ARC Day 2020

ARC所蔵資料データベースの評価データセット作成プロジェクト

前田　亮（情報理工学部）
LI Kangying（情報理工学研究科 D2）
WANG Jiayun（情報理工学研究科 D3）
SONG Yuting（情報理工学部 特任助教）
BATJARGAL Biligsaikhan（衣笠総合研究機構 専門研究員）
佐藤 英男（情報理工学研究科 M2）

1

研究の概要

- ARC所蔵の日本文化資源データベースのオープンデータ化および容易な多言語情報アクセスの実現に向けて研究を行っている

ARC浮世絵ポータルデータベース　　ARC古典籍ポータルデータベース

2

2

現状の課題

- 情報系の研究では，提案手法の評価が重要
- 浮世絵や古典籍のデータについて，評価用のデータセットが存在しない
  - 現時点では，研究室内で小規模なデータセットを作成して使用
  - 小規模なデータセットでは，信頼性に欠ける
  - 研究室内で大規模なデータセットの構築は難しい
- クラウドソーシング，ゲーミフィケーションを用いた評価用データセットの作成手法を検討中

3

3

紹介する研究

1. **Multimodal Representation learning for Ukiyo-e records retrieval and analysis**
   LI Kangying（情報理工学研究科 D2）
2. **ARC所蔵浮世絵資料のための作品推薦・同定手法**
   WANG Jiayun（情報理工学研究科 D3）
3. **Finding Identical Ukiyo-e Prints across Databases in Japanese, English and Dutch**
   SONG Yuting（情報理工学部 特任助教）

4

4

Exploring information of Ukiyo-e records by using Multi-source Data

Multimodal Representation Learning for Ukiyo-e Records Retrieval and Analysis

LI Kangying

5

Outline

- **Motivation**
- **Some progress：**
  Pre-processing of titles
  - **Compare some morphological analyzers**
  - **Build some user dictionaries**
  - **Token embedding from BERT's pre-trained model**
  - **Visualization of the results**

6

6

Motivation

- If we only use a single database to search some items：

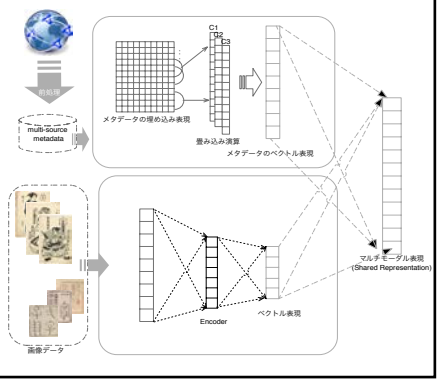
incomplete text information

metadata without image information

7

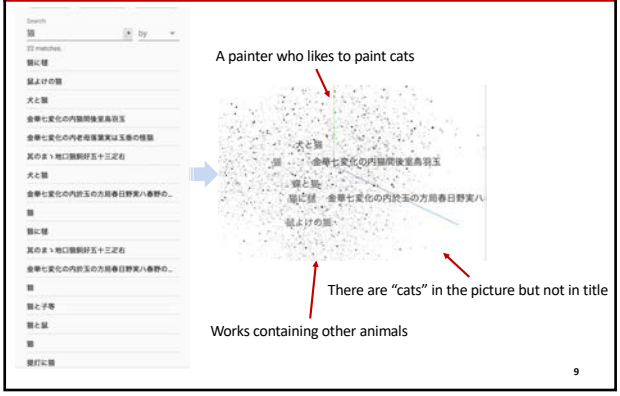
Motivation

- Now there are many multilingual open source databases, what we can do is...
  - ARC database
  - WikiData
  - Wikipedia
  - WikiArt
  - ...



8

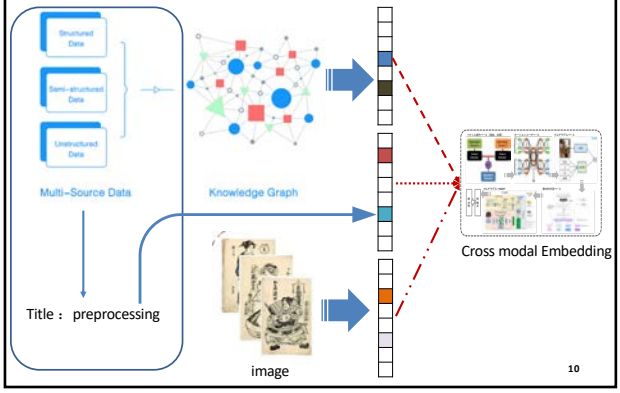
If we can build a system that...



