A3.

1. (a). i. Using momentum means to keep track of previous gradients, so that the descent direction varies less than SGD.

This low variance helps to reach a steady optimization route, will be faster to learn.

ii. larger elements, aligns with the optimization direction? (faster convergence?)

Adam: decrease the velocity for a careful search

converge faster; (initial progress faster)

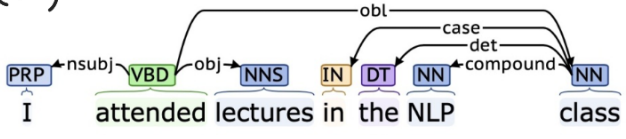
(b). i. $E_{p_{drop}}[h_{drop}]_i = E_{p_{drop}}[\gamma d \odot h_i] = \gamma(1-p_{drop})h_i = h_i$

$$\gamma = \frac{1}{1-p_{drop}}.$$

ii. during train: avoid overfitting, for better generalization

during eval: no need to erase possibly useful information

2. (a)



Stack	Buffer	New dependency	Transition
[ROOT]	[I, attended, lectures, in, the, NLP, class]		Initial Configuration
[ROOT, I]	[attended, lectures, in, the, NLP, class]		SHIFT
[ROOT, I, attended]	[lectures, in, the, NLP, class]		SHIFT
[ROOT, attended]	[lectures, in, the, NLP, class]	attended→I	LEFT-ARC
[ROOT, attended, lectures]	[in, the, NLP, class]		SHIFT
[ROOT, attended]	[in, the, NLP, class]	attended→lectures	RIGHT-ARC
[ROOT, attended, in]	[the, NLP, class]		SHIFT
[ROOT, attended, in, the]	[NLP, class]		SHIFT
[ROOT, attended, in, the, NLP]	[class]		SHIFT
[ROOT, attended, in, the, NLP, class]	[]		SHIFT
[ROOT, attended, in, the, class]	[]	class→NLP	LEFT-ARC
[ROOT, attended, in, class]	[]	class→the	LEFT-ARC

[ROOT, attended, class]	[]	class → in	LEFT-ARC
[ROOT, attended]	[]	attended → class	RIGHT-ARC
[ROOT]	[]	ROOT → attended	RIGHT-ARC

(b) 1 (initial config) + n (shifts) + n (arcs) = 2n+1 (steps)

(c) DONE

(d) DONE

(e) Best dev UAS: 88.85, corresponding loss: 0.044

Test UAS: 89.30

(b, ?) (b, f) (b, fxembed)

Sentence → features W → input vector X

↓

$\hat{y}$ = softmax (l) ← logits: $l = hU + b_2$ ← hidden: $h = \text{ReLu}(xW + b_1)$

(b, c) (b, c)

batch-size = b.

n-features = f.

n-classes = c

hidden-size = h.

b_1: (1, h).

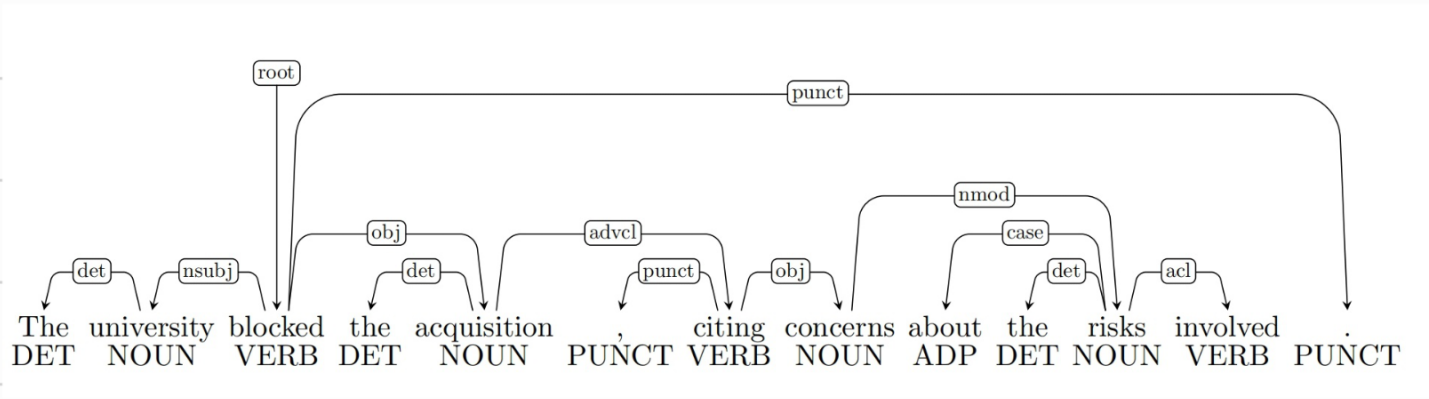
b_2: (1, c)

W: (fxembed, h)

h: (1, h)

U: (h, c)

