Vol.2019-MPS-123 No.62
Vol.2019-BIO-58 No.62
2019/6/19

地球環境の定量的記述のためのWikipediaデータ活用の試み

大林武†

概要：ゲノム解読技術の低価格化と情報処理技法の発展により、集団間のゲノム多型の差異を容易に検出できるようになってきた。その一方で、地域間の環境の違いは、気象学的データや土壌学的データに限られてしまい、定量的な情報が不足しており、このことが集団間のゲノム多型の違いの解釈可能性のボトルネックの一つとなっている。本研究では、ゲノム多型情報のメタ解析の有効性を向上を目的に、地球環境の地域差を機械的に導出するため、世界各地で使用される言語の類縁関係を用いたアプローチを試みる。特に、多数の言語が網羅されているWikipediaの言語間リンクの活用について検討する。

キーワード：進化，比較言語学，集団遺伝学

1. はじめに

ゲノム解読技術の進展により、世界規模での一塩基多様性[1]、構造多様性[2-4]ならびにマイクロバイオームの多様性[5]が、様々な生物種で報告されるようになった。特に民族移動の歴史[6]、あるいは動物の家畜化、植物の栽培化といった人類学的に重要な知見も多くもたらされている。ゲノムと環境の相互作用についても、医学的観点ならびに人類進化の観点から重要なトピックである。しかし、ゲノム配列多様性の偏りの同定が容易になり、データベースの整備が進んでいる[7]のに対して、それの解釈に必要な全世界規模での環境のデータは限られている。気象学的データ、土壌学的データ、さらに近年では人工衛星による夜間光のデータ[8]など、広範囲に測定できる地域環境データの利用価値は高く、今後の各種測定技術の進展が期待される。

さて、人類は数万年前より世界中に居住域を広げており、地域環境における生活に支障がないように各言語の語彙が整備されている。例えば、米食が重要な地域ではコメの状態に関する語彙が増える（英語の「rice」に対して、日本語では「イネ」「コメ」「ご飯」を使い分ける）などである。別の言い方をすれば、世界中に存在する様々な言語の語彙を地域環境のセンサーをして捉えることが可能かもしれない。また、複数の人類集団が接触するときには、片方の言語の語彙がもう片方の言語の語彙として取り込まれる（借用）ことがある。語彙の借用は、政治的な強弱関係のような社会的背景に他に、地域に特徴的な動植物や文化が輸出される場合にも観察され、借用語の借用パターンを解析することで、地域環境の特徴付けができる可能性も考えられる。

異なる言語において同じ意味をなす言葉の対応づけは、コメとriceの意味範囲が異なるように簡単にはいかない。これまでに各種言語を専門とする言語学者チームによって多様な言語データが収集、解析されている。54名の言語学者の結果であるWorld Loanword Databaseは、41言語の約2000語の言語的由来を記述しており、その解析から、動詞よりも名詞の方が借用が起きやすい、身体に関わる語彙は借用が起きにくいなどの結果が示されている[9]。このような専門家による緻密な解析は、言語学には不可欠であるが、データの収集に多大な労力がかかる点が研究の大きな律速にある。Ethnologue[10]は、その7000言語を網羅する巨大なデータベースを、非営利目的も含めて有償とすることで継続開発する体制を構築している。近年の様々な領域でデータの大規模化オープン化が進む中で、自然言語解析に関しても、フリーアクセスで集合知としてデータが蓄積するタイプのデータベースをもとにした解析が可能になれば、専門家によるドメイン特化型の研究アプローチを補完するアプローチとなる可能性がある。Wikipediaは世界最大級の集合知データベースであり、比較言語学的解析においても、幅広い自然言語における単語の出現頻度データのソースとして用いるなど、目的に応じて研究に利用されつつある[11]。

本研究では各言語が有する語彙が、環境や文化ならびにその歴史とどのように関連するかを網羅的に探索する手法の開発を行う。特にWikipediaを代表とする集合知データベースの活用について検討する。

2. 結果

2.1. 任意の言語ペアの綴り字の対応

異なる言語は異なる文字セットを用いて音素を文字として表現する。語彙の借用は、本質的には音声的な借用であり、借用元と借用先の言語における文字の対応関係は自明ではない。そこでまず、任意の言語間の文字の対応関係を導出するために、Wikipediaで最大のページ数を持つ英語ページのタイトル語をもとに、その多言語リンク関係に基づき、複数の言語における意義的等価語句セットを作成した。異なる言語では一般に文字数が異なるため、文字の対応を調べるためには文字列を整列する必要があるが、文字の対応が未知の言語間では文字列整列の評価ができないため、整列もできない。この問題を回避するため、各単語の先頭文字は整列していると仮定し、異なる言語の綴り字の対応関係を導出した。例えば、元素のコバルトは、

