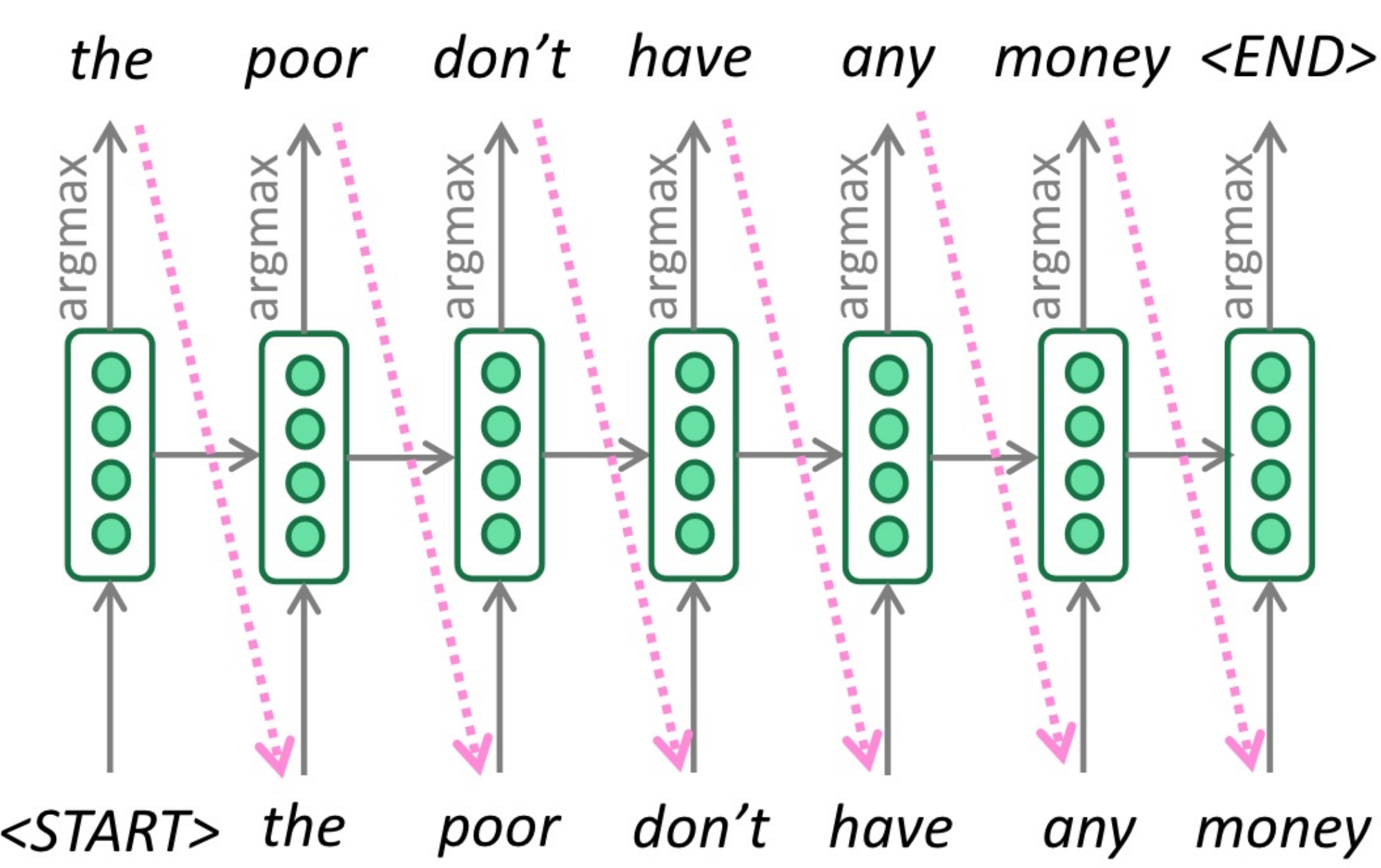
Machine Translation

Jordan Boyd-Graber

University of Maryland

Decoding

Adapted from material by Mohit Iyyer, Luke Zettlemoyer, Kalpesh Krishna, Karthik Narasimhan, Greg Durrett, Chris Manning, Dan Jurafsky



