Extractive vs. Abstractive Text Summarization

Michelle Zhao

Motivation

- Summarization applied to scientific writing
- Intent: Effective summarization for scientific papers
- For now: Abstracts to titles

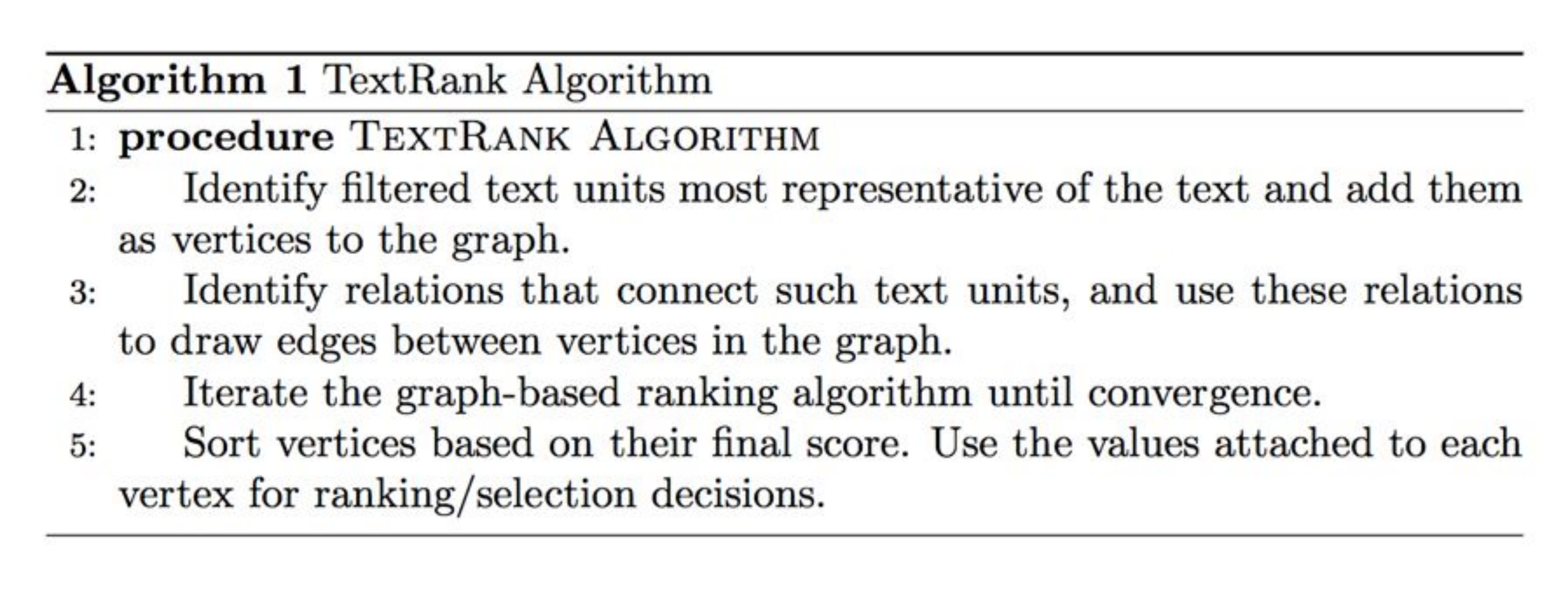
Materials and Methods (Dataset)

- NSF Research Award Abstracts 1990-2003 Data Set from the UCI machine learning repository.
- Abstracts that had won the NSF research awards from 1990 to 2003, along with the title of the paper.
- Abstractive learning
  - training input X: abstract
  - training input Y: title

Extractive Text Summarization

- Does not use words aside from the ones already in the text
- Selects some combination of the existing words most relevant to the meaning of the source.
- Ranking sentences and phrases in order of importance and selecting the most important components of the document to construct the summary.
- Robust because they use existing phrases
- Lack flexibility since they cannot use new words or paraphrase.

