苗字・国籍ビッグデータによる民族の空間的特徴の把握

○全珠美[1]　水野貴之[1,2]（総合研究大学院大学[1]，国立情報学研究所[2]）

Spatial Characteristics of Ethnic Groups using Surname and Nationality Big Data

* Joomi Jun[1] and Takayuki Mizuno[1,2] (SOKENDAI[1], National Institute of Informatics[2])

概要： Recurrent Neural Network に，全世界 3700 万の経済人（会社役員，株主，公的人物など）の苗字と国籍を学習させることにより，国別に苗字生成の言語モデルを構築し，各国の経済人の人数を反映させることによって，苗字から民族（その苗字のルーツとなる国）を推定する分類器を作成する．分類器により各国の民族分布が推定でき，分布のエントロピーによって各国の民族多様性を定量評価できる．国間での民族分布の比較により，国間の民族の類似度をネットワークで可視化することで，アフリカ大陸で空間的に広がる民族の特徴を把握できる．

キーワード: Name-Ethnicity classification, Name-Nationality classification, RNN, Ethnicity network

1　はじめに

名前は人間の最も基本的な情報インデックスである．名前は社会や文化の特徴的なシステムによって作られて利用されるため，名前からは，その人物の性別，地域，文化的な背景を読み取ることができる．ファーストネームからは，地域，性別、時代の流行などが読み取れ，一族が代々継承するファミリーネームからは，民族など人物のルーツが読み取れる．

例えばアメリカ人で「Sally」というファーストネームを持つ人物のフールネームが「Sally Zhang」の場合，その人物が中国系だと容易に推測できる．画家の名で有名な「Van Gogh」は，「from/of Gogh」を意味するオランダ特有のファミリーネームで，彼の一族のルーツがGogh地域であることが分かる．このように名前，特にファミリーネームからは，しばしば，ルーツとなる民族を推定することが可能である．

本稿では，ファミリーネーム(以下：苗字)を用いて国籍を分類する分類器を作成し，苗字から得られる各国の民族（ルーツとなる国）の確率分布を求める．Jensen–Shannon divergenceを用いて，各国間の確率分布の類似度を定義し，MapEquationにより類似する国をクラスタリングする．これにより，空間的に広がる民族による国間の繋がりを可視化した．

2　RNNによる苗字の言語モデル

名前データを用いて国籍や民族を分類する研究は生物医学[1)]，社会学[2,3)]，人口統計学[4,5)]，マーケティング[6,7)]など様々な分野でおこなわれてきた．しかし，名前の種類は膨大であり，名前と民族のリストデータを，そのまま利用する単純な手法[8)]では，リストにある名前しか分類できないという限界があった．

近年，HMMsとDecision Treeを利用して名前から民族を分類する手法[9)]，Bayesian Approachによる分類[10)]，SVMを基盤とする性別の分類[11)]，機械学習を利用する手法[12)]などが開発された．Naïve Bayes手法でフランス社会グループの苗字の起源を分析した研究[13)]もあるが，本稿とは手法と分類範囲で違いがある．本稿では，自然言語処理で使われる再帰型ニューラルネットワーク(Recurrent Neural Network，以下RNN)を用いて，国籍別に名前生成の言語モデルを構築することにより，名前から国籍を推定する手法[14)]を採用する．

ニューラルネットワークは，脳神経が学習によりネットワークを生成する形を真似した手法である．RNNでは，多層レイヤーの循環構造により時系列のような連続性を持つデータを学習することができ，文章，音声認識などで用いられる．

本稿では，名前のテキストを文字に分解して，RNNにより学習をおこなった．例えば，「Garcia」を「G」+「a」+「r」+「c」+「i」+「a」の連続するアルファベット文字に分解して学習をする．この学習によってRNNは，「G」の後に「a」が，「a」の後に「r」が，また，「Ga」のあとに「r」がといった，名前における文字列の生成確率を覚える．同時に，名前に紐づけて国籍を一緒に学習させることにより，この生成確率が国籍ごとに異なることも覚える．従って，十分に学習させたRNNに，ある名前を入力すると，その名前の出現確率を国籍別に返してくれる．この手法の長所は，学習させた名前リストにはない名前であっても，出現確率を算出できることである．

