國立臺灣大學
National Taiwan University

JOHNS HOPKINS
UNIVERSITY

THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Scene Graph Parsing as Dependency Parsing

Author: **Yu-Siang Wang**, Chenxi Liu, Xiaohui Zeng, Alan Yuille

Conference: North American Chapter of the Association for Computational Linguistics, 2018

Outline

- Introduction
- Method
- Experiments
- Conclusion

Introduction

- **Introduction**
- Method
- Experiments
- Conclusion

Introduction

- Many multimodal tasks fit into this picture

A young boy wearing black shirt is in front of a goal

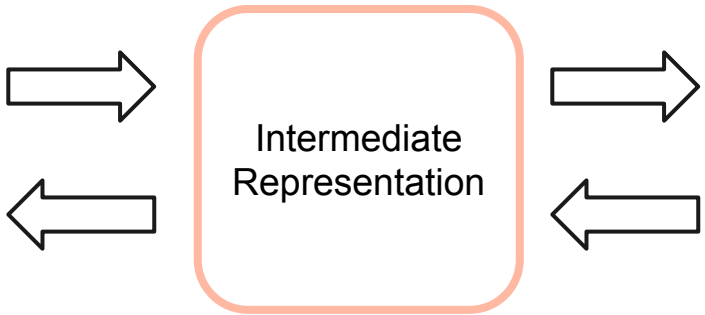


Image Generation from Text

A young boy wearing black shirt is in front of a goal

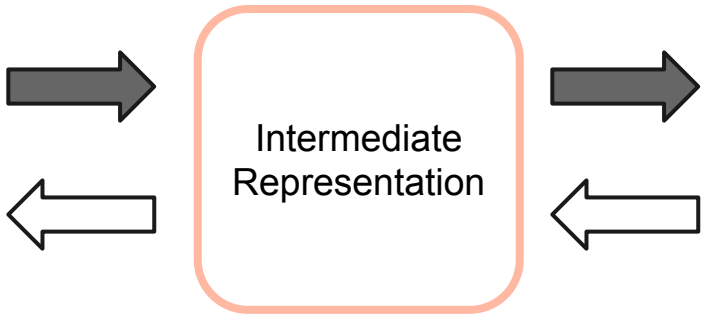


Image Captioning

A young boy wearing black shirt is in front of a goal

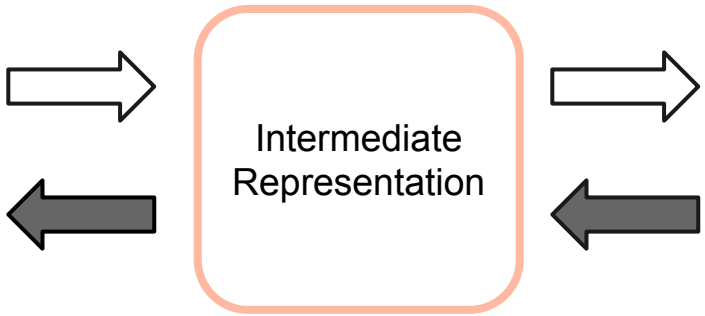
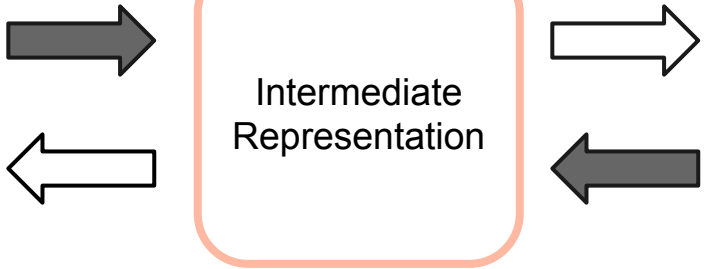
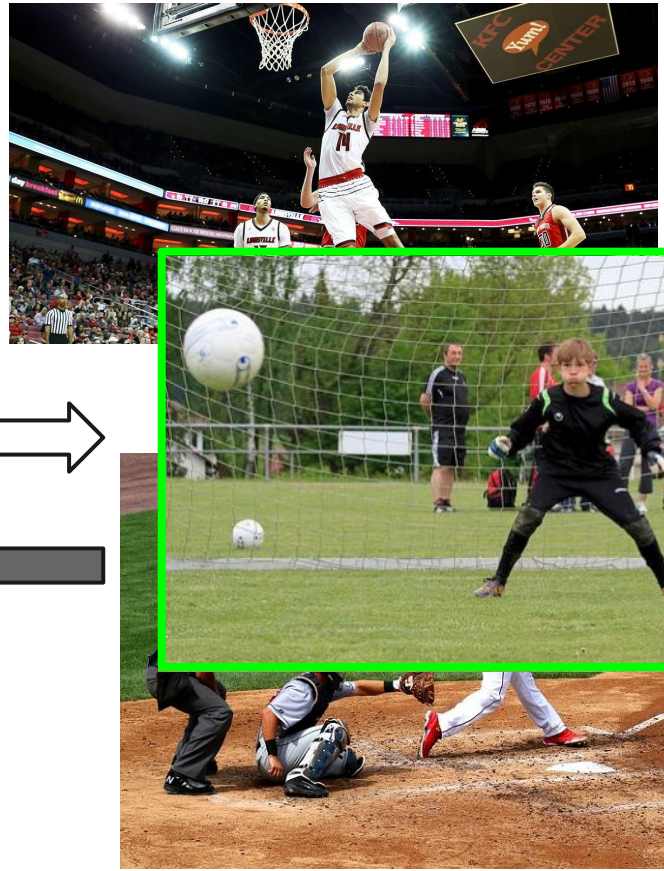


Image Retrieval

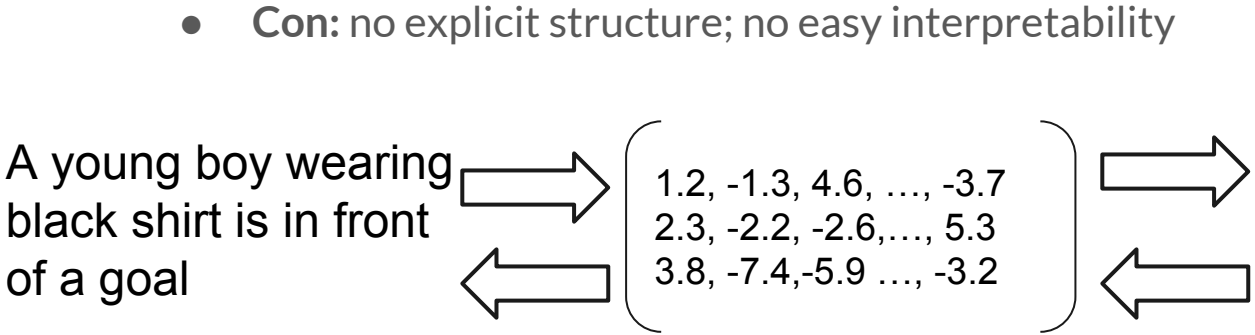
A young boy wearing black shirt is in front of a goal





Neural Network Embedding

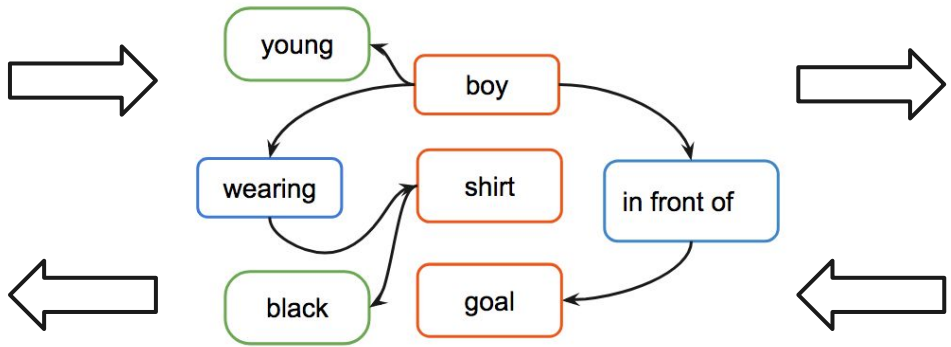
- Neural network embeddings often used as the intermediate representation
- **Pro:** easy training; similarity with cosine distance
- **Con:** no explicit structure; no easy interpretability



Scene Graph

- More recently, people start exploring a more explainable representation
- Has 3 types of nodes: object, attribute, relation

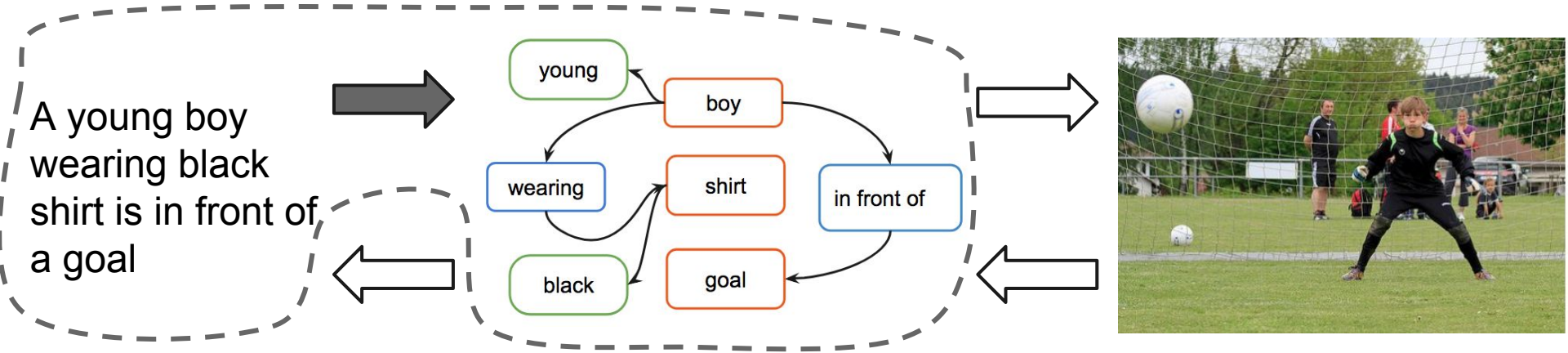
A young boy wearing black shirt is in front of a goal