9

9

Can we build it by ....?



10

10

Outline

- Motivation
- Some progress：

  Pre-processing of titles
  - Compare some morphological analyzers
  - Build some user dictionaries
  - Token embedding from BERT's pre-trained model
  - Visualization of the results

11

11

Pre-processing of titles: morphological analyzer (proper noun extraction)

- Input sentence:この時期は役者絵や合巻の挿絵などを描いていたが、あまり人気が出ず作品も僅かであった。また、勝川春亭にも学んでおり、さらに葛飾北斎の影響も受け、後に3代堤等琳に学んで、雪谷とも号した

```
この    連体詞,*,*,*,*,この,コノ,コノ
時期    名詞,副詞可能,*,*,*,*,時期,ジキ,ジキ
は      助詞,係助詞,*,*,*,*,は,ハ,ワ
役者    名詞,一般,*,*,*,*,役者,ヤクシャ,ヤクシャ
絵      名詞,接尾,一般,*,*,*,絵,エ,エ
や      助詞,並立助詞,*,*,*,*,や,ヤ,ヤ
合      名詞,一般,*,*,*,*,合,ゴウ,ゴー
巻      名詞,一般,*,*,*,*,巻,マキ,マキ
の      助詞,連体化,*,*,*,*,の,ノ,ノ
...
勝川    名詞,固有名詞,地域,一般,*,*,勝川,カチガワ,カチガワ
春      名詞,一般,*,*,*,*,春,ハル,ハル
亭      名詞,接尾,一般,*,*,*,亭,テイ,テイ
に      助詞,格助詞,一般,*,*,*,に,ニ,ニ
も      助詞,係助詞,*,*,*,*,も,モ,モ
```

By: mecab-unidic-neologd

12

12

Pre-processing of titles: morphological analyzer

- Input sentence:この時期は役者絵や合巻の挿絵などを描いていたが、あまり人気が出ず作品も僅かであった。また、勝川春亭にも学んでおり、さらに蔦飾北斎の影響も受け、後に3代堤等琳に学んで、雪谷とも号した

この	コノ	この	連体詞
時期	ジキ	時期	名詞-副詞可能
は	ハ	は	助詞-係助詞
役者	ヤクシャ	役者	名詞-一般
絵	エ	絵	名詞-接尾-一般
や	ヤ	や	助詞-並立助詞
合	ゴウ	合	名詞-一般
巻	マキ	巻	名詞-一般
の	ノ	の	助詞-連体化
挿絵	サシエ	挿絵	名詞-一般
…			
勝川	カチガワ	勝川	名詞-固有名詞-地域-一般
春	ハル	春	名詞-一般
亭	テイ	亭	名詞-接尾-一般
に	ニ	に	助詞-格助詞-一般

By:茶釜Ochasen

13

13

Outline

- Motivation
- A book review
- Some progress：

  Pre-processing of titles
  - Compare some morphological analyzers
  - Build some user dictionaries
  - Token embedding from BERT's pre-trained model
  - Visualization of the results

14

14

User dictionary

- Items, Events, Person's names and Stories Contained in Ukiyo-e：

人物画、風景画、伝説と神話、広告等の情報…

reference： https://www.kumon-ukiyoe.jp/history.html

15

15

浮世絵用語辞典&日本人名辞書 (JBDB)

