

て各地域、各言語へと広がっていく。主要な伝播経路の種類が有限だとすれば、その伝播経路の種類に対して十分多くの「語彙の地理的起源」と「各言語における呼び名」がペアとなったデータを用いることで、その関係を学習することが可能である。

我々はこれまでに、83の陸上哺乳類の生息域と、表1の20言語におけるそれらの呼び名の関係から、語彙の多様性と単語長が生息域の広さと相関することを示してきた[3]。本稿では、同じデータを用いて、野生動物の呼び名から生息域を予測する予測器を構築し、その特性を解析する。

表1 解析対象言語

ロマンス語派		ゲルマン語派		スラブ語派	
ca	Catalan	de	German	bg	Bulgarian
eu	Basque	en	English	ru	Russian
fr	French	nl	Dutch	pl	Polish
pt	Portuguese	sv	Swedish		

ウラル語派		西アジアの言語		東アジアの言語	
fi	Finnish	ar	Arabic	ko	Korean
		fa	Persian	id	Indonesian
		he	Hebrew	ms	Malay
		tr	Turkish	vi	Vietnamese

2. 手法

2.1. 多言語名称データ、生息域データ

多言語名称データ、生息域データ共に、我々の以前の解析[3]を踏襲する。すなわち、多言語名称データとしてWikipediaの言語間リンク[8]を用い、生息域データとして、国際自然保護連合にて公開されている陸上哺乳類の生息域データ[1]を用いる。特に、2箇所以上の生息報告があり、かつ表1の20言語中で18言語以上にWikipediaの項目がある(当該生物種の名称がある)83生物種を解析対象とした[3]。生物種による観測総数の違いを軽減するため、生息地のデータは図3の8地域に粗視化し、生息の有無を0/1で表現している[3]。

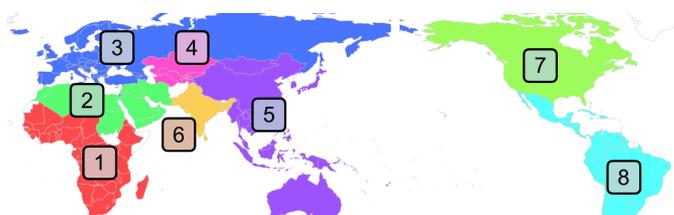


図3 生息域の粗視化

2.2. 言語間の単語の類似度

言語間の単語の類似度は以前の手法[3]を踏襲した。すなわち、

- 異なる言語の文字の対応：Wikipediaの見出し語の多言語対応から、異なる言語における同一項目の先頭文字の対応関係を集計し、ランダム類似度との違いを基に、文字の対応の情報量を求める。
- 異なる言語の文字列の対応：文字の対応の情報量を基

に、動的計画法を用いて文字列(単語)の整列と一致度スコアを算出する。最大可能スコアとランダムスコアに基づき、値が0から1になるように補正する(0:ランダムレベル、1:完全一致)。

2.3. 多層パーセプトロンによる生息域の学習

単語から生息域を学習するモデルとして、100次元の中間層を持つ多層パーセプトロン(MLP)を用いた。入力には表2に示す22次元ベクトル(全言語平均類似度と平均単語長が含まれているので実質20次元)であり、出力層は8つの地域それぞれについて、生息しているかの有無である。例として、ライオン、コアラ、フクロテナガザルの入出力値を表2に示す。言語別の平均類似度と単語長は平均0と標準偏差1に正規化している。全言語の平均類似度は0(ランダムレベル)から1(完全一致)までの値をとる。

中間層の活性化関数にはReLU、出力層の活性化関数にはシグモイド関数を用いた。出力層がシグモイド関数なのは、8地域各々に対して生息確率を予測するためである。損失関数は平均二乗誤差で、ドロップアウト(50%)により正則化を行った[9]。

表2 学習の入出力ベクトル

		ライオン	コアラ	フクロテナガザル		
入力	単語類似度	全言語の平均類似度	0.248	0.723	0.623	
		言語1(ca)の平均類似度	1.438	0.678	2.300	
		⋮	⋮	⋮	⋮	
		言語20(vi)の平均類似度	1.243	4.982	-0.648	
	単語長	平均単語長	-1.299	-1.086	-0.47	
		言語1(ca)の単語長	-1.233	-1.053	-0.694	
		⋮	⋮	⋮	⋮	
		言語20(vi)の単語長	-1.31	-1.31	-0.762	
	出力	生息域	領域1での生息	1	0	0
			領域2での生息	1	0	0
領域3での生息			0	0	0	
領域4での生息			0	0	0	
領域5での生息			0	1	1	
領域6での生息			1	0	0	
領域7での生息			0	0	0	
領域8での生息			0	0	0	