Argmax at every time step

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\left\{\beta \cdot \vec{f}(w)\right\}}{\sum_{w'} \exp\left\{\beta \cdot \vec{f}(w')\right\}} \tag{1}$$

- top-k
- Nucleus / top-p
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\left\{\beta \cdot \vec{f}(w)\right\}}{\sum_{w'} \exp\left\{\beta \cdot \vec{f}(w')\right\}} \tag{1}$$

- top-k: Only sample from k items with highest probability
- Nucleus / top-p
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\left\{\beta \cdot \vec{f}(w)\right\}}{\sum_{w'} \exp\left\{\beta \cdot \vec{f}(w')\right\}} \tag{1}$$

- top-k: Only sample from k items with highest probability
- Nucleus / top-p: Only sample from highest items with at least p probability
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\left\{\beta \cdot \vec{f}(w)\right\}}{\sum_{w'} \exp\left\{\beta \cdot \vec{f}(w')\right\}} \tag{1}$$

- top-k: Only sample from k items with highest probability
- Nucleus / top-p: Only sample from highest items with at least p probability
- Temperature

$$p(w) = \frac{\exp\left\{\frac{\beta \cdot \vec{f}(w)}{T}\right\}}{\sum_{w'} \exp\left\{\frac{\beta \cdot \vec{f}(w')}{T}\right\}} \tag{2}$$

Josiah Willard Gibbs, From Wikipedia

From Wikipedia

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\left\{\beta \cdot \vec{f}(w)\right\}}{\sum_{w'} \exp\left\{\beta \cdot \vec{f}(w')\right\}} \tag{3}$$

- top-k
- Nucleus / top-p
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\left\{\beta \cdot \vec{f}(w)\right\}}{\sum_{w'} \exp\left\{\beta \cdot \vec{f}(w')\right\}} \tag{3}$$

- top-k: Only sample from k items with highest probability
- Nucleus / top-p
- Temperature

Sampling Methods

Softmax distribution

$$p(w) = \frac{\exp\left\{\beta \cdot \vec{f}(w)\right\}}{\sum_{w'} \exp\left\{\beta \cdot \vec{f}(w')\right\}} \tag{3}$$

- top-k: Only sample from k items with highest probability
- Nucleus / top-p: Only sample from highest items with at least p probability
- Temperature

