Multiple News Articles Summarization based on Event Reference Information

Masaharu YOSHIOKA Makoto HARAGUCHI Hokkaido University
{yoshioka,makoto}@db-ei.eng.hokudai.ac.jp

Background

- Multiple News Articles Summarization
  - Text: Multiple news articles about particular events
  - Characteristics:Not a single document summarization
    - Redundant description
    - Important events might be referred several times in different articles

Objective

- Proposal of a method for Multiple News Articles Summarization based on Event Reference Information

Extraction of Events from a Sentence

- We apply Cabocha to obtain dependency analysis tree.
- We select verbs and nouns that have modification words as candidates of "Root" for events.
- We extract "Modifier" information from dependency analysis tree. At this time, we classify types of modifier by using POS tag and postpositional particle.
- When we can extract date information from the sentence, we set this date as "Date" for events that has dependency with date words.
- "ArticleDate" is obtained from article information.
- "Depth" and "Chunks" are calculated by comparing event information with the dependency analysis tree.

Event Reference

- Similar Events
  - Compare "Date" and "ArticleDate"
  - Compare "Root" and Corresponding element of "Root" and "Modifiers"

Important Sentence Extraction by using PageRank Algorithm

- PageRank : Calculate importance of pages by using link structure

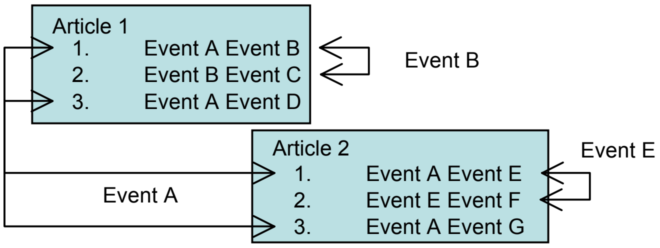
$$\vec{r}_{i+1} = M \times \vec{r}_i \qquad \vec{r}_\infty = \lim_{i \to \infty} \vec{r}_i$$

- Our algorithm
  - Node: a sentence
  - Link: a bidirectional link exists when two nodes shares same events or words
  - Transition matrix is calculated by a combination of event reference information and word reference information
    - Me: transition probability based on event reference biased with event importance.
    - Mw: transition probability based on word reference biased with IDF.

$$m_{ij} = \beta \times me_{ij} + (1-\beta) \times mw_{ij}$$

- Topic-Sensitive PageRank (Calculate importance of pages biased with initial importance vector)
  - Initial importance vector: Sentence position in an article

$$\vec{r}_{i+1} = (1-\alpha) * M \times \vec{r}_i + \alpha * \vec{v}$$



Article 1
1. Event A Event B
2. Event B Event C
3. Event A Event D

Event B

Article 2
1. Event A Event E
2. Event E Event F
3. Event A Event G

Event E

Event A

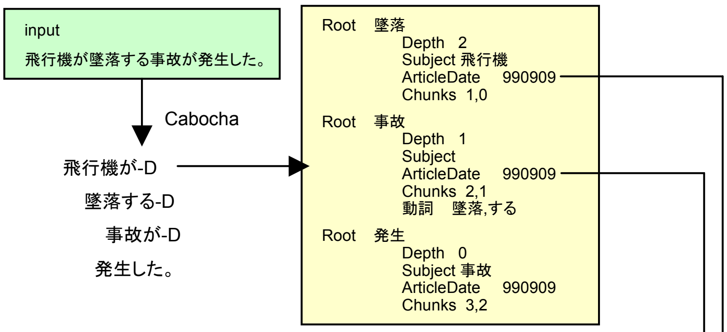
Link Structure based on Event Reference Information

Evaluation Results of Importance Sentence Extraction

		Short	Long
Event only	Coverage	0.325	0.313
	Precision	0.491	0.540
Event & Word	Coverage	0.323	0.341
	Precision	0.523	0.592
Word only	Coverage	0.313	0.344
	Precision	0.521	0.593

Event

- Information that describes facts and related information on particular date.
  - Root is a word that dominates an event (verb that represents action or noun that represents subject or object)
  - Modifier is words that modify root word. Words are categorized into several groups, such as subject and object words for verbs and adjective and adnominal words for nouns.
  - Negative represents modality of expression.
  - Depth is a path length between Root of the event and root of the sentence in dependency analysis tree.
  - Date is a date that characterize the event. This slot is not a required slot to define an event.
  - ArticleDate is a date that the article was published.
  - Chunks represents list of word positions in a sentence.



input
飛行機が墜落する事故が発生した。

Cabocha

飛行機が-D
墜落する-D
事故が-D
発生した。

Root 墜落
Depth 2
Subject 飛行機
ArticleDate 990909
Chunks 1,0

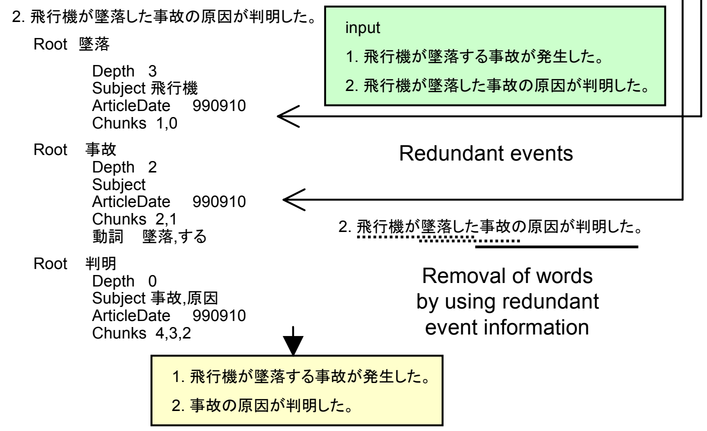
Root 事故
Depth 1
Subject
ArticleDate 990909
Chunks 2,1
動詞 墜落,する

Root 発生
Depth 0
Subject 事故
ArticleDate 990909
Chunks 3,2

An Example of Events Extraction from a Sentence

Text Reordering and Compaction by using Event Reference Information

- Keep sentence order in an article and chronological order of articles
  - When sentences comes from different articles in a same date, find similar sentence in its original article and set order based on it.
- Remove redundant event description from a sentence
  - An event that has similar events in the extracted event set is selected as a candidate one to remove.
  - Keep words, which is a "Root" element of an event and also belong to other non-redundant events.



2. 飛行機が墜落した事故の原因が判明した。

Root 墜落
Depth 3
Subject 飛行機
ArticleDate 990910
Chunks 1,0

Root 事故
Depth 2
Subject
ArticleDate 990910
Chunks 2,1
動詞 墜落,する

Root 判明
Depth 0
Subject 事故,原因
ArticleDate 990910
Chunks 4,3,2

input
1. 飛行機が墜落する事故が発生した。
2. 飛行機が墜落した事故の原因が判明した。

Redundant events

2. 飛行機が墜落した事故の原因が判明した。

Removal of words by using redundant event information

1. 飛行機が墜落する事故が発生した。
2. 事故の原因が判明した。

An Example of Text Compaction

Evaluated Result

- Importance sentence extraction
  - Usage of event reference information only is not good compared with usage of word information.
- Abstraction
  - Positive
    - q00: Our method for removing redundant description works well.
    - q08,q02: Our sentence reordering methods works well.
    - q04: Our system tends to select most frequent description because of word reference information.
  - Negative
    - q01: Removal of redundant information is too naïve. We assume that usage of anaphoric word is necessary to solve this problem.
    - q10: Our method of removing words from a sentence is too naive