鵬春水	らいしゅんすい	Rai Shunsui
阿波屋仁左衛門	あわやにざえもん	Awaya Nizaemon
永富独嘯庵	ながとみどくしょうあん	Nagatomi Dokushōan
福原五岳	ふくはらごがく	Fukuhara Gogaku
泉屋庄助	いずみやしょうすけ	Izumiya Shōsuke
しほ屋仁左衛門	しほやにざえもん	Shioya Nizaemon
はりまや権兵衛	はりまやごんべえ	Harimaya Gonbe'e
林幸斎	はやしこうさい	Hayashi Kōzai
留守退蔵	るすたいぞう	Rusu Taizō
林景介	はやしけいすけ	Hayashi Keisuke
愛宕里坊	あたごりぼう	Atagori Bō
芥川義軒	あくたがわよしけん	Akutagawa Yōken
高志喜治	たかしようこう	Takashi Yōkō
丹後屋藤兵衛	たんごやとうべえ	Tangoya Tōbe'e
超倫	ちょうりん	Chōrin
開泉書市	かいせんきいち	Kaisen Kiichi
青石大順	あおいしだいじゅん	Aoishi Daijun
葛子琴	かつしきん	Katsu Shikin
木村喜殿堂	きむらけんかどう	Kimura Kenkadō
上林吉次郎	かんばやしきちじろう	Kanbayashi Kichijirō
竜草廬	りゅうそうろ	Ryū Sōro
無染和尚	むせんおしょう	Musen Oshō

16

16

歴史地名データ: 139,454 words

- For knowledge extraction: 地名ー別名ー属性ー上位地名ー上位地名属性ー上位地名別名
- For user dictionary: 地名＋別名＋上位地名＋上位地名属性 (.dic)

https://www.nihu.jp/ja/publication/source_map

17

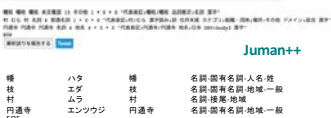
17

Morphological analysis

- 「幡枝村円通寺」

幡	名詞,固有名詞,人名,姓,*,*,幡,ハタ,ハタ
枝	名詞,固有名詞,地域,一般,*,*,枝,エダ,エダ
村	名詞,接尾,地域,*,*,*,村,ムラ,ムラ
円通寺	名詞,固有名詞,地域,一般,*,*,円通寺,エンツウジ,エンツージ

mecab-unidic-neologd



Juman++

幡	ハタ	幡	名詞-固有名詞-人名-姓
枝	エダ	枝	名詞-固有名詞-地域-一般
村	ムラ	村	名詞-接尾-地域
円通寺	エンツウジ	円通寺	名詞-固有名詞-地域-一般
EOS			

chasen

幡枝村円通寺
幡枝　名詞,固有名詞,地域,一般,*,*,幡枝,,
村　名詞,接尾,地域,*,*,*,村,ムラ,ムラ
円通寺　名詞,固有名詞,地域,一般,*,*,円通寺,,
EOS

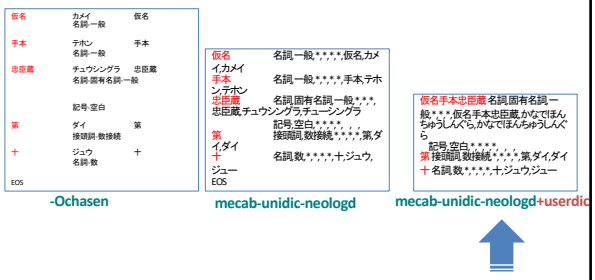
mecab-unidic-neologd+userdic

18

18

Morphological analysis

- INPUT title：仮名手本忠臣蔵　第十



19

19

Outline

- Motivation
- A book review
- Some progress：

  Pre-processing of titles
  - Compare some morphological analyzers
  - Build some user dictionaries
  - Token embedding from BERT's pre-trained model
  - Visualization of the results

20

20

Token embedding from BERT's pre-trained model

MODEL (Google): BERT-Base, Multilingual Cased (New, recommended): 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters

➢ **First step：clean the data ,Because it contains a lot of symbols**
e.g. '〔〕★●･[]「」？＃･「」『』()()<>[]{}'　「阿羅々仙人」「☆量汝弥」「耶倫陀羅姫」「羅☆羅太子」「車匿舎人」