RNNは学習率，レイヤー層，使う関数の設定などよって，学習効率が変わる．我々は，RNNにLSTM(Long Short Term Memory)[15)]のレイヤーを重ねて学習をさせた．LSTMは，情報の距離が長くなることによってRNNの学習能力が落ちる問題(Vanishing Gradient Problem)を解決するために使われる．本研究における名前の学習においても，LSTMのレイヤーを重ねて学習させることにより，学習能力が向上した．

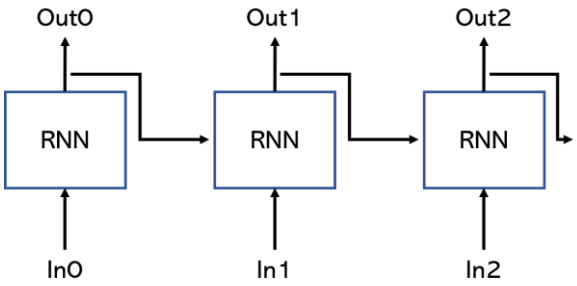

Fig. 1: RNNの概念図

3　苗字・国籍ビッグデータ

我々は，２つのデータセットを混ぜ合わせて利用する．１つは，Bureau van Dijk 社が提供する 203 ヶ国の

会社役員や個人株主の（ローマ字表記の）名前と国籍が網羅的に収録された 2016 年の ORBIS データセットである．約 3,500 万人が収録されている．もう 1 つは，Dow Jones 社が提供する 217 ヶ国の公的人物および，その親密な関係者（例えば，大統領夫人）の（ローマ字表記の）名前と国籍が網羅的に収録された 2018 年の Watch list データセットである．約 210 万人が収録されている．

我々は，この２つのデータセットから，収録人数上位 77 ヶ国（37,081,935 人）と，アフリカ大陸の４８ヶ国（894,115 人）を用いて RNN による学習とその後の分析をおこなった．

4　苗字の言語モデルを用いた国籍分類器

収録人数上位 77 ヵ国の 37,081,935 人について，国籍と苗字を RNN で学習した．学習における loss 値は Cross entropy

$$L(y,\hat{y}) = -\sum_{i=1}^{N} y_i log\, \hat{y}_i$$

によって測ることができる．ここで，y と$\hat{y}$は，教師データから得られる真の国籍分布と，分類器から推定された国籍分布で，N は国籍の数，すなわち 77 である．我々は，学習回数（Iteration 数）を 200 万回に設定して学習した．Fig.2 は，RNN における学習回数と loss 値の関係である．学習回数が 400,000 回（＝20×20,000 回）以上から loss 値の下降は緩やかになり学習が十分に完了していることが分かる．

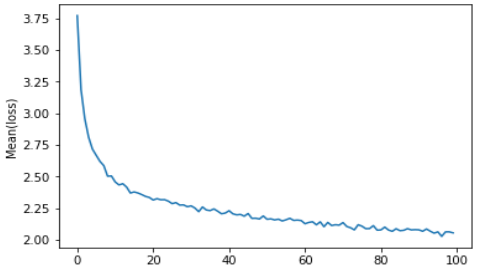

Fig. 2：学習と loss 値の変化（loss 値は 2 万回学習するごとにその平均を求めている）

学習を終えた RNN による苗字からの国籍分類器は，苗字を入力することで，各国籍における入力苗字の生成確率から算出される対数尤度を返してくれる．例えば，日本人名を入力すると，対数尤度は日本（国籍）で最も高い値を示す．一方で，トルコ人名を入力すると，対数尤度は，トルコと北キプロスで高い値を示す．北キプロスの全人口の 99%がトルコ系住民であるから，トルコと北キプロスの判別が本 RNN では不可能であるからである．分類器の精度を上げるために，我々は経済人（役員や株主，公的人物など）が世界レベルで網羅的に収録された本データセットの特徴を生かして，対数尤度を各国の人数の対数値で割って規格化する．つまり，現実問題，新聞等でトルコ人名が現れたら北キプロスではなく，ほぼトルコである．このような性質を規格化により反映させた．

