異言語の映画データベース間における同一作品の言語横断レコード同定手法

佐藤 英男† Yuting Song‡ Biligsaikhan Batjargal†‡ 前田 亮‡‡

†立命館大学大学院情報理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1
‡立命館大学総合科学技術研究機構 〒525-8577 滋賀県草津市野路東 1-1-1
†‡立命館大学衣笠総合研究機構 〒603-8577 京都府京都市北区等持院北町 56-1
‡‡立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1
E-mail: †is0309xr@ed.ritsumei.ac.jp, ‡ytsong@gst.ritsumei.ac.jp,
†‡biligee@fc.ritsumei.ac.jp, ‡‡amaeda@is.ritsumei.ac.jp

あらまし 近年，インターネット上には映画，音楽，浮世絵といった文化資源など，様々な対象について多くのデータベースが存在する．これらのデータベースでは，データのレコードが様々な言語や表記で記述される．それらの異言語データベース間において，同一の事柄を指すレコードを同定し，結びつけを行うことは重要である．そこで本研究では，異言語で記述された映画データベースを対象とし，データベース間において，同一作品を指すレコードを同定する手法を提案する．提案手法として，比較対象となるそれぞれのレコードのメタデータを同一の表記に変換し比較することで,レコード間の類似度を算出し,同定結果をランキング形式で出力する手法を提案する．また，本論文では日本語と英語で記述されたデータベースを対象とし，その中でも西洋で製作された外国映画である洋画を対象として扱う．

キーワード レコード同定，データ分析，言語横断，映画，テキスト処理

1. はじめに

近年，インターネット上には映画，音楽，浮世絵といった文化資源など，様々な対象について多くのデータベースが存在する．また，異なるデータベースにおいて同一の事柄を指すデータのメタデータを比較することで,データの修正や補完を行うことが可能となり，データ品質の向上が望める．そのため，異なるデータベース間で同一の事柄を指すレコードを同定し，結びつけを行うことは重要である．しかし現状では，同一の事柄を指すレコードでもデータベースによってメタデータの言語や表記が異なる場合が多々あり，同一の事柄を指すレコードを検索する事は容易ではない．

そこで，本研究では，異言語で記述された映画データベースを対象とし，データベース間において，同一作品を指すレコードを同定する手法を提案する．

また，本論文では日本語と英語で記述されたデータベースを対象とし，その中でも西洋で製作された外国映画である洋画を対象として扱う．

2. 関連研究

著者が所属する研究室では，世界中のデータベースに存在する浮世絵を検索する多言語に対応した言語横断浮世絵検索システムを構築しており，このシステムを支援する研究が行われている．Song ら[1]は，異言語浮世絵データベース間において同一作品を指すレコードを同定する手法として，メタデータを辞書ベースで機械翻訳しメタデータ間の類似性を推定する手法，メタデータを機械学習で機械翻訳しメタデータ間の類似性を推定する手法，メタデータを Word embedding を用いてベクトル変換し比較することでメタデータ間の類似性を推定する手法を提案している．また，メタデータ間の不一致や，メタデータに含まれる固有名詞が重要だと捉え，これらを考慮した手法を提案することで言語横断レコード同定の精度を向上させている．

また，アルファベット表記とカタカナ表記の結びつけを行う研究は数多く行われている．

尾上ら[2]は，アルファベット表記とカタカナ表記が対応した外国人名辞書に収録されている表記を分割し，その出現頻度を算出することによって，アルファベット表記とカタカナ表記を対応させた規則を自動生成するアルゴリズムを提案した．この手法は辞書データが存在するならば，特定の言語に依存せず表記の対応規則の生成が可能であることを述べている．

服部ら[3]は，情報検索を目的とし，英語音韻に着目することで 1 つのカタカナ表記から多様なカタカナ異表記を生成する手法を提案している．その手法では，あるカタカナ文字列をカタカナ文字・ローマ字対応表を使用してアルファベット変換し，それを更にカタカナ音，英音素へと変換した後にカタカナ音へと逆変換することによって，多様なカタカナ表記を生成している．本研究の提案手法では，服部らが使用したカタカナ文字・ローマ字対応表を基に改良を加えた対応表を使用している．

