Critics of "Estimating Document Focus Time"

Adam Jatowt, Ching-Man Au Yeng, Katsumi Tanaka

November 7, 2016

Overview

1. Introduction
2. Estimation of Focus Time
3. Limitations
4. Conclusion

Introduction

- **What** is Focus Time of a Document?

Introduction

- **What** is Focus Time of a Document?
  - ⋆ Time period to which document **content's** refers, i.e, the relation of document **content** with a time period

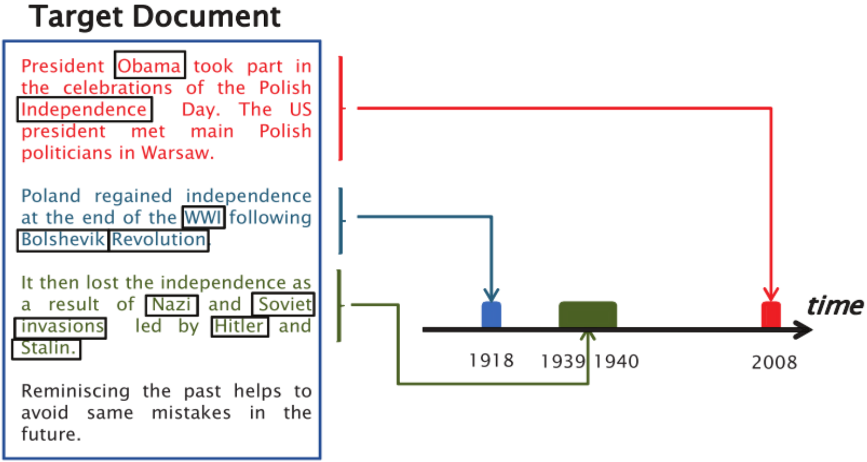
Introduction

- Consider a hypothetical Document

[1] Ack: Generic Method for calculating document focus time , Jatowt et al.

Introduction

- Consider a hypothetical Document



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.

Fig. Mapping content of an example document onto timeline. [1]

[1] Ack: Generic Method for calculating document focus time , Jatowt et al.

Introduction

- It contains sentences referring to different time points.
- It also contains sentence which is atemporal

Introduction

- It contains sentences referring to different time points.
- It also contains sentence which is atemporal
- But none of the sentence contains any temporal expression

Introduction

- It contains sentences referring to different time points.
- It also contains sentence which is atemporal
- But none of the sentence contains any temporal expression
- As a human we can position its content onto time line (as indicated)using temporal clue word "Obama", "Nazi", "Soviet" and "Stalin"
- Moreover, we can some what infer that the document is mainly focusing on **Polish Independence day**

Background

- The concept of focus time is defined as.

Defination

A temporal document, d, has the **focus time** τ if its content refers to τ

Background

- The concept of focus time is defined as.

Defination

A temporal document, d, has the **focus time** τ if its content refers to τ

→ The document describes events which has occurred in the given time period τ.

Background

- This corpus-based statistical method is composed of three steps

Background

- This corpus-based statistical method is composed of three steps
  1. **Word-Time Association**
     - ⋆ Associating each word with a time point
     - ⋆ "Nazi" and "Hitler" are strongly related to time period 1939 to 1945

Background

- This corpus-based statistical method is composed of three steps
  1. **Word-Time Association**
     - ⋆ Associating each word with a time point
     - ⋆ "Nazi" and "Hitler" are strongly related to time period 1939 to 1945
  2. **Estimating Temporal Weight** of words
     - ⋆ Some words have more temporal discrimination than others.
     - ⋆ Example, "Hitler", "Nazi", "Sun", "Tree"
     - ⋆ We need to give more weightage to the words which are discriminative of the focus time.

Background

- This corpus-based statistical method is composed of three steps
  1. **Word-Time Association**
     - ⋆ Associating each word with a time point
     - ⋆ "Nazi" and "Hitler" are strongly related to time period 1939 to 1945
  2. **Estimating Temporal Weight** of words
     - ⋆ Some words have more temporal discrimination than others.
     - ⋆ Example, "Hitler", "Nazi", "Sun", "Tree"
     - ⋆ We need to give more weightage to the words which are discriminative of the focus time.
  3. Calculating Text Focus Time
     - ⋆ Final estimation is done by extrapolating term focus point to document focus time by set of combination methods.
     - ⋆ Finding synchronicity between different temporal pointers

Word-Time Association

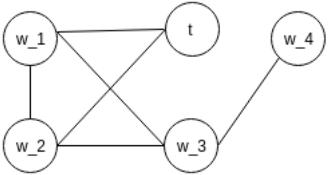
- For this, we use an external knowledge base which contains reference to past events with absolute dates.
- Dataset: Large collection news article on diverse topic is used as resource

