Intuit - Text Summarization

Brenton Chu, Shubhra Ganguly,
Dee Guo, Mudit Gupta, Jinlin He,
Arjun Sripathy

Summary of Summarization

- Extractive - take parts of the source text to use as the summary
  - Less flexible but easier to do
  - Currently has overall better results (based on ROUGE scores) than abstractive
  - Can be a combination of statistical, linguistic, and traditional ML methods
- Abstractive - generate completely new text
  - Theoretically allows for better (more concise, more accurate, etc.) summarization
  - Area of active research
  - State of the art almost exclusively uses deep neural methods

Extractive Labels

- Human-annotated summaries given by the data are abstractive
- Generate extractive labels by greedily maximizing ROUGE score between proposed subset of sentences and abstractive summary
  - Iterate through sentences in the text and build a working summary
  - For each each sentence that, when appended to the working summary, improves the ROUGE between the working summary and the human-annotated summary, add that sentence as an abstractive label
  - Continue until all sentences have been visited
- Issue in such a method for extractive summaries: pronouns and references to entities outside of the extracted summary
  - Proposed solution: Coreference resolution

Using Coreference Resolution

```
###
Before:
It comes one day after authorities confirmed the first case

After:
A second case of the potentially deadly MERS virus comes one day after authorities confirmed the first case
###
Before:
Bobby Jindal on Monday stood by his criticism of so-called``no-go''zones in Europe,where sovereign nations allege
dly cede authority to Muslim immigrants,a controversial idea that many critics say is overblown

Bobby Jindal on Monday stood by Louisiana Gov. Bobby Jindal's criticism of so-called``no-go''zones in Europe,wher
e sovereign nations allegedly cede authority to Muslim immigrants,a controversial idea that many critics say is o
verblown
###
Before:


``As a result,they believe they don'thave to answer to any government authority,including courts,taxing entities,
motor vehicle departments or law enforcement,''the agency says

After:


``As a result,sovereign citizens believe they don'thave to answer to any government authority,including courts,ta
xing entities,motor vehicle departments or law enforcement,''the agency says
```

Using Coreference Resolution

- spaCy's implementation:
  - Reference clusters

- Resolution Rules:
  1. Replace the reference with the 'identity' that occured before it
  2. Avoid multiple resolutions in the same sentence

- Specifics/corner cases:
  1. Possessive pronouns
  2. Nested references

Data

- CNN/Dailymail news articles
  - Large dataset consisting of CNN articles and multiple “highlights”
  - “highlights” are sentences that capture something important from the article
    - Not a sentence directly taken from article.
- New York Times news articles
  - Dataset containing a decade’s worth of NYT news articles
  - Dataset contains corresponding multi-sentence human written summaries for news articles.

Baseline Model

- Extractive: determine which sentences from original document to keep
- Features
  - Sentence position
  - Sentence length
  - Number of named entities and proper nouns
  - SumBasic word frequency
    - Word probability is number of occurrences divided by number of words in doc
    - Sentence weight is sum of word probabilities divided by number of words in doc
- Logistic Regression

Neural Extractive Model

- Implementation of SummaRuNNer Model based on research paper.
  - Created using Tensorflow
- Ultimate goal is to optimally assign a p to each sentence in text, where p represents probability that model uses sentence in a generated summary.
- First uses a Dynamic Bidirectional RNN at the word level to generate inputs for another BiRNN at the sentence level which is used in creating sentence representations which are used in assigning probabilities.
- Network is trained using abstractive labels by using an auxiliary sequential model that uses the words of the label summary as inputs and also utilizes the summary representation generated by the main network.