3. 結果

3.1. 多層パーセプトロンによる学習

全データセットをランダムに60%(50生物種)、20%(16生物種)、20%(17生物種)に分割し、それぞれ学習セット、バリデーションセット、テストセットとして用いた。図4に示す学習セットとバリデーションセットの損失傾向をもとに、40エポックで学習を打ち切った。40エポックでの損失は学習セットで0.084、バリデーションセットで0.153であった。

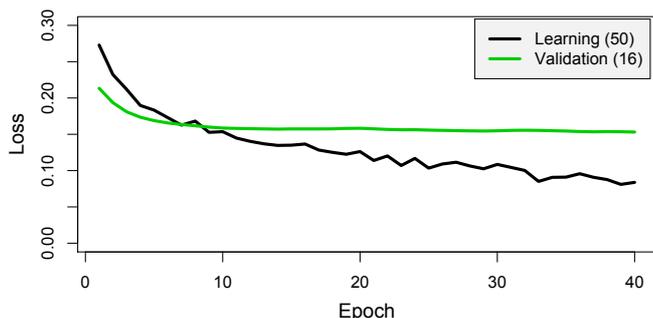


図4 学習曲線

3.2. 予測結果

テストデータセット (17生物種) に対する予測結果を表3に示す。各領域の生息確率を10倍して四捨五入することで0から10までの数値として表している。また、実際の生息域を黄色で示し、生物種は二乗誤差の小さい順に並べている。もっとも高精度に生息域を予測できたのはマレーシアとインドネシアのスマトラ島に生息するフクロテナガザル (*Symphalangus syndactylus*) であり、2番目はインドネシアからミャンマーに至るアジア地域に生息するバンテン (*Bos javanicus*, 牛の一種) であった。上記2種に、中国に生息するジャイアントパンダ (*Ailuropoda melanoleuca*) とベンガルヤマネコ (*Prionailurus bengalensis*) を含め、アジ

表3 MLPによる予測

	R1	R2	R3	R4	R5	R6	R7	R8	Error
<i>Symphalangus syndactylus</i>	1	0	0	0	9	1	0	1	0.008
<i>Bos javanicus</i>	1	0	0	0	5	0	0	2	0.041
<i>Antilope cervicapra</i>	1	0	0	0	2	6	4	1	0.050
<i>Ursus americanus</i>	0	0	1	1	1	2	4	3	0.064
<i>Ailuropoda melanoleuca</i>	5	1	0	1	4	2	1	1	0.097
<i>Bassariscus astutus</i>	0	0	0	0	0	1	1	1	0.106
<i>Hippotragus niger</i>	2	1	1	0	0	3	3	2	0.115
<i>Cebuella pygmaea</i>	1	0	0	1	2	2	2	1	0.118
<i>Leptailurus serval</i>	7	1	0	0	2	2	0	2	0.123
<i>Crocota crocuta</i>	1	1	1	2	3	2	1	2	0.129
<i>Prionailurus bengalensis</i>	1	1	1	2	2	4	2	2	0.129
<i>Diceros bicornis</i>	1	0	0	1	2	2	2	1	0.130
<i>Ondatra zibethicus</i>	3	5	4	8	4	3	2	2	0.198
<i>Felis chaus</i>	2	2	3	7	4	3	1	1	0.243
<i>Ceratherium simum</i>	1	0	0	1	2	2	3	2	0.248
<i>Gulo gulo</i>	1	6	5	7	7	3	1	1	0.269
<i>Felis margarita</i>	2	2	3	6	5	2	3	2	0.333
平均	2	1	1	2	3	2	2	1	0.141

表4 平均生息率のみによる予測

	R1	R2	R3	R4	R5	R6	R7	R8	Error
<i>Symphalangus syndactylus</i>	2	2	2	3	4	2	2	1	0.080
<i>Bos javanicus</i>	2	2	2	3	4	2	2	1	0.080
<i>Ailuropoda melanoleuca</i>	2	2	2	3	4	2	2	1	0.080
<i>Prionailurus bengalensis</i>	2	2	2	3	4	2	2	1	0.080
<i>Antilope cervicapra</i>	2	2	2	3	4	2	2	1	0.137
<i>Hippotragus niger</i>	2	2	2	3	4	2	2	1	0.137
<i>Crocota crocuta</i>	2	2	2	3	4	2	2	1	0.137
<i>Diceros bicornis</i>	2	2	2	3	4	2	2	1	0.137
<i>Ursus americanus</i>	2	2	2	3	4	2	2	1	0.145
<i>Bassariscus astutus</i>	2	2	2	3	4	2	2	1	0.156
<i>Cebuella pygmaea</i>	2	2	2	3	4	2	2	1	0.156
<i>Ondatra zibethicus</i>	2	2	2	3	4	2	2	1	0.183
<i>Gulo gulo</i>	2	2	2	3	4	2	2	1	0.183
<i>Leptailurus serval</i>	2	2	2	3	4	2	2	1	0.221
<i>Felis chaus</i>	2	2	2	3	4	2	2	1	0.221
<i>Ceratherium simum</i>	2	2	2	3	4	2	2	1	0.221
<i>Felis margarita</i>	2	2	2	3	4	2	2	1	0.330
平均	2	2	2	3	4	2	2	1	0.158

