The approximation is provided in the equation right before Section 4 in Clark and Manning (2016)[1]. It is rather obscure but when we take into account the surprising fact that actions are really probabilistically independent (what's the point of reinforcement learning then?), we could (sort of) derive the formula.

Consider a simplified setting where number of actions $T = 2$:

$$\begin{aligned}
J(\theta) &= \mathbb{E}_{[a_{1:2} \sim p_\theta]} R(a_{1:2}) \\
\Rightarrow J(\theta) &= \sum_{a'_1 \in \mathcal{A}} \sum_{a'_2 \in \mathcal{A}} p_\theta(a'_1) p_\theta(a'_2) R(a'_1, a'_2) \\
\Rightarrow J(\theta) &= \sum_{a'_1 \in \mathcal{A}} p_\theta(a'_1) \left[\sum_{a'_2 \in \mathcal{A}} p_\theta(a'_2) R(a'_1, a'_2) \right] \\
\Rightarrow J(\theta) &= \mathbb{E}_{a'_1 \in \mathcal{A} \sim p_\theta} \left[\sum_{a'_2 \in \mathcal{A}} p_\theta(a'_2) R(a'_1, a'_2) \right] \\
\xRightarrow{\text{Monte Carlo}} J(\theta) &\approx \sum_{a'_2 \in \mathcal{A}} p_\theta(a'_2) R(a_1, a'_2) \\
\Rightarrow \nabla_\theta J(\theta) &\approx \sum_{a'_2 \in \mathcal{A}} \nabla_\theta p_\theta(a'_2) R(a_1, a'_2)
\end{aligned} \tag{1}$$

where a_1 is any sampled action.

Similarly, we have:

$$\nabla_\theta J(\theta) \approx \sum_{a'_1 \in \mathcal{A}} \nabla_\theta p_\theta(a'_1) R(a'_1, a_2) \tag{2}$$

So, in the end:

$$\nabla_\theta J(\theta) \approx \frac{1}{2} \sum_{i=1}^{2} \sum_{a'_i \in \mathcal{A}} \nabla_\theta p_\theta(a'_i) R(\dots, a'_i, \dots) \tag{3}$$

We can generalize:

$$\nabla_\theta J(\theta) \approx \frac{1}{T} \sum_{i=1}^{T} \sum_{a'_i \in \mathcal{A}} \nabla_\theta p_\theta(a'_i) R(\dots, a'_i, \dots) \tag{4}$$

In the paper, they omitted the $1/T$ factor. This will surely skews the objective in favor of longer documents.

References

[1] Kevin Clark and Christopher D. Manning. Deep Reinforcement Learning for Mention-Ranking Coreference Models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, pages 2256–2262, 2016.