Table 1 は，入力された苗字に対する，出力された上位３ヶ国の規格化された対数尤度の値である．Smith は，イギリスで最も生成されやすいが，オーストラリア，ニュージーランドでも生成されやすい苗字であることが分かる．　また，Mori は，圧倒的に他国に比べて日本で生成されやすい苗字であることがわかる．日本人名のような単一民族国家に紐づく苗字では，このような傾向が現れる．

Table 1：予測結果表

Input : Smith	Output (Top3): (-0.264) United Kingdom (-0.320) Australia (-0.401) New Zealand
Input : Obama	Output (Top3): (-0.246) Kenya (-0.424) Nigeria (-0.470) Japan
Input : Mori	Output (Top3): (-0.042) Japan (-0.614) Papua New Guinea (-0.656) Italy

RNN による学習と対数尤度の規格化により構築した分類器の精度を，ランダムに選択したテストデータセットを用いて検証する．ここでは，入力した名字に対して最も高い規格化対数尤度を示す国を，推定された国籍とした．分類結果を混同行列で表したのが Fig. 3 である．Table 2 は各国の検証結果を表している．正答率が高かった上位の国籍はアイスランド(91.1%)，韓国(89.2%)，ベトナム(88.1%)で，下位の国籍はニュージーランド(0.8%)，アルゼンチン(1.4%)，カナダ(1.6%)であった．民族の構成が単純化している国ほど正答率が高いことが分かる．

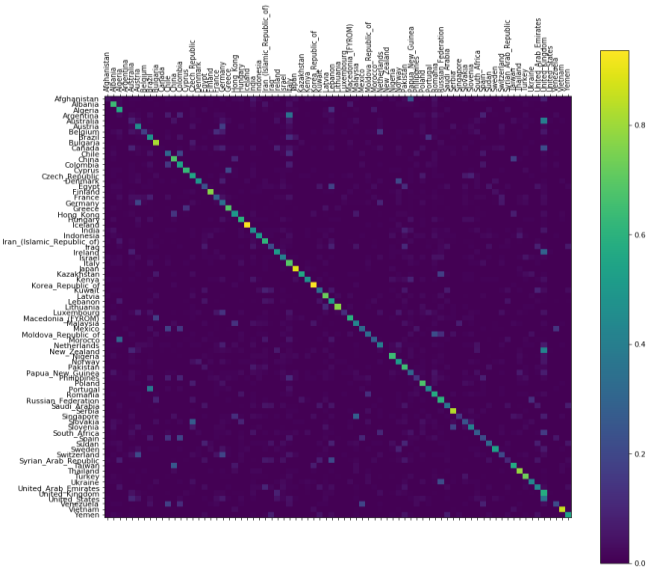

Fig. 3：予測テストの混同行列

Table 2 : 各国の予測精度(Top/Bottom 10)

順位	国名	正答率(%)
1	Iceland	91.1
2	Rep. of Korea	89.2
3	Vietnam	88.1
4	Japan	85.6
5	Thailand	83.5
6	Bulgaria	79.4
7	Serbia	77.7
8	Poland	75.7
9	Latvia	74.6
10	Macedonia	74.5
(中略)		
68	Philippines	13.5
69	Syrian Arab Republic	11.9
70	Australia	10.9
71	Singapore	9.4
72	Luxembourg	5.2
73	Afghanistan	2.4
74	United States	1.6
75	Canada	1.6
76	Argentina	1.4
77	New Zealand	0.8

5 各国の民族分布の推定

一般的に，ニューラルネットワークで構築した分類器は，イレギュラーな特徴を無視して一般的な特徴を学習する．そのほうが，推定精度が高くなるからである．例えば，稀にしか現れない韓国国籍や中国国籍のの「Mori」を学習して，どちらの国籍か推定する精度を高めるよりも，頻繁に現れる日本国籍を間違えなく日本であると推定できるようにするほうが，全体の推定精度は向上する. つまり, 分類器が推定する国籍は，どの国籍の苗字の特徴を最も反映しているかを表しており，その苗字のルーツとなる国籍を返している．例えば，米国国籍の中国系移民の名前を入力すると，中国（国籍）で, 規格化された対数尤度は最大値を示す．この分類器の特徴を利用して，各国における民族（ルーツとなる国）の分布を推定する．