3. 提案手法

3.1.使用する用語の定義

以下に本論文で用いる用語を定義する．

- 起点言語：ある文字列について，翻訳前の原文の言語を指す．本論文において，入力するレコードの言語が起点言語に当たる．
- 目標言語：ある文字列について，翻訳後の訳文の言語を指す．本論文において，入力するレコードとの比較対象であるレコードの言語が目標言語に当たる．
- 起点レコード：各フィールドが起点言語で記述されたレコードを指す．提案手法では入力するレコードが起点レコードに当たる．
- 目標レコード：各フィールドが目標言語で記述されたレコードを指す．提案手法では起点レコードとの比較対象となるレコードに当たる．

3.2.使用するデータ

本研究において使用する日本語レコードは，映画についての日本のデータベースである「映画 DB」[4]より収集した，1990 年以降に製作された洋画の日本語レコード 6,107 件を使用する．各レコードは「タイトル」，「監督名」，「脚本家名」，最多で 5 名分の「俳優名」の 8 個のフィールドで構成される．

また，本研究において使用する外国語レコードは，映画，俳優，テレビ番組などについてアルファベットで記述されたデータベースである「IMDb」[5]より収集した，1990 年以降に製作された洋画の外国語レコード 55483 件を使用する．各レコードは「タイトル」，「監督名」，「脚本家名」，最多で 5 名分の「俳優名」の 8 個のフィールドで構成される．

3.3.提案手法の概要

本研究における提案手法の処理の流れを図 1 に示す．

提案手法では，まずユーザがいずれかのデータベースから選択したレコードを入力とする．

入力レコードの「監督名」，「脚本家名」，「俳優名」のフィールドをそれぞれ，文字列の綴りを発音に変換するアルゴリズムである Metaphone[6]を使用して記号化を行う．また，目標言語のデータベースの各レコードにおいても同様に記号化を行う．

次に，それぞれの記号同士で類似度計算を行い，起点レコードと各目標レコードとの類似度を求め，類似度の高い上位 10 件の目標レコードを同定結果として出力する．

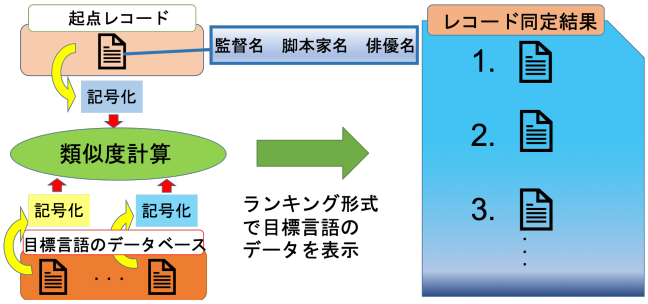

図 1:提案手法の処理の流れ

3.4.前処理

外国人名について，外国語レコードではファーストネームとラストネームの間は空白で記述されることが多いが，日本語データベースでは「・」で記述されることが多い．このように，データベース間で名前の間に挿入される文字は異なる場合があるため，本研究では，前処理として名前の間の文字を空白に置き換える処理を行う．

3.5.フィールドの記号化

Metaphone は文字列の綴りを発音に変換するアルゴリズムである．綴りから母音を取り除き，16 種類の子音のみで発音を近似させる．

提案手法では日本語と英語を対象とし，それぞれの言語で表記された人名の表記は原文の音訳を元に表されることが多いため，各フィールドを発音に基づいて記号化し，比較する手法が有効であると考えられる．

外国人名に対して Metaphone を適用した例を表 1 に示す．

表 1:Metaphone を適用した例

原文	Metaphone適用
Christopher Nolan	KRSTFRNLN
Will Smith	ALSM0

3.6.アルファベット変換

レコードのフィールドが日本語で記述されている場合，Metaphone を適用する前にアルファベット表記に変換する必要がある．提案手法において，アルファベット表記への変換はローマ字変換表を使用して行う．ローマ字変換表は Knight ら[7]，服部らが作成したものに著者が一部の変換規則を追加して作成した．追加した変換規則を表 2 に示す．

