異言語の浮世絵データベースにおける描写的作品名に対応した同一作品の同定手法の提案

木村 泰典（立命館大学 情報理工学研究科）
Yuting Song（立命館大学 情報理工学研究科）
Biligsaikhan Batjargal（立命館大学 総合科学技術研究機構）
木村 文則（尾道市立大学 経済情報学部）
前田 亮（立命館大学 情報理工学部）

浮世絵は明治時代に多くの複製や異版が海外に散逸したが，現在は世界中の美術館や博物館のデータベースに公開されている．また，浮世絵は版画であるため，同一作品が多数存在しており，浮世絵のメタデータはデータベース毎に異なる言語や形式で表記されている．メタデータの言語や形式が異なる場合，同一の作品を発見するのに困難な場合がある．そこで，本論文ではn-gramを用いた文字列分割や複数の辞書・シソーラスを用いたマッチングのための訳語候補抽出を行うことで，異言語の浮世絵データベース間での同一作品の同定において大きな問題となる描写的な作品名に対応した手法を提案する．

Identifying the Same Artworks among Depictive Titles in Multiple Ukiyo-e Databases in Different Languages

Taisuke Kimura (Graduate School of Information Science and Engineering, Ritsumeikan University)
Yuting Song (Graduate School of Information Science and Engineering, Ritsumeikan University)
Biligsaikhan Batjargal (Research Organization of Science and Engineering, Ritsumeikan University)
Fuminori Kimura (Faculty of Economics, Management, and Information Science, Onomichi City University)
Akira Maeda (College of Information Science and Engineering, Ritsumeikan University)

Many copies and variants of Ukiyo-e prints had been scattered around the world in the 19th century, and now they have been digitized and available in databases of museums and galleries in many countries. As being a woodblock print, many copies are printed for an Ukiyo-e. Nowadays, many copies of the same Ukiyo-e prints exist in different database and metadata of these prints are diverse and written in different languages. Therefore, it is difficult to find the same Ukiyo-e prints due to the notable differences in languages and formats of metadata. In this paper, we propose a method to identify the same Ukiyo-e prints from multiple databases in different languages by their descriptive titles, which tackles a big challenge in Cross-language record linkage We identify the same prints by comparing the translation candidates extracted by using dictionaries, thesaurus and character n-gram.

１．はじめに

浮世絵は江戸時代に成立した絵画のジャンルであり，人々の日常の生活や風物などを題材として描かれている．現代では浮世絵の芸術性の高さが見直され，美術品としての価値が高まっており，海外でも注目されている．近年，美術品などの文

表 1：同一作品の表記の違いの例

作品名	画像	データベース
冨嶽三十六景 神奈川沖浪裏（原題）		江戸東京博物館
Under the Wave off Kanagawa（英訳）		メトロポリタン美術館
Sous la vague au large de Kanagawa dite aussi La Grande vague（フランス語訳）		French Photo Agency

化資源をディジタル化し，ディジタルアーカイブとして保存する動きが進んでおり，世界中の各美術館・博物館で浮世絵データベースが公開されている．日本のデータベースは日本語で，海外のデータベースは主に英語でメタデータが記述されているが，フランス語，ロシア語，オランダ語その他様々な言語で記述されているデータベースも少なからず存在する．表 1 に浮世絵の同一作品のデータベースによる表記の違いの例を示す．このような状況において，これらのデータベースを利用する浮世絵研究者から，浮世絵の画像やメタデータを網羅的に検索したいとの要望がある．また，異なるデータベース間で同じ作品のメタデータを比較することで，データの修正や補完などを行いたいという要望もある．しかし，現状では，同じ作品でもデータベースによって言語や表記の違いがあるため，同一作品を見つけるのは容易ではない．このような問題を解決するために，我々はこれまでに異言語の浮世絵データベース間における同一作品の同定手法を提案している

