Lecture 15 – Text Summarization

Overview

1. Summarization as Clustering/Optimization
    a. Find words and sentences that best represent a document
    b. Many approaches…
        i. Centroid based (k-means)
        ii. ILP
        iii. …
        iv. These approaches are lightly supervised (need to provide a function that captures how well a piece of text is summarized)
2. Summarization using statistical techniques
    a. Didn't really catch on
3. Easier problem – sentence summarization
    a. Summarization phenomena
        i. Generalization: Russian Defense Minister → Russia
        ii. Deletion (removing details)
        iii. Paraphrase: for combatting → against
    b. Types
        i. Compressive (delete only)
        ii. Extractive (deletion and reordering)
        iii. Abstractive (any transformation)
4. What do humans do?
    a. Mostly compression and extraction? (needs more studies)
5. What type of problem is summary?
    a. Clustering
    b. Neural
        i. Supervised – where do we get the data?
        ii. Can only be done in some domains…where there is a lot of data.
        iii. Data sources for sentence summarization
            1. Newspapers
                a. Some papers are organized so that the first sentence is basically contains all the content of the article. So, the title is a summary of the first sentence of the article
        iv. Data sources for document summarization
            1. Newspapers
                a. Use story highlights feature of some newspapers
            2. Research articles
                a. Abstract summarizes the article
        v. One issue with these data sources is that they are naturally abstractive, but we want systems that are extractive

Neural Sentence Summarization

- Source document $x_1, \dots, x_M$
- Target document $w_1, \dots, w_N$
- Assume both from vocabulary V, generally $M \gg N$
- We want to learn the distribution P
- Compressive Model 1
    - Encoder/Decoder
        - $h_m^x \to RNN(h_{m-1}^x, x_m)$, $h_n \to RNN(h_{n-1}, w_n)$
    - Context
        - $c_m = h_M^x$
    - Large improvement over previous models
    - Surprisingly, it is not necessary to know things such as syntax
- Abstractive Model 2
    - Same encoder, decoder, and context
- Model 3
    - Included attention mechanism
        - Take the dot product between the encoder hidden states and the current decoder hidden state, and take the softmax
        - The context vector is the weighted average of the encoder hidden states

Abstractive Summarization Beyond Translation

- RNN fix many issues related to length, but introduces new problems
- Issues
    - Need a huge vocabulary, so it is computationally expensive
    - Summary should not repeat itself
    - Hard to determine which parts of the original document are important
- Model 4
    - We predict whether we should copy a word or generate a new word
- Model 5
    - Method 1: Penalize model for attention to the same words too many times
    - Method 2: Enforce during beam search at test time

Current Challenges

- How can we apply these models to other domains such as social media?
- There should be many good summaries.
- Convert data into a document