また，吃音「ッ」はその後ろに来る文字の子音部分を重複させ，長音「ー」はその前に来る文字の母音部分を重複させることで変換している．

表 2:カタカナ文字の追加変換規則

カタカナ文字の追加変換規則			
ヲ:wo			
複合カタカナ文字の追加変換規則			
ファ:fa フャ:fya	フィ:fi フュ:fyu	フェ:fe フョ:fyo	フォ:fo

3.7.レコード間の類似度計算

起点レコードと目標レコードそれぞれで記号化された「監督名」,「脚本家名」,「俳優名」フィールドは，それぞれ Jaro-Winkler 距離[8]を計算し，フィールド毎の類似度とする．

本研究では，使用するレコードにおける「監督名」，「脚本家名」フィールドはそれぞれ一名で構成されるため，起点レコードと目標レコードとで対応するフィールドで類似度を計算する．

「俳優名」フィールドは，それぞれ最多で 5 名ずつで構成される．起点レコードと目標レコードはそれぞれ異なるデータベースから収集したため，それぞれのレコードで同一人物のフィールドを含むとは限らず，それぞれのレコードでの「俳優名」フィールド数は異なる場合がある．そのため，比較するレコード間で「俳優名」フィールドについて総当たりで類似度を計算し，その中で類似度が最大となる組を採用する．次にその組を除いた中での類似度が最大となる組を採用する．この処理を残る組が存在しなくなるまで繰り返す．

このようにして計算した「監督名」,「脚本家名」,「俳優名」の類似度を合計した値をレコード間の類似度とする．

4. 予備実験

本章では，評価実験の準備として行なった予備実験の結果を示す．

4.1.使用したテストデータ

予備実験で使用したテストデータは，収集した日本語レコードと外国語レコードから，入力レコードとして用いる 7 件の日本語レコードと，目標レコードとして用いる 55483 件の外国語レコードを使用した．

今回正解レコードとして選んだ 7 組のレコードの一例を表 3 に示す．

表 3:正解レコードとして選んだ 2 組のレコード

	映画DB	IMDb
正解レコード	レオン	Léon: The Professional
	ダークナイト	The Dark Knight

また，「ダークナイト」，「The Dark Knight」の同シリーズの映画のレコードを表 4 に示す．

表 4:「ダークナイト」，「The Dark Knight」と同シリーズの映画のレコード

	映画DB	IMDb
同シリーズ	バットマン　ビギンズ	Batman Begins
	ダークナイト　ライジング	The Dark Knight Rises

4.2.実験結果

実験は 7 通り行い，入力レコードをそれぞれ「レオン」，「ダークナイト」とした場合の実験結果を表 5，表 6 に，その他 5 件の入力レコードで同定を行なった場合の正解データの出力順位を表 7 に示す．

「レオン」を入力レコードとした場合，正解データである「Léon: The Professional」が出力結果の最上位に位置する結果となった．

また，「ダークナイト」を入力レコードとした場合，正解データである「The Dark Knight」が最上位に位置する結果となり，次いで同シリーズのレコードである「The Dark Knight Rises」，「Batman Begins」が位置する結果となった．「The Dark Knight Rises」と「The Dark Knight」とのフィールドの違いは「タイトル」を除いて，1 名分の「俳優名」のみであり，実験結果の類似度の差は 0.296 であった．また，「Batman Begins」と「The Dark Knight」とのフィールドの違いは「タイトル」を除いて，「脚本家名」のみであり，類似度の差は 0.616 であった．

また，その他 5 件を入力レコードとした場合では、5 件中 3 件において正解データが最上位に位置する結果となった．「ホビット 竜に奪われた王国」，「ミッション：インポッシブル/ゴースト・プロトコル」を入力レコードとした場合では，正解データが 2 位に位置する結果となった．これらの場合では，同シリーズの作品レコードが最上位に位置する結果となった．

