IOWA STATE UNIVERSITY

To Attend or not to Attend:
A Case Study on Syntactic Structures for Semantic Relatedness

Authors

Amulya Gupta
Email: guptaam@iastate.edu

Zhu (Drew) Zhang
Email: zhuzhang@iastate.edu

https://github.com/amulyahwr/acl2018

Agenda

- Introduction
- Classical world
- Alternate world
- Our contribution
- Summary

Problem Statement

Given two sentences, determine the semantic similarity between them.

Tasks

- Semantic relatedness for sentence pairs.
  1. Predict relatedness score (real value) for a pair of *sentences*
  2. Higher score implies higher semantic similarity among sentences
- Paraphrase detection for question pairs.
  1. Given a pair of *questions*, classify them as paraphrase or not
  2. Binary classification
     1. 1 : Paraphrase
     2. 0: Not paraphrase

Essence: Given two sentences, determine the semantic similarity between them.

Datasets used

- Semantic relatedness for sentence pairs.
  1. SICK (Marelli et al., 2014)
     1. Score range: [1, 5]
     2. Dataset: 4500/500/4927(train/dev/test)
  2. MSRpar (Agirre et al., 2012)
     1. Score range: [0, 5]
     2. Dataset: 750/750 (train/test)
- Paraphrase detection for question pairs.
  1. Quora (Iyer et al., Kaggle, 2017)
     1. Binary classification
        1. 1 : Paraphrase
        2. 0: Not paraphrase
     2. Dataset: Used 50,000 data-points out of 400,000

     80%(5%) /20% (train(dev)/test)

Examples

SICK

The badger is burrowing a hole	A hole is being burrowed by the badger	4.9

MSRpar

The reading for both August and July is the best seen since the survey began in August 1997.	It is the highest reading since the index was created in August 1997.	3

Quora

What is bigdata?	Is bigdata really doing well?	0

Linear

Generally, a sentence is read in a *linear* form.

English (Left to Right):

The badger is burrowing a hole.

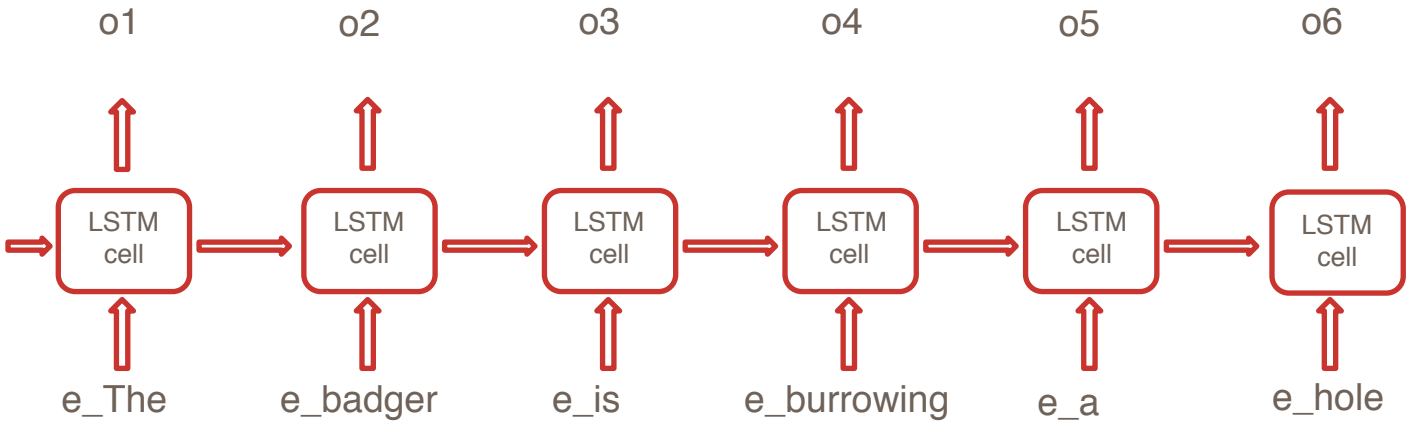
Urdu (Right to Left):