➢ **Second step:**

Grouping the experimental subject：

小倉擬百人一首
木津勘助
仮名手本忠臣蔵
古今忠孝伝
木津関兵衛　中村駒之助
市川団十郎
八犬伝犬乃そうしの内
黒谷金戒光明寺
西山月輪寺
幡枝村円通寺

['小倉','擬百人一首'],
['木津','勘助'],
['仮名手本忠臣蔵'],
['古今','忠孝','伝'],
['木津','関兵衛','中村','駒','之','助'],
['市川団十郎'],
['八犬伝','犬','乃','そう','し','の','内'],
['黒谷','金戒光明寺'],
['西山','月輪寺'],
['幡枝','村','円通寺']

['小','倉','擬','百','人','一','首'],
['木','津','勘','助'],
['仮','名','手','本','忠','臣','蔵'],
['古','今','忠','孝','伝'],
['木','津','関','兵','衛','中','村','駒','之','助'],
['市','川','団','十','郎'],
['八','犬','伝','犬','乃','そ','う','し','の','内'],
['黒','谷','金','戒','光','明','寺'],
['西','山','月','輪','寺'],
['幡','枝','村','円','通','寺']

Sentence　Words-level　Characters-level

21

21

Compareing the performance

Query: 幡枝村円通寺　Database: [小倉擬百人一首,木津勘助,仮名手本忠臣蔵,古今忠孝伝,木津関兵衛,中村駒之助,市川団十郎,八犬伝犬乃そうしの内,黒谷金戒光明寺,西山月輪寺]

Sentence　Words-level　Characters-level

22

22

compare the performance

Query: 木津勘助　Database: [小倉擬百人一首, 仮名手本忠臣蔵,古今忠孝伝,木津関兵衛,中村駒之助,市川団十郎,八犬伝犬乃そうしの内,黒谷金戒光明寺,西山月輪寺,幡枝村円通寺]

Sentence　Words-level　Characters-level

23

23

Outline

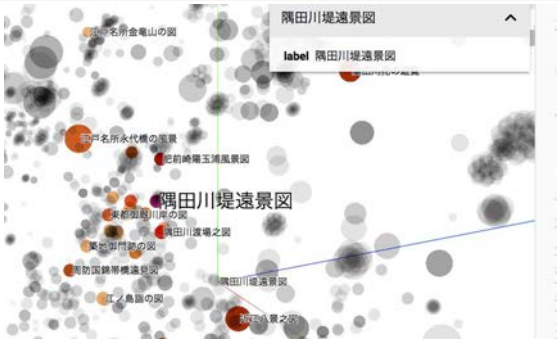
- Motivation
- A book review
- Some progress：

  Pre-processing of titles
  - Compare some morphological analyzers
  - Build some user dictionaries
  - Token embedding from BERT's pre-trained model
  - Visualization of the results