Ref: Johnson et al., Image Retrieval Using Scene Graph, CVPR 2015

Our Goal

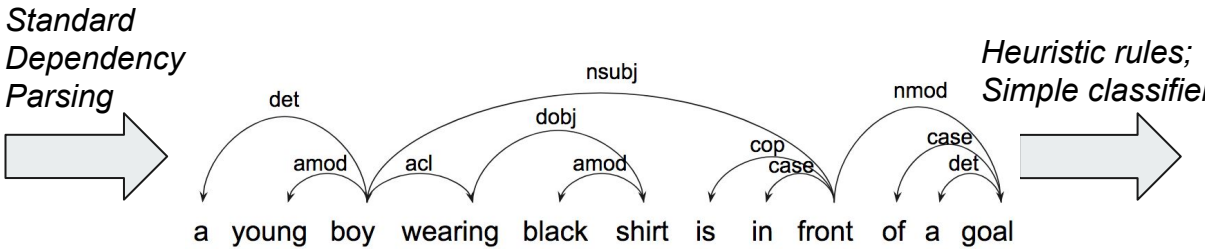
- Parsing from sentence to scene graph (i.e., scene graph parsing)



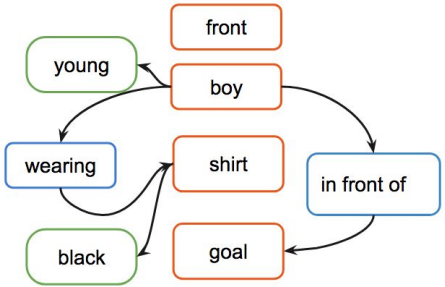
Previous Work: Separated Two-stage

a young boy wearing black shirt is in front of a man

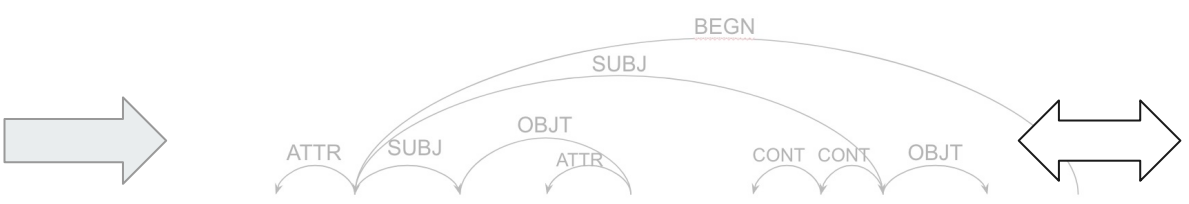
Standard Dependency Parsing

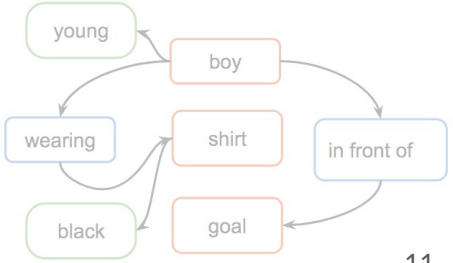


Heuristic rules; Simple classifier



Ref: Anderson et al., SPICE: Semantic Propositional Image Caption Evaluation, ECCV 2016



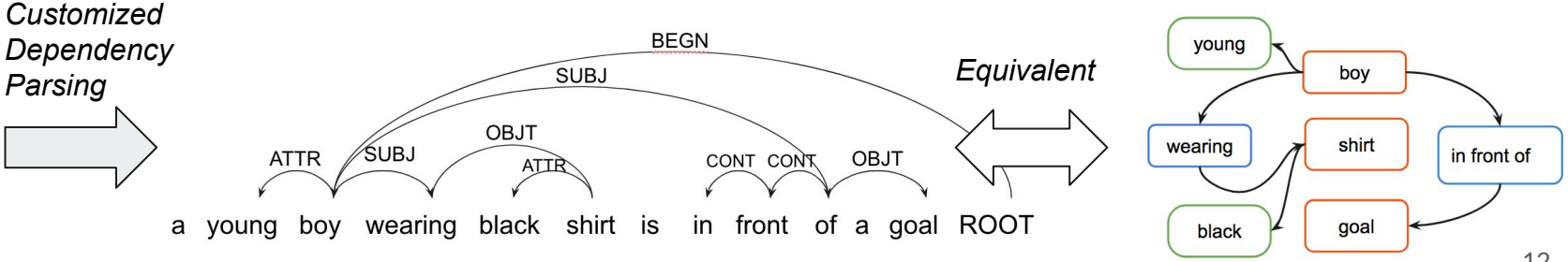


Our Work: End-to-end One-stage

a young boy wearing black shirt is in front of a man

Ref: Anderson et al., SPICE: Semantic Propositional Image Caption Evaluation, ECCV 2016

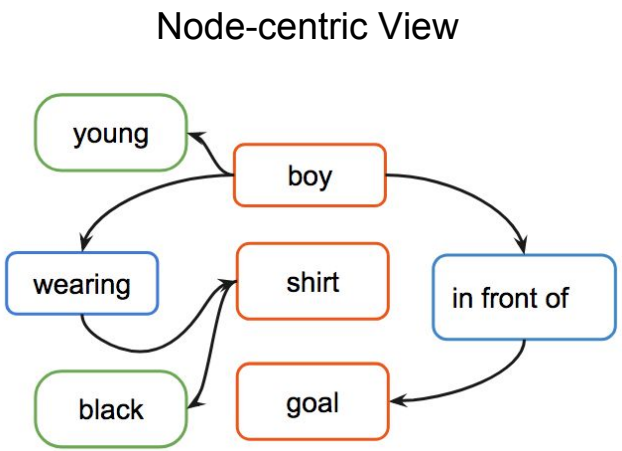
Customized Dependency Parsing



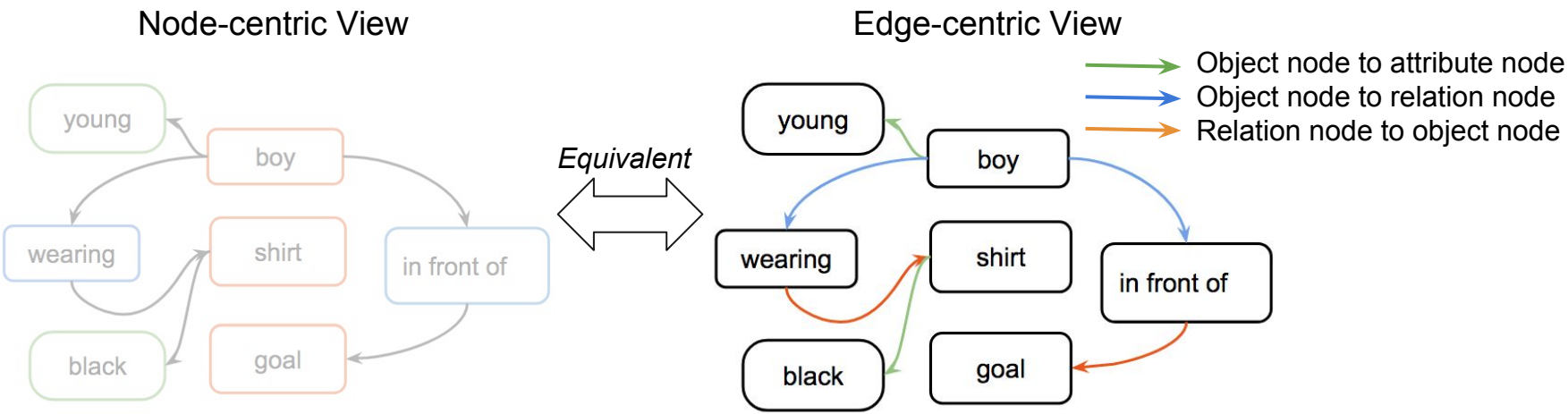
Method

- Introduction
- **Method**
- Experiments
- Conclusion

Scene Graph


Node-centric View
young
boy
wearing
shirt
in front of
black
goal

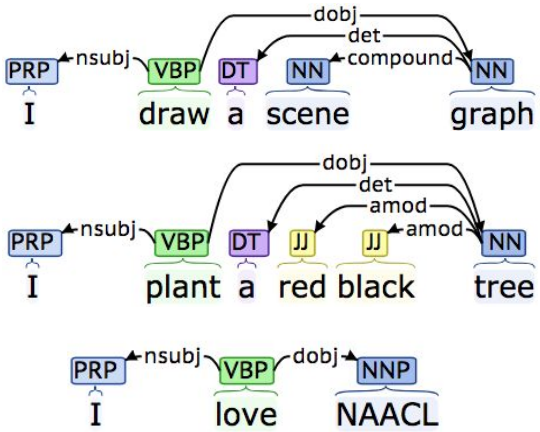
Pushing Labels from Node to Arc



- Different colors are different arc labels
- Under the edge-centric view, scene graphs begin to look like dependency parses

Review of Dependency Parsing

1. Get a Corpus!



2. Define a Label Space!

NSUBJ
NMOD
CASE
DET
...

