
		Utterance level perturbations ($\Delta PPL_{[\sigma]}$)					Word level perturbations ($\Delta PPL_{[\sigma]}$)				
DailyDialog											
seq2seq_lstm	32.90$_{[1.40]}$	1.70$_{[0.41]}$	**3.35**$_{[0.38]}$	**4.04**$_{[0.28]}$	0.13$_{[0.04]}$	**5.08**$_{[0.79]}$	1.58$_{[0.15]}$	0.87$_{[0.08]}$	1.06$_{[0.28]}$	**3.37**$_{[0.33]}$	3.10$_{[0.45]}$
seq2seq_lstm_att	29.65$_{[1.10]}$	**4.76**$_{[0.39]}$	2.54$_{[0.24]}$	3.31$_{[0.49]}$	**0.32**$_{[0.03]}$	4.84$_{[0.42]}$	**2.03**$_{[0.25]}$	**1.37**$_{[0.29]}$	2.22$_{[0.22]}$	2.82$_{[0.31]}$	**3.29**$_{[0.25]}$
transformer	**28.73**$_{[1.30]}$	3.28$_{[1.37]}$	0.82$_{[0.40]}$	1.25$_{[0.62]}$	0.27$_{[0.19]}$	2.43$_{[0.83]}$	1.20$_{[0.69]}$	0.63$_{[0.17]}$	**2.60**$_{[0.98]}$	0.15$_{[0.08]}$	0.26$_{[0.18]}$
Persona Chat											
seq2seq_lstm	43.24$_{[0.99]}$	3.27$_{[0.13]}$	6.29$_{[0.48]}$	**13.11**$_{[1.22]}$	0.47$_{[0.21]}$	**6.10**$_{[0.46]}$	1.81$_{[0.25]}$	0.68$_{[0.19]}$	0.75$_{[0.15]}$	1.29$_{[0.17]}$	1.95$_{[0.20]}$
seq2seq_lstm_att	42.90$_{[1.76]}$	**4.44**$_{[0.81]}$	**6.70**$_{[0.67]}$	11.61$_{[0.75]}$	**2.99**$_{[2.24]}$	5.58$_{[0.45]}$	**2.47**$_{[0.67]}$	**1.11**$_{[0.27]}$	**1.20**$_{[0.23]}$	**2.03**$_{[0.46]}$	**2.39**$_{[0.31]}$
transformer	**40.78**$_{[0.31]}$	1.90$_{[0.08]}$	1.22$_{[0.22]}$	1.41$_{[0.54]}$	$-0.1$$_{[0.07]}$	1.59$_{[0.39]}$	0.54$_{[0.08]}$	0.40$_{[0.00]}$	0.32$_{[0.18]}$	0.01$_{[0.01]}$	0.00$_{[0.06]}$
MutualFriends											
seq2seq_lstm	14.17$_{[0.29]}$	1.44$_{[0.86]}$	**1.42**$_{[0.25]}$	1.24$_{[0.34]}$	0.00$_{[0.00]}$	0.76$_{[0.10]}$	0.28$_{[0.11]}$	0.00$_{[0.03]}$	0.61$_{[0.39]}$	0.31$_{[0.25]}$	0.56$_{[0.39]}$
seq2seq_lstm_att	**10.60**$_{[0.21]}$	**32.13**$_{[4.08]}$	1.24$_{[0.19]}$	1.06$_{[0.24]}$	0.08$_{[0.03]}$	**1.35**$_{[0.15]}$	**1.56**$_{[0.20]}$	0.15$_{[0.07]}$	**3.28**$_{[0.38]}$	**2.35**$_{[0.22]}$	**4.59**$_{[0.46]}$
transformer	10.63$_{[0.03]}$	20.11$_{[0.67]}$	1.06$_{[0.16]}$	**1.62**$_{[0.44]}$	**0.12**$_{[0.03]}$	0.81$_{[0.09]}$	0.75$_{[0.05]}$	**0.16**$_{[0.02]}$	1.50$_{[0.12]}$	0.07$_{[0.01]}$	0.13$_{[0.04]}$
bAbi dailog: Task5											
seq2seq_lstm	1.28$_{[0.02]}$	1.31$_{[0.50]}$	**43.61**$_{[15.9]}$	**40.99**$_{[9.38]}$	0.00$_{[0.00]}$	4.28$_{[1.90]}$	0.38$_{[0.11]}$	0.01$_{[0.00]}$	0.10$_{[0.06]}$	0.09$_{[0.02]}$	0.42$_{[0.38]}$
seq2seq_lstm_att	**1.06**$_{[0.02]}$	**9.14**$_{[1.28]}$	41.21$_{[8.03]}$	34.32$_{[10.7]}$	0.00$_{[0.00]}$	**6.75**$_{[1.86]}$	**0.64**$_{[0.07]}$	0.03$_{[0.03]}$	0.22$_{[0.04]}$	**0.25**$_{[0.01]}$	**1.10**$_{[0.80]}$
transformer	1.07$_{[0.00]}$	4.06$_{[0.33]}$	0.38$_{[0.02]}$	0.62$_{[0.02]}$	0.00$_{[0.00]}$	0.21$_{[0.02]}$	0.36$_{[0.02]}$	**0.25**$_{[0.06]}$	**0.37**$_{[0.06]}$	0.00$_{[0.00]}$	0.00$_{[0.00]}$

