ACM/IEEE Joint Conference on Digital Library (JCDL) 2011 Report

Kazunari Sugiyama (杉山一成)

National University of Singapore

Information Access Symposium, SIG-IFAT, IPSJ

14th September, 2011
Outline of JCDL11

• Venue
  – Ottawa University, Ottawa, Ontario, Canada
Outline of JCDL11

• Acceptance rate
  – 23.5% [57 / 243]
    • Submitted papers: 243 papers
    • Accepted papers: 57 papers
      – 28 full papers and 29 short papers

• Future JCDL
  – 2012: Washington DC, US
  – 2013: Indianapolis, US
  – 2014: ???
    • Candidates: Argentina, Italy, UK
    • Joint with TPDL (Theory and Practice of Digital Libraries)
      – “ECDL” -> “TPDL” (since 2011)
Research Topics in JCDL

- Collaborative and participatory information environments
- Cyberinfrastructure architectures, applications, and deployments
- Data mining/extraction of structure from networked information
- Digital library and Web Science curriculum development
- Distributed information systems
- Evaluation of online information environments
- Impact and evaluation of digital libraries and information in education
- Information and knowledge systems
- Information policy and copyright law
- Information visualization
- Interfaces to information for novices and experts
- Personal digital information management
- Retrieval and browsing
- Scientific data curation, citation and scholarly publication
- Social networks, virtual organizations and networked information
- Social-technical perspectives of digital information
- Studies of human factors in networked information
- Systems, algorithms, and models for data preservation
- Theoretical models of information interaction and organization
- User behavior and modeling
- Visualization of large-scale information environments

(Cited from “http://jcdl2011.org/ExtendedCallForPapers”)
Presented Research Topics

• **Content analysis** (18 papers)
  – information extraction, plagiarism detection, topic coherence, etc.
• Education
• Information policy, rights
• Infrastructure
• Interfaces
• **Metadata, Annotation** (8 papers)
• Mobile applications
• Preservation, Archive
• **User’s information needs** (8 papers)
• Visualization
• WWW
Presented Research Topics

• Content analysis (18 papers)
  – information extraction, plagiarism detection, topic coherence, etc.

• Education

• Infrastructure

• Interfaces

• Metadata, Annotation (8 papers)

• Mobile applications

• Preservation

• User’s information needs (8 papers)

• Visualization

• WWW

“Measuring Historical Word Sense Variation”

“SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination”
Content Analysis

“Measuring Historical Word Sense Variation”

David Bamman (Tufts University, US),
Gregory Crane (Tufts University, US)
Outline

• Automatically classify Latin word senses

Track the historical variation of these senses more than 2,000 years span

  – Example: “radical”
    “Oxford English Dictionary”
    (1) Political meaning (1783 - )
    “advocating thorough or far-reaching political or social reform”
    (2) Slang term
    “Excellent, fantastic”

• Dataset
  – 83,892 words from the aligned parallel corpus
  – Manually annotated sample of 525 words
Proposed Approach

• Constructing Latin corpus
• Inducing Latin senses in English
• Word sense disambiguation
• Tracking sense variation over 2,000 years
Constructing Latin Corpus

• Collect Latin books from Internet Archive (http://www.archive.org/)
  – 7,055 books
  – 389 million words

(Cited from D. Bamman and G. Crane: “Measuring Historical Word Sense Variation,” JCDL2011)
Inducing Latin senses

129 translation book pairs

40,323 sentence pairs

MGIZA GIZA++

Alignments at the level of individual words

Clean alignments for 504,857 words

Aggregate English Translations for each Latin Lemma

3,412 Latin-English translation pairs
(Training set for word sense disambiguation)
Word Sense Disambiguation

• Classifiers
  – Language model classifier
    • Trained on Uni-gram, bi-gram, 5-gram, 6-gram
  – Naïve Bayes
    • Trained on uni-gram
  – TF-IDF
    • Uni-gram
  – K-nearest neighbor

• Features
  – 20 words around each target word

• Baseline
  – Simply select the most frequent sense from the lexicon
Tracking sense variation over 2,000 years

• Apply 6-gram language model classifier to 389 million words

Latin word “oratio” (“prayer,” “speech” in English)

(Cited from D. Bamman and G. Crane: “Measuring Historical Word Sense Variation,” JCDL2011)
Metadata, Annotation

• “SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination”

Robert Sanderson (Los Alamos National Laboratory), Benjamin Albritton (Stanford University), Rafael Schwemmer (e-codices), Herbert Van de Sompel (Los Alamos National Laboratory)
Outline

• Model interrelationships of images, texts, and other resources for medieval manuscripts or culturally important hand written documents
• Smooth interoperability between manuscripts
• Construct links to DL resources
Proposed Approach

• Construct data model of medieval manuscripts to satisfy the following requirements:
  – Images and their relationship with the physical object
  – Texts and their relationship with the images
  – Sequencing of the images and texts
  – Rendering of images and texts
Images and their Relationship with the Physical Object

Important to have mappings from the parts of the image to another page

Images and their Relationship with the Physical Object

Related images

Collected in 15th century

Collected in 4th century

Assembled in 15th century

Texts and their Relationship with the Images

In page 10, “Figure 3 illustrates ...”

In page 11, Figure 3
Sequencing of the Images and Texts

- When users are interested in particular contents, systems must provide textual sections correctly.
Rendering of Images and Texts

- Collaborative environment for manuscript description between different communities and individuals (like social tagging services)

Canvas:
A kind of medium that links text and image

(2) Sequence and Range of Canvases

Multiple canvas to be presented to users

- Model textual or physical boundaries
- “Rng” is a part of “Seq”

SharedCanvas

(3) Group of Sequences

TextOrdr:
Organize small annotations to satisfy reading order requirement

ImgList:
Organize image annotations

ZoneList:
Maintain all of their associated annotations by delimiting Text and Images

Manifest:
Manage multiple sequences for a single manuscript

(See the paper about instantiation of SharedCanvas)
Doctoral Consortium

• 10 PhD students
  – Germany (1), Norway (2), Portugal (1), UK (1), US (5)

• Topics
  – Metadata
  – IR and NLP in Semi-structured data
  – Applications
  – Information Discovery

• Travel support for students