3. Pick a System (e.g. Arc-Hybrid) and its Actions!

LEFT
RIGHT
SHIFT
...

How we do Scene Graph Parsing?

1. Get a Corpus!

?

2. Define a Label Space!

?

3. Pick a System (e.g. Arc-Hybrid) and its Actions!

?

How we do Scene Graph Parsing?

1. Get a Corpus!

?

2. Define a Label Space!

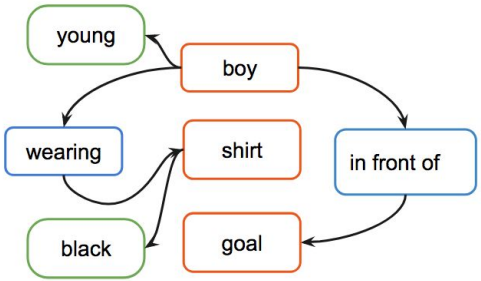
?

3. Pick a System (e.g. Arc-Hybrid) and its Actions!

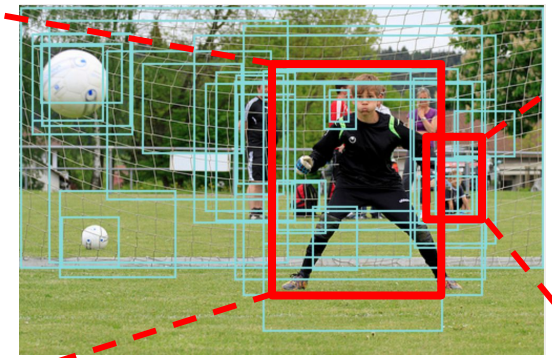
?

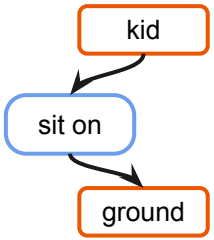
Visual Genome

- In Visual Genome, every image is annotated with 30 regions on average
- Every region is annotated with a (region) description and a (region) scene graph



A young boy wearing black shirt is in front of a goal





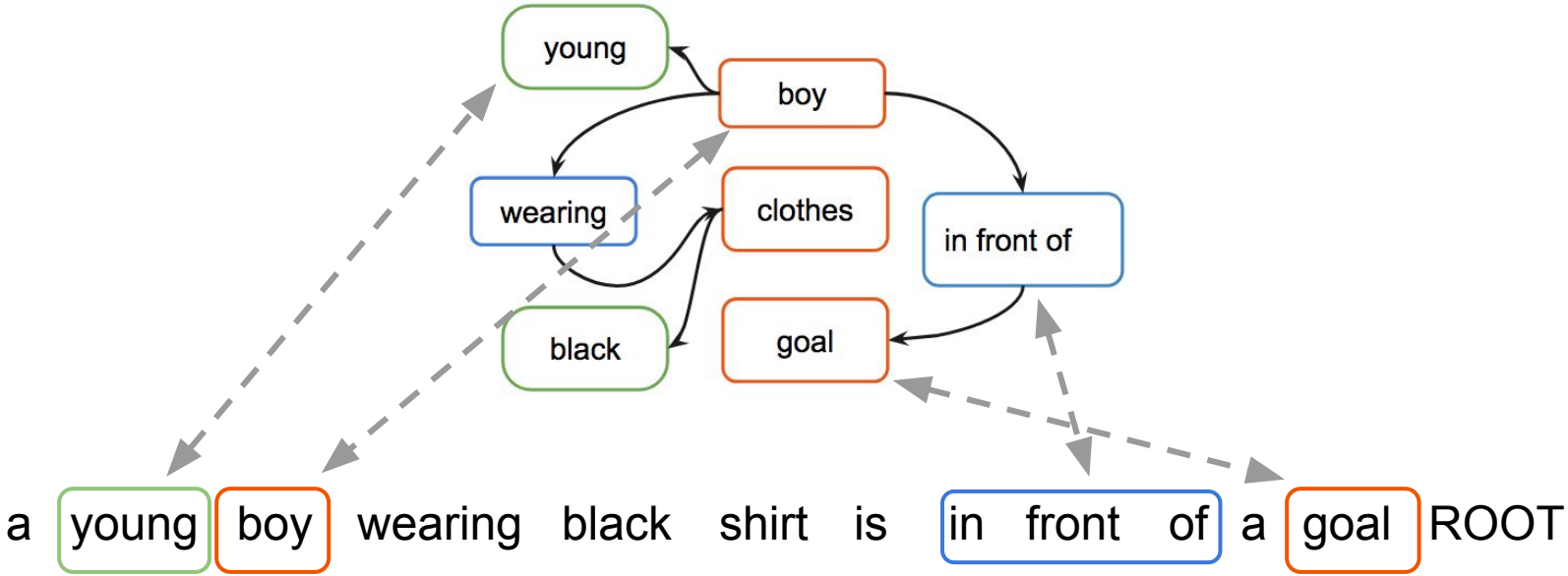
A kid is sitting on the ground

Ref: Krishna et al., Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, IJCV 2017

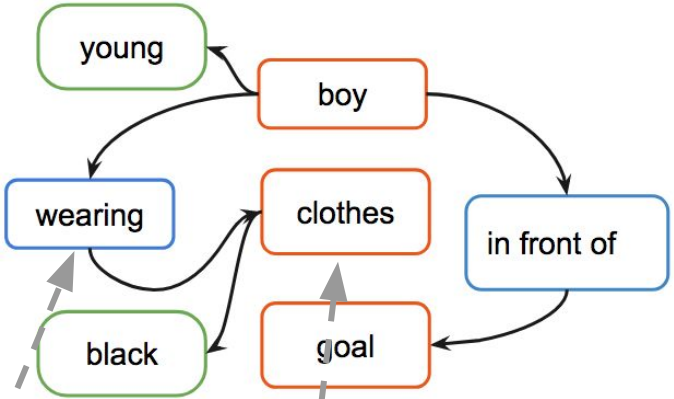
Alignment Strategy

- To mimic a dependency parsing training corpus, we need alignment between **nodes in the scene graph** and **words in the sentence**
- We propose a two-round alignment strategy:
  - Within each round, object, attribute, relation nodes are aligned in this order
  - First round is more “conservative” (word-by-word match)
  - Second round is more “aggressive” (synonyms match)

Alignments made in Round 1

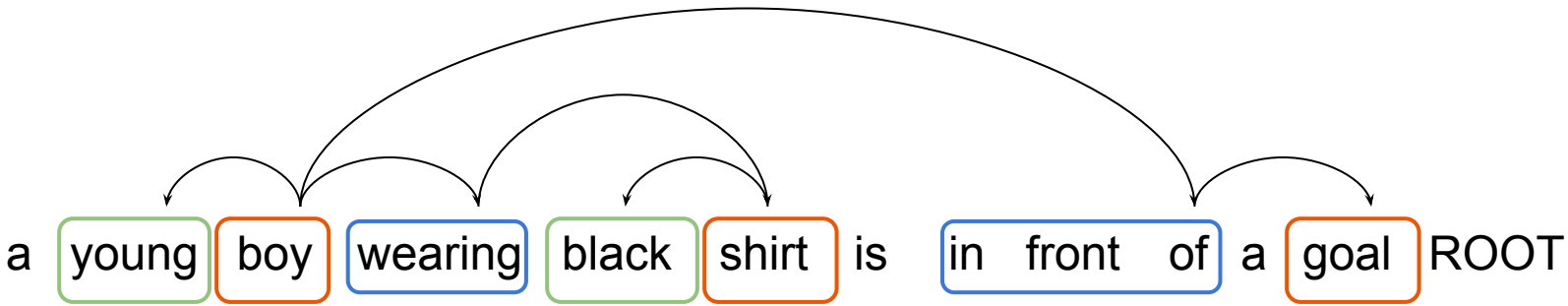


Alignments made in Round 2



a young boy wearing black shirt is in front of a goal ROOT

Alignment Result



How we do Scene Graph Parsing?

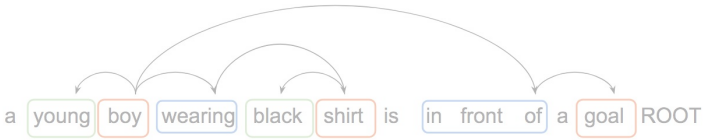
1. Get a Corpus!

2. Define a Label Space!

?

3. Pick a System (e.g. Arc-Hybrid) and its Actions!

?