24

24

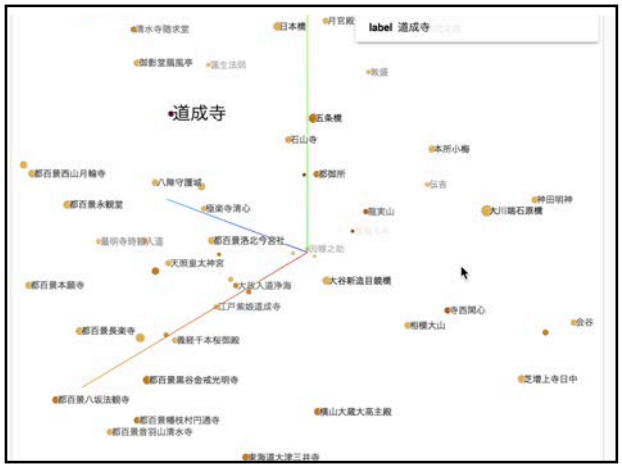
Visualization



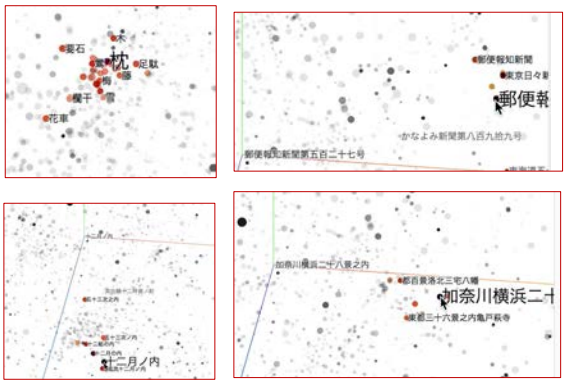
2,0000 title vecs:1:748

25

25



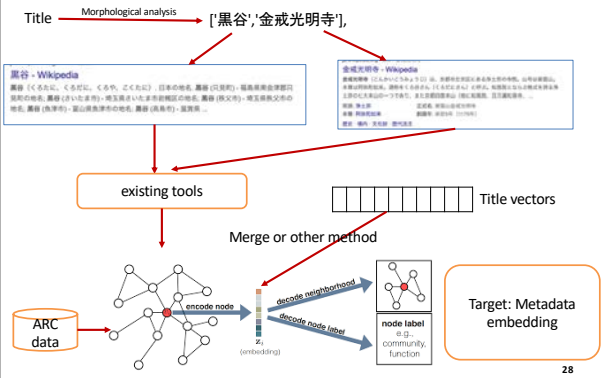
26



27

27

Plans



28

28

Plans

- Vector data analysis
- Extend user dictionary
- About the new research of seals retrieval: in the experimental stage..

29

29

ARC所蔵浮世絵資料のための作品推薦・同定手法

WANG Jiayun

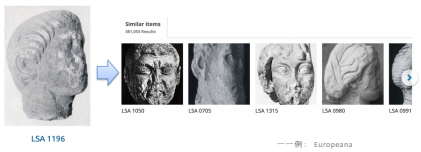
RITSUMEIKAN

DIGITAL LIBRARY LAB

30

デジタルアーカイブの推薦システムの現状

1. 個人的な興味で文化遺産のデジタルアーカイブを閲覧・探索するユーザーが増えている[1]が、既存の推薦システムでは、同じタグ・カテゴリー以外のユーザが興味のあるアイテムを見つけるのは困難である.



一一例： Europeana

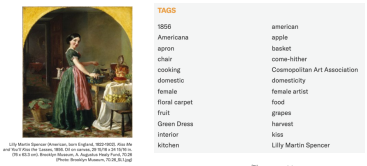
[1] Clough, Paul, et al. "Europeana: What users search for and why." *International Conference on Theory and Practice of Digital Libraries*. Springer, Cham, 2017.

31

31

デジタルアーカイブの推薦システムの現状

2. 一般的推薦システムとは異なり、タグが正確さや専門知識を欠いている場合、folksonomyベースの文化遺産の推薦システム[2]はユーザが誤解する可能性がある.



一一例： Brooklyn Museum

[2] Semeraro, Giovanni, et al. "A folksonomy-based recommender system for personalized access to digital artworks." *Journal on Computing and Cultural Heritage (JOCCH)* 5.3 (2012): 11.

32

32

研究目的

- 画像とメタデータ（現在はメタデータのみ）を用い、デジタルアーカイブのための推薦アルゴリズムを開発する。
- アイテムの基本的情報だけではなく、ユーザの閲覧行動も考慮した推薦結果を算出する。
- Folksonomyベースの方法ではなく、専門家が付けたメタデータを用いて正確に推薦を行う。

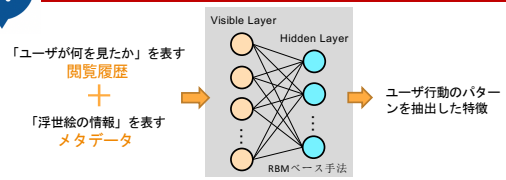
問題点

- 前処理を行った後、ほとんどのユーザが一枚の浮世絵しか閲覧していなかったため、ユーザごとに興味のあるアイテムを推測することは困難である。
- メタデータには空いている所があるため、疎なデータは「推薦結果が見つからない」、「多数のアイテムは区別できない」などの問題を引き起こす。