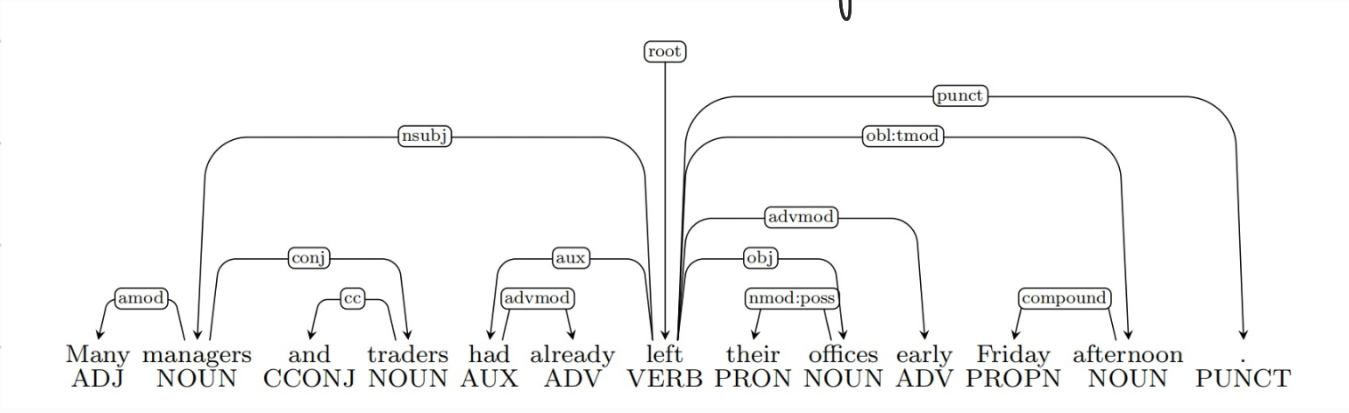
(f)



error type: verb attachment error.

wrong arc: acquisition → citing

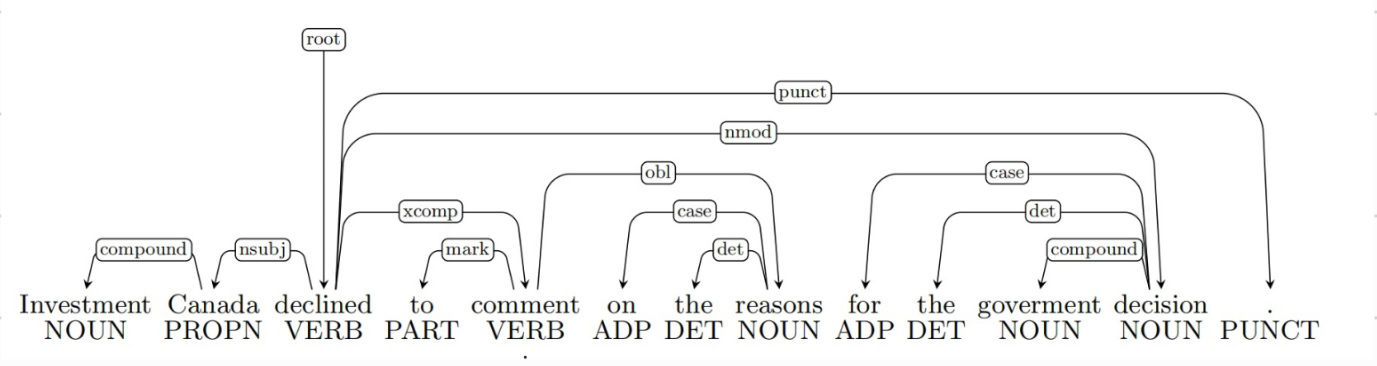
correct arc: blocked → citing



error type: modifier attachment error

wrong arc: had → already

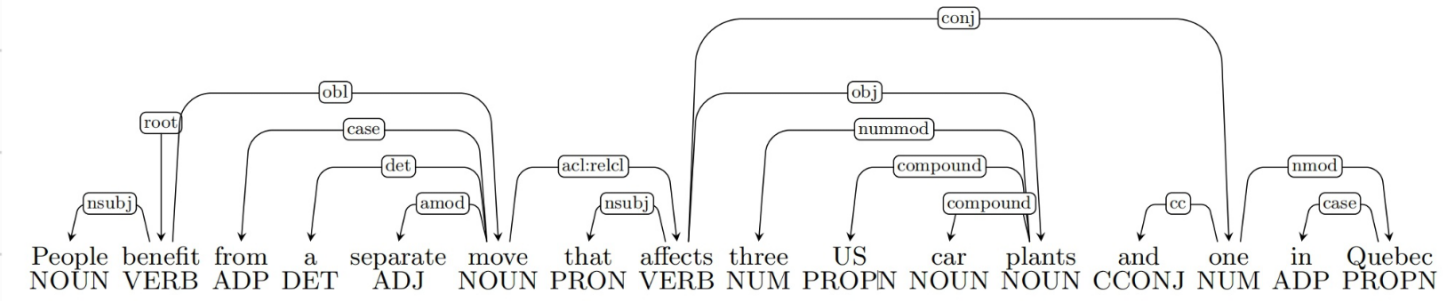
correct arc: left → already



error type: prepositional phrase attachment error.

wrong arc: declined → decision

correct arc: comment → decision



error type: coordination attachment error

wrong arc: affects → one

correct arc: plants → one

(g) POS tags share significant similarities with arc labels.