Fig.4 は，推定された民族（ルーツとなる国）の分布が，特に散らばった国を表す．移民大国の米国では，英国やオーストラリア，ドイツ，イタリア，アイルランドをルーツとする人々が多いことが観測される．また，英国からの移民の多い，オーストラリア，ニュージーランドでも，同じ傾向が観測される．地理的には米国に近いメキシコであるが，民族構成は米国よりもベネズエラに近いことが分かる．チリは，ベネズエラやメキシコとは若干異なり，イタリア系の民族性が現れている．

単一民族国家である日本では，最大民族 Japanese が人口の 98.5%を，一方で，多民族国家であるケニアでは, 最大民族 Kikuyu が占める人口は僅か 17%である．我々は，民族多様性が，推定された民族分布によって計測できることを，各国の分布のエントロピーと，各

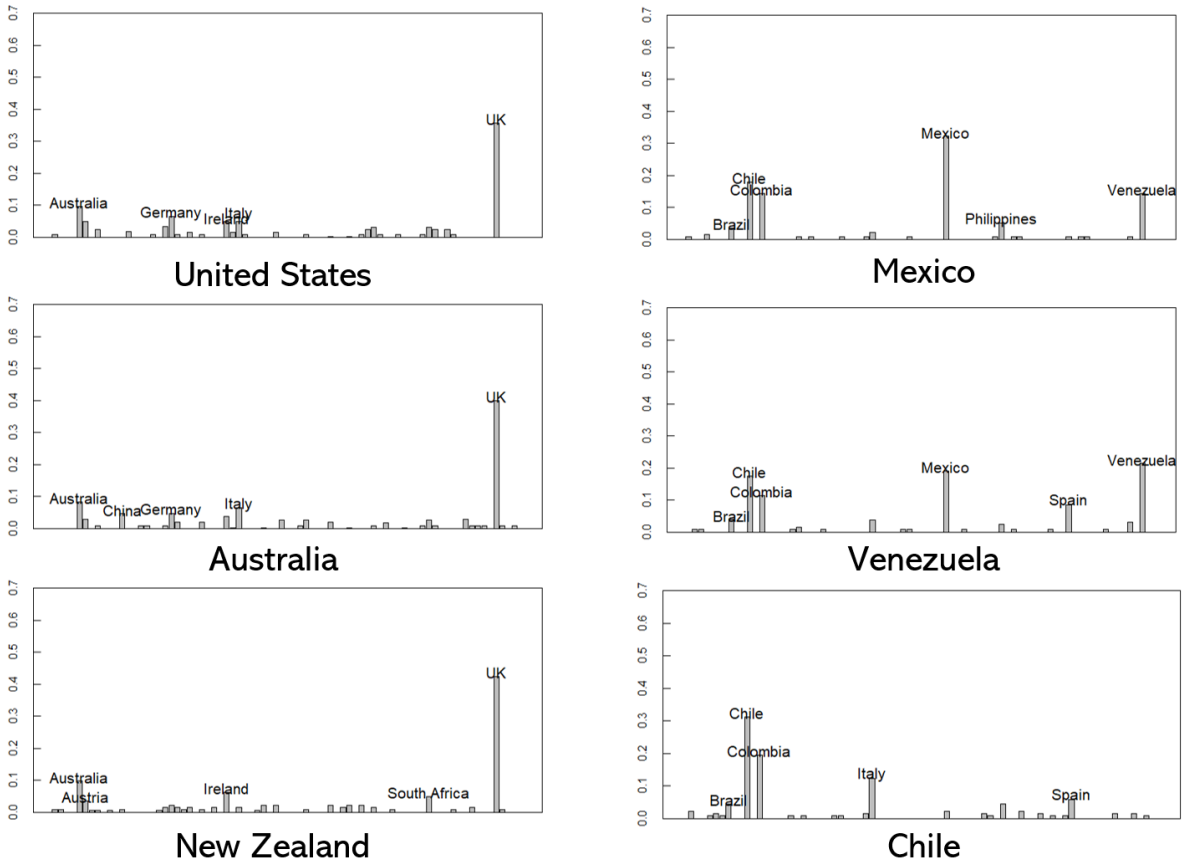

Fig. 4 :予測結果の分布から見える民族の構成

国における最大民族の割合 [16,17]との関係から示す．Fig.5 は，その関係である．エントロピーが高いほど，多民族国家，民族多様性が高いことが分かる．

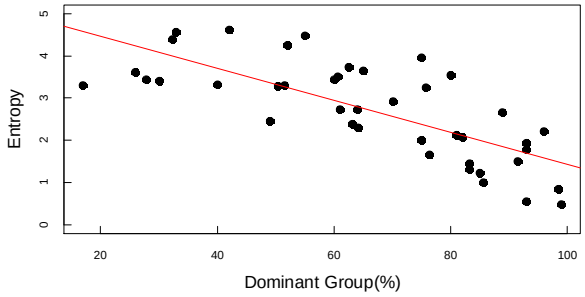

