類型論的分析のための言語特徴値の推定

高村大也
東京工業大学
takamura@pi.titech.ac.jp

永田亮
甲南大学
nagata-nlp2015@hyogo-u.ac.jp

1 はじめに

地球上には非常に多くの言語が存在する．それらは，語彙だけでなく，統語規則，発音規則，書記体系など様々な点で互いに異なっている．そのような様々な観点から，言語の分類体系や言語間の類似性を議論する学問分野が言語類型論である [1]．言語類型論で基になるのは各言語の種々の特徴であり，それら一つ一つはこの分野の研究者によるフィールドワークを含む調査と分析により明らかにされている．蓄積された知見の一部は，データベースという形で集約され，さらなる研究に役立てられている．そのようなデータベースのうち大規模なものとして The World Atlas of Language Structures (WALS)[2] が挙げられる．WALS には，破裂音の有無，単語の性の数，支配的な語順 (SVO か，SOV か，など)，基本色の数など，多種多様な特徴 (feature) が登録され，各言語に対するその特徴値が記載されている．しかし，2 節で詳しく述べるように，WALS はデータとして疎である．すなわち，上記のすべての言語についