The NTCIR-18 **SUSHI** Pilot Task: **S**earching **U**nseen **S**ources for **H**istorical **I**nformation

Tokinori Suzuki,[1] Douglas W. Oard,[2] Emi Ishita,[1] Yoichi Tomiura[1]

[1]Kyushu University, [2]University of Maryland

ntcir-sushi@googlegroup.com

Kickoff Meeting: March 29, 2024

The SUSHI Pilot Task

- Goal:
  - Find undigitized documents in archival repositories

- Subtasks:
  - A: Container Ranking
  - B: Archival Reference Detection
    - B1: Citation classification
    - B2: Reference boundary detection

Archival containers in the US National Archives

* https://bit.ly/3tV3eNd

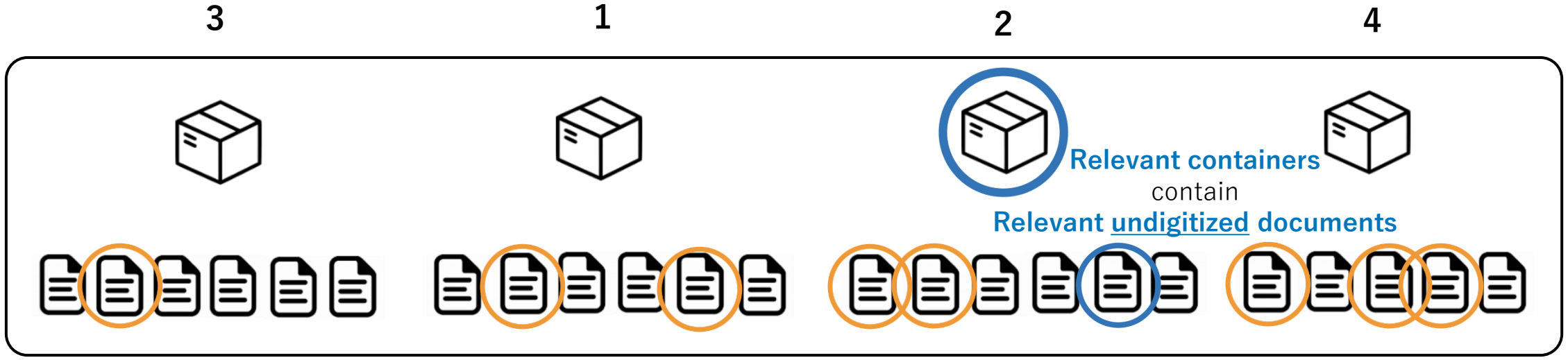
Subtask A: Container Ranking

Container Ranking System

Ranked Containers

Evaluation measures are computed from a system's **ranking** of containers

Average Precision: 0.5

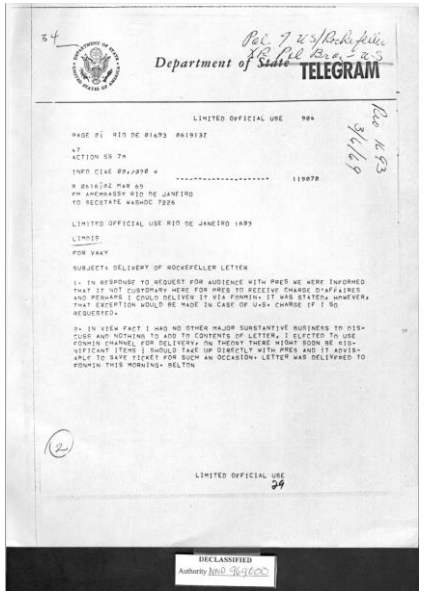


Previously digitized documents can be used to learn a model of container content

Container Ranking: Document Collection

Such dense digitization is rare!

- ~23,000 fully digitized documents from the US National Archives
- Each was originally stored in one of ~100 boxes
- A small predefined sample is defined for use by teams to train systems
- Relevance judgments are made using digitized documents **not** in the sample