Fig. 5 : エントロピーと民族多様性

6　アフリカの民族ネットワーク

アフリカは，多様な民族と言語，宗教が混在する大陸である．そして，多くの国が人口構成に関する正確な統計がない. 我々は, アフリカ大陸の 48 ヶ国（894,115 人）の苗字と国籍データを RNN により学習し，対数尤度を規格化することで，アフリカ各国の民族構成を推定する．そして，推定された国別の民族分布を比較することによって，国間の民族の類似度をネットワークで可視化する．このネットワークのコミュニティを抽出することで，アフリカの国を跨いだ民族の空間分布を明らかにする．

Table 3 は，作成した分類器により苗字から推定された各国の民族（ルーツとなる国）分布のエントロピーを表す．エントロピーの低い国はコンゴ民主共和国，コンゴ,コートジボワールで，エントロピーの高い国はカメルーン，チャド，モロッコであった．この分類器でも，最大民族の人口比率が低い多民族国家ほどエントロピーが高く，うまく民族（ルーツとなる国）分布，民族多様性をとらえている．

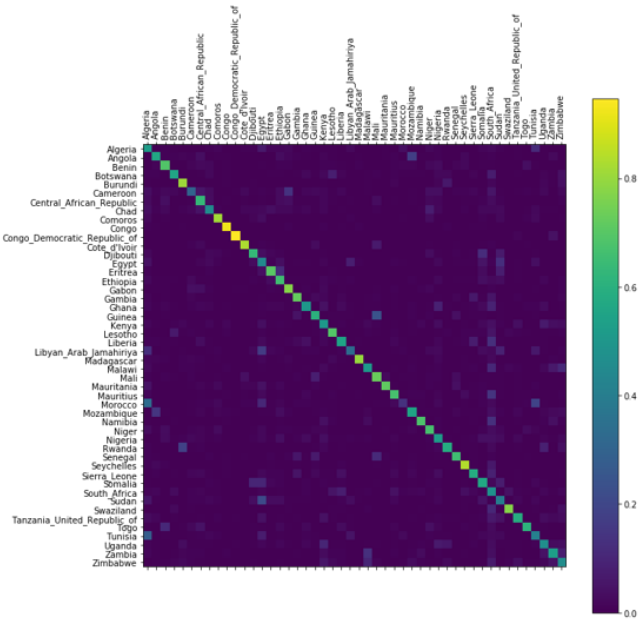

Fig. 6 : アフリカ大陸の予測テスト混同行列

Table 3 : 各国のエントロピー(Top/Bottom 10)