Sampling Methods

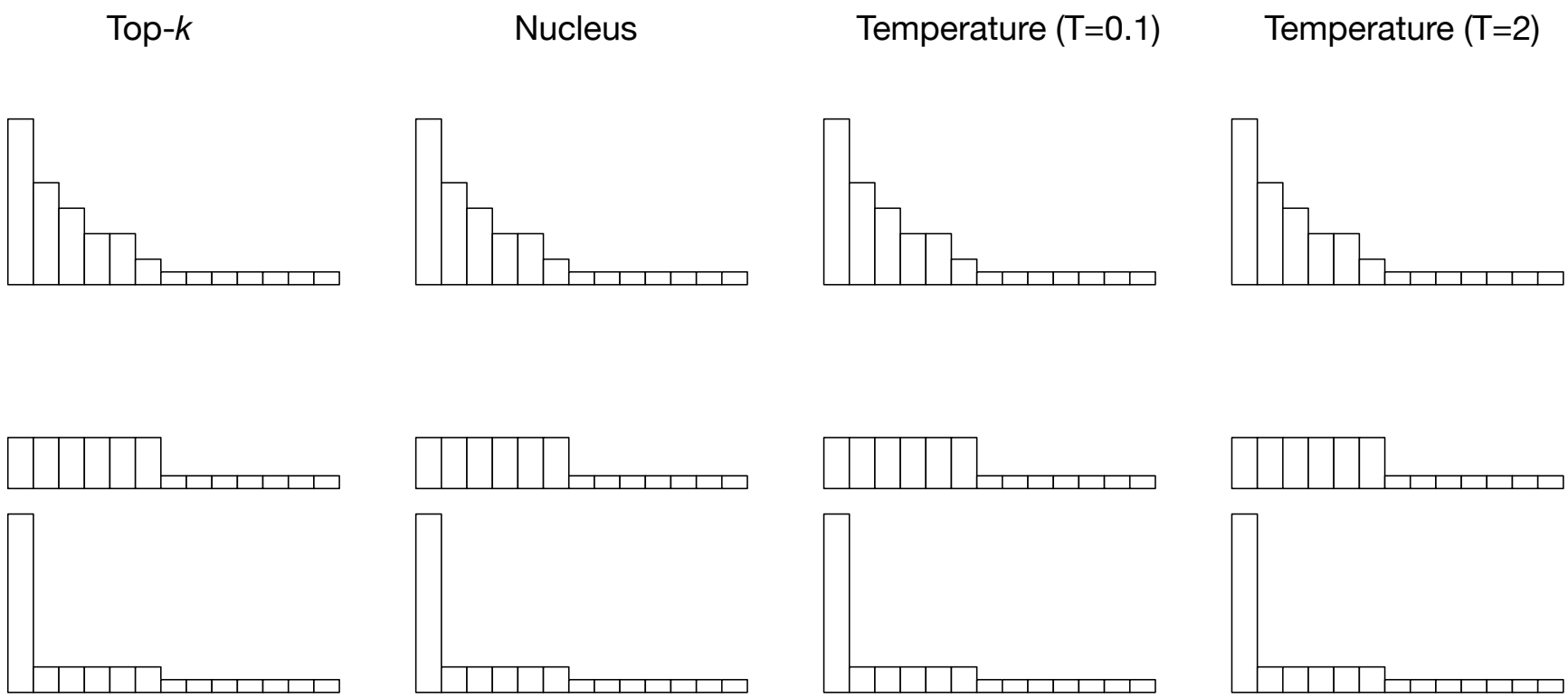
Softmax distribution

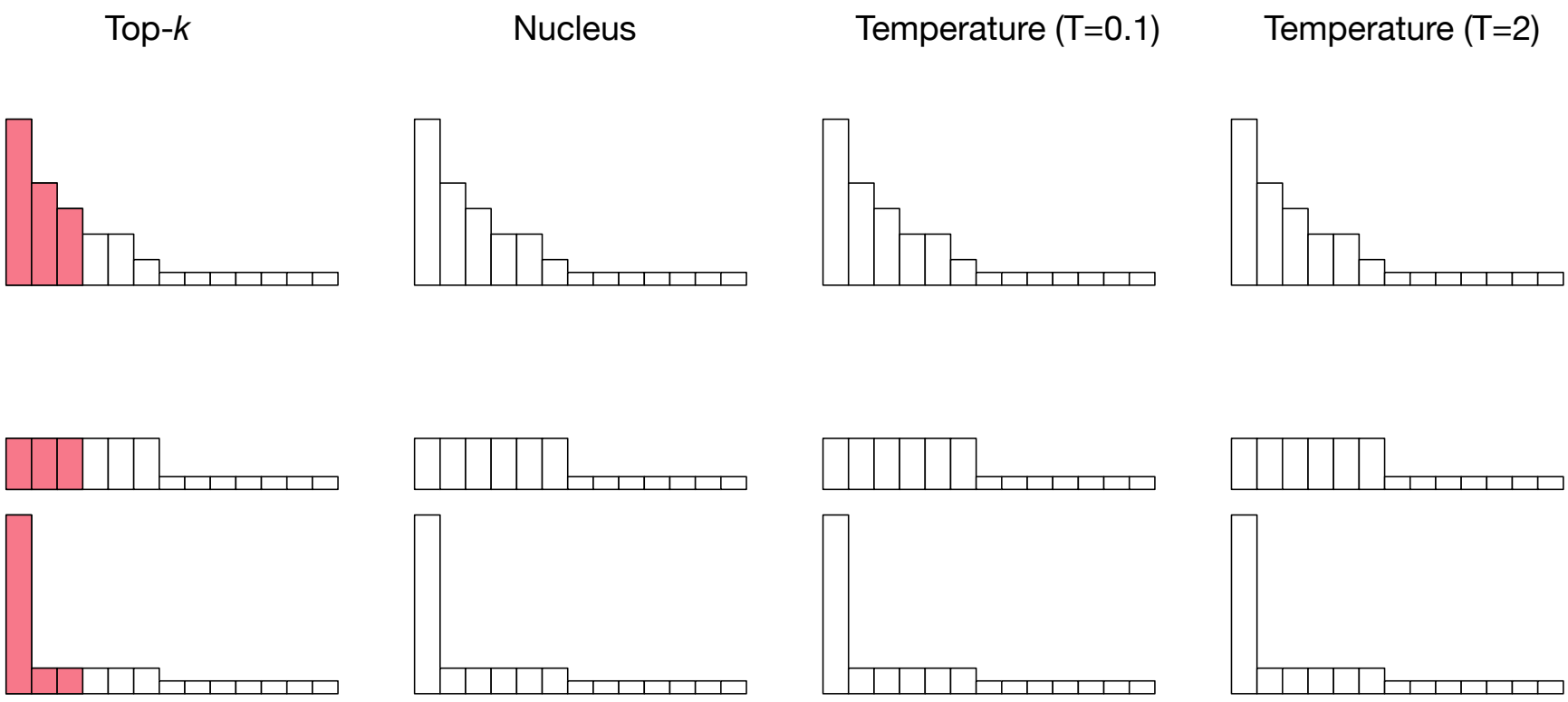
$$p(w) = \frac{\exp\left\{\beta \cdot \vec{f}(w)\right\}}{\sum_{w'} \exp\left\{\beta \cdot \vec{f}(w')\right\}} \tag{3}$$