بیج ایک سوراخ پھینک دیتا ہے۔

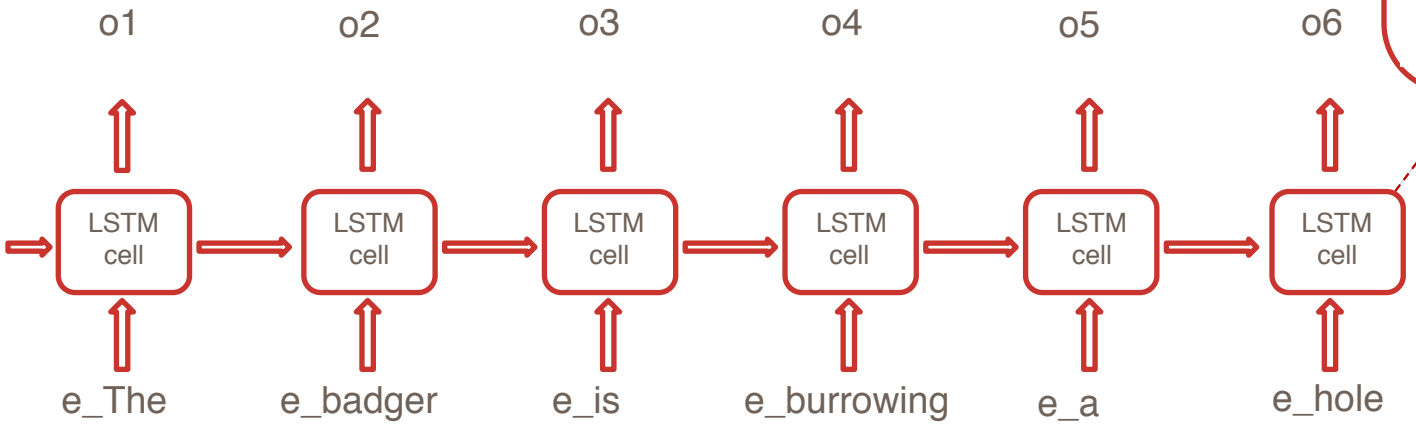
(Google Translate)

Traditional Chinese (Top to Bottom):

Long Short Term Memory (LSTM)



Long Short Term Memory (LSTM)



LSTM cell

$$i_t = \sigma(w_t W^i + h_{t-1} R^i + b^i) \quad (1)$$

$$f_t = \sigma(w_t W^f + h_{t-1} R^f + b^f) \quad (2)$$

$$o_t = \sigma(w_t W^o + h_{t-1} R^o + b^o) \quad (3)$$

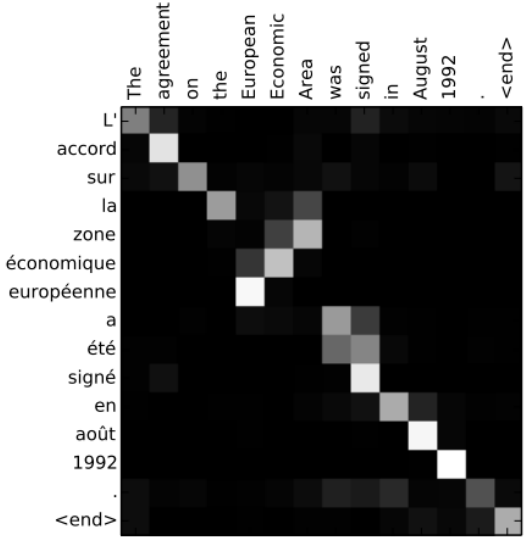
$$u_t = \tanh(w_t W^u + h_{t-1} R^u + b^u) \quad (4)$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1} \quad (5)$$

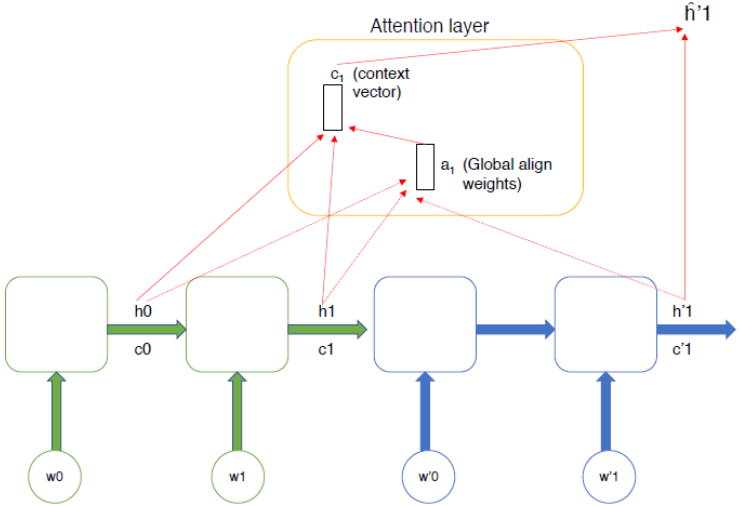
$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

Attention mechanism

Neural Machine Translation (NMT)
(Bahdanau et al., 2014)

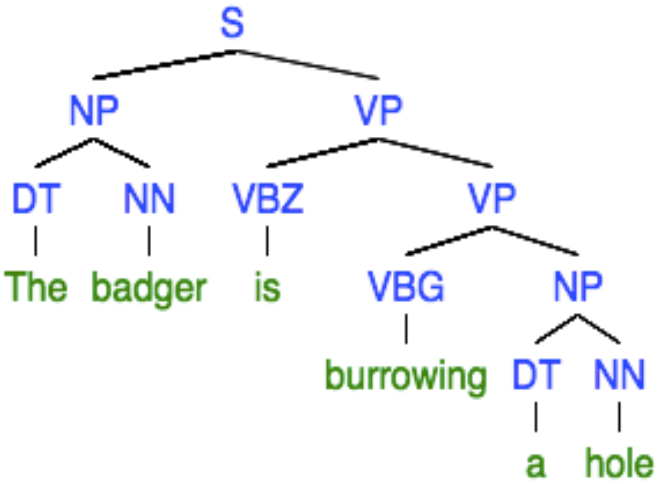


Global Attention Model (GAM)
(Luong et al., 2015)