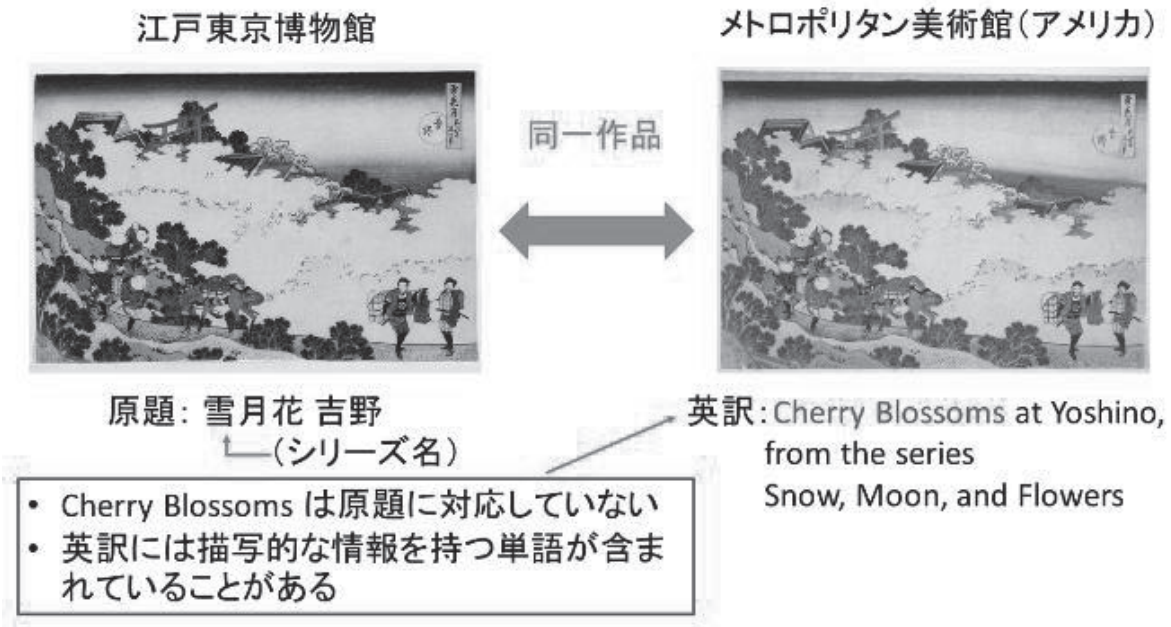

図 1：描写的な作品名の例

図 2：Ukiyo-e.org

[1][2]. これらの手法では, 浮世絵の作品名を用いて, 音訳（ローマ字）同士など同表記同士の比較や音訳と英訳の作品の比較, 原題と英訳の作品の比較が行われ, 実験の結果高い精度が得られているが, 原題と蘭訳を用いた手法に関しては, 原題と英訳を用いた手法に比べて精度が劣っていた. その主な原因として, 日本語以外の浮世絵の作品名には, 原題の翻訳ではなく, 浮世絵が描写している事物などを表現している単語が含まれる場合があるためである. この場合, 作品名を直訳しても単語がマッチしない問題が発生する. 図 1 にその例を示す. 本論文では, 日本語による作品名を文字 n-gram に分割し翻訳することで, その浮世絵に関連している訳語候補を取得できる可能性を高め, さらにその訳語候補の類義語も取得することで, 描写的な作品名に対して同一作品の同定精度を向上する手法について述べる.

2. 関連研究

本研究で扱うような, 異なるデータベースに存在する同一実体を表すレコードを自動的に見つけ出す問題は,「レコード同定」「レコード照合」などと呼ばれ, 古くから研究が行われている. レコード同定に関する研究動向については, 相澤ら[3]によるサーベイ論文がある. この論文では同言語データベース間でのレコード同定について様々な手法が紹介されているが, 本研究では異言語間のデータベースでのレコード同定となるため, 従来の研究とは大きく異なる. 同言語同士で比較を行う場合は, たとえば編集距離などの文字列照合関数を用いて類似度を算出することができる. Monge ら[4]は, メタデータ項目ごとの類似度を算出する際の編集距離の拡張として, Smith-Waterman アルゴリズムの適用を提案している. これは編集距離の置換操作によるコストを文字ペアごとに定義することで大文字/小文字の違いを無視するなど柔軟な対応が可能になるものである. また文字列の類似度だけでは同一か否かの判定が行えない場合に, 略称や異表記などの変換ルールを事例から自動学習する手法も提案されている[5]. しかし, 異言語同士で比較を行うには, 一方の言語を他方の言語に翻訳する必要がある. 本研究ではこの課題の解決に取り組んでいる.