Word-Time Association

- For this we use graph based method
- Graph is constructed with the occurrence of dates and words in the sentence
- We construct a weighted, undirected graph $G(V, E)$, where:
  - ⋆ V denotes the set of vertices being the vocabulary of the news
  - ⋆ E is set of edges representing their co-occurrences

Word-Time Association

- For this we use graph based method
- Graph is constructed with the occurrence of dates and words in the sentence
- We construct a weighted, undirected graph $G(V, E)$, where:
  - ⋆ V denotes the set of vertices being the vocabulary of the news
  - ⋆ E is set of edges representing their co-occurrences



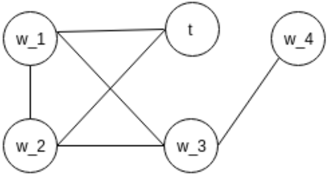
- The main focus here is to give the edge weight.

Word-Time Association: Edge weighing method

- Jaccard similarity
  - Uses number of co-occurrence of dates and words.
  - Number of times they have occurred individually.

Word-Time Association: Edge weighing method

- Jaccard similarity
  - Uses number of co-occurrence of dates and words.
  - Number of times they have occurred individually.



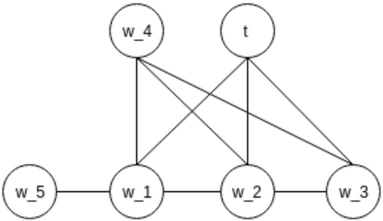
- Limitation
  - Occurrence of dates in text are sparse.
  - All words do not have occurrence with dates.

Word-Time Association: Edge weighing method

- To tackle the sparsity problem we consider co-occurrence among words.

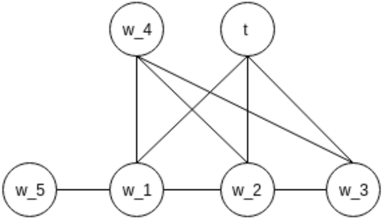
Intuition

Word w is strongly associated with time point t if many other words that strongly co-occur with w are also strongly associated with t



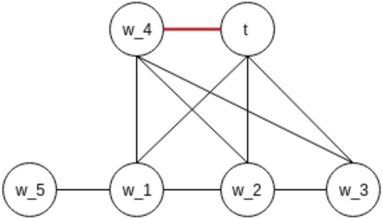
Word-Time Association: Edge weighing method

- Context Based Association.



Word-Time Association: Edge weighing method

- Context Based Association.



Estimating Temporal Weight

- We need to give some measure for the word importance with respect to time
  - ⋆ Not all words will be equally useful to determine focus time
  - ⋆ A term such as "earthquake" or "war" would have higher score than that of "tree" or "sun"
  - ⋆ To effectively determine the usefulness of word in discriminating time

Estimating Temporal Weight

- We need to give some measure for the word importance with respect to time
  - ⋆ Not all words will be equally useful to determine focus time
  - ⋆ A term such as "earthquake" or "war" would have higher score than that of "tree" or "sun"
  - ⋆ To effectively determine the usefulness of word in discriminating time
- So we follow the following assumption after analyzing the word association with time

Assumption

A word has high discriminative capability for determining document focus time if it has strong association with only few time points and weak association with other time points

Estimating Temporal Weight

- To rank term according to their discriminating capability we compute **temporal entropy** over association score of word with all time points
  - ⋆ We normalize the association scores to obtain the probability distribution over time
  - ⋆ Given a word, divide its association score with particular time point by sum of word's association score with all time points.

Estimating Temporal Weight

- It favors words that have non-uniform probability distribution over association with time
  - Because word having uniform association with different years are clearly not useful

Estimating Temporal Weight

- It favors words that have non-uniform probability distribution over association with time
  - Because word having uniform association with different years are clearly not useful
- But! this measure does not consider the distance between peaks in word-time associations
  - ⋆ "earthquake" or "war" would have long distance between their peak in word-time probability distribution
- Relying on them may bring confusion and hinder the performance of focus time

Estimating Temporal Weight

- We need to find terms having strong association with few nearby years or only one year (e.g, "Einstein" or "Hurricane Katrina")

Estimating Temporal Weight

- We need to find terms having strong association with few nearby years or only one year (e.g, "Einstein" or "Hurricane Katrina")
- To reflect this, as second term measure, **temporal kurtosis** is introduced
- It favors the words having distribution with one high peak ("Kurtosis is a measure of tailedness of probability distribution ")

Calculating Document-Time Association

- The basic intuition for the calculating document-time association is

Intuition

The more words strongly associated with time point t are contained in a document d, the more it is likely that t belongs to the focus time of d

Calculating Document-Time Association

- Computes weighted average over scores of terms present in the document
- Combining the association score with the temporal weightage

Calculating Document-Time Association

- Computes weighted average over scores of terms present in the document
- Combining the association score with the temporal weightage
- Method 1:**Unique Words**
  - considering each words only once

Calculating Document-Time Association

- Computes weighted average over scores of terms present in the document
- Combining the association score with the temporal weightage
- Method 1:**Unique Words**
  - considering each words only once
- Method 2: **Term Frequency**
  - The frequent the word is, more central it is to document.