33

33

提案手法（１）：RBM*ベース手法



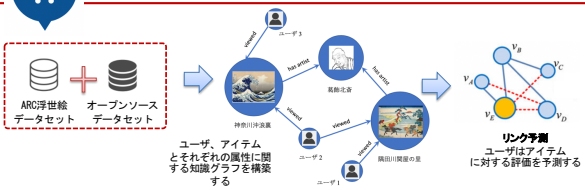
- 原理：入力データの分布を学習することにより、入力データの特徴を抽出する。
- メリット：一人一人の行動より、ユーザ全体の行動を分析できる。
- デメリット：よく閲覧された絵・メタデータが多い絵・メタデータがよく出現する絵（「役者絵」のメタデータを持つ絵）が推薦されやすい。

* RBM: restricted Boltzmann machine

34

34

提案手法（２）：KG*ベース手法



- 原理：浮世絵の情報とその関係をユーザの閲覧行動として扱う。
- メリット：浮世絵の情報を重視するため、RBMの一部のデメリットを回避できる。
- デメリット：グラフの構築が容易ではない。（データノイズ、メタデータの欠如、ユーザごとの平均閲覧数が少ないなどが原因）

* KG: knowledge graph

35

35

KGベース手法による作品同定の考案

- 研究目的：浮世絵の研究者は同版・異版作品を比較することが多いが、ARC浮世絵DBの中で、それを網羅的に抽出するアルゴリズムは知る限りではない。
- 原理：浮世絵アイテムの画像・メタデータ特徴をグラフ構造に入れ、アイテムの近隣ノードの情報を用いて同版・異版の浮世絵を推測する。同版の浮世絵は画像の類似度が高いので、画像の特徴を重視することにし、異版の浮世絵は画像の類似度はより低いので、メタデータを重視することが必要になる。

36

36

Finding Identical Ukiyo-e Prints across Databases in Japanese, English and Dutch

SONG Yuting

RITSUMEIKAN

DIGITAL LIBRARY LAB

37

37

Ukiyo-e

- Japanese traditional woodblock print
- One of the popular arts of the Edo period (1603–1868)
- Many libraries, museums and galleries in Western countries have digitalized ukiyo-e woodblock prints

Image source:The Metropolitan Museum of Art, https://www.metmuseum.org/art/collection/search/45434

Japan	• Edo-Tokyo Museum • Ritsumeikan University • …
United States	• Metropolitan Museum of Art • Museum of Fine Arts • …
United Kingdom	• British Museum • Ashmolean Museum • …
Netherlands	• Rijksmuseum • …

38

38

Identical ukiyo-e prints in different languages

	Metadata		Language	
	Title	Artist		
江戸東京博物館 (Japan)	 雪月花 淀川	葛飾北斎		Identical ukiyo-e prints
	 凱風快晴	葛飾北斎		
Metropolitan Museum of Art (United States)	 Moonlight on the Yodo River, from the series Snow, Moon, and Flowers	Katsushika Hokusai	English	
	 Morning Mist at Mishima	Utagawa Hiroshige (I)		Identical ukiyo-e prints
Rijksmuseum (Netherlands)	 Helder weer en een zuidelijke wind	Katsushika Hokusai		
	 Mishima in ochtendmist	Hiroshige (I), Utagawa		

4

39

Motivation

- To find identical ukiyo-e prints in different languages
  - by comparing metadata

作品名	作者
富嶽三十六景 神奈川沖浪裏	葛飾北斎
富嶽三十六景 深川万年橋下	葛飾北斎
日本橋 朝之景	歌川広重(初代)
雪月花 隅田	葛飾北斎

Title	Artist
Under the Wave off Kanagawa, from the series Thirty-six Views of Mount Fuji	Katsushika Hokusai
Snow on the Sumida River, from the series, Snow, Moon, and Flowers	Katsushika Hokusai
Morning View of Nihonbashi	Utagawa Hiroshige

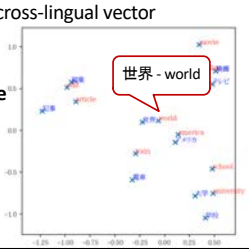
- Challenges
  - Metadata values are in different languages
    - Language barriers: similarity measures cannot be employed directly

40

40

Our proposed method

- To calculate metadata similarity in different languages by using bilingual word embeddings
  - Bilingual word embeddings
    - Word embeddings: dense, low-dimensional and real-valued vectors for representing words
    - Bilingual word embeddings: cross-lingual vector representations of words
    - **Advantage: similar words in different languages can have similar vector representations**