一方, 多数の浮世絵データベース中から同一作品を見つけることができる Web サイトとして Ukiyo-e.org[1]がある. このサイトで用いている手法と本論文の提案手法との違いは, Ukiyo-e.org では画像の類似度を用いた同定を行っているのに対し, 本提案手法では浮世絵作品のメタデータによる同定を行っている点である. Ukiyo-e.org の画像検索結果を図 2 に示す. データベースによっては, 一部のレコードに画像が無くメタデータのみ存在する場合があり, 提案手法により Ukiyo-e.org では同定できなかった作品を抽出できる可能性がある. それに加えて, 浮世絵には原画を修正し出版された異版が存在する. 異版は画像の特徴差が大きい場合があるため, 類似画像検索の手法では区別することができない作品がある. 上記のように, 本提案手法は画像の類似度を用いた手法で同定できない一部の作品を同定し, 補うことができる.

[1] http://ukiyo-e.org/

我々は，世界中のデータベースにある浮世絵を検索する複数言語に対応したシステムとして FeSSU [1]（Federated Searching System for Ukiyo-e prints）[6]を構築している．本システムでは，ユーザが作者名や作品名などのクエリを入力すると，システムは各データベースに対して SRU（Search Retrieve via URL）またはスクレイピングを用いて検索を行い，クエリに関する浮世絵作品をユーザに提示する．SRU とは，横断検索用プロトコルのことで，検索要求情報を含んだ URL をサーバへ送り，その検索結果を XML 形式で返すものである．スクレイピングとはウェブサイトから必要な情報を自動で収集する処理のことを指す．また，このシステムは検索の際メタデータを多言語に翻訳し検索を行っており，世界中の多くの浮世絵データベースに対応した多言語横断アクセスを可能にしている．本論文で提案する同一作品の同定手法は，将来的に本システムに組み込み，複数データベースから同一作品を提示する機能として実装する予定である．

3．提案手法

提案手法全体の流れを図 3 沿って説明する．なお，図 3 は対象のデータベースが英訳表記の場合の一例を示している．初めにユーザは日本語表記のデータベースから浮世絵作品を 1 つ選択する．そして，選択した作品名を文字 n-gram に分割する（図 3①）．その後，分割した全 n-gram を対訳辞書や固有名詞辞書を用いて英語に翻訳し，訳語を取得する（図 3②）．さらに，取得した訳語全てに対してシソーラス（類語辞典）を用いて類義語を取得する（図 3③）．取得した全ての訳語と類語語を，同定対象データベースの浮世絵レコード群の作品名と単語単位でマッチングを行い，類似度を計算する（図 3④，⑤）．ここで，同定対象データベースのレコード群は最初にユーザが選択した作者のレコードのみに絞り込んでいる．最終的に，閾値以上のものをユーザに同一作品として提示する．

3.1 n-gram による作品名の分割

n-gram とは文字列を連続する n 個の文字で分割するテキスト分割方法のことである．例えば，「山下白雨」を 2-gram（bigram）で分割すると，「山下」，「下白」，「白雨」となる．本提案手法では 1 から n（文字列の最長数）まで全ての個数で分割を行い，より多くの訳語候補を取得するようにしている．