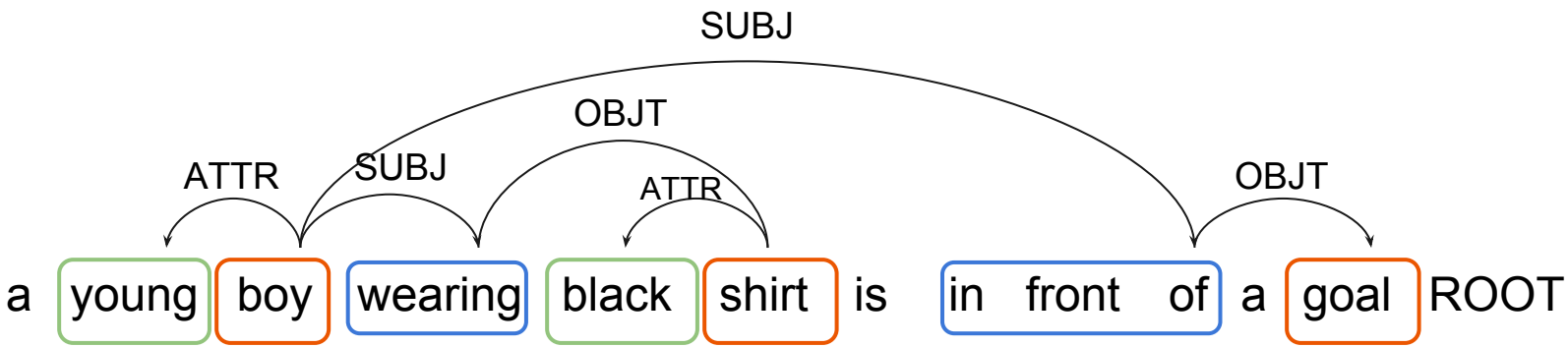


Regular Labels

1. ATTR
Object to Attribute

2. SUBJ
Object to Relation

3. OBJT
Relation to Object



Auxiliary Labels

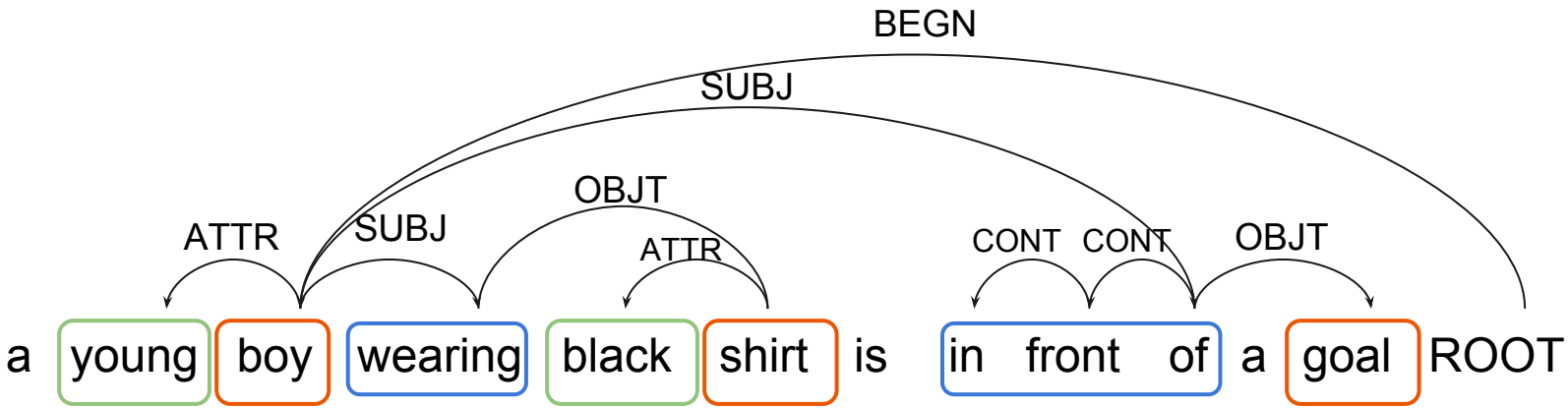
1. ATTR
Object to Attribute

2. SUBJ
Object to Relation

3. OBJT
Relation to Object

4. CONT
Phrase

5. BEGN
ROOT to Obj without Head

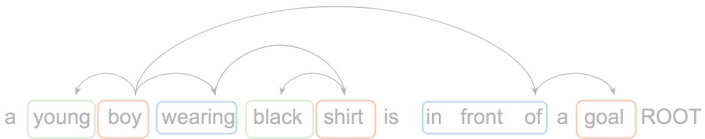


How we do Scene Graph Parsing?

1. Get a Corpus!

2. Define a Label Space!

3. Pick a System (e.g. Arc-Hybrid) and its Actions!



BEGN
SUBJ
OBJT
CONT
ATTR

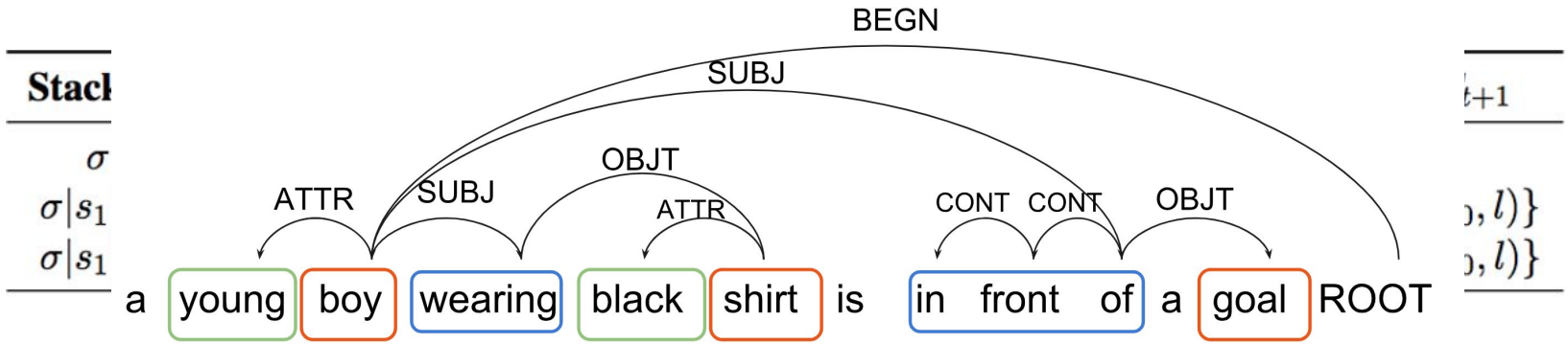
?

Transition-Based Arc-Hybrid System

Stack σ_t	Buffer β_t	Arc set T_t	Action	Stack σ_{t+1}	Buffer β_{t+1}	Arc set T_{t+1}
σ	$b_0\|\beta$	T	SHIFT	$\sigma\|b_0$	β	T
$\sigma\|s_1\|s_0$	$b_0\|\beta$	T	LEFT(l)	$\sigma\|s_1$	$b_0\|\beta$	$T \cup \{(b_0, s_0, l)\}$
$\sigma\|s_1\|s_0$	β	T	RIGHT(l)	$\sigma\|s_1$	β	$T \cup \{(s_1, s_0, l)\}$

Ref: Kuhlmann et al., Dynamic programming algorithms for transition-based dependency parsers, ACL 2011

Transition-Based Arc-Hybrid System



Ref: Kuhlmann et al., Dynamic programming algorithms for transition-based dependency parsers, ACL 2011

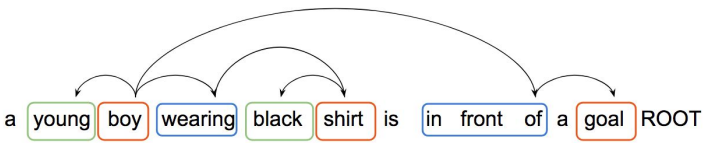
Augmented Arc-Hybrid

- We augment Arc-Hybrid with one more action that is REDUCE
- This is because we don't require every word to have a head (e.g. "is")

Stack σ_t	**Buffer** β_t	**Arc set** T_t	**Action**	**Stack** σ_{t+1}	**Buffer** β_{t+1}	**Arc set** T_{t+1}
σ	$b_0\|\beta$	T	SHIFT	$\sigma\|b_0$	β	T
$\sigma\|s_1\|s_0$	$b_0\|\beta$	T	LEFT(l)	$\sigma\|s_1$	$b_0\|\beta$	$T \cup \{(b_0, s_0, l)\}$
$\sigma\|s_1\|s_0$	β	T	RIGHT(l)	$\sigma\|s_1$	β	$T \cup \{(s_1, s_0, l)\}$
$\sigma\|s_0$	β	T	REDUCE	σ	β	T

How we do Scene Graph Parsing?