41

41

Metadata similarity calculation (1/2)

- We measure the metadata similarities through word-to-word matching between metadata in different languages
  - Notation
    - Let b be the vector space of Japanese-English bilingual word embeddings
    - Each word w_m^J in Japanese metadata is represented as $b(w_m^J)$ by using the bilingual word embeddings.
    - Similarly, each word w_n^E in English metadata can be represented as $b(w_n^E)$
  - Word similarity
    - $roma(w_m^J)$ is the romanization of the Japanese word w_m^J

$$Similarity(w_m^J, w_n^E) = \max[cos\_sim\left(b(w_m^J), b(w_n^E)\right),\ cos\_sim\left(b(roma(w_m^J)), b(w_n^E)\right)]$$

The cosine similarity between w_m^J and w_n^E

The cosine similarity between the romanization of w_m^J and w_n^E

42

42

Metadata similarity calculation (2/2)

- **We measure the metadata similarities through word-to-word matching between metadata in different languages**
  - **Metadata similarity**
    - Each word w_m^J in the Japanese metadata is used to compare with all the words in English metadata
    - The maximum similarity score is used as the contribution of w_m^J to the metadata similarity

$$Metadata\ Similarity = \sum_{m=1}^{N^J} \max_{n \in \{1,2,\ldots,N^E\}} [Similarity(w_m^J, w_n^E)]$$

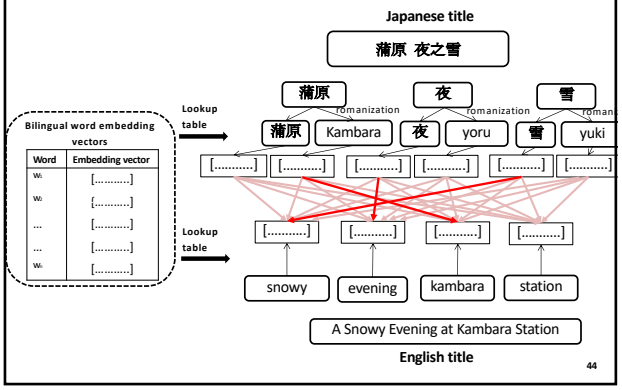
All the words in an English metadata

The word similarity between w_m^J and w_n^E

43

43

Example



44

44

Experiments

- **Finding identical metadata records of ukiyo-e prints between Japanese, English and Dutch databases**
- **Experimental data**
  - Ukiyo-e prints dataset
    - Metadata: title, artist name

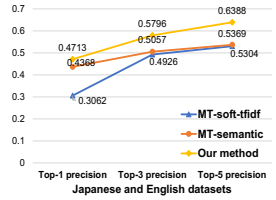

Language	Data source	Number of Records
Japanese	Edo-Tokyo Museum	203
English	Metropolitan Museum of Art	3,398
Japanese	Edo-Tokyo Museum	53
Dutch	Rijksmuseum	614

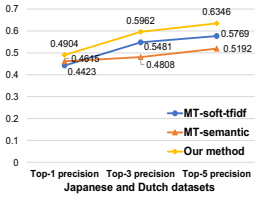
  - Each Japanese ukiyo-e metadata record has at least one corresponding ukiyo-e metadata records in the English or Dutch dataset

45

45

Experimental results





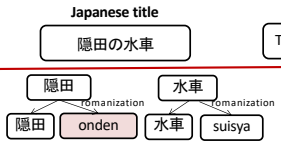
- **Compared with the methods that relies on machine translation(MT)**
  - **Soft-tfidf similarity (*MT-soft-tfidf*)**
    - A string-based similarity metric and showed the best performance in title matching against 20 other commonly used string-based similarity metrics
  - **Semantic matching (*MT-semantic*)**
    - This method is based on monolingual word embeddings

46

46

Discussion (1/2)

- **Example: bilingual word embedding is better than MT**



  - MT-based method
    - MT results of 「隠田の水車」
      - Hidden Water Wheel (Microsoft Translator)
      - Hidden Waterwheel (Google Translator)
      - water wheel in a hidden land (DeepL Translator)
      - The Water Wheel of Hidden Field (みらい翻訳)

47

47

Discussion (2/2)

- **Limitation of our approaches**
  - **The corresponding English title is inadequate translation**

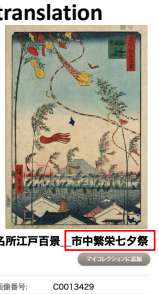


Image source: Tokyo National Museum, https://webarchives.tnm.jp/imgsearch/show/C0013429

The Tanabata Festival from the series One Hundred Famous...