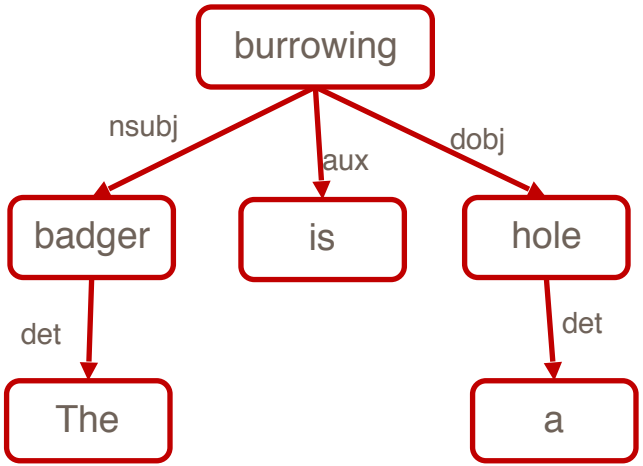


Tree

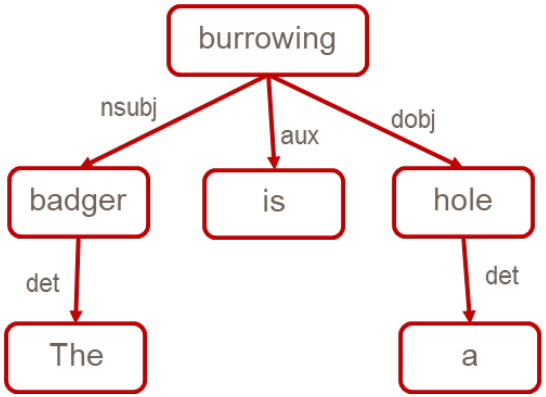
Constituency

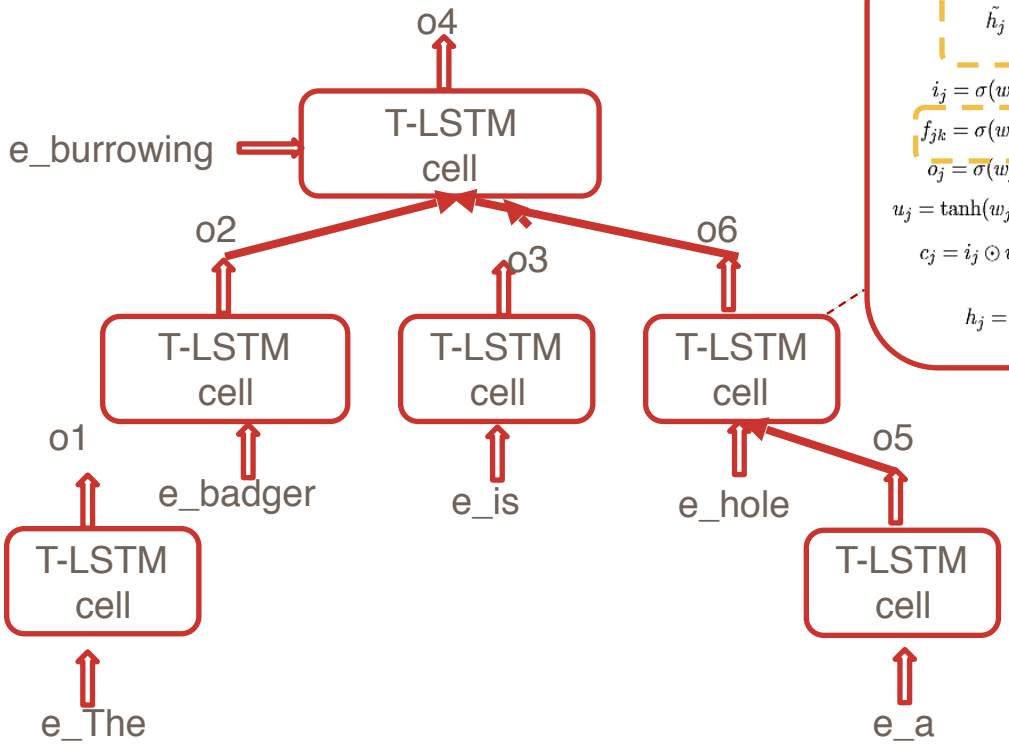


Dependency



Tree-LSTM (Tai et al., 2015)