順位	国名	エントロピー
1	DR. Congo	0.34
2	Congo	0.64
3	Cote d'Ivoire	1.12
4	Seychelles	1.14
5	Burundi	1.27
6	Comoros	1.30
7	Swaziland	1.37
8	Madagascar	1.52
9	Eritrea	1.58
10	Gabon	1.65
(中略)		
39	Sudan	3.15
40	Zimbabwe	3.17
41	Nigeria	3.19
42	Uganda	3.22
43	Kenya	3.24
44	Libyan Arab Jamahiriya	3.28
45	South Africa	3.28
46	Morocco	3.29
47	Chad	3.46
48	Cameroon	3.99

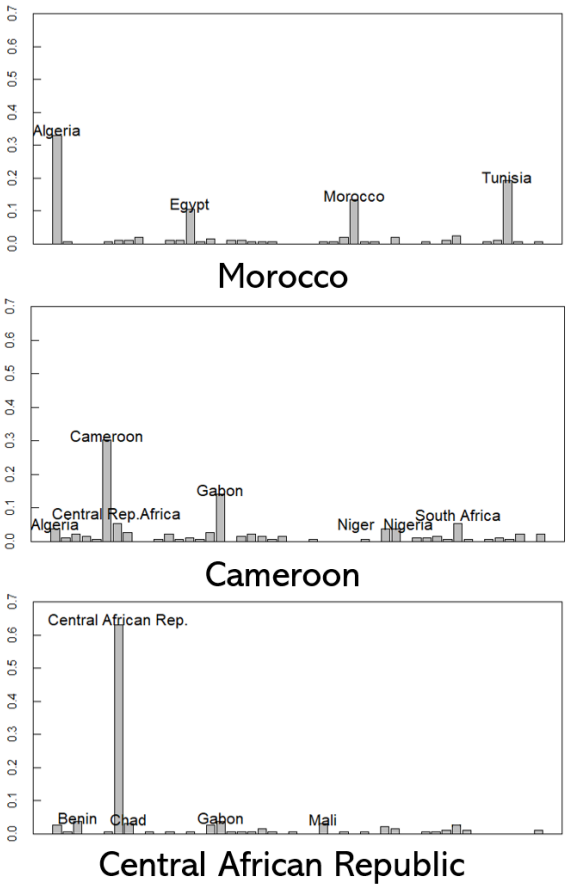

Fig. 7 : 予測結果値から見える民族構成

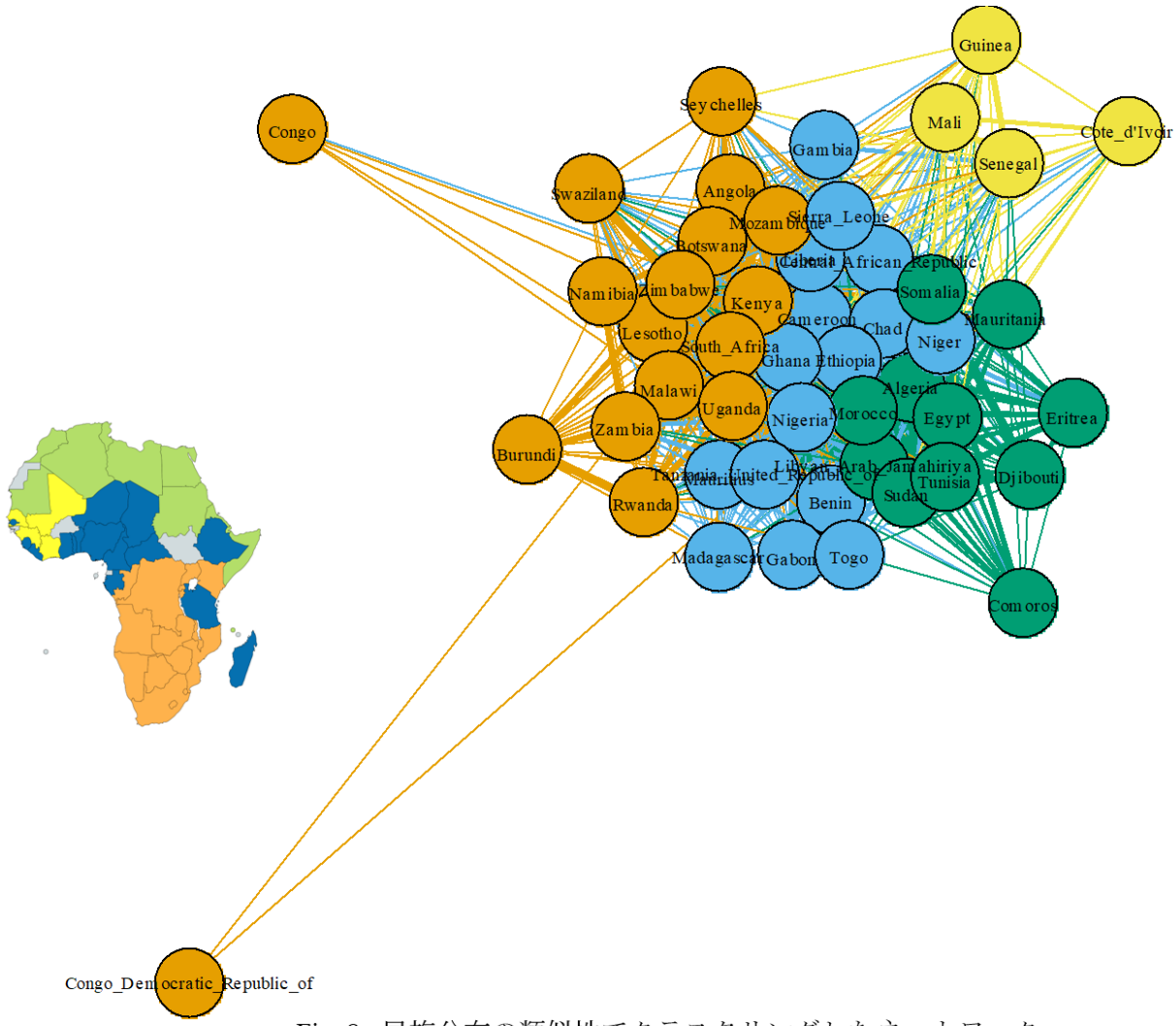

Fig. 8 : 民族分布の類似性でクラスタリングしたネットワーク

Fig. 7 は, 民族多様性の高いモロッコとカメルーン, 民族多様性が比較的低い中央アフリカ共和国の推定された民族分布を表す. モロッコには, モロッコ以外に, 地理的に近いアルジェリア, チュニジア, エジプト系の民族がいることが分かる. アフリカの中央に位置するカメルーンには, 地理的に近い国々をルーツにする民族がいる. 中央アフリカ共和国でも, 近隣国であるガボン, ベニン, チャド系の民族が観測される. これらの結果は, 近隣国からの移住だけではなく, アフリカの国境が民族ではなく大国の思惑で引かれてしまった歴史的背景も表している.

次に, 我々は国間の民族のつながりを, 民族の確率分布の国間の類似度を Jensen-Shannon Divergence (JSD) により測ることで定量化する. Fig. 8 は, 国をノード, JSD をノード間のリンク及びリンクの太さとしたネットワークである. ここで, JSD が 0.09 以下のリンクは削除した. 民族構成が似ている国々が近くに集まっていることが分かる.

我々は, 空間的に広がる民族による国間の繋がりをとらえるために, MapEquation[18]を用いて, ネットワークにおけるクラスタを検出した. 検出された 4 つのクラスタを色分けし, Fig. 8 のアフリカ地図にマッピングした. ここで, グレーは十分なデータ数がなく分析から除いた国である. 地理的に近い国がクラスタとしてまとまっていることが読み取れる. このクラスタの形状は, アフリカにおける言語や宗教の空間分布に類似している.