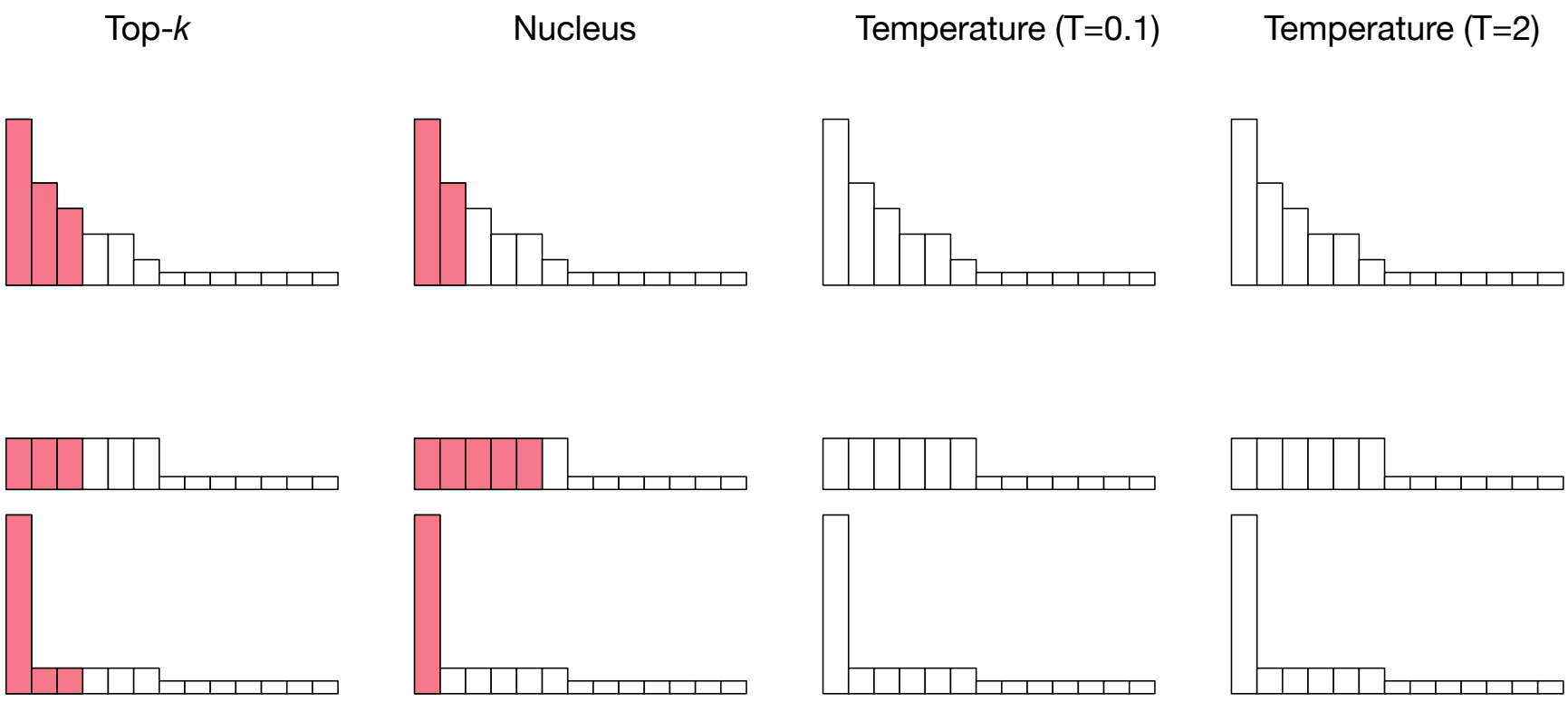
- top-k: Only sample from k items with highest probability
- Nucleus / top-p: Only sample from highest items with at least p probability
- Temperature