T-LSTM cell

$$\tilde{h}_j = \sum_{k \in C(j)} h_k \quad (7)$$

$$i_j = \sigma(w_j W^i + \tilde{h}_j R^i + b^i) \quad (8)$$

$$f_{jk} = \sigma(w_j W^f + h_k R^f + b^f) \quad (9)$$

$$o_j = \sigma(w_j W^o + \tilde{h}_j R^o + b^o) \quad (10)$$

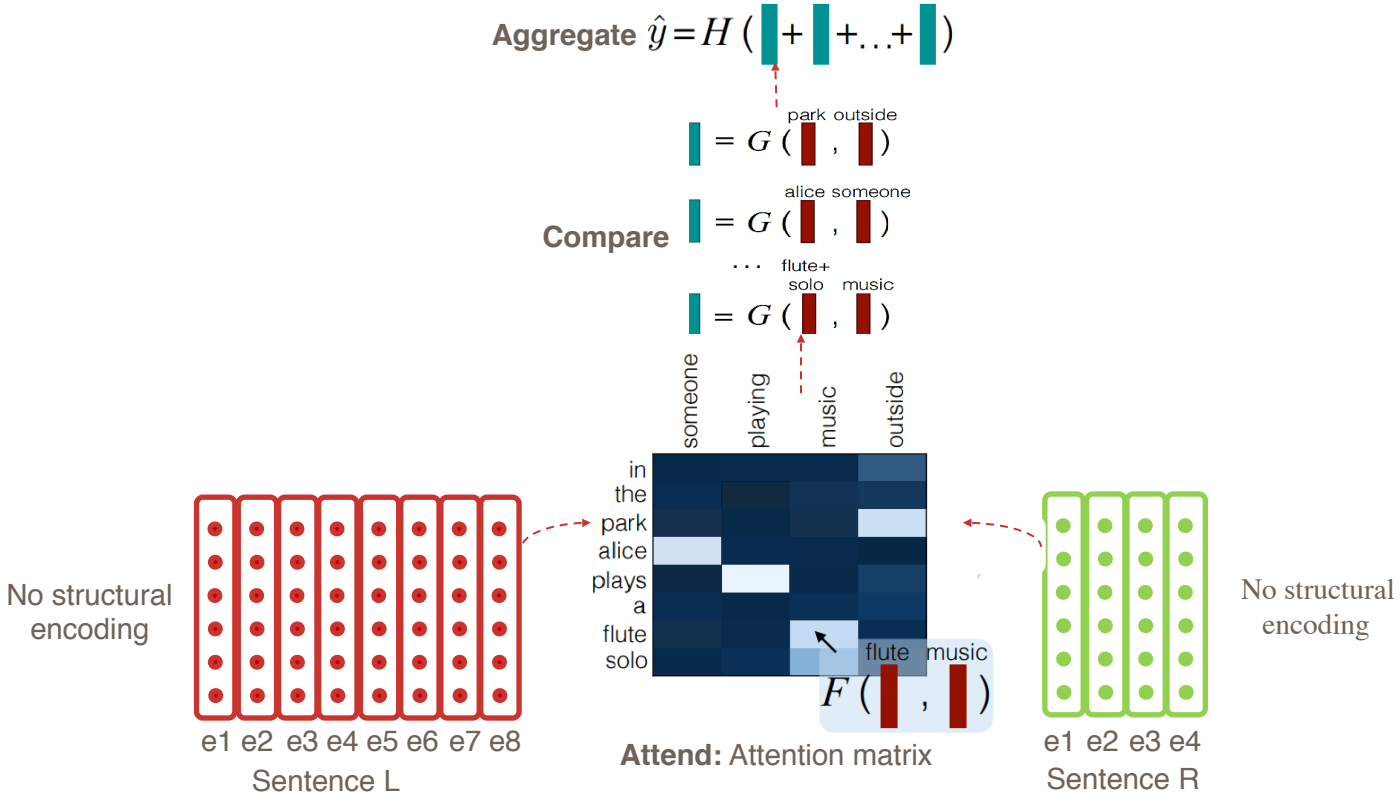
$$u_j = \tanh(w_j W^u + \tilde{h}_j R^u + b^u) \quad (11)$$