Department of State TELEGRAM

LIMITED OFFICIAL USE

DECLASSIFIED

Example document

```
Department o
PAGE 01
47
ACTION SS 7*
INFO CIAE 00,/070 W RIO DE 01693 06I913Z LIMITED OFFICIAL USE
I I 9070
R 06I6T0Z MAR 69
FM AMEMPASSY RIO HE JANFIRO
TO SECSTATE WASHDC 7226
LIMITED OFFICIAL USE RIO DE JANEIRO t693
L I MO IS
FOR VAKY
SUBJECTi DELIVERY OF ROCKEFELLER LETTER
I, IN RESPONSE TO REQUEST FOR AUDIENCE WITH PRES WE WERE INFORMED
THAT IT NOT CUSTOMARY HERE FOR PRES TO PECEIVF CHARGE D'AFFAIRES
AND PERHAPS I COULD DELIVER IT VIA FONMIN. IT WAS STATED* HOWEVER.
THAT EXCEPTION WOULD BE MADE IN CASE OF U.S. CHARGE IF I So
REQUESTED.
p. IN VIEW FACT I HAD NO OTHER MAJOR SUBSTANTIVE BUSINESS TO DIS
CUSS AND NOTHING T0 ADD TO CONTENTS OF LETTER, I ELFCTED Tn USE
FONMIN CHANNEL FOR DELIVERY* ON THEORY THERE MIGHT SOON BE SIG
NIFICANT ITEMS i SHOULD TA<E UP DIRECTLY WITH PRES AND IT ADVIS
ABLE TO SAVr TICKET FOR SUCH AN OCCASION. LETTER WAS DELIVFPED TO
FONMIN THIS MORNING* BELTON
LIMITED OFFICIAL USE
```

Uncorrected OCR text

Metadata Item	Example
Title	Delivery of Rockefeller Letter
Institution	United States. National Archives and Records Administration
Folder name	POL BRAZ 1/1/69
Date	1969-03-6
Genre	Telegram
Topic	Rockefeller, Fonmin
Related entity	State (addressee), Rio de Janeiro (Brazil)

Metadata

~25 Topics for Container Ranking Pilot Task

Title: Meetings with Brazil's President

Description: Meetings between US diplomats and the President of Brazil

Narrative: Documents discussing in person meetings between representatives of the United States government and the President of Brazil are relevant. These may be past, present, or future meetings, and documents discussing a possible meetings would be relevant regardless of whether the meeting actually occurs. Meetings conducted by telephone, or meetings that involve only Brazilian government representatives other than the President would not be relevant.

Container Ranking: Evaluation Measures

- Mean Container-nDCG [Main measure]
  - Each relevant document is assigned a relevance grade
    - Highly relevant [score: 3]
    - Somewhat relevant [score: 1]
    - No relevant [score: 0]
  - Raw score (CRel) for a container is sum of scores for its undigitized documents
    - Limited to the 10 most relevant documents per container

$$Container - DCG = \sum_{i=1}^{k} \frac{2^{CRel_i} - 1}{\log_2 i + 1}$$

- Mean Success@1
  - Count as a success if the most relevant container is at first of system's rank

Subtask B: Archival Reference Detection

Footnotes, Endnotes, References ("Citations")

8. Slayton DK. Gemini Extravehicular Activity Program: Mission National Archives at Fort Worth, TX 1964 (30 January), RG 255, E.75, Box 383, file 2: Extravehicular Planning, Crew Procedures and Training.

Wheeler, D., and R. Garcıa-Herrera, 2008: Ships' logbooks in climatological research: Reflections and prospects. Ann. New York Acad. … Several archive sources have been used in the preparation of this paper, including the following: Log-book of HMS Richmond. The U.K. National Archives. ADM/51/3949

Archival Reference Detection System

B1) Citation classification

Archival Reference

8. Slayton DK. Gemini Extravehicular Activity Program: Mission National Archives at Fort Worth, TX 1964 (30 January), RG 255, E.75, Box 383, file 2: Extravehicular Planning, Crew Procedures and Training.

Not an Archival Reference

2. Thomas P. Hughes. The evolution of large technological systems. In The social construction of technological systems. Bijker, W., Hughes, T. P., and Pinch, T. Eds. Cambridge, MA: MIT Press, 1987:50-82.

B2) Reference boundary detection

Part of Archival Reference

Ketting, De Amsterdamse chirurgijns, unpu… masters' thesis, 39, 41, 44. After 1760, the pupils take abou… eleven to twelve years to graduate their master's examination. **NA, VOC 14.204, 14.209, 14.217**. **Municipal archive of Schiedam: chirurgijns gilde OAA 3077** Date of birth was retrieved by J. Leenders in the municipal archive of Rotterdam, 1-6-1742.

Evaluation of Archival Reference Detection

- Citations:
  - ~1,000,000 citations automatically extracted from scholarly publications
- Ground truth judgement created by trained annotators
- Evaluation Measures:
  - Citation classification
    - Precision
    - Recall
    - F_1 [Main measure]
  - Reference boundary detection
    - Jaccard coefficient, computed on characters [Main measure]

How To Participate

- Sign up with NTCIR!
- Join the mailing list
  - Send an email to ntcir-sushi@googlegroup.com
- Download the collections and baseline systems
  - Available May 1, 2024
- Try something!
  - You can participate in any or all of the subtasks
- Submit to the Dry Run for early feedback
  - Due August 31, 2024
- Submit your official runs
  - Due January 31, 2025
- Come to Tokyo for NTCIR-18!
  - June 10-13, 2025 (online participation will be possible)