ア地域 (R5) に生息する動物4種の予測精度が軒並み高い (表3)。これが個々の動物の「呼び名」に基づく予測になっているか判断するため、個々の動物について予測するのではなく、各領域の平均生息率を用いてどの生物種にも同じ予測をする予測器を作成した (表4)。全体の4割の動物がアジア地域 (R5) に生息するため、アジア地域の生息予測は他の地域に比べて当てやすいことがわかる。アジア地域4種の平均二乗誤差は、ランダム予測でも0.080であることを考慮すると、東南アジアに生息するフクロテナガザルとバンテンでは、MLPにより予測精度が向上したのに対して、中国に生息するジャイアントパンダとベンガルヤマネコではランダムよりも悪い予測精度しか得られていない。後者2種はアジア地域 (R5) 中でも、ユーラシア地域 (R3) に近いことが、精度が悪い一因だと考えられる。平均二乗誤差のテストデータセット全体における平均値はMLPによる予測で0.141、平均生息率による予測で0.158であり、平均値としては大きな改善はないと言える。

より詳細に比較するために、MLPによる予測と、コントロールである平均生息率による予測の平均二乗誤差を生物種ごとに比較を行った (図5)。各生物種を示す各点は大方対角線に乗っており、ランダム予測とMLP予測の平均二乗誤差は大差ないことがわかる。8領域に粗視化した領域において、単一の領域に生息している生物種 (黒)、複数の領域に生息している生物種 (赤) を色で区別して示している。全体では予測性能の改善がないのに対して、単一の生物種に対しては、ランダム予測における平均二乗誤差を改善していることがわかる ($p < 0.05$, 対応のあるt検定)。一方で、複数の地域に生息している生物種の予測はランダ

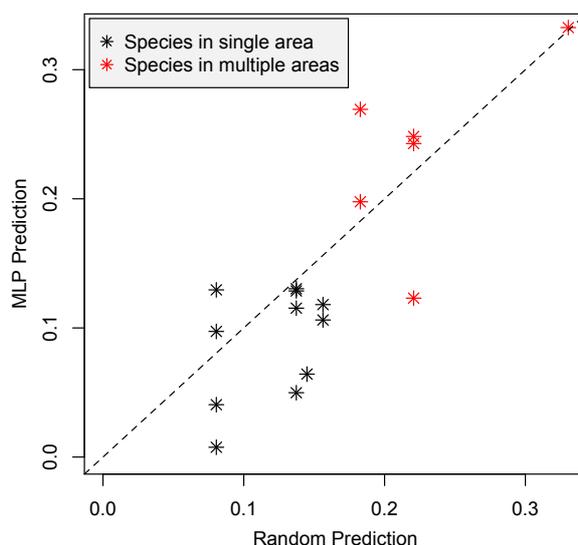


図5 平均二乗誤差の比較

表5 学習に用いた66生物種の生息域パターン

領域数	1	2	3	4	5	6	7	8
生物種数	34	16	8	6	1	1	0	0
観察された組合せ数	7	10	6	5	1	1	0	0
理論的な組合せ数	8	28	56	70	56	28	8	1

ム予測と変わらない大きさの平均二乗誤差となっている。これは生息域が複数地域に渡る場合、その各々の地域から呼び名が伝播するため、伝播経路が複雑になるのに対して、それを学習するだけのデータがないことが原因であろう。表5は、学習セットとバリデーションセットに含まれる生物種の生息域の領域パターンを示したものである。8領域中1領域のみに生息する生物種は34種あり、平均して1領域あたり5生物種を学習に利用することができる（今回のデータセットには領域2のみに生息する生物種が含まれていない）。一方で、2領域以上に生息する生物種の場合には、領域の組み合わせパターンが多くなり、想定されるパターンを十分に網羅できていないと考えられる。各パターンあたりの生物種数もほぼ1種であり、普遍的なルールを学習するのは困難な状況である。単一の領域に生息する動物のみに絞った予測問題にするなど、データサイズに見合った問題で検証実験も有用だと思われる。

表6 入力ベクトルの寄与（単語長の影響）

	R1	R2	R3	R4	R5	R6	R7	R8
French	-0.3	-0.3	-0.3	-0.7	0.2	-0.3	0.1	0.1
Dutch	0.0	-0.4	-0.7	-0.7	-0.8	0.1	0.1	0.3
Russian	0.5	0.1	0.1	0.0	-0.6	-0.3	0.0	-0.1
Hebrew	-0.3	-0.7	0.3	-0.2	-0.6	-0.4	0.0	0.2
Vietnamese	0.8	0.3	-0.1	-0.2	-0.8	0.3	-0.1	0.2
Indonesian	-0.2	0.3	0.5	0.7	0.6	0.4	0.9	0.1