$$c_j = i_j \odot u_j + \sum_{k \in C(j)} f_{jk} \odot c_k \quad (12)$$

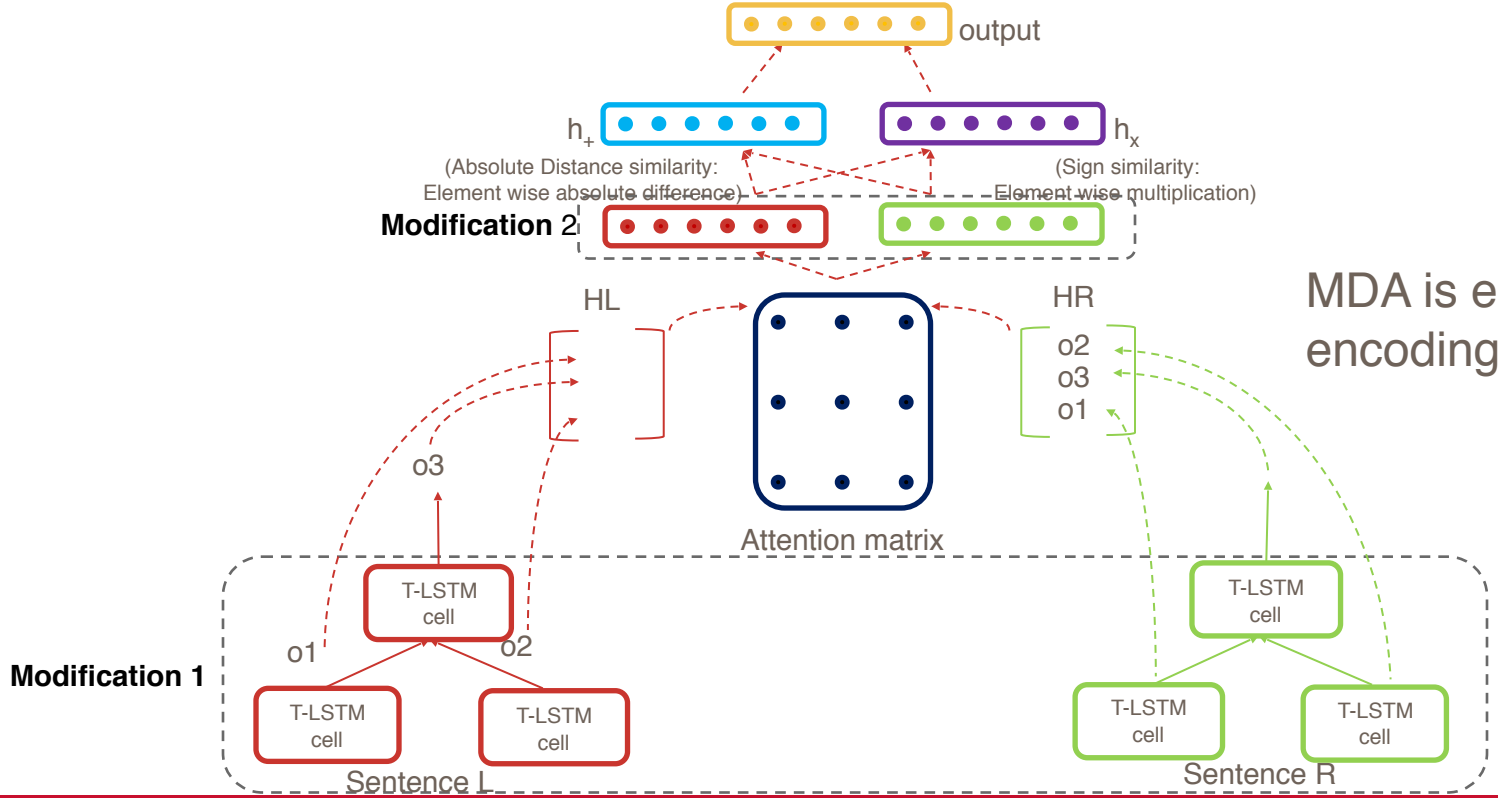
$$h_j = o_j \odot \tanh(c_j) \quad (13)$$

Attention mechanism

Decomposable Attention (Parikh et al., 2016)



Modified Decomposable Attention (MDA)



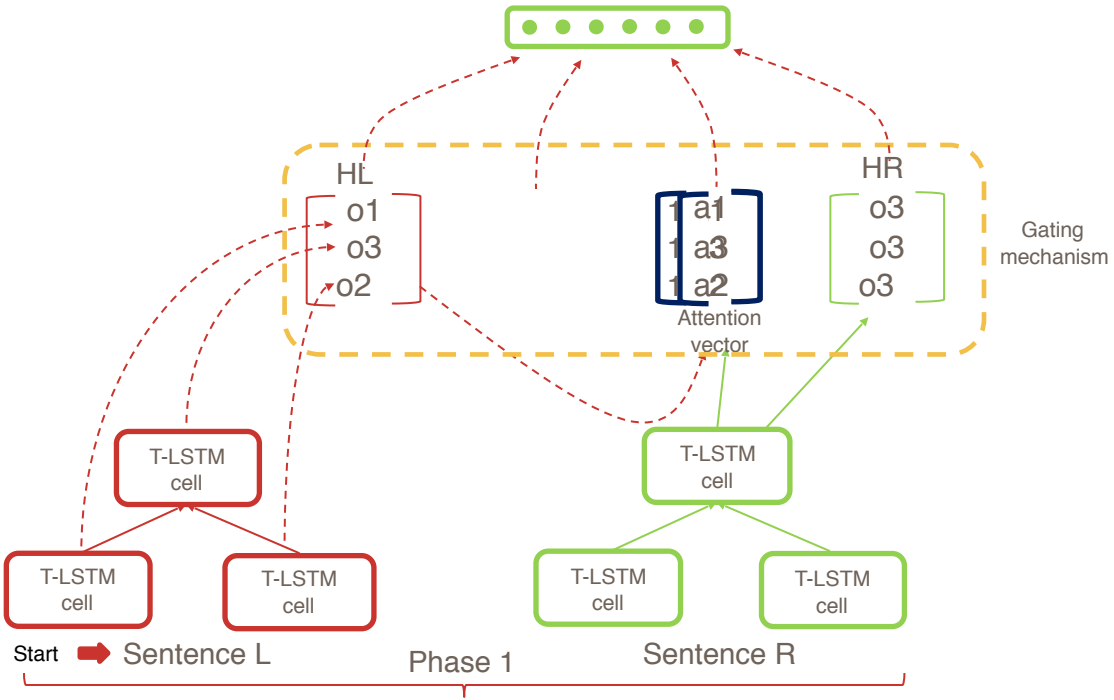
MDA is employed **after** encoding sentences.

