Representations for Sentences

- Vector addition/multiplication
  - bag of words models
  - no tuning of representations
- Convolutional Neural Networks (CNNs)
  - feature learning over word subsequences
  - not strictly compositional
- Sequential language models (RNNs, LSTMs)
  - structure $\approx$ linear order
  - compositional representations for length $n$ sequences
- Recursive neural networks
  - structure $\approx$ binary trees
  - learn representations for linguistic units

Representations for Documents

- We have to first decide how to represent sentences.
- Then, we compose sentences into a document representation.
- Many classes of models are possible!
Representations for Documents

- We have to first decide how to represent sentences.
- Then, we compose sentences into a document representation.
- Many classes of models are possible!

\[ S \rightarrow NP \rightarrow \text{VP} \rightarrow \text{ADJP} \]

Mary was very hungry.
She didn’t find any food.
She went to the restaurant.

- Model is trained with a classification task in mind.
- Hidden layer learns document representation.
- We will see how this can be used for coherence modeling.

More Formally

- Let clique \( C \) denote a window of sentences.
- Each clique has a label \( y_C \in 1 \) if \( C \) coherent and 0 otherwise.
- How are labels generated?


**More Formally**

*d* = 10

Mary was hungry
She didn’t find any food at home
So she went to the restaurant
She had a prawn cocktail
It smelled funny
When she went home, she felt sick
She had food poisoning!

---

**Training Objective**

\[
J(\Theta) = \frac{1}{M} \sum_{C \in \text{trainset}} \{-y_C \log[p(y_C = 1)]
\]

\[-(1 - y_C) \log[1 - p(y_C = 1)]\}

\[+ \frac{Q}{2} \sum_{\theta \in \Theta} \theta^2 \]

\-

- Collection of *M* labeled training cliques *C*
- Error for one clique *C* with label *y_C*
- *\Theta* = [\text{W}_{\text{Recurrent}}, \text{W}_{\text{sen}}, \text{U}_{\text{sen}}]
- Regularization parameter

---

**Coherence Rating**

- Let \(h_C = [h_{s1}, h_{s2}, \ldots, h_{sL}]\) denote concatenation of sentences.
- Each clique \(C\) takes as input a \(L \times K\) vector \(h_C\) (\(L\) is the sentence window size, \(K\) dimensionality input sentence)
- Hidden layer \(H\) takes \(h_C\) as input and performs convolution using non-linear function:
  \[q_C = \tanh(W_{\text{sen}} \times h_C + b_{\text{sen}})\]
- Output layer takes \(q_C\) and generates a scalar using linear function; sigmoid projects value to [0,1] probability space:
  \[p(y_C = 1) = \text{sigmoid}(U^T q_C + b)\]
Results: Ordering

Hierarchical RNNs

Hierarchical CNNs

Conclusions

- Techniques for sentence modeling transfer to documents
- Different classes of models depending on choice of composition model for sentences/documents
- Is it reasonable to compress the meaning of a document in a single vector?
- Choice is motivated by computational reasons.