表7 入力ベクトルの寄与（平均類似度の影響）

	R1	R2	R3	R4	R5	R6	R7	R8
Bulgarian	0.2	-0.1	-0.6	-0.2	-0.3	0.0	0.4	0.2
Polish	0.1	-0.1	-0.4	-0.6	0.2	-0.3	-0.2	0.0

3.3. MLPの重み

学習がうまく進んでいるかを判断する別の観点として、学習後の重みについて吟味する。今回用いた3層パーセプトロンでは、1層目と2層目を繋ぐ重み行列（42行100列）と2層目と3層目を繋ぐ行列（100行8列）が重みとして学習される。実際には活性化関数による非線型変換が含まれるが、これが線形変換だったとすると、この2つの行列を乗算することで、入力ベクトルの各要素の出力結果への寄与を42行8列の行列として推定することができる。特徴的な入力項目に注目するため、寄与の絶対値が0.6を超える入力項目について、各地域への影響度を表6、表7に示す。

単語長には選択圧がかかっており、その地域（言語）に重要である単語ほど、単語長が短くなる傾向がある[10]。ここから現地に生息する生物種は、生息していない生物種よりも重要であり、そのため単語長が短くなりやすいと期待できる。今回のMLPの学習では、その傾向が確認できる（表6）。すなわち、各言語における単語長が短いと（すなわち重要な単語だと）、その言語の地域に生息する確率が上がる（重みが負になる）。逆に単語長が長いと、遠方に地域に生息する確率が上がる（重みが正になる）。

表7は平均類似度の影響を示す。単語長より影響の度合い

が小さいものの、東ヨーロッパのブルガリアやポーランドにおける呼び名が没個性である（平均類似度が高い）と、ヨーロッパからユーラシア（R3, R4）に生息している可能性は減る（重みが負になる）ことを示しており、こちらも想定通りに学習されていることがわかる。呼び名の類似度はペアで定義されるため、20言語、190言語ペアの情報として計算しているが、今回MLPの入力として使用するにあたって、各言語について平均類似度を取ることで、20次元に縮約している。呼び名の伝播経路は線で表現されるものであり、言語ごとに縮約する今回の次元圧縮方法についても今後検討を進める。

4. まとめ

我々はこれまでに呼び名から生息域の広さを推定できることを示してきた[3]。本報告では、生息域の広さだけでなく、生息域そのものの予測を行った。学習の重み行列（表6, 表7）は想定通りに学習されていることを示している一方で、全体的な予測性能の改善に関しては、単一地域に生息する生物種に対する予測精度向上にとどまった（図5）。今後、複数地域に生息する生物種についての予測精度向上のためには、何よりデータ数の不足を解消することが課題となる。地域区分の見直しとも合わせて、呼び名と生息域の関係、延いては任意の語彙とその地理的起源の関係を通じて、人類による地球規模の情報伝播の全体像を明らかにしていく。

参考文献

1. IUCN 2019. The IUCN Red List of Threatened Species. Version 2019-1. <http://www.iucnredlist.org>. Downloaded on 3 June 2019.
2. 大林武, 「地球環境の定量的記述のためのWikipediaデータ活用の試み」, 情報処理学会研究報告バイオ情報学, vol. 2019-BIO-58, no. 62, p.1-2, ISSN:2188-8590.
3. 大林武, 山田和範, 長野明子, 「言語間語彙比較に基づく野生動物の生息域推定の試み」, 情報処理学会研究報告人文科学とコンピュータ, vol. 2019-CH-121, no. 7, p1-5, ISSN:2188-8957.
4. Oxford English Dictionary: <https://www.oed.com/aboutthisentry/108800>
5. American Heritage Dictionary: <https://www.ahdictionary.com/word/search.html?q=lion>
6. John Simmons. Twenty-six Ways of Looking at a BlackBerry: How to Let Writing Release the Creativity of Your Brand. A&C Black. p. 173. ISBN:9781408105962
7. Oxford English Dictionary: <https://www.oed.com/aboutthisentry/104226>
8. Wikipedia Database Backup Dump: <https://dumps.wikimedia.org/enwiki/20190401/enwiki-20190401-langlinks.sql.gz>
9. Ian Goodfellow and Yoshua Bengio and Aaron Courville. (2016) Deep Learning, Chapter 7, MIT Press.
10. Mahowald K. et al. Word Forms Are Structured for Efficient Use. Cogn Sci. 2018, vol. 42, p. 3116-3134.