Testset Results

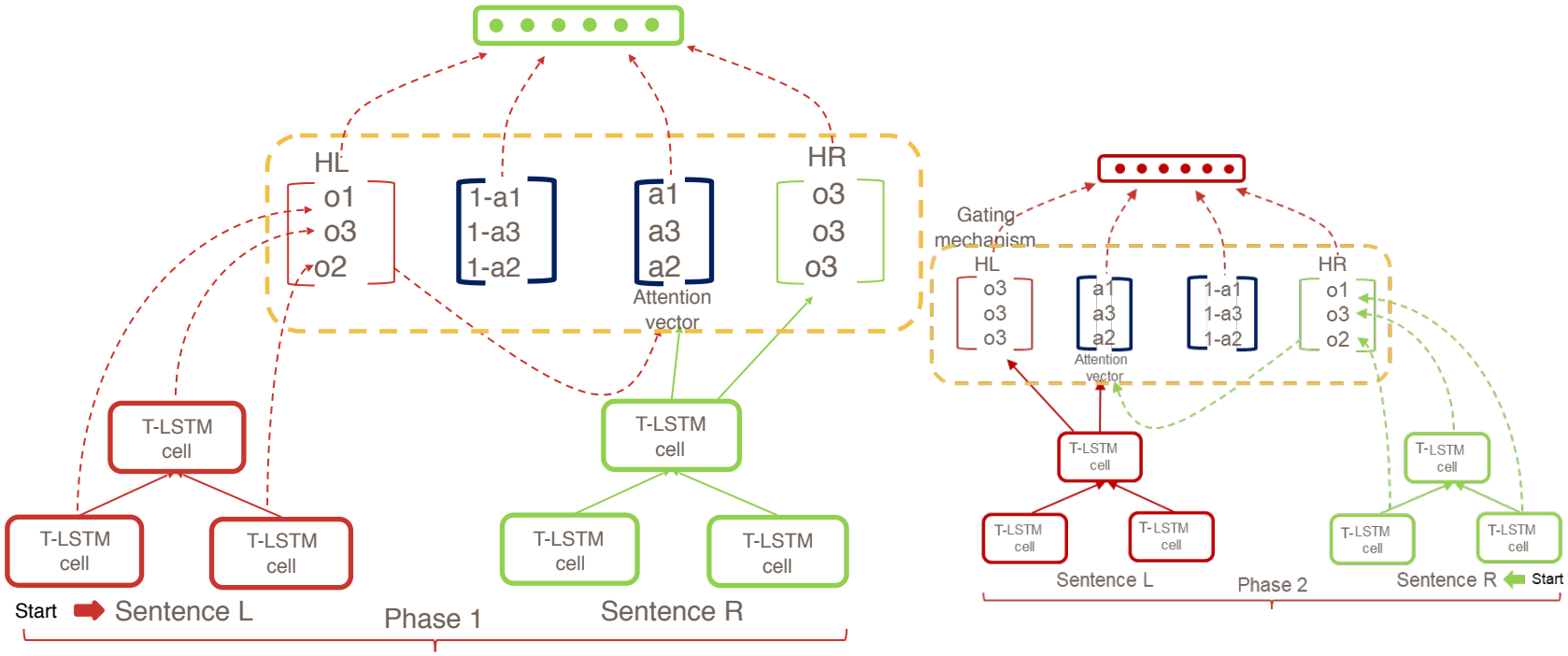
		Linear		Constituency		Dependency	
		w/o Attention	MDA	w/o Attention	MDA	w/o Attention	MDA
MSRpar	Pearson's r	0.327	0.3763	0.3981	0.3991	0.4921	0.4016
	Spearman's ρ	0.2205	0.3025	0.315	0.3237	0.4519	0.331
	MSE	0.8098	0.729	0.7407	0.722	0.6611	0.7243