7 おわりに

本研究では, 自然言語処理における言語モデルで利用される 3 層のLSTM レイヤーを重ねたRNNに, 全世界3700万人の苗字と国籍を学習させることにより, 国別に苗字生成の言語モデルを構築した. この言語モデルと各国の経済人（会社役員, 株主, 公的人物など）の人数をもとに, 苗字から民族（その苗字のルーツとなる国）を推定する分類器を作成した. 分類器を用いて, 各国の民族分布を推定し, この分布のエントロピーによって民族多様性を定量化できることを示した. さらに, 推定された国別の民族分布を国間で比較することによって, 国間の民族の類似度をネットワークで可視化し, クラスタリングをおこなうことで, アフリカ大陸では, 地理的に近い国がクラスタ化され, 空間的に広がる民族による国間の繋がり把握できた. 推定されたアフリカにおける民族の空間分布, 言語や宗教の空間分布に定性的に類似しており, 今後, 空間分布

の精度検証が課題である．

本研究では，ある一時点の民族の空間的な特徴を可視化したが，時間情報を持つデータを加えることによって，民族の時空間的な特徴をとらえることができる．これにより，民族多様性や多文化共生社会の視点から，歴史イベントや公共政策を振り返ることが可能になる．現在，世界中で移民問題や民族多様性に関して議論されている．本研究が，これらの議論の土台となる科学的エビデンスの１つになることを期待する．

謝辞

本研究の一部は，科学研究費補助金17H05123，17KT0034，及び，JSTさきがけネットワーク「人流ビッグデータによる地球規模の課題解決のための情報学と社会科学の融合基盤構築」の助成を受けている．

参考文献

1) Esteban González Burchard, Elad Ziv, Eliseo J Pérez Stable, and Dean Sheppard : The importance of race and ethnic background in biomedical research and clinical practice. The New England journal of medicine 348, 12, 1170/1175, (2003)
2) Donald A Barr : Health disparities in the United States: Social class, race, ethnicity, and health, 2nd ed., Johns Hopkins University Press (2014)
3) James Quesada, Laurie Kain Hart, and Philippe Bourgois : Structural vulnerability and health : Latino migrant laborers in the United States, Medical Anthropology 30, 4, 339/362 (2011)
4) Diane S. Lauderdale and Bert Kestenbaum : Asian American ethnic identification by surname, Population Research and Policy Review 19, 3, 283/300 (2000)
5) Pablo Mateos : A review of name-based ethnicity classification methods and their potential in population studies, Population, Space and Place 13, 4, 243/263 (2007)
6) Osei Appiah : Ethnic identification on adolescents' evaluations of advertisements, Journal of Advertising Research 41, 5, 7/22 (2001)
7) Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow : Ethnicity on Social Networks, ICWSM 10, 18/25 (2010)
8) Andrew J. Coldman, Terry Braun, Richard P. Gallagher : The classification of ethnic status using name information, Journal of Epidemiology and Community Health, 42,4, 390/395 (1988)
9) Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, Steven Skiena : Name-ethnicity classification from open sources, ACM SIGKDD, 49/58 (2009)
10) Jonathan Chang, Itamar Rosenn, Lars Backstrom, Cameron Marlow : Ethnicity on social networks, ICWSM, 10, 18/25 (2010)
11) Wendy Liu and Derek Ruths : What's in a name? using first names as features for gender inference in twitter, Analyzing microtext :AAAI spring symposium, (2013)
12) Marco Pennacchiotti and Ana-Maria Popescu : A machine learning approach to twitter user classification, ICWSM, 11, 1, 281/288 (2011)
13) Antoine Mazières and Camille Roth : Large-scale diversity estimation through surname origin inference, Bulletin of Sociological Methodology, 139, 1, 59/73 (2018)
14) Jinhyuk Lee, Hyunjae Kim, Miyoung Ko, Donghee Choi, Jaehoon Choi, Jaewoo Kang: Name Nationality Classification with Recurrent Neural Networks,IJCAI, 2081/2087 (2017)
15) Sepp Hochreiter, Jürgen Schmidhuber : Long short-term memory, Neural Computation, 9, 8, 1735/1780 (1997)
16) http://worldpopulationreview.com/
17) https://www.cia.gov/library/publications/the-world-factbook/
18) Martin Rosvall, Daniel Axelsson, Carl T. Bergstrom : The map equation, Eur. Phys. J. Special Topics 178, 13 (2009)