1. Get a Corpus!



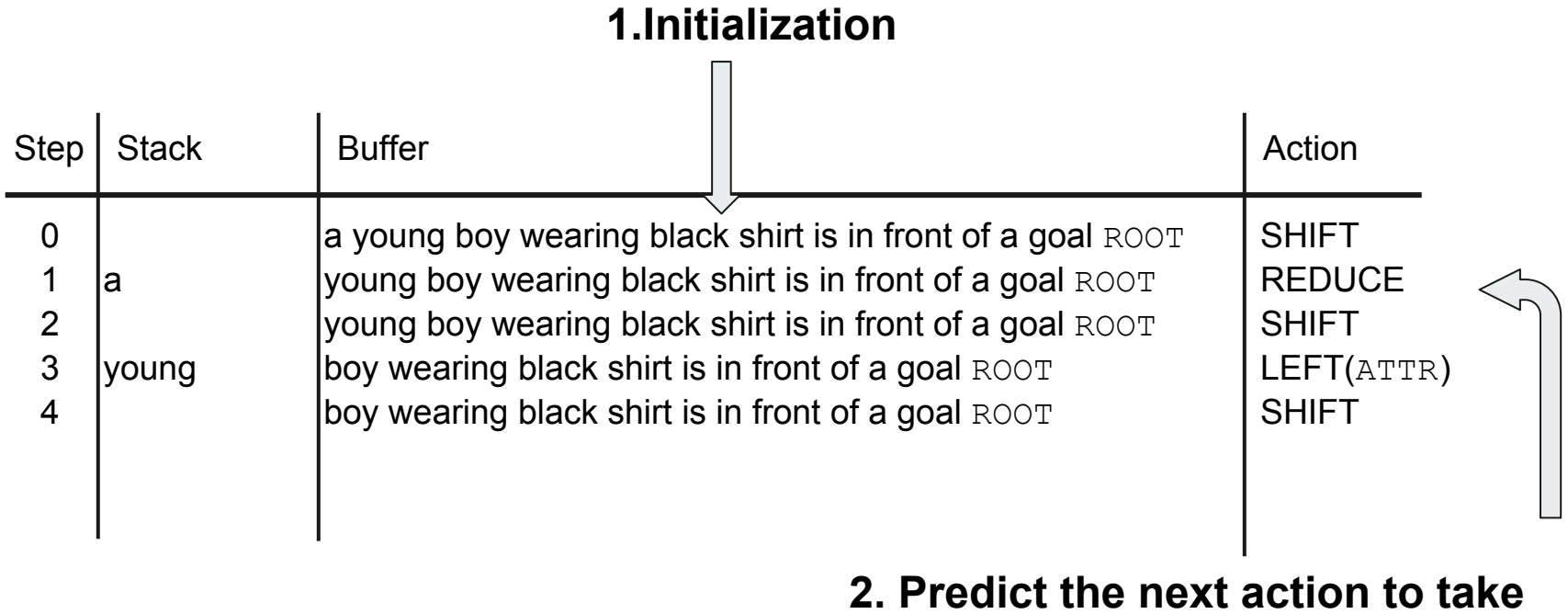
2. Define a Label Space!

BEGN
SUBJ
OBJT
CONT
ATTR

3. Define Actions in a System (e.g. Arc-Hybrid)!

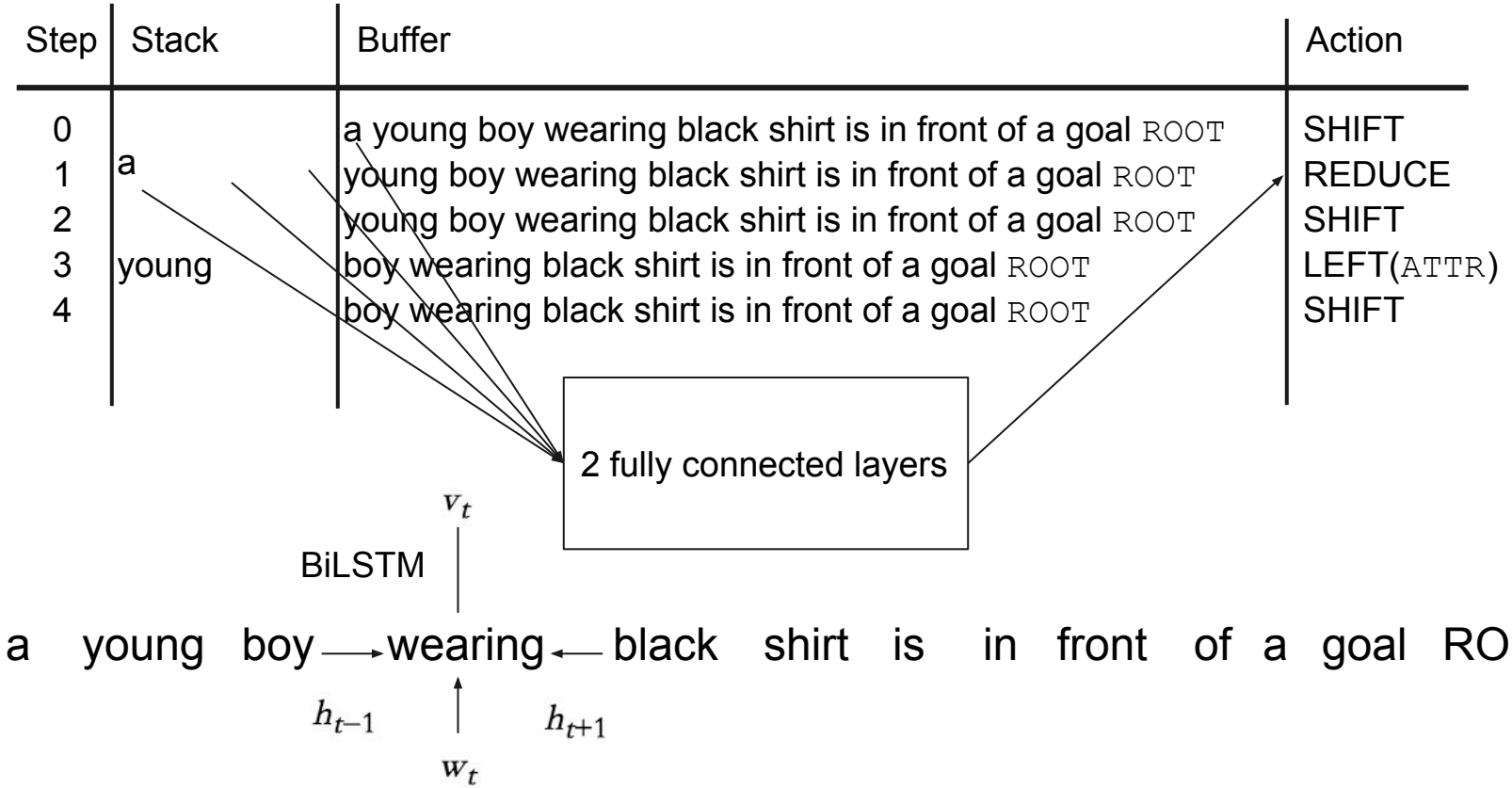
LEFT
RIGHT
SHIFT
REDUCE

Detailed Architecture



Ref: Kiperwasser and Goldberg, Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations, TACL 2016

Detailed Architecture



Ref: Kiperwasser and Goldberg, Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations, TACL 2016

Step	Stack	Buffer	Action
0		a young boy wearing black shirt is in front of a goal ROOT	SHIFT

a young boy wearing black shirt is in front of a goal ROOT

Step	Stack	Buffer	Action
0		a young boy wearing black shirt is in front of a goal ROOT	SHIFT
1	a	young boy wearing black shirt is in front of a goal ROOT	REDUCE

a young boy wearing black shirt is in front of a goal ROOT

Step	Stack	Buffer	Action
0		a young boy wearing black shirt is in front of a goal ROOT	SHIFT
1	a	young boy wearing black shirt is in front of a goal ROOT	REDUCE
2		young boy wearing black shirt is in front of a goal ROOT	SHIFT

a young boy wearing black shirt is in front of a goal ROOT

Step	Stack	Buffer	Action
0		a young boy wearing black shirt is in front of a goal ROOT	SHIFT
1	a	young boy wearing black shirt is in front of a goal ROOT	REDUCE
2		young boy wearing black shirt is in front of a goal ROOT	SHIFT
3	young	boy wearing black shirt is in front of a goal ROOT	LEFT(ATTR)

a young boy wearing black shirt is in front of a goal ROOT

Step	Stack	Buffer	Action
0		a young boy wearing black shirt is in front of a goal ROOT	SHIFT
1	a	young boy wearing black shirt is in front of a goal ROOT	REDUCE
2		young boy wearing black shirt is in front of a goal ROOT	SHIFT
3	young	boy wearing black shirt is in front of a goal ROOT	LEFT(ATTR)
4		boy wearing black shirt is in front of a goal ROOT	SHIFT

a young boy wearing black shirt is in front of a goal ROOT

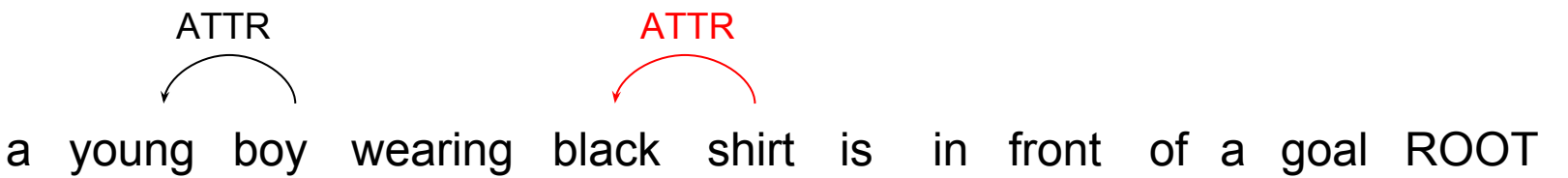
Step	Stack	Buffer	Action
1	a	young boy wearing black shirt is in front of a goal ROOT	REDUCE
2		young boy wearing black shirt is in front of a goal ROOT	SHIFT
3	young	boy wearing black shirt is in front of a goal ROOT	LEFT(ATTR)
4		boy wearing black shirt is in front of a goal ROOT	SHIFT
5	boy	wearing black shirt is in front of a goal ROOT	SHIFT