表 5:「レオン」を入力にした実験結果

目標レコードのタイトル	類似度
Léon: The Professional	5.395
The Dwarves of Demrel	5.015
Cabin Fever 3: Patient Zero	4.412
Head Case	4.376
Two Tickets to Paradise	4.29
Psycho Killer Bloodbath	4.286
Break	4.283
Tony 'n' Tina's Wedding	4.22
Neptune	4.205
Dick Baby	4.167

表 6:「ダークナイト」を入力にした実験結果

目標レコードのタイトル	類似度
The Dark Knight	5.869
The Dark Knight Rises	5.573
The Prestige	5.414
Batman Begins	5.253
The Confabulators	5.179
Lost	5.151
Animal Among Us	5.146
Woman Wanted	5.112
Trapped	5.097
Friends and Family	5.092

表 7:その他の入力レコードと正解レコードの順位

起点レコードのタイトル	正解レコードの順位
アナと雪の女王	1
スター・ウォーズ エピソード3 シスの復讐	1
スパイダーマン3	1
ホビット 竜に奪われた王国	2
ミッション：インポッシブル／ゴースト・プロトコル	2

5. 考察

予備実験の結果では，7 件中 5 件の実験パターンにおいて正解データが最上位に位置する結果となり，予備実験の段階での同定精度は 71%となった．「ダークナイト」を入力レコードとした場合について，正解データと同シリーズ作品との類似度の差はそれぞれ 0.296，0.616 であり，「ホビット 竜に奪われた王国」，「ミッション：インポッシブル/ゴースト・プロトコル」を入力レコードとした場合において，同シリーズの別作品のレコードが最上位に位置する結果となったことから，これらの場合のように類似度計算に使用したフィールドの大部分が共通し，類似度の差が小さくなってしまうことにより，同定結果の精度が悪くなってしまう可能性がある．そのため，例えば，今回使用したフィールドに加えて「タイトル」フィールドも考慮するなど，このような場合への対処が可能となる改良が必要であると考えられる．

6. おわりに

本研究では，異言語で記述された映画データベースを対象とし，データベース間において，同一作品を指すレコードを同定するために，人名が記述されたフィールドを使用する手法を提案した．その中でも，日本語と英語で記述されたデータベースを対象とし，洋画を対象とした場合の同定手法を提案した．

予備実験では，7 パターンの入力レコードで実験を行なった．結果としては，7 作品のうち 5 作品のレコードの正解レコードが出力結果のランキングの最上位に位置する結果となった．しかし，同シリーズの映画でフィールドの大部分が共通する場合があり，入力レコードによっては精度が悪くなってしまう可能性あることが明らかになった．そのための改善案としてタイトルを考慮する手法が考えられる．

本予備実験での結果を踏まえ，明らかになった部分を改善しつつ，対象とする言語を追加する，洋画だけでなく邦画も対象に含めるなど，より汎用性の高い提案手法へと改良を重ねたいと考えている．

参 考 文 献

[1] Yuting Song, Biligsaikhan Batjargal, and Akira Maeda. Cross-Language Record Linkage based on Semantic Matching of Metadata. 日本データベース学会英文論文誌, Vol. 17, No. 1, pp. 1-8, Mar. 2019.

[2] 尾上 徹, 梅村 恭司, 岡部 正幸, “アルファベット表記とカタカナ表記の対応規則の生成”, 第 52 回プログラミングシンポジウム, 2011.

[3] 服部 弘幸, 関 和広, 上原 邦昭, “英語音韻を考慮した情報検索のための多様なカタカナ異表記生成”, 情報処理学会論文誌 数理モデル化と応用 Vol.2 No.1 144-155, 2009

[4] 映画 DB, https://eigadb.com (参照 2020/1/9)

[5] IMDb, https://www.imdb.com (参照 2020/1/9)

[6] Metaphone, https://docs.oracle.com/cd/B72202_01/Content/processor_library/transformation/metaphone.htm (参照 2020/1/9)

[7] Knight K., Graehl J., “Machine Transliteration”, Computational Linguistics, Vol.24, No.4, pp.599-612, 1998

[8] Winkler, William E., "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.", Proceedings of the Section on Survey Research Methods. American Statistical Association: 354–359, 1990