		Linear		Constituency		Dependency	
		w/o Attention	MDA	w/o Attention	MDA	w/o Attention	MDA
SICK	Pearson's r	0.8398	0.7899	0.8582	0.779	0.8676	0.8239
	Spearman's ρ	0.7782	0.7173	0.7966	0.7074	0.8083	0.7614
	MSE	0.3024	0.3897	0.2734	0.4044	0.2532	0.3326

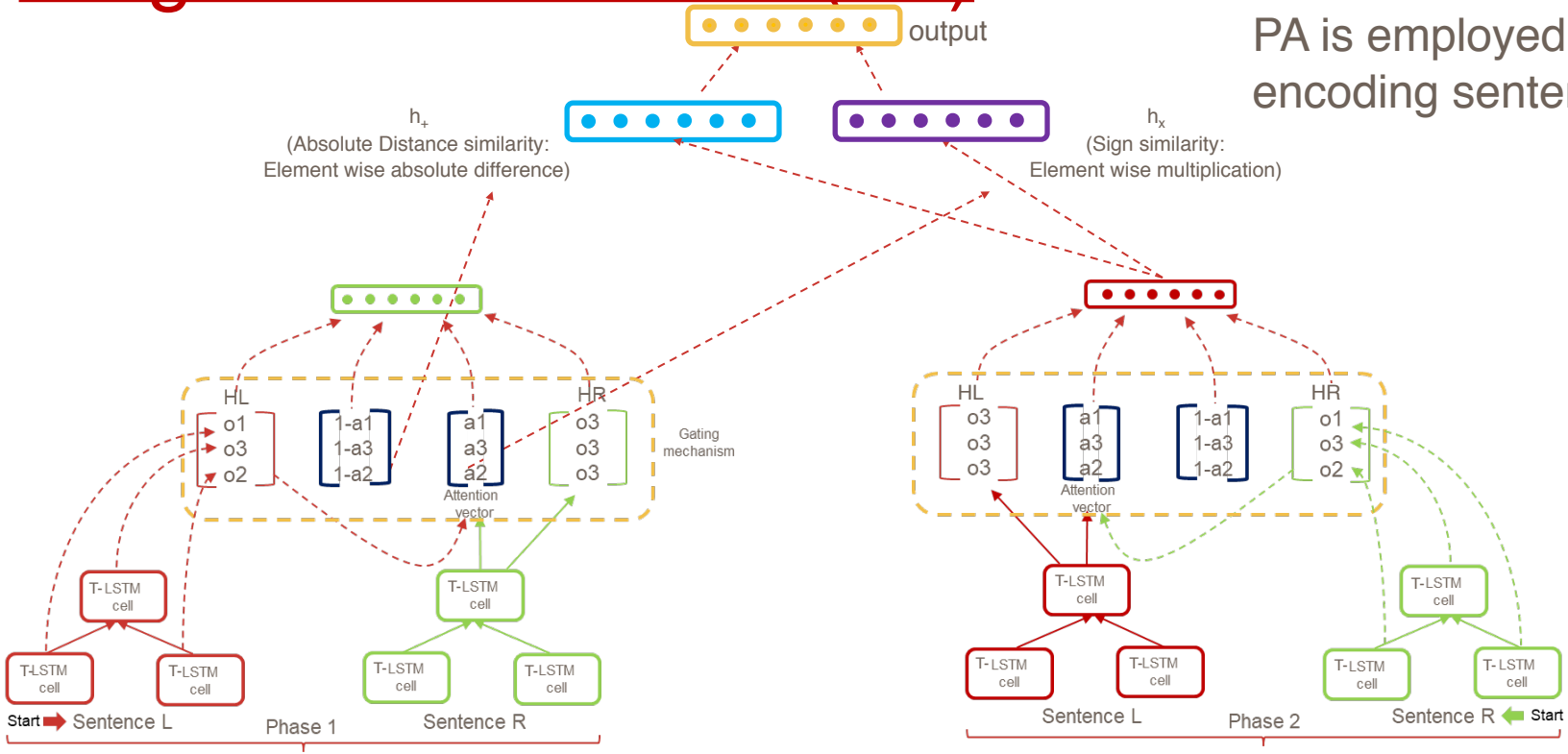
Progressive Attention (PA)



Progressive Attention (PA)



Progressive Attention (PA)



PA is employed **during** encoding sentences.

Effectiveness of PA

ID	Sentence 1	Sentence 2	Gold	Linear		Constituency		Dependency	
				No attn	PA	No attn	PA	No attn	PA
1	The badger is burrowing a hole	A hole is being burrowed by the badger	4.9	2.60	3.02	3.52	4.34	3.41	4.63

Testset Results

		Linear			Constituency			Dependency		
		w/o Attention	MDA	PA	w/o Attention	MDA	PA	w/o Attention	MDA	PA
MSRpar	Pearson's r	0.327	0.3763	**0.4773**	0.3981	0.3991	**0.5104**	0.4921	0.4016	0.4727
	Spearman's ρ	0.2205	0.3025	**0.4453**	0.315	0.3237	**0.4764**	0.4519	0.331	0.4216
	MSE	0.8098	0.729	**0.6758**	0.7407	0.722	**0.6436**	0.6611	0.7243	0.6823