a young boy wearing black shirt is in front of a goal ROOT

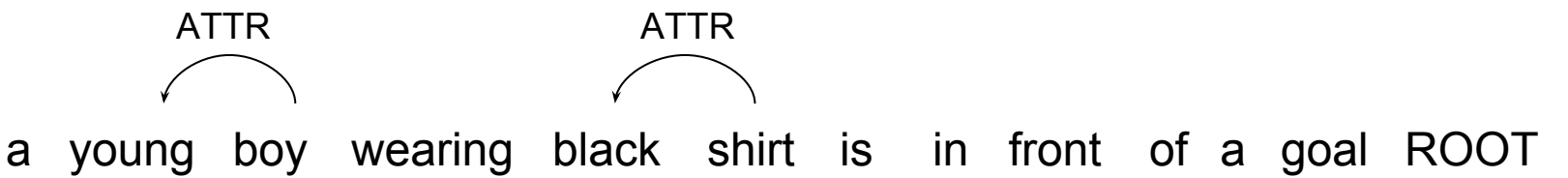
Step	Stack	Buffer	Action
2		young boy wearing black shirt is in front of a goal ROOT	SHIFT
3	young	boy wearing black shirt is in front of a goal ROOT	LEFT(ATTR)
4		boy wearing black shirt is in front of a goal ROOT	SHIFT
5	boy	wearing black shirt is in front of a goal ROOT	SHIFT
6	boy wearing	black shirt is in front of a goal ROOT	SHIFT

a young boy wearing black shirt is in front of a goal ROOT

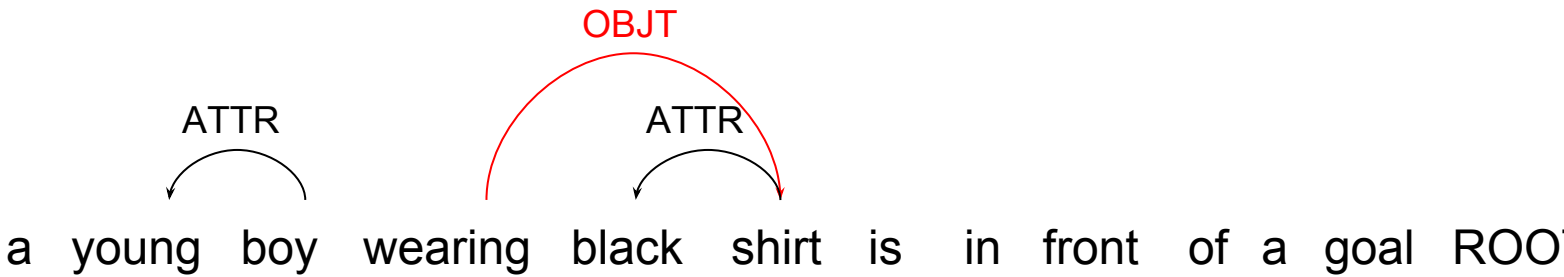
Step	Stack	Buffer	Action
3	young	boy wearing black shirt is in front of a goal ROOT	LEFT(ATTR)
4		boy wearing black shirt is in front of a goal ROOT	SHIFT
5	boy	wearing black shirt is in front of a goal ROOT	SHIFT
6	boy wearing	black shirt is in front of a goal ROOT	SHIFT
7	boy wearing black	shirt is in front of a goal ROOT	LEFT(ATTR)



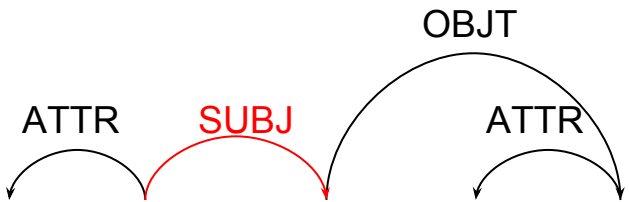
Step	Stack	Buffer	Action
4		boy wearing black shirt is in front of a goal ROOT	SHIFT
5	boy	wearing black shirt is in front of a goal ROOT	SHIFT
6	boy wearing	black shirt is in front of a goal ROOT	SHIFT
7	boy wearing black	shirt is in front of a goal ROOT	LEFT(ATTR)
8	boy wearing	shirt is in front of a goal ROOT	SHIFT



Step	Stack	Buffer	Action
5	boy	wearing black shirt is in front of a goal ROOT	SHIFT
6	boy wearing	black shirt is in front of a goal ROOT	SHIFT
7	boy wearing black	shirt is in front of a goal ROOT	LEFT(ATTR)
8	boy wearing	shirt is in front of a goal ROOT	SHIFT
9	boy wearing shirt	is in front of a goal ROOT	RIGHT(OBJT)

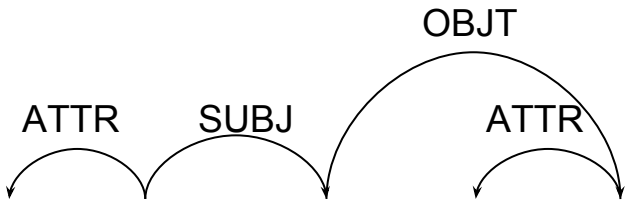


Step	Stack	Buffer	Action
6	boy wearing	black shirt is in front of a goal ROOT	SHIFT
7	boy wearing black	shirt is in front of a goal ROOT	LEFT(ATTR)
8	boy wearing	shirt is in front of a goal ROOT	SHIFT
9	boy wearing shirt	is in front of a goal ROOT	RIGHT(OBJT)
10	boy wearing	is in front of a goal ROOT	RIGHT(SUBJ)



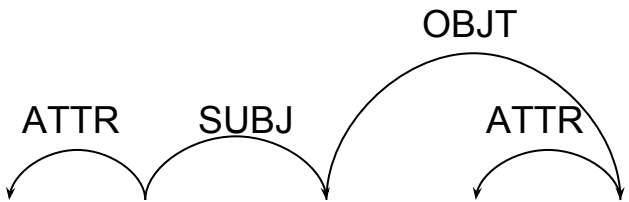
a young boy wearing black shirt is in front of a goal ROOT

Step	Stack	Buffer	Action
7	boy wearing black	shirt is in front of a goal ROOT	LEFT(ATTR)
8	boy wearing	shirt is in front of a goal ROOT	SHIFT
9	boy wearing shirt	is in front of a goal ROOT	RIGHT(OBJT)
10	boy wearing	is in front of a goal ROOT	RIGHT(SUBJ)
11	boy	is in front of a goal ROOT	SHIFT



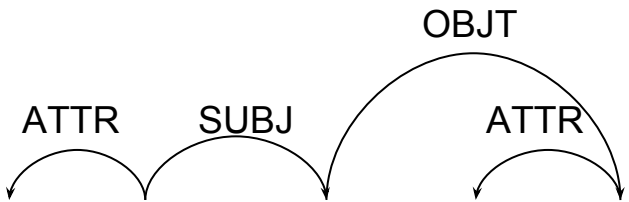
a young boy wearing black shirt is in front of a goal ROOT

Step	Stack	Buffer	Action
8	boy wearing	shirt is in front of a goal ROOT	SHIFT
9	boy wearing shirt	is in front of a goal ROOT	RIGHT(OBJT)
10	boy wearing	is in front of a goal ROOT	RIGHT(SUBJ)
11	boy	in front of a goal ROOT	SHIFT
12	boy is	in front of a goal ROOT	REDUCE



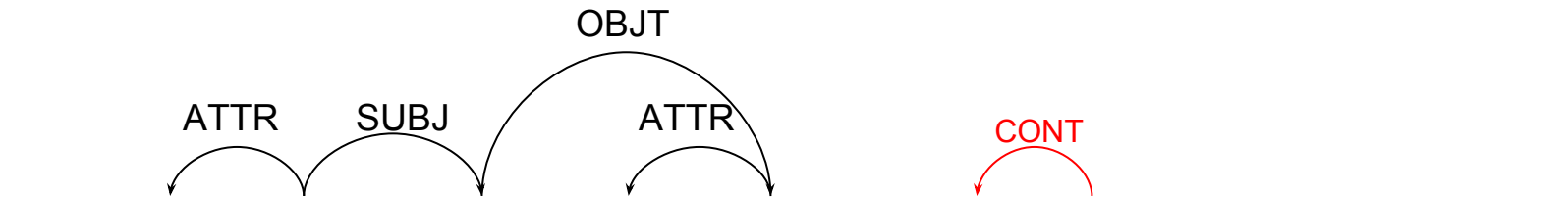
a young boy wearing black shirt is in front of a goal ROOT

Step	Stack	Buffer	Action
9	boy wearing shirt	is in front of a goal ROOT	RIGHT(OBJT)
10	boy wearing	is in front of a goal ROOT	RIGHT(SUBJ)
11	boy	in front of a goal ROOT	SHIFT
12	boy is	in front of a goal ROOT	REDUCE
13	boy	in front of a goal ROOT	SHIFT