Algorithm



Algorithm 1 TextRank Algorithm

```
procedure TEXTRANK ALGORITHM
    Identify filtered text units most representative of the text and add them as vertices to the graph.
    Identify relations that connect such text units, and use these relations to draw edges between vertices in the graph.
    Iterate the graph-based ranking algorithm until convergence.
    Sort vertices based on their final score. Use the values attached to each vertex for ranking/selection decisions.
```

Algorithm Continued

1. First, we take the input text and split the entire text down to individual words.
2. Using a list of stop words, words are filtered so that only nouns and keywords are considered.
3. Then a graph of words is created where the words are the nodes/vertices. Each vertex’ edges are defined by connections of a word to other words that are close to it in the text. Each node is given a weight of 1.

Algorithm Continued

4. Then, we go through the list of nodes and collect the number of edges and connections the word has, which is essentially the influence of the connected vertex.

5. The scores are computed for every node, and the algorithm takes the top-scoring words that have been identified as important keywords.

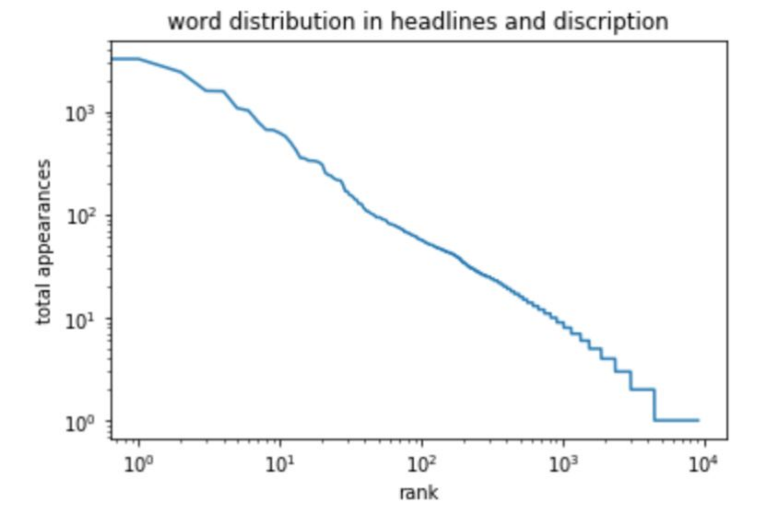
6. The algorithm sums up the scores for each of the keywords in all of the sentences, and ranks the sentences in order of score and significance. Finally, the top K sentences are returned to become the TextRank generated summary.

Abstractive Text Summarization

- Generating entirely new phrases and sentences to capture the meaning of the text.
- Tend to be more complex
- Learn to construct some cohesive phrasing of the relevant concepts.
- Most similar to how humans summarize, as humans often summarize by paraphrasing.

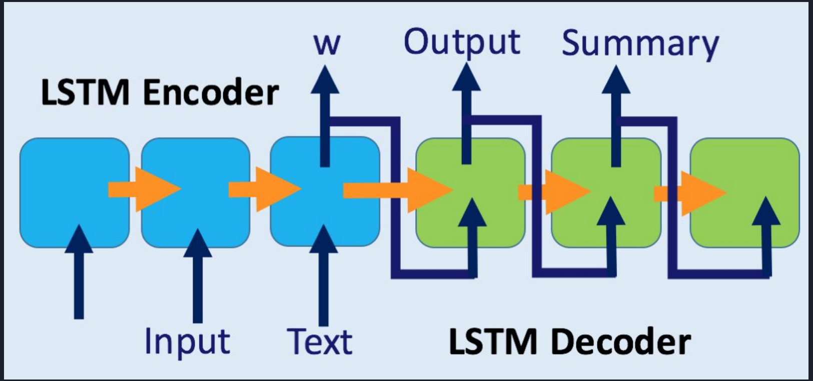
Algorithm

- Vectorize using Global Vectors: a count-based embedding



Algorithm

- Encoder converts an input document into a latent representation (a vector)
- Decoder reads the latent input, generating a summary as it decodes.



Encoder

- The encoder is a bidirectional LSTM recurrent neural network (RNN).
- RNNs can use their internal state (memory) to process sequences of inputs.
- LSTMs are capable of learning long term dependencies by storing long-term states and inputs in gated cell memory.
- The tokenized words of the text are fed one-by-one into the encoder, a single-layer bidirectional LSTM, producing a sequence of hidden states, which is a latent representation of the input.

Decoder

- The decoder is a single-layer unidirectional LSTM, which receives the word embedding of the previous word
- The embedding is transformed into a word representation, the summary.

