Descriptive Metadata Framework and Taxonomy to Organize Topic-Specific Collections:

Text-mining for No Gun Ri Collections

Donghee Sinn (University at Albany)
SAA Research Forum, Aug. 23 2011

Topic-Specific Collections

- Various types of resources that are pertinent to a topic
- A typical descriptive metadata standard (MARC, Dublin Core) may not be useful.

- Topical approach to collect information
  - Library pathfinders
  - Digital libraries
  - Individual web sites

No Gun Ri

- No Gun Ri Massacre during the Korea War (July 1950)
  - Mass killing of South Korean refugees under a railroad overpass at No Gun Ri
  - By 7th Regiment soldiers in 1st Cavalry Division
  - Harsh refugee policies appeared in military documents from neighbor army units (25th Infantry, etc)
  - First reported in the US by AP in 1999
  - Controversies over testimonies of veterans (Edward Daily) and US No Gun Ri Review

A 1960 photo of the bridge taken by and obtained from the villagers who brought the claim against the United States. (AP Photo)

NGR Collection

- Materials from survivors' community, archival documents, journalistic publications, academic research studies, legal documents, government reports, media broadcasting, etc.
- A variety of types in format and nature
  - Hard to organize effectively, using an existing descriptive standard

Text-Mining

- Finding representative patterns from unstructured textual data
- Analyzing the contents in the collection to find how the contents represent the collection itself
- Text analysis tool: TAPoR (Text Analysis Portal for Research)
  - Keywords Finder
    - top 20 words; top 10 word pairs; and top 10 word triplets
    - recommended keywords/phrases

Text-Analysis for NGR Collection (Preliminary)

- 31 Archival Materials
  - 27 military documents
  - 4 survivors and the AP reporters documents
- 23 Academic publications
  - in fields of history, law, media studies, Asian studies, military, etc.
    - Journal articles, thesis and dissertations, chapters of books
- 55 Journalistic publications
  - news and magazine articles in US, UK, and Korea
- 1 government report
- 1 web package

Text-Analysis for NGR Collection

- All text, including captions, citations, footnotes
- Excludes images, audio, and multimedia
- Only English materials analyzed
  - TAPoR Keywords Finder does not support other languages

Problem

- “No” in No Gun Ri
  - Stopword, not counted: “Gun,” “Ri,” “Gun Ri” appeared as keywords
  - The chances that the term “No Gun Ri” is used for searching is assumed to be low.

Findings: Text Frequency

- Taxonomy Creation
  - Top 20 keywords
  - Top 10 word pairs
  - Top 10 word triplets

- Descriptive Data Categories
  - Recommended Keywords and Phrases
    - 175 terms: 143 words after eliminating repetitive terms (refugee and refugees) and meaningless words (pg, mr).

Top 20 keywords	Top 10 word pairs	Top 10 word triplets
Korean	Gun Ri	1st Cavalry Division
Ri	Korean war	Gun Ri Massacre
Gun	South Korea	Gun Ri incident
War	South Korean	7th Cavalry regiment
Korea	North Korean	Gun Ri researchers
South	Review team	Gun Ri research
1950	Jul-50	Double railroad overpass
Refugees	1st Cavalry	Gun Ri Review
Army	7th Cavalry	World war II
Soldiers	Cavalry division	25th Infantry division
Military		
American		
Civilians		
July		
Team		
North		
Cavalry		
Archival		
States		
Review		

Keywords by Type: Top 10 Word Pairs

Archival	Academic	Journalistic	Government	Web
Gun Ri	Gun Ri	Gun Ri	Gun Ri	Gun Ri
7th Cavalry	Korean War	South Korean	Review Team	South Korean
Railroad overpass	South Korea	Korean War	Jul-50	1st Cavalry
Double railroad	North Korean	North Korean	1st Cavalry	South Korea
Cavalry regiment	South Korean	South Korea	Cavalry Division	Cavalry Division
2nd battalion	Archival materials	American soldiers	7th Cavalry	Korean War
South Korean	North Korea	Korean civilians	Air Force	North Korean
Korean Report	Archival documents	7th Cavalry	2nd Battalion	Air Force
North Korean	Anti-Americanism	Jul-50	Aug-50	Ex GIs
Cav 590	Jul-50	Air Force	Eighth Army	Korean Refugees

Red: specific terms Blue: generic terms

No Gun Ri Taxonomy (from all word combinations)

General Background	NGR History	NGR Research/ controversy	Politics/ Diplomatic
• War • Army • Soldiers • Military • Cavalry • Division • Koreans • Korean War • World War II • Air Force • American Soldiers • Eighth Army	• 1950 • South Korea • Refugees • Civilians • 1st Cavalry Division • 7th Regiment • Railroad • Railroad bridge • Railroad overpass • Double railroad overpass • Order • 25th Infantry Division • North Korean soldiers • Im Gae Ri • Joo Gok Ri • Fighter Bomber Squadron	• Review • Research • Report • Law • Researchers • AP • Daily • Entry • Veterans • Author • Comment • Archival Materials • Review Team • Korean Report • No Gun Ri Review • Korean Witness Statements • Periodic Intelligence Report	• Anti-Americanism • International Humanitarian law • Customary International Law • South Korean Government

Data Categories
(identified from recommended keywords/phrases)

- People
  - Organization (1st Cavalry Division); group of persons (veteran, refugees); occupation; nationality (South Korean, American)
- Place
  - Geographic name; landmark (railroad bridge)
- Time (1950, World War II)
- Activities
  - Functions (evidence); process and technique (research, analysis, operations)
- Topic (law, anti-Americanism)
- Genre
  - Resource type (documents, war diary, articles, reports, imagery); media/format (film); nature (journalistic)
- Object (railroad, tunnel)
- Event (Korean War, Massacre)

- Proper names
  - Personal names(Daily), geographical names (Joo Gok Ri), event names, titles (NY times), Organization names (AP, Eighth Army)

Discussions

- Simple keyword extraction can be a useful tool
  - Inexpensive method for hard data
  - Relatively effective for creating taxonomy and analyzing properties of contents for data categories
- Different results for different types of texts
  - Archival documents vs. academic publications: specific vs. generic keywords
- Amount of text matters
  - The keyword extraction based on frequency
  - The amount of text in archival documents vs. that of academic publications
- Text Analysis can be done by type, then results be aggregated

Thank you.

Donghee Sinn
(dsinn@albany.edu)