Calculating Document-Time Association

- Computes weighted average over scores of terms present in the document
- Combining the association score with the temporal weightage
- Method 1:**Unique Words**
  - considering each words only once
- Method 2: **Term Frequency**
  - The frequent the word is, more central it is to document.
- Method 3: **TextRank**
  - Graph based key word extraction technique

Calculating Document-Time Association

- Note: all the association do not explicitly use temporal expression
- Temporal expression give important signals of document focus time

Calculating Document-Time Association

- Note: all the association do not explicitly use temporal expression
- Temporal expression give important signals of document focus time
- Extracts all the dates in document.
- Generate Gaussian distribution centered at the extracted date
- Add this weight to the previously calculated weights.

	Method
Word Time Association	Jaccard Coefficient
	Context Based Association
Temporal Weightage	Entropy
	Kurtosis
Document Time Association	Unique Words
	Term Frequency
	Text Rank
	Explicit Dates

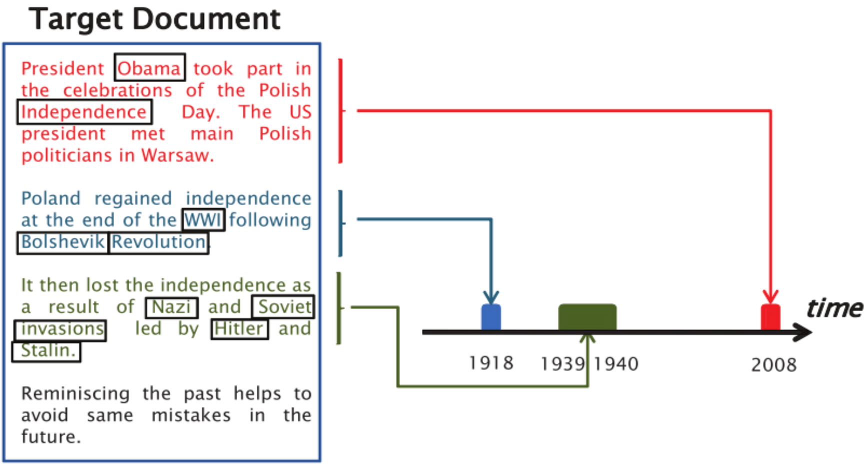
Limitation: Word Level Search

Limitation: Word Level Search

- World level search
  - both at W-T association and D-T association
  - associate each word to a time

Limitation: Word Level Search

- A hypothetical Document



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.

Fig. Mapping content of an example document onto timeline. [2]

[2] Ack: Generic Method for calculating document focus time , Jatowt et al.

Limitation:Problems associated with Word Level Search

- Includes non-relevant words (False positive)
  - words that are not central to the core theme of the document
  - eg, word Stalin(WW-II) present at the document of Berlin wall collapse(1989)

Limitation:Problems associated with Word Level Search

- Problem also occur due to polysemous and synonymous words
- Berlin wall collapse/fall of the wall/German reunification or White revolution/Operation flood
  1. It does not consider phrases
  2. It does not consider similar words while assigning temporal weightage

Limitation: Word-Time Association

- Idea of word co-occurring with date is sparse.
- In text dates does not occur as frequently.

Limitation:Temporal Modifiers

- Temporal modifiers give important clues on the order of the event
  - there was peace **after** the cold war era

Limitation:Temporal Modifiers

- Temporal modifiers give important clues on the order of the event
  - there was peace **after** the cold war era
  - other words are after, following, before, during, from, and between x and y

Limitation: Time granularity

- Only year is taken as granularity, document have finer granularity(month, day, day-time)
- Only single time point has been assigned to the text, where as it is possible to have multiple time points or document time can span over range of years

Limitation: Event Identification

- Events are always time specific, i.e.they have associated time
- It do not consider event.

Conclusion

- Described the concept of document focus time and provide a range of methods for its estimation
- Approach uses corpus statistics, especially it uses absolute references to past years in news articles
- This method also works for documents which do not contain any temporal expressions

Conclusion

- Described the concept of document focus time and provide a range of methods for its estimation
- Approach uses corpus statistics, especially it uses absolute references to past years in news articles
- This method also works for documents which do not contain any temporal expressions
- Central limitations are
  - Word Level Search
  - Time granularity
  - Temporal Modifiers
  - Event identification

References

Estimating Document Focus Time (CIKM 2013), Adam Jatowt, Ching-Man Au Yeng, Katsumi Tanaka
Generic method for detecting focus time of a documents.(Journal:IPM 2015), Adam Jatowt, Ching-Man Au Yeng, Katsumi Tanaka.

Thanks