†東北大学情報科学研究科

Vol.2019-MPS-123 No.62
Vol.2019-BIO-58 No.62
2019/6/19

Cobalt(英語、フランス語)、Kobalt(ドイツ語)、кобальт(ロシア語)のように各言語で綴るため、その先頭文字の比較から、各言語ペアにおける綴りパターンを学習できる。学習に用いた単語ペアは、最も多い英語 - フランス語間で1359540単語ペア。最も少ないカタロニア語-クロアチア語間で60272単語ペアであった。本研究では、ランダム頻度の4倍より多く観察された綴り字の対応を、その言語ペアにおける対応のある綴り字とした。

2.2. 単語の一致率

次に単語の形態的一致率を算出した。言語Aの文字列a $[a_0, a_1, \ldots, a_n]$と、言語Bの文字列b$[b_0, b_1, \ldots, b_m]$の相同性は、2.1で求めた(a_i, b_j)の対応スコアをもとに動的計画法により決定する。ここでは、対応のある文字ペアのスコアを1、ギャップスコアを-0.2とし、最終的なスコアをその文字列ペアの理論最大値で割った値を、単語ペア(a,b)の一致率とした（一致なし0、完全一致1）。動的計画法を行う際に周期性を考慮することで構文の異なる言語の一致率も算出できるようにした。単語の一致率の例を表1に示す。

言語1	言語2	一致率
(英) Lepton_number	(仏) Nombre leptonique	0.71
(英) Lepton_number	(独) Leptonenzahl	0.63
(英) Smooth_number	(仏) Entier friable	0.25
(英) Smooth_number	(独) Glatte Zahl	0.18

表 1　単語一致率の例

2.3. 単語一致率に基づく言語類似性

2.2で計算した綴り字一致率の妥当性を検証するために、言語間の平均綴り字一致率をもとに言語類似性を算出した。百科事典であるWikipediaのタイトル語は固有名詞に偏る傾向があるため、ここでは、Wikipedia、Wiktionary、WordNetの共通語であり、解析対象とした33言語全てに対応するWikipediaページが存在する199語の平均一致率を用いて、距離を1 - (平均単語一致率)とし、群平均法で階層的クラスタリングを行った結果を示す（図1）。この言語類似性はゲルマン語族、ロマンス語族、スラブ語族、アラビア系、アジア系の言語分類とよく一致しており、本研究での単語一致率が概ね妥当であることを示している。

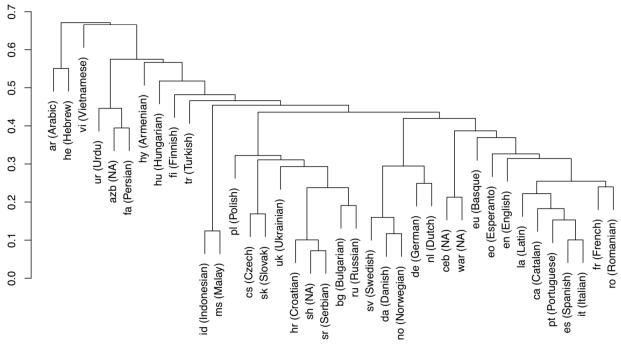

図 1　単語一致率に基づく言語類似性

平均単語一致率は言語間の距離をよく反映している一方で、単語一致率をもとにした単語の類似パターンは、単語によって様々である。例えば、戦争(war）、剣(sword)という単語において、アジア系言語がロマンス語を採用している傾向があるなど、単語一致率のパターン（借用パターン）と単語のカテゴリに関係性を示唆する例が見つかった。

3. まとめ

本研究では、世界最大の集合知データベースであるWikipediaを用いて、綴り字の一致度を算出し、そのパターンと概念カテゴリに関係を示唆する結果を得た。今後、既存データベースとの無矛盾性を検証していくとともに、解析手法の高度化を進め、世界中に広まっている自然言語に埋め込まれた人類史や地域環境の情報を抽出、ならびに集団遺伝学データとの比較解析へと発展させていきたい。

参考文献

1. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010, vol. 467, p1061-73.
2. Sudmant P.H., et al. Diversity of human copy number variation and multicopy genes. Science. 2010, vol. 330, p. 641-6.
3. Sudmant P.H., et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2018, vol. 349, p. aab3761-1.
4. Narang A., et al. Extensive copy number variations in admixed Indian population of African ancestry: potential involvement in adaptation. Genome Biol Evol. 2014, vol. 6, p. 3171-3181.
5. Pasolli E. et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell 2019, vol. 176, p. 649-662.
6. Pagani L. et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature 2016 vol. 538, p. 238-242.
7. Murga-Moreno J. et al. PopHumanScan: the online catalog of human genome adaptation. Nucleic Acids Res. 2019, vol 47, p. D1080-D1089.
8. Kyba C. et al. High-Resolution Imagery of Earth at Night: New Sources, Opportunities and Challenges. Remote Sens. 2015, vol. 7, p. 1-23.
9. Haspelmath P., et al. (eds.) 2009. World Loanword Database. Leipzig: Max Planck Institute for Evolutionary Anthropology.
10. Lewis M. P. et al. (eds.) 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International.
11. Mahowald K. et al. Word Forms Are Structured for Efficient Use. Cogn Sci. 2018, vol. 42, p. 3116-3134.