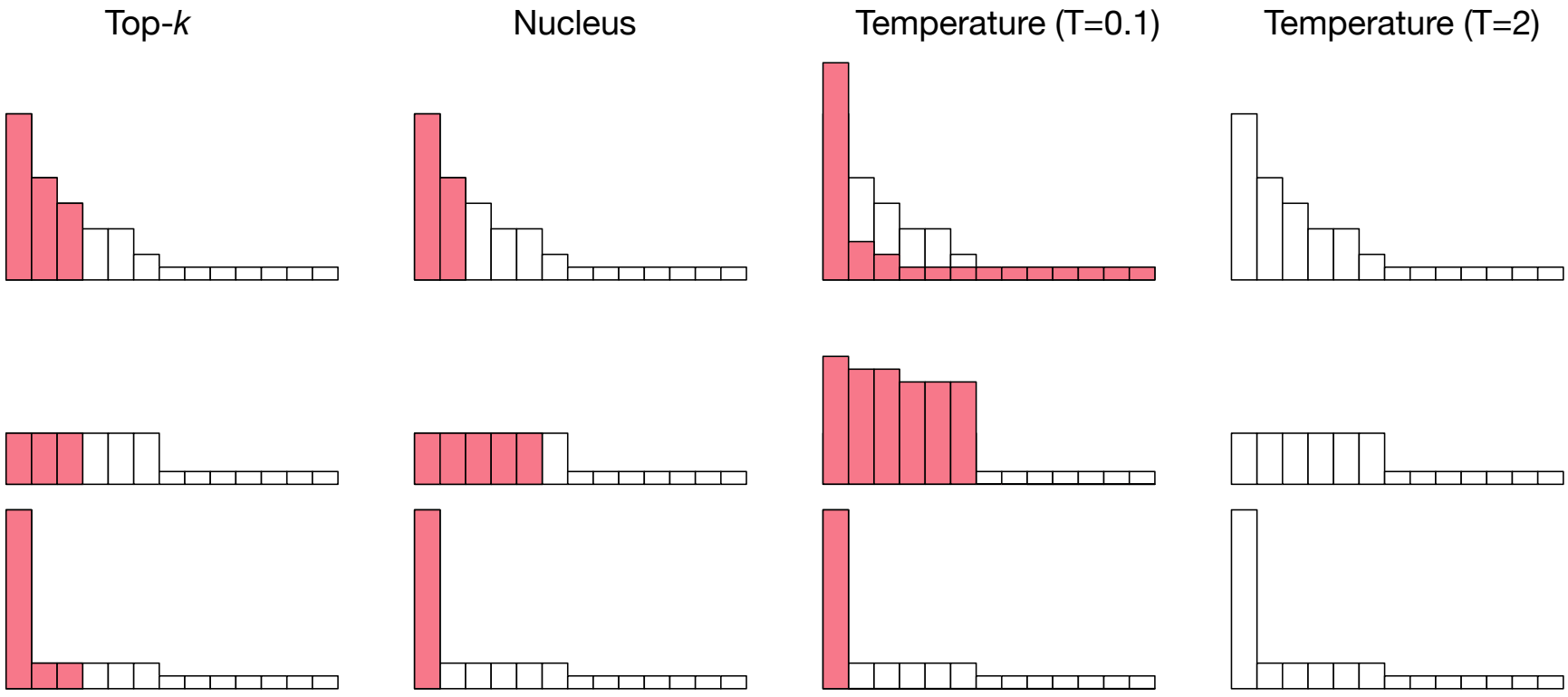
$$p(w) = \frac{\exp\left\{\frac{\beta \cdot \vec{f}(w)}{T}\right\}}{\sum_{w'} \exp\left\{\frac{\beta \cdot \vec{f}(w')}{T}\right\}} \tag{4}$$