Utagawa Hiroshige, 1857

Image source: The Metropolitan Museum of Art, https://www.metmuseum.org/art/collection/search/36541

48

48

Translation patterns of Japanese titles

1) **Transliteration:** all the words in an original Japanese title are converted to the other language one by one via transliteration
   - Example: Okabe; Utsu no Yama（岡部 宇津之山）
2) **Literal translation:** all the words in Japanese title are converted to their equivalents in the other language, which remains the content of original Japanese title as much as possible
   - Example: Surugadai in Edo（東都駿台）

     The Waterwheel at Onden (隠田の水車)
3) **Free translation:** The titles in the target language are created based on the content of Japanese objects without sticking to the original Japanese titles
   - Adding some information
     - Example: Snow on the Sumida River（隅田）
   - The Japanese title and corresponding English title have no meaning overlap

49

Translation patterns (cont')

1) **Transliteration: 64**
   - **Adequate transliteration:** 49
     - Example: Okabe; Utsu no Yama（岡部 宇津之山）
   - **Inadequate transliteration:** 15
     - Example: Yanagishima no Zu (柳嶋之図　橋本)
2) **Literal translation: 135**
   - **Adequate literation translation:** 119
     - Example: The Waterwheel at Onden (隠田の水車)
   - **Inadequate literation translation:** 16
     - Example: The Tanabata Festival (市中繁栄七夕祭)

50

Experiments

- **Datasets**

Language	Databases	Number of ukiyo-e prints
Japanese	Edo-Tokyo Museum	203
English	Metropolitan Museum of Art	3,398

- **Baseline methods**
  - **Transliteration + softtfidf (new)**
  - MT + softtfidf
  - MT + semantic
  - MT + mix
  - BiWE
- **Proposed method: BiWE+Roma**

51

Experimental setup

- **Transliteration + softtfidf**
  - **Mecab + IPADic**
  - **Hepburn romanization system**
  - **Softtfidf**
    - Secondary similarity function: edit distance
    - Similarity threshold: 0.9
- **MT + softfidf**
  - **Microsoft Translator Text API** (translation results on 2020/06/28)
    - ~~Statistical MT model (st)~~
    - ~~Neural network MT model (nn)~~
- **MT + semantic / MT+ mix**
  - **English word embeddings (Wikipedia articles + word2vec)**
- **BiWE / Wib**
  - **Japanese-English word embeddings**
    - Monolingual word embeddings (Wikipedia articles + word2vec)
    - 9000 Japanese-English word pairs

52

Experimental results (MAP)

Table 1: Results (MAP)

	All	Transliteration		Literal translation		Free translation
		Adequate	Inadequate	Adequate	Inadequate	
Transliteration	0.4555	**0.5818**	0.3434	0.4518	0.375	0.1723
MT+softtfidf	0.4949	0.3931	**0.4413**	0.5474	**0.3877**	0.3147
MT+semantic	0.4696	0.375	0.37	0.5431	0.3085	0.2887
MT+mix	0.4693	0.4114	0.3618	0.5488	0.3177	0.2045
Bi-WE	0.2698	0.1009	0.0097	0.353	0.1719	0.2483
Bi-WE+Roma	**0.5338**	0.5671	0.1351	**0.6046**	0.2867	**0.3288**

- **All datasets, adequate literal translation, free translation**
  - Proposed method is better than baseline methods
- **Adequate transliteration**
  - Transliteration+softtfidf performs best

53

Conclusion

- We presented a metadata similarity calculation method by using bilingual word embeddings
- We showed our method's ability to find identical metadata records of ukiyo-e prints across Japanese, English and Dutch databases

Future work

- To further improve the performance by using other metadata
- To apply our methods on other humanities databases

54