Character-level Language Models With Word-level Learning

Arvid Frydenlund

March 16, 2018

Character-level Language models

- Want language models with an open vocabulary
  - Character-level models give this for free
- Treat the probability of a word as the product of character probabilities

$$P_w(w = c_1, ..., c_m | h_i) = \prod_{j=0}^{m} \frac{e^{s_c(c_{j+1}, j)}}{\sum_{c' \in \mathbb{V}_c} e^{s_c(c', j)}} \quad (1)$$

- Where $\mathbb{V}_c$ is the character 'vocabulary'
- Models are trained to minimize per character cross entropy
- **Issue:** Training focuses on how words look and not what they mean
- **Solution:** Do not define the probability of a word as the product of character probabilities

Globally normalized word probabilities

- Conditional Random Field objective

$$P_w(w = c_1, ..., c_m | h_i) = \frac{e^{s_w(w=c_1,...,c_m,h_i)}}{\sum_{w' \in \mathbb{V}} e^{s_w(w',h_i)}} \quad (2)$$

- normalizing partition function over all words in the (open) vocabulary
- **Issue:** Partition function is intractable
- **Solution:** Use beam search to limit the scope of the elements comprising the partition function.
  - This can be seen as approximating $P(w)$ by normalizing over the top most probable candidate words.
- **Issue:** Elements of partition are words of different length.
  - Score function and beam search need to be length agnostic.

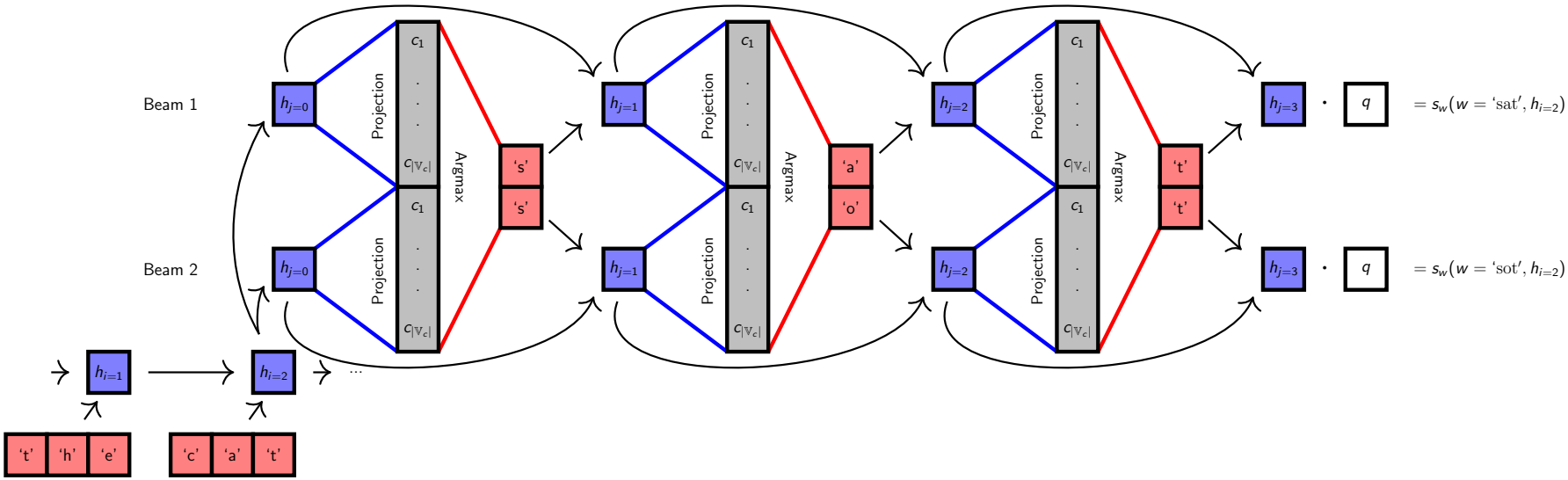

Figure: Predicting the next word in the sequence 'the cat'. The beam search uses two beams over three steps and produces the words 'sat' and 'sot' in the top beams.

- Beam search in back pass as well

$$J = \sum_{i=1}^{n} \left(-s_w(w_i, h_i) + \sum_{w' \in \mathbb{B}_{\text{top}}(i)} s_w(w', h_i) \right) \qquad (3)$$

Experiments

- Toy problem of generating word-forms given word embeddings
  - Compare to LSTM baseline
  - Test accuracy across different score functions (average character score, average character probability, hidden-state score)
  - Test accuracy across different beam-sizes
- Eventually a full language model
  - This model has dynamic vocabulary at every step
  - New evaluation metric for open vocabulary language models