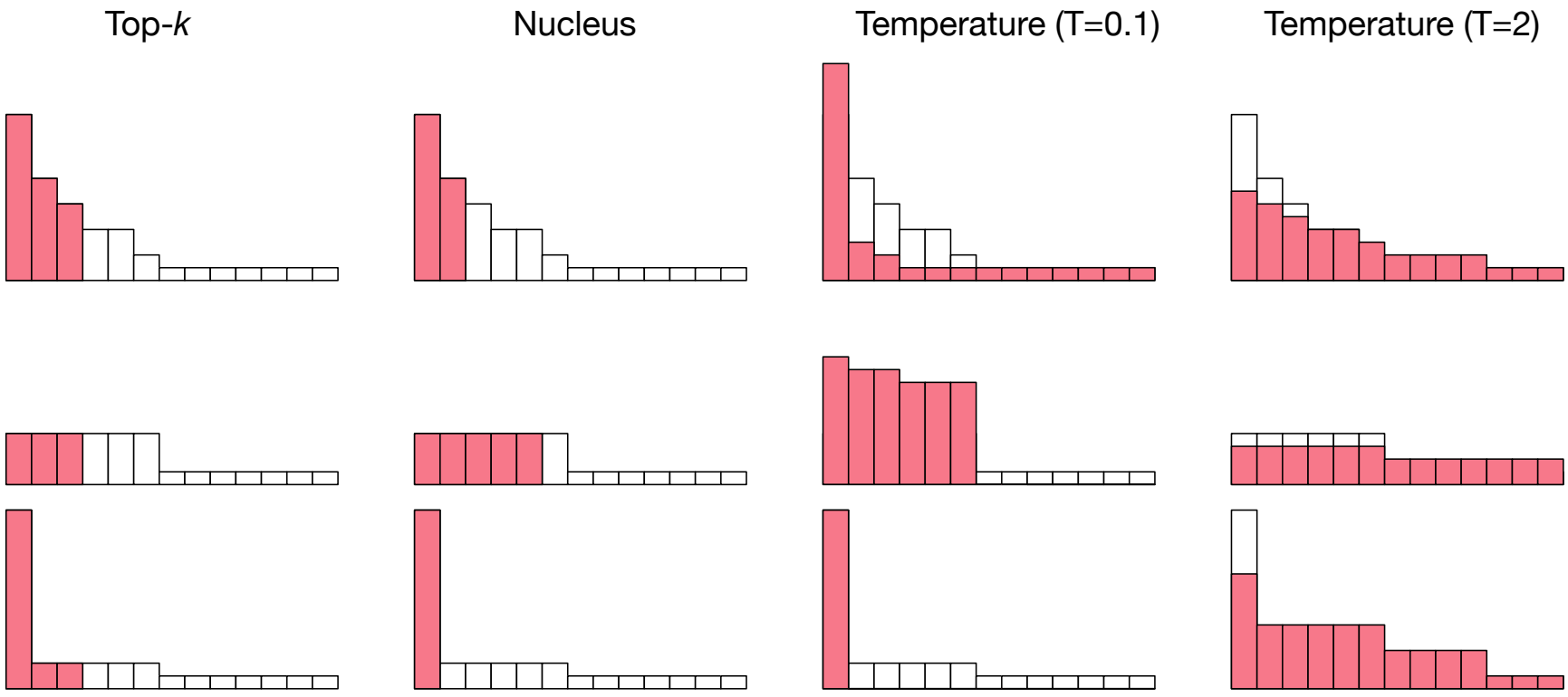
Top-*k* Nucleus Temperature (T=0.1) Temperature (T=2)


Top-*k*
Nucleus
Temperature (T=0.1)
Temperature (T=2)


Top-*k*
Nucleus
Temperature (T=0.1)
Temperature (T=2)


Top-*k*
Nucleus
Temperature (T=0.1)
Temperature (T=2)


Top-*k*
Nucleus
Temperature (T=0.1)
Temperature (T=2)


Top-*k*
Nucleus
Temperature (T=0.1)
Temperature (T=2)

What do you do with samples?