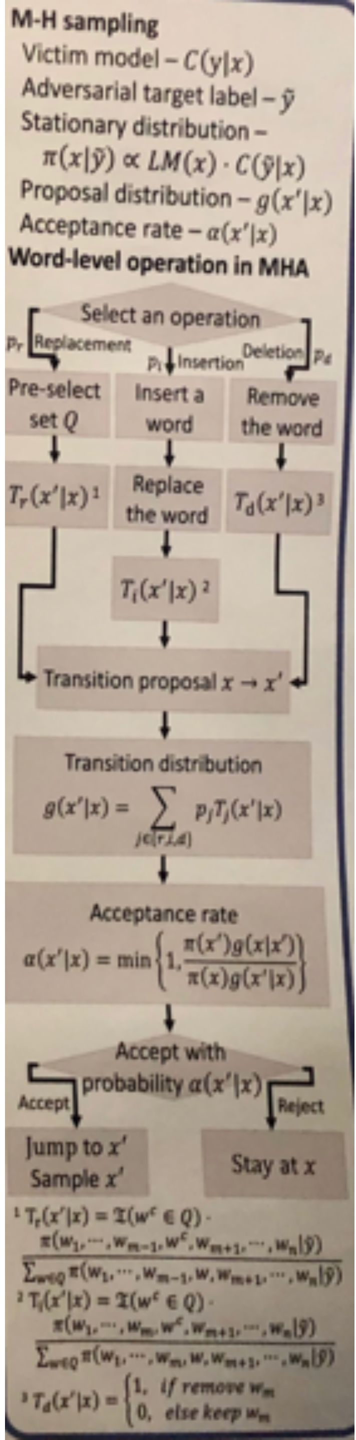
Generating Fluent Adversarial Examples for Natural Languages

- Motivation & Tasks
  - Efficiently building an adversarial attacker for natural language processing is challenging.
    - Sentence space is discrete and it is difficult to make small perturbations along the direction of gradients.
    - The fluency of the generated examples cannot be guaranteed.
- Solution
  - Black-box / White-box Attact
    - Overall Structure
    - Different lies in the pre-selector
      - For Black-box

$$S^B(w|x) = LM(w|x_{[1:m-1]}) \cdot LM_b(w|x_{[m+1:n]})$$

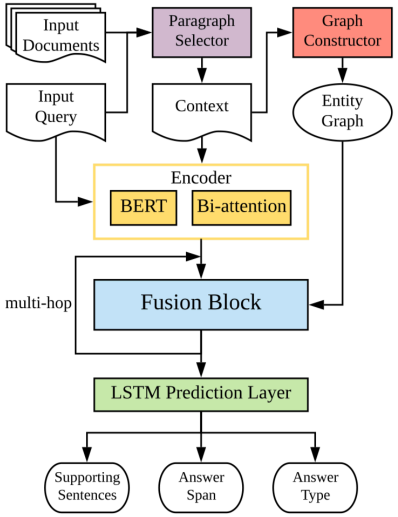
      - For White-box

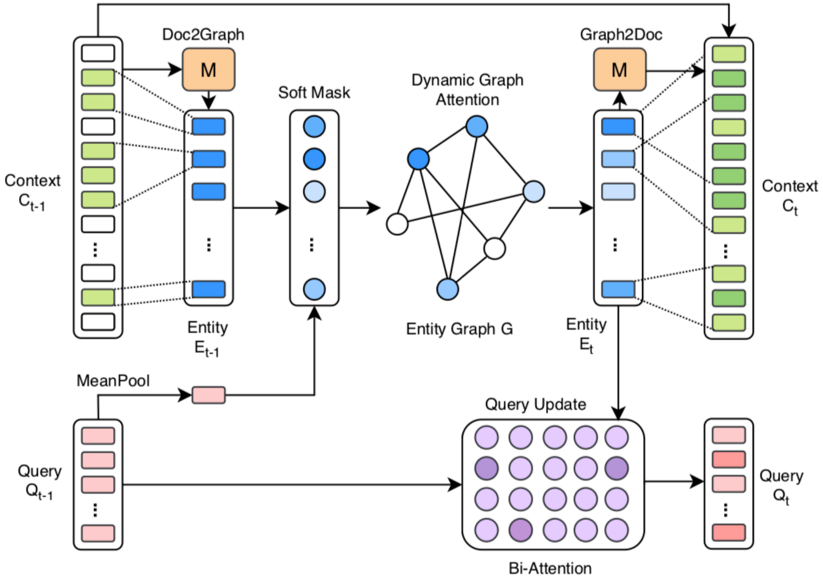
$$S^W(w|x) = S^B(w|x) \cdot S(\frac{\partial \tilde{\mathcal{L}}}{\partial e_m}, e_m - e)$$