Results

Extractive Results

Text Input	TextRank Summary
Commercial exploitation over the past two hundred years drove the great Mysticete whales to near extinction. Variation in the sizes of populations prior to exploitation, minimal population size during exploitation and current population sizes permit analyses of the effects of differing levels of exploitation on species with different biogeographical distributions and life-history characteristics. Dr. Stephen Palumbi at the University of Hawaii will study the genetic population structure of three whale species in this context, the Humpback Whale, the Gray Whale and the Bowhead Whale. The effect of demographic history will be determined by comparing the genetic structure of the three species. Additional studies will be carried out on the Humpback Whale. The humpback has a world-wide distribution, but the Atlantic and Pacific populations of the northern hemisphere appear to be discrete populations, as is the population of the southern hemispheric oceans. Each of these oceanic populations may be further subdivided into smaller isolates, each with its own migratory pattern and somewhat distinct gene pool. This study will provide information on the level of genetic isolation among populations and the levels of gene flow and genealogical relationships among populations. This detailed genetic information will facilitate international policy decisions regarding the conservation and management of these magnificent mammals.	**90% Reduction:** Stephen Palumbi at the University of Hawaii will study the genetic population structure of three whale species in this context, the Humpback Whale, the Gray Whale and the Bowhead Whale.

Extractive Results

Text Input	**TextRank Summary**
Commercial exploitation over the past two hundred years drove the great Mysticete whales to near extinction. Variation in the sizes of populations prior to exploitation, minimal population size during exploitation and current population sizes permit analyses of the effects of differing levels of exploitation on species with different biogeographical distributions and life-history characteristics. Dr. Stephen Palumbi at the University of Hawaii will study the genetic population structure of three whale species in this context, the Humpback Whale, the Gray Whale and the Bowhead Whale. The effect of demographic history will be determined by comparing the genetic structure of the three species. Additional studies will be carried out on the Humpback Whale. The humpback has a world-wide distribution, but the Atlantic and Pacific populations of the northern hemisphere appear to be discrete populations, as is the population of the southern hemispheric oceans. Each of these oceanic populations may be further subdivided into smaller isolates, each with its own migratory pattern and somewhat distinct gene pool. This study will provide information on the level of genetic isolation among populations and the levels of gene flow and genealogical relationships among populations. This detailed genetic information will facilitate international policy decisions regarding the conservation and management of these magnificent mammals.	**70% Reduction:** Variation in the sizes of populations prior to exploitation, minimal population size during exploitation and current population sizes permit analyses of the effects of differing levels of exploitation on species with different biogeographical distributions and life-history characteristics. Stephen Palumbi at the University of Hawaii will study the genetic population structure of three whale species in this context, the Humpback Whale, the Gray Whale and the Bowhead Whale. This study will provide information on the level of genetic isolation among populations and the levels of gene flow and genealogical relationships among populations.

Abstractive Results

Text Input	**Summary**
Proposal seeks to demonstrate a technique for observing ocean currents by electric field measurements using a towed instrument of recent design measurements will be made in conjunction with a cruise across the in which several additional observational techniques will be employed several data types will be to improve the accuracy of the methods	**Summary #1** Drum frame multidisciplinary **Summary #2** Extension solver bearing.

Abstractive Results

Text Input	**Summary**
Proposal seeks to demonstrate a technique for observing ocean currents by electric field measurements using a towed instrument of recent design measurements will be made in conjunction with a cruise across the in which several additional observational techniques will be employed several data types will be to improve the accuracy of the methods	**Summary #3** Exceptional geology goal visited

Conclusions: So what happened?

- TextRank selected the K most significant sentences in the text.
- E-D generated two different three-word summaries, using words not present in the text

Speed Comparison

- The extractive summaries were generated much quicker than the abstractive ones.
- The TextRank algorithm took about 2 seconds to generate a summary, while the encoder-decoder network took about 15 minutes to train.

Quality Comparison

- The summaries generated by ED were not representative of the text and did not make logical sense.
- The extractive summarizer worked better than the abstractive text summarizer.
- This may have been because the encoder-decoder network didn’t have enough training.
- If the encoder-decoder network perhaps have had more epochs of training, it would have performed better. The training input may have also been too small.

Next Steps

- Train on a larger training set
- Experiment with model hyper-parameters
- Use beam search
- Explore different preprocessing methods.

Thanks