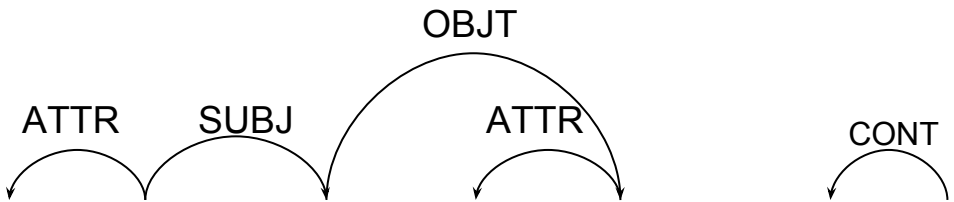


a young boy wearing black shirt is in front of a goal ROOT

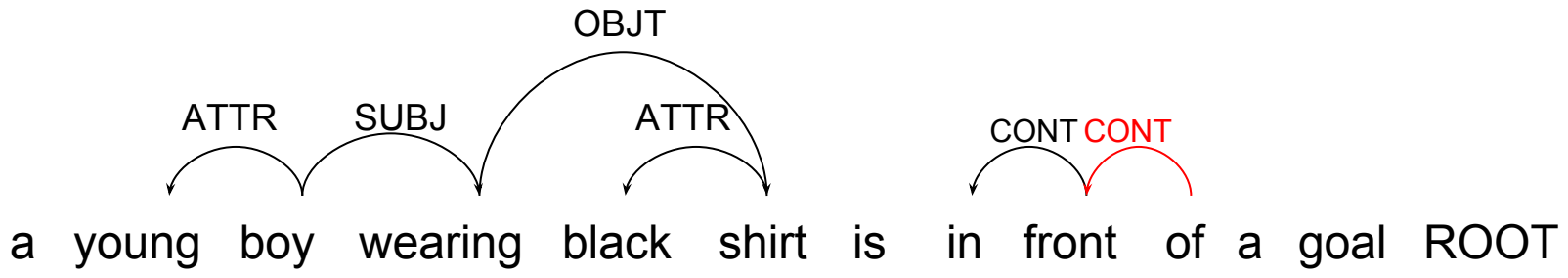
Step	Stack	Buffer	Action
10	boy wearing	is in front of a goal ROOT	RIGHT(SUBJ)
11	boy	in front of a goal ROOT	SHIFT
12	boy is	in front of a goal ROOT	REDUCE
13	boy	in front of a goal ROOT	SHIFT
14	boy in	front of a goal ROOT	LEFT(CONT)



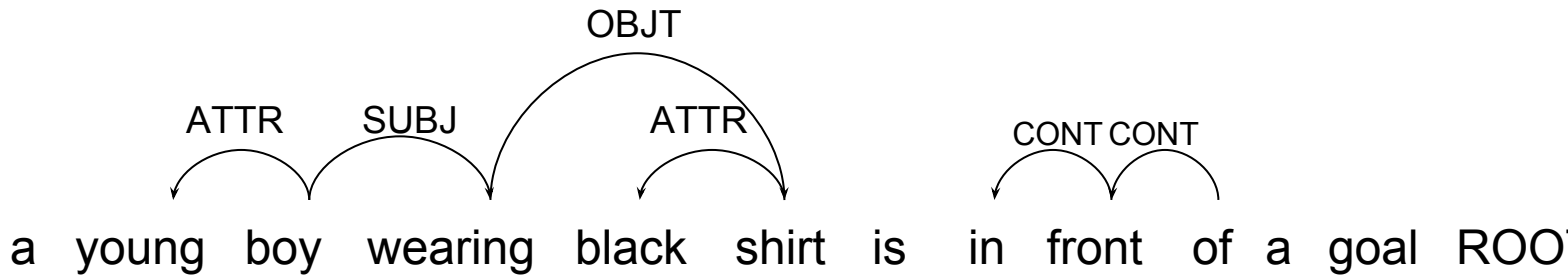
Step	Stack	Buffer	Action
11	boy	in front of a goal ROOT	SHIFT
12	boy is	in front of a goal ROOT	REDUCE
13	boy	in front of a goal ROOT	SHIFT
14	boy in	front of a goal ROOT	LEFT(CONT)
15	boy	front of a goal ROOT	SHIFT



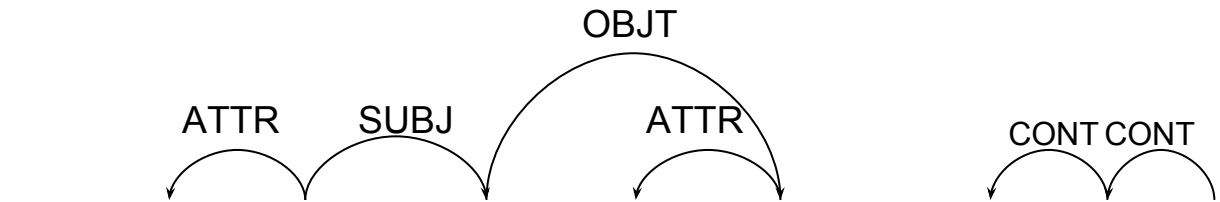
Step	Stack	Buffer	Action
12	boy is	in front of a goal ROOT	REDUCE
13	boy	in front of a goal ROOT	SHIFT
14	boy in	front of a goal ROOT	LEFT(CONT)
15	boy	front of a goal ROOT	SHIFT
16	boy front	of a goal ROOT	LEFT(CONT)



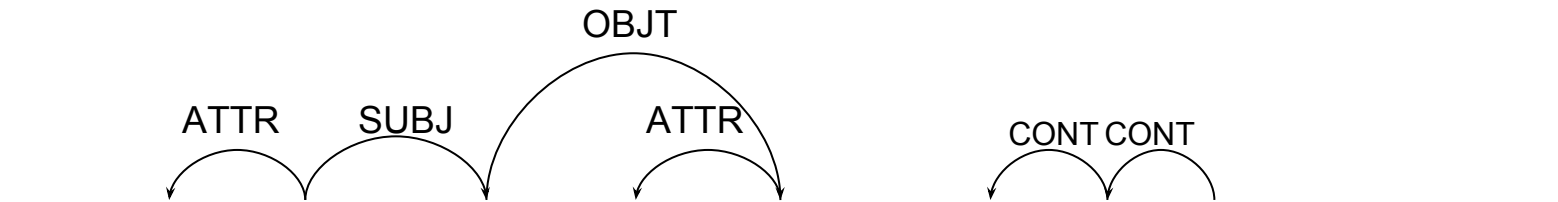
Step	Stack	Buffer	Action
13	boy	in front of a goal ROOT	SHIFT
14	boy in	front of a goal ROOT	LEFT(CONT)
15	boy	front of a goal ROOT	SHIFT
16	boy front	of a goal ROOT	LEFT(CONT)
17	boy	of a goal ROOT	SHIFT



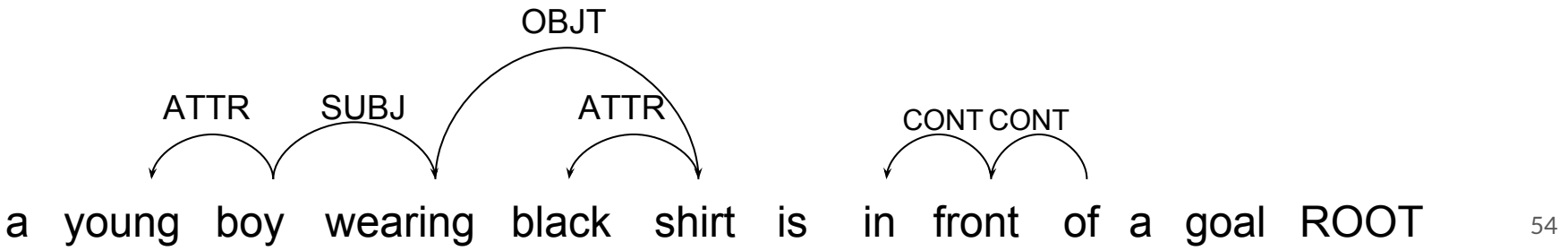
Step	Stack	Buffer	Action
14	boy in	front of a goal ROOT	LEFT(CONT)
15	boy	front of a goal ROOT	SHIFT
16	boy front	of a goal ROOT	LEFT(CONT)
17	boy	of a goal ROOT	SHIFT
18	boy of	a goal ROOT	SHIFT



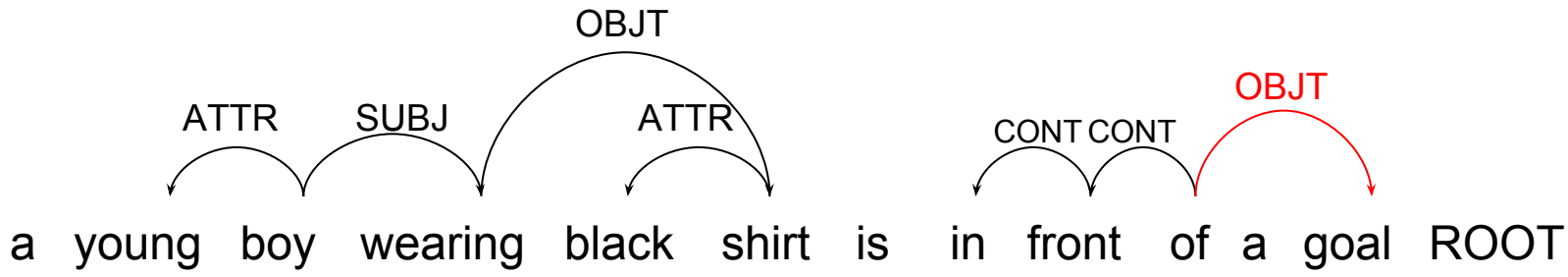
Step	Stack	Buffer	Action
15	boy	front of a goal ROOT	SHIFT
16	boy front	of a goal ROOT	LEFT(CONT)
17	boy	of a goal ROOT	SHIFT
18	boy of	a goal ROOT	SHIFT
19	boy of a	goal ROOT	REDUCE