単語として明らかに不適切な n-gram からは訳語候補が得られることはないため，そのような n-gram も使用することの悪影響は小さい．また，

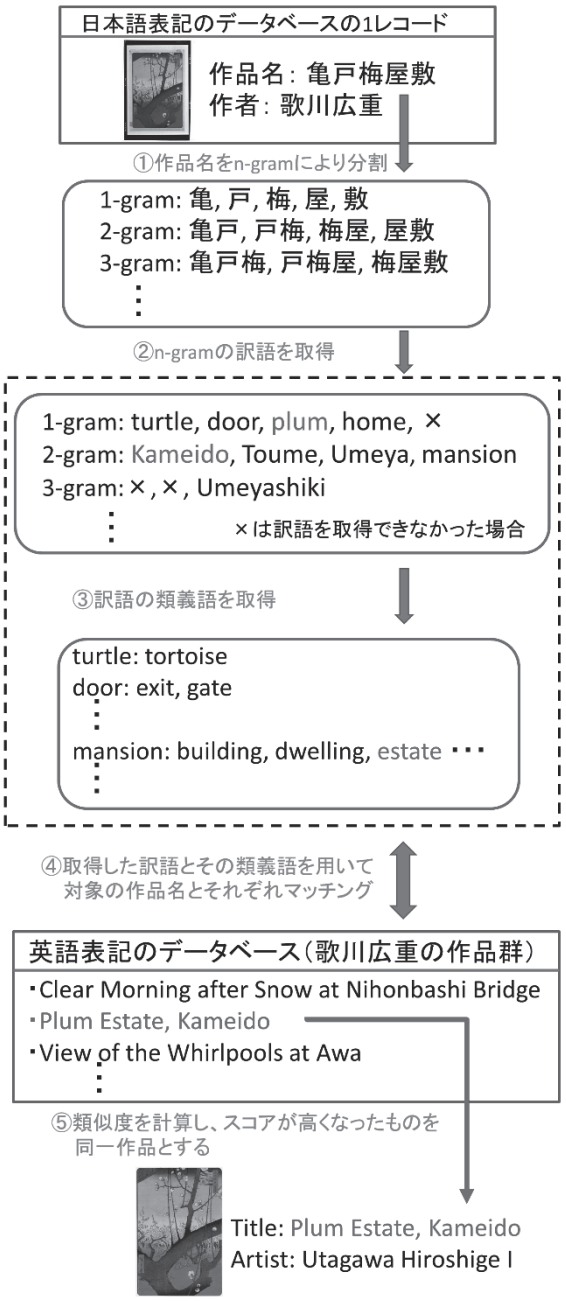

図 3：提案手法全体の流れ

得られた n-gram，特に 1-gram の中には，元の作品とは関係ない意味の単語を表す場合もあるが，本論文では浮世絵のタイトルのみを対象としていることから，そのような n-gram は訳語候補が得られたとしても浮世絵のタイトルとマッチしない可能性が高い．場合によってはある n-gram の訳語候補がマッチする可能性もあるが，複数の訳語候補がマッチしなければ対象作品タイトルのスコアが高くならないことを考慮すると，このような場合の悪影響もそれほど大きくな

[1] http://www.dl.is.ritsumei.ac.jp/fessu/ukiyoe.html#lang=ja

らないと考えられる.

3.2 訳語の取得

訳語の取得には複数の辞書を用いて行う. 作品名に含まれる固有名詞に対応するために浮世絵関連語(『日本演劇辞典』,『浮世絵大辞典』など浮世絵関連の辞書を電子化したもの)と地名辞書(旧国名とその略称のペアを, Web サイトの情報を参考に作成したもの)を使用する. また, 一般名詞等に対応するために日英対訳辞書を使用する. それぞれの辞書において, 訳語が複数あった場合すべてマッチングに使用する.

3.3 訳語の類義語の取得

シソーラスを用いて, 分割した n-gram の訳語の類義語を取得する. 類義語を取得する理由として, 日英対訳辞書に含まれていない訳語を抽出し, より多くの訳語候補を得ることで, 先述した描写的な意味を含む単語や原題から微妙に変化した単語に対応できる可能性を高め, 同定精度の向上に繋がる.

3.4 作品名同士のマッチング・スコア計算

英語表記の作品名と訳語・類義語とのマッチングは単語(文字列)単位で行う. そして, マッチした単語の種類によってスコアの重みを変えるものとする. 翻訳による影響が少ない固有名詞は固有名詞以外の単語よりも重みを大きく設定する. なお, “a”や “the”等の冠詞や接続詞などは多くの作品名に含まれている可能性が高いため, 機能語(代名詞・前置詞・接続詞・助動詞・限定詞)はマッチング対象外とする.

以下にスコアの算出式, 図 4 に算出式の説明を示す.

$$S = \frac{\text{実際に一致した値}}{\text{英訳の全単語が一致した場合の値}}$$

$$= \frac{w_p M_p + w_{np} M_{np} + w_s M_s}{w_p N_p + w_{np} N_{np} + w_s N_s}$$

S : スコア
Wp : 訳語(固有名詞)の重み
Wnp : 訳語(固有名詞以外)の重み
Ws : 類義語の重み
Mp : 訳語(固有名詞)の一致数
Mnp : 訳語(固有名詞以外)の一致数
Ms : 類義語の一致数
Np : 英訳に含まれる訳語(固有名詞)の数
Nnp : 英訳に含まれる訳語(固有名詞以外)の数
Ns : 英訳に含まれる類義語の数

図 4 : スコアの計算式

英訳表記の作品名の単語にマッチした訳語・類義語の数に重みをかけ, それぞれを足し合わせる. そこからスコアの範囲を 0(最少値)から 1(最大値)に調整するために, 英訳表記の作品名の単語が全て訳語・類義語にマッチすると仮定した場合に取りうる値で割っている.

4. 実験

第 3 章で述べた提案手法による浮世絵作品の同一レコードの同定の精度を確認するために実験を行った.