Dynamically Fused Graph Network for Multi-hop Reasoning

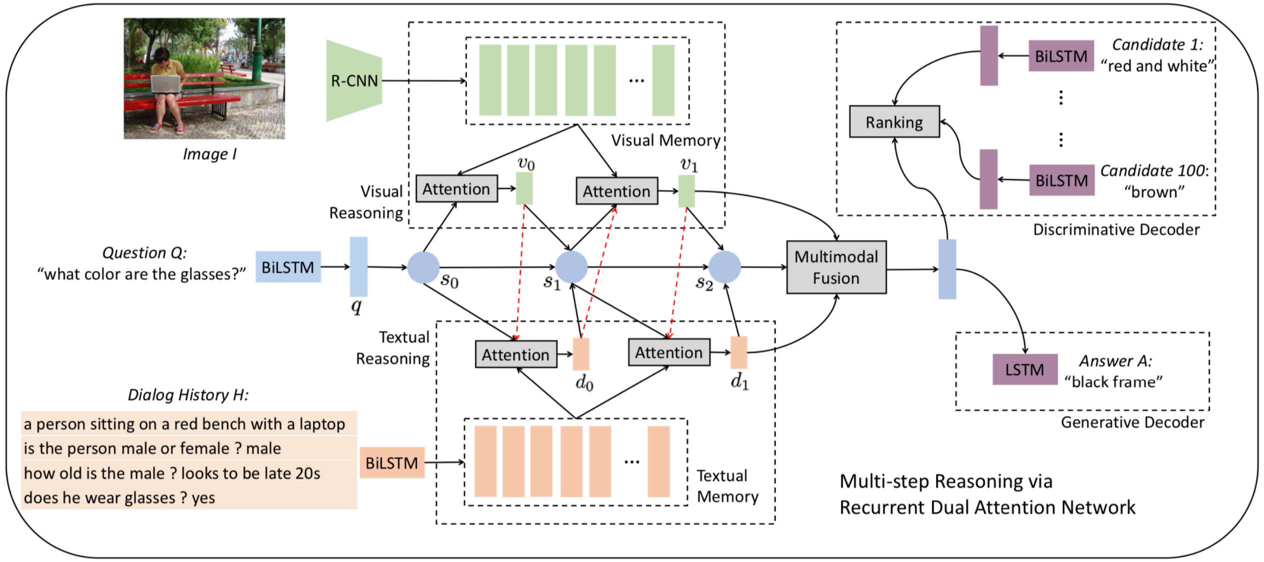
- Motivation & Tasks
  - A query and a set of accompanying document are given, the answer can only be obtained by selecting two or more evidence from the documents.
  - Since not every document contain relevant information, multi-hop text-based QA requires filtering out noises from multiple paragraphs and extracting useful information.
  - Previous work on multi-hop QA usually aggregates document information to an entity graph, and answers are then directly selected on entities of the entity graph.
- Solution





Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog

- Motivation
  - After taking a first glimpse of the image and the dialog history, readers often re- visit specific sub-areas of both image and text to obtain a better understanding of the multimodal context.
- Solution



Model	NDCG	MRR	R@1	R@5	R@10	Mean
MN-D (Das et al., 2017a)	55.13	60.42	46.09	78.14	88.05	4.63
HCIAE-D (Lu et al., 2017)	57.65	62.96	48.94	80.50	89.66	4.24
CoAtt-D (Wu et al., 2018)	57.72	62.91	48.86	80.41	89.83	4.21
ReDAN-D (T=1)	58.49	63.35	49.47	80.72	90.05	4.19
ReDAN-D (T=2)	59.26	63.46	49.61	80.75	89.96	4.15
ReDAN-D (T=3)	59.32	64.21	50.60	81.39	90.26	4.05
Ensemble of 4	**60.53**	**65.30**	**51.67**	**82.40**	**91.09**	**3.82**

Table 1: Comparison of ReDAN with a discriminative decoder to state-of-the-art methods on VisDial v1.0 validation set. Higher score is better for NDCG, MRR and Recall@k, while lower score is better for mean rank. All these baselines are re-implemented with bottom-up features and incorporated with GloVe vectors for fair comparison.

Model	NDCG	MRR	R@1	R@5	R@10	Mean
MN-G (Das et al., 2017a)	56.99	47.83	38.01	57.49	64.08	18.76
HCIAE-G (Lu et al., 2017)	59.70	49.07	39.72	58.23	64.73	18.43
CoAtt-G (Wu et al., 2018)	59.24	49.64	40.09	59.37	65.92	17.86
ReDAN-G (T=1)	59.41	49.60	39.95	59.32	65.97	17.79
ReDAN-G (T=2)	60.11	49.96	40.36	59.72	66.57	17.53
ReDAN-G (T=3)	60.47	50.02	40.27	59.93	66.78	17.40
Ensemble of 4	**61.43**	**50.41**	**40.85**	**60.08**	**67.17**	**17.38**

Table 2: Comparison of ReDAN with a generative decoder to state-of-the-art generative methods on VisDial val v1.0. All the baseline models are re-implemented with bottom-up features and incorporated with GloVe vectors for fair comparison.

Thanks & QA