		Linear			Constituency			Dependency		
		w/o Attention	MDA	PA	w/o Attention	MDA	PA	w/o Attention	MDA	PA
SICK	Pearson's r	0.8398	0.7899	**0.8550**	0.8582	0.779	**0.8625**	0.8676	0.8239	0.8424
	Spearman's ρ	0.7782	0.7173	**0.7873**	0.7966	0.7074	**0.7997**	0.8083	0.7614	0.7733
	MSE	0.3024	0.3897	**0.2761**	0.2734	0.4044	**0.2610**	0.2532	0.3326	0.2963

Discussion

		Linear			Constituency			Dependency		
		w/o Attention	MDA	PA	w/o Attention	MDA	PA	w/o Attention	MDA	PA
MSRpar	Pearson's r	0.327	0.3763	**0.4773**	0.3981	0.3991	**0.5104**	0.4921	0.4016	0.4727
	Spearman's ρ	0.2205	0.3025	**0.4453**	0.315	0.3237	**0.4764**	0.4519	0.331	0.4216
	MSE	0.8098	0.729	**0.6758**	0.7407	0.722	**0.6436**	0.6611	0.7243	0.6823

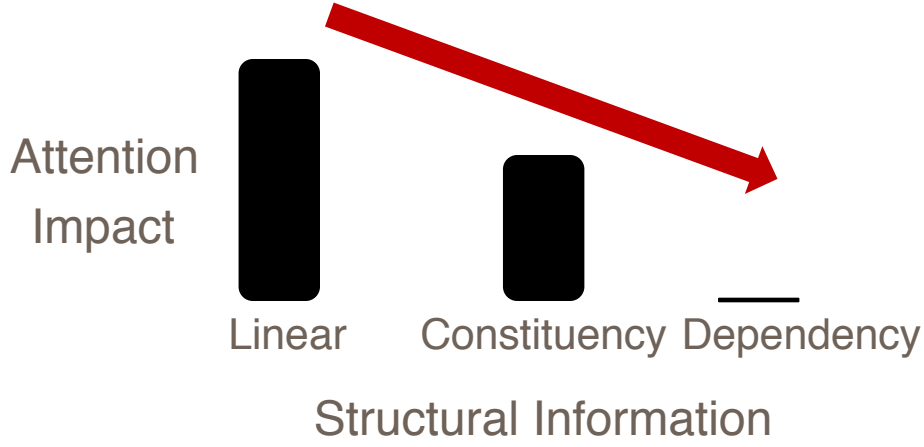
		Linear			Constituency			Dependency		
		w/o Attention	MDA	PA	w/o Attention	MDA	PA	w/o Attention	MDA	PA
SICK	Pearson's r	0.8398	0.7899	**0.8550**	0.8582	0.779	**0.8625**	0.8676	0.8239	0.8424
	Spearman's ρ	0.7782	0.7173	**0.7873**	0.7966	0.7074	**0.7997**	0.8083	0.7614	0.7733
	MSE	0.3024	0.3897	**0.2761**	0.2734	0.4044	**0.2610**	0.2532	0.3326	0.2963

Discussion

- **Gildea (2004):** Dependencies vs. Constituents for Tree-Based Alignment
- Is it because attention can be considered as an implicit form of structure which complements the explicit form of syntactic structure?
- If yes, does there exist some tradeoff between modeling efforts invested in syntactic and attention structure?
- Does this mean there is a closer affinity between dependency structure and compositional semantics?
- If yes, is it because dependency structure embody more semantic information?

		Linear			Constituency			Dependency		
		w/o Attention	MDA	PA	w/o Attention	MDA	PA	w/o Attention	MDA	PA
MSRpar	Pearson's r	0.327	0.3763	**0.4773**	0.3981	0.3991	**0.5104**	0.4921	0.4016	0.4727
	Spearman's ρ	0.2205	0.3025	**0.4453**	0.315	0.3237	**0.4764**	0.4519	0.331	0.4216
	MSE	0.8098	0.729	**0.6758**	0.7407	0.722	**0.6436**	0.6611	0.7243	0.6823

		Linear			Constituency			Dependency		
		w/o Attention	MDA	PA	w/o Attention	MDA	PA	w/o Attention	MDA	PA
SICK	Pearson's r	0.8398	0.7899	**0.8550**	0.8582	0.779	**0.8625**	0.8676	0.8239	0.8424
	Spearman's ρ	0.7782	0.7173	**0.7873**	0.7966	0.7074	**0.7997**	0.8083	0.7614	0.7733
	MSE	0.3024	0.3897	**0.2761**	0.2734	0.4044	**0.2610**	0.2532	0.3326	0.2963



Summary

- Proposed a modified decomposable attention (MDA) and a novel progressive attention (PA) model on tree based structures.

- Investigated the impact of proposed attention models on syntactic structures in linguistics.