- Getting out of being stuck in a garden path
- Getting diverse outputs
- Combining multiple models together
- Rescoring by a non-probability metric

From Zhang et al. (Dive into Deep Learning)

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Beam Search Decoding for: Die Arme haben kein Geld

Using multiple sources

- Generate from multiple models
- Generate from multiple directions
- Generate from multiple data
- Generate from multiple temperatures

How to pick?

- Show to a user

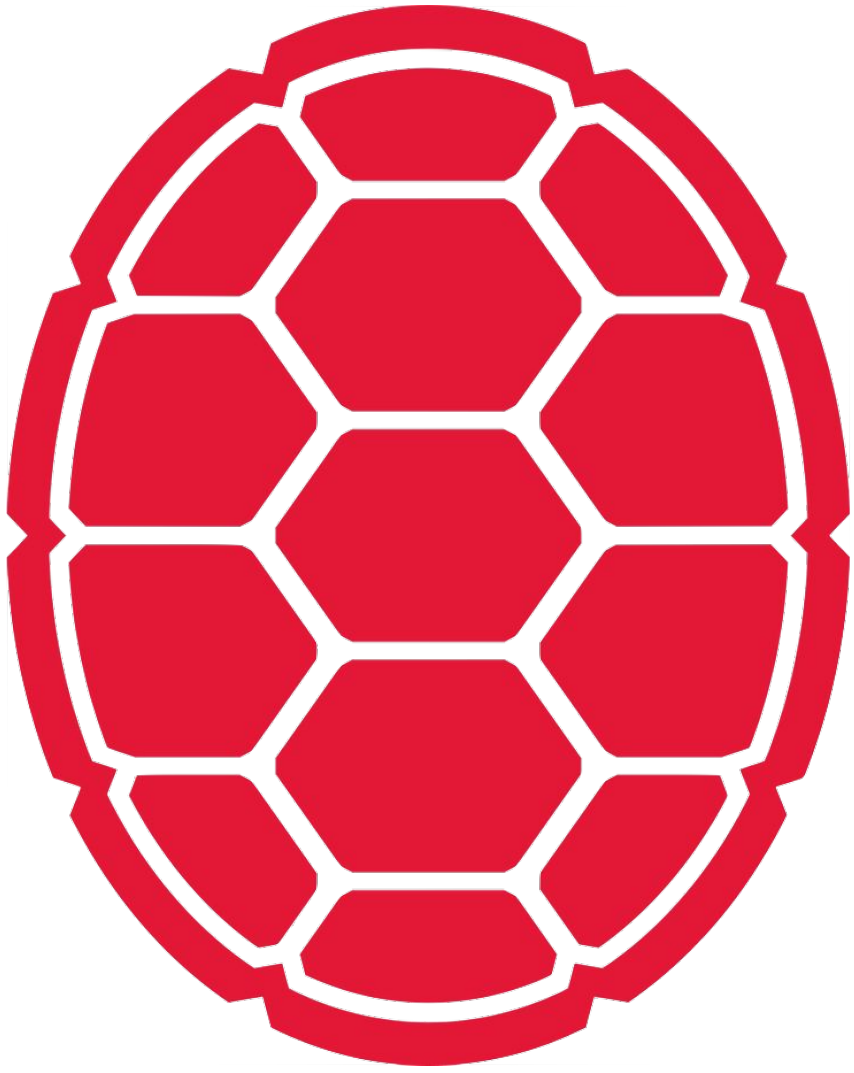
How to pick?

- Show to a user
- Take highest probability

How to pick?

- Show to a user
- Take highest probability
- Rerank

Tones in Chinese (for “shu”, not “ma” like I said)

Misheard lyrics when the tones are wrong

Decoding song translations with tones in decoder