4.1 実験方法

4.1.1 使用するデータ

実験の準備として, 江戸東京博物館のデータベース[3]にある浮世絵 76 件の作品名の原題と, メトロポリタン美術館のデータベース[4]にある浮世絵 450 件の作品名の英訳を用意した. 英訳 450 件の中には原題 76 件の同一作品が含まれている.

4.1.2 翻訳に使用した辞書

本実験では日英対訳辞書に「EDICT[5]」, シソーラスに「WordNet[6]」を使用する. それぞれ無償で提供されている.

4.1.3 スコア計算における重みの設定

本実験ではマッチング時の固有名詞の重み(Wp)を 2,それ以外の訳語(Wnp)・類義語(Ws)については 1 としてスコアを計算している.

4.2 実験結果

提案手法の同定精度評価の実験結果を表 2 に示す. 正解データとは, 原題に対して英訳表記のデータベースに含まれている同一作品のことを指す.

表 2 : 提案手法の同定精度評価の実験結果

	件数	割合
正解データがスコア 1 位	51/76	0.6705
正解データがスコア 10 位以内	59/76	0.7763

4.3 実験結果の考察

描写的な単語を作品名に含む作品をうまく同定できた例を図 5 に示す. 図 5 では歌川広重の作品である「名所江戸百景 : 深川木場」の “深川木場” の部分を n-gram に分割したものから訳語・類義語に変換しているが, その中の単語が同一作品の作品名である「The Lumber Yard at Fukagawa」の “Lumber” と “Fukagawa” にマッチした. “Lumber”(日本語訳で材木)は日

[3] http://digitalmuseum.rekibun.or.jp/app/collection/
[4] http://www.metmuseum.org/art/collection/search/
[5] http://www.edrdg.org/jmdict/edict.html
[6] http://wordnet.princeton.edu/wordnet/

英対訳辞書には無く，シソーラスによって抽出された．“木場”はこの作品の場合，地名と貯木場としての名詞と両方の意味で捉えることができるが，本提案手法ではどちらにも対応していることがわかる．

一方で，図 5 の英訳にある “Yard” にマッチする単語が訳語・類義語の中に含まれていないことが分かる．このように同義語や類義語の域を超えた単語が英訳の中に含まれている場合はマッチングが困難となってしまう．

図 5：実験結果の例

5．おわりに

本論文では，異言語の浮世絵データベース間における同一作品の同定において，描写的作品名に対応する新たな手法を提案した．n-gram による分割と訳語・類義語の取得を行うことで，描写的な作品名も同定することができたと考えられる．また，各言語に対応した対訳辞書やシソーラスさえあれば，本提案手法は言語を問わず適用することが可能である．

今後の課題としては，浮世絵のデータベースだけではなく，書籍など画像検索では同一レコード同定が難しいジャンルのデータベースに対して有効な手法を提案したいと考えている．

6．謝辞

本研究の一部は，日本学術振興会科学研究費補助金基盤研究 (C)「多言語ディジタルアーカイブにおける言語横断レコード同定手法の研究」（研究代表者：前田亮，課題番号：16K00452）の支援を受けている．

参考文献

1） 久山岳夫， B. Batjargal，木村文則，前田亮：複数の異種浮世絵データベース間における同一浮世絵の同定手法の提案，人文科学とコンピュータシンポジウム 2013 論文集，pp.225-232(2013).

2）木村泰典，B. Batjargal，木村文則， 前田亮：多言語の浮世絵データベース間における同一作品の同定手法の提案，人文科学とコンピュータシンポジウム 2015 論文集，pp.117-124(2015).

3）相澤彰子，大山敬三，高須淳宏，安達淳：レコード同定問題に関する研究の課題と現状，電子情報通信学会論文誌，DI，Vol.J88-DI, No.3，pp.576-589 (2005).

4）A. E. Monge and C. P. Elkan : An efficient domain-independent algorithm for detecting approximately duplicate database records. In *Proceedings of the ACM-SIGMOD Workshop on Research Issues on Knowledge Discovery and Data Mining*，pp.23-29, (1997).

5）S. Tejada，C. A. Knoblock，and S. Minton : Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.350-359，(2002).

6）B. Batjargal，F. Kimura，and A. Maeda : Metadata-related Challenges for Realizing Federated Searching System for Japanese Humanities Databases. In *Proceedings of the 11th International Conference on Dublin Core and Metadata Applications (DC-2011)*, pp.80-85, (2011).