Estimating Document Focus Time

Rahul Vadaga, Dikesh Chaturgoshthi

Course project for CS6370:
Natural Language Processing

Work Done

- Read the following papers —
  - *“TextRank: Bringing Order into Texts”*, Rada Mihalcea, Paul Tarau
    - Discusses graph based ranking algorithm for text processing applications.
    - Applies it for 2 problems - key phrase and sentence extraction.
  - *“Estimating Document Focus Time”*, Adam Jatowt et al., CIKM’13
    - Proposes a method to find the time to which a document content is referring to.
    - Learns word time associations from a giant corpus.
    - Uses this acquired information on a given document by combining the word time associations, giving its focus time.
  - *“An overview of graph based keyword extraction methods and approaches”*, S Beliga et al.

Limitations

- Word level search
  - Trying to associate each word to time
  - Both at W-T and D-T associations
- Doesn’t relate polysemous and synonymous words
  - eg. White Revolution (or) Operation Flood; Berlin Wall Collapse (or) Fall of the Wall (or) German Reunification
  - Doesn’t combine words to form phrases like Berlin Wall Collapse, etc.
- Words occurring with dates in documents is quite infrequent.
  - relying on dates in text isn’t a good idea.
- Temporal modifiers are not taken into account.
  - words like before, after, between.. and.. , from, during, since
  - eg. Before Lincoln took office in 1861, … ; 5 years before the …

Limitations

- Considering words that are not relevant to the document theme

Target Document

President Obama took part in the celebrations of the Polish Independence Day. The US president met main Polish politicians in Warsaw.

Poland regained independence at the end of the WWI following Bolshevik Revolution.

It then lost the independence as a result of Nazi and Soviet invasions led by Hitler and Stalin.

Reminiscing the past helps to avoid same mistakes in the future.

— Looking at it, we can say that the document is talking about Polish Independence.

— But this method doesn't eliminate Obama, Soviet invasions, etc., and it does W-T association on these words too.

Limitations

- Attributing document to a single time point is not always accurate.
  - Trying to associate each word to time
  - Documents can have finer granularity - month or day level

Ideas to explore

- Can use TextRank for sentence summarisation and find document focus time just from the summary.
  - Eliminates sentences not central to the document theme.
- Consider temporal modifiers (adverbs) within the vicinity (say, 2 or 3 words within the key phrase) to narrow down on the focus time.
  - Eg. Before the war…; 2 hours before the accident…
- Somehow find a way (like the sentence summarisation) to find the topic of the document, and perform word time association on that.
  - Helps in eliminating the noise in the document (Polish Independence example).
  - Intuitively, this is how humans tend to figure out what the document is about.
- Focus on finding sentence time associations and combine them to get the document focus time.

Problems to be addressed

- Constructing knowledge base from corpus that extends over several years (10 to 20 years) is time consuming.
  - Will be using TOI data over 2 years.
- For testing, Wikipedia articles would be used since they have accurate information about when an event occurred (date level granularity).
  - What time period should we choose?
- If we are able to identify what entities pronouns are referring to, it will help us know the document topic better.