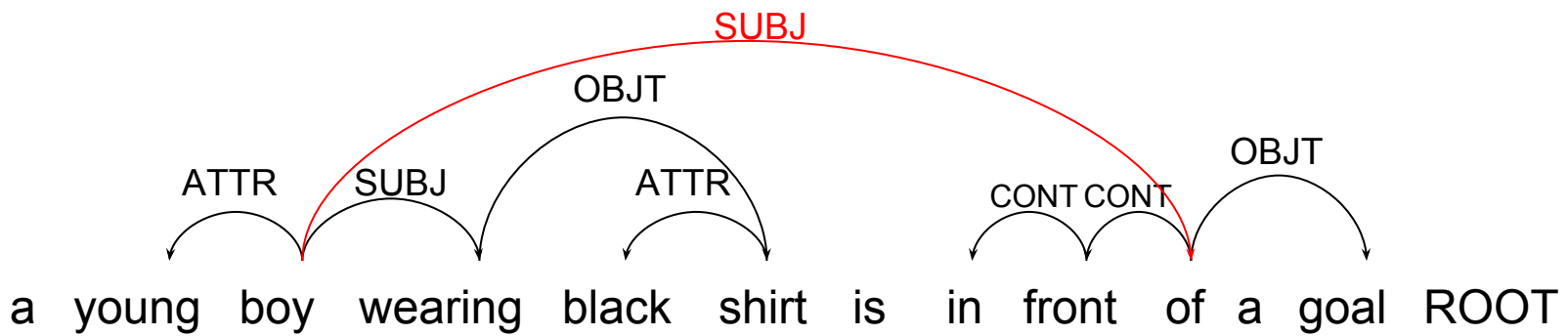
Step	Stack	Buffer	Action
16	boy front	of a goal ROOT	LEFT(CONT)
17	boy	of a goal ROOT	SHIFT
18	boy of	a goal ROOT	SHIFT
19	boy of a	goal ROOT	REDUCE
20	boy of	goal ROOT	SHIFT



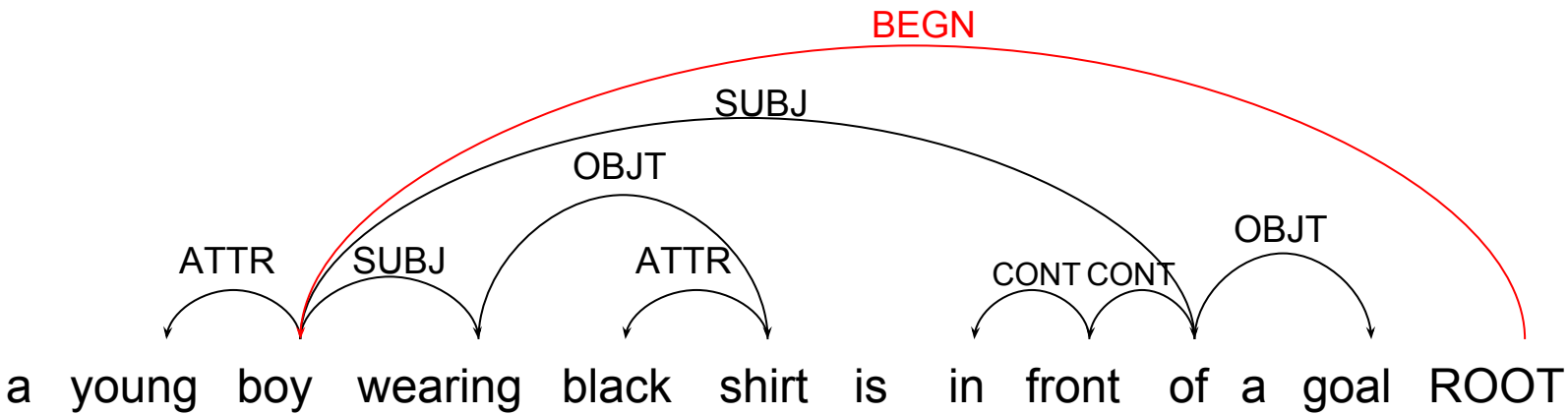
Step	Stack	Buffer	Action
17	boy	of a goal ROOT	SHIFT
18	boy of	a goal ROOT	SHIFT
19	boy of a	goal ROOT	REDUCE
20	boy of	goal ROOT	SHIFT
21	boy of goal	ROOT	RIGHT(OBJT)



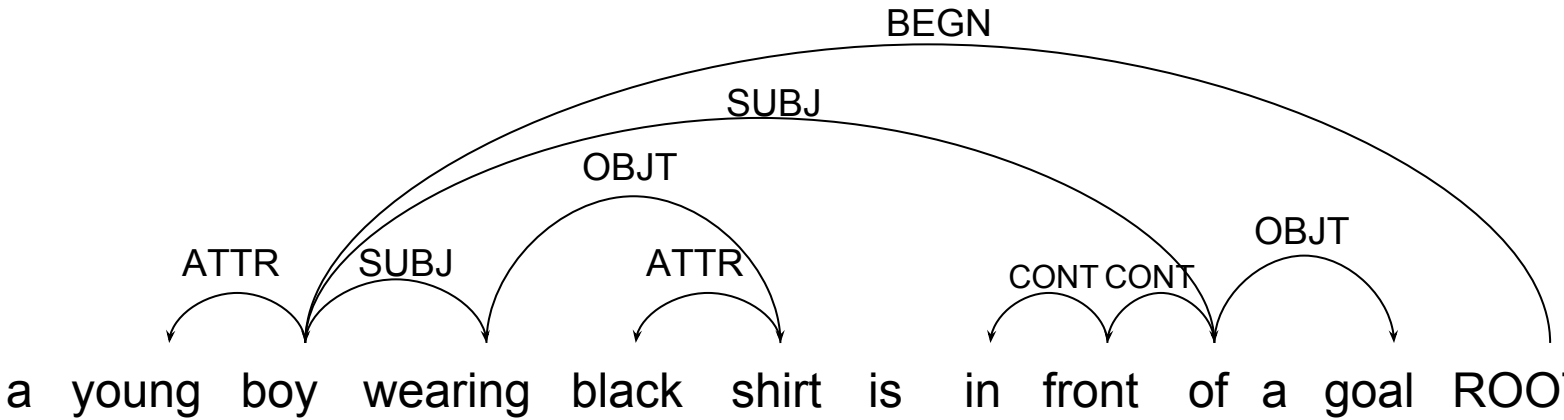
Step	Stack	Buffer	Action
18	boy of	a goal ROOT	SHIFT
19	boy of a	goal ROOT	REDUCE
20	boy of	goal ROOT	SHIFT
21	boy of goal	ROOT	RIGHT(OBJT)
22	boy of	ROOT	RIGHT(SUBJ)



Step	Stack	Buffer	Action
19	boy of a	goal ROOT	REDUCE
20	boy of	goal ROOT	SHIFT
21	boy of goal	ROOT	RIGHT(OBJT)
22	boy of	ROOT	RIGHT(SUBJ)
23	boy	ROOT	LEFT(BEGN)



Step	Stack	Buffer	Action
20	boy of	goal ROOT	SHIFT
21	boy of goal	ROOT	RIGHT(OBJT)
22	boy of	ROOT	RIGHT(SUBJ)
23	boy	ROOT	LEFT(BEGN)
24		ROOT	



Experiments

- Introduction
- Method
- **Experiments**
- Conclusion

Experiment 1: Scene Graph Parsing

- Dataset statistics (intersection of MS COCO and Visual Genome):

	Training	Validation
# of Images	34027	17471
# of Sentences/Scene Graphs	**1070145**	**547795**

- Evaluated by **F-score** between parsed scene graph and ground truth scene graph

Scene Graph Parsing Results

Parser	*F-Score*
Stanford (Schuster et al,. 2015) [Separated Two-stage]	0.3549
SPICE (Anderson et al., 2016) [Separated Two-stage]	0.4469
Ours [End-to-end One-stage]	**0.4967**

Scene Graph Parsing Oracle

Parser	*F-Score*
Stanford (Schuster et al,. 2015) [Separated Two-stage]	0.3549
SPICE (Anderson et al., 2016) [Separated Two-stage]	0.4469
Ours [End-to-end One-stage]	**0.4967**
Oracle	0.6985

Scene Graph Parsing Ablation Studies

Parser	*F-Score*
Ours (CONT pointing left)	**0.4967**
Ours (CONT pointing right)	0.4952
Ours (1 round aggressive alignment)	0.4877
Ours (1 round conservative alignment)	0.4538

Experiment 2: Image Retrieval

- **Task:** Rank images based on relevance to the input sentence/query.
- **Dataset:** Same as (Schuster et al., 2015); a smaller version of Visual Genome.
- **Experiment:**
  - Directly apply the parser learned in the previous experiment to parse the query into scene graph
  - Compute the F-score between the parsed scene graph and ground truth scene graph obtained from image
  - Rank the images based on this F-score similarity
- **Evaluation metric**: Recall@5; Recall@10; Median rank.

Ref: Schuster et al., Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval, EMNLP 2015 Workshop

Image Retrieval Dataset Statistics

	Development Set	Test Set
Intersection of YFCC100m and MS COCO		
# of Images	454	456
# of Regions	4953	5180

Image Retrieval Results

	Development set			Test set		
	R@5	**R@10**	**Med. rank**	**R@5**	**R@10**	**Med. rank**
(Schuster et al., 2015)	33.82%	45.58%	6	34.96%	45.68%	**5**
Ours	**36.69%**	**49.41%**	**4**	**36.70%**	**49.37%**	**5**

Conclusion

- Introduction
- Method
- Experiments
- **Conclusion**

Conclusion

- Scene graph is a structured, explainable intermediate representation connecting image and text
- By taking the edge-centric view of scene graphs, we adapt techniques from dependency parsing to train a scene graph parser end-to-end
- We outperform previous works by a large margin, and efficacy is evaluated in terms of both F-score similarity and image retrieval experiments
- Code is released at https://github.com/Yusics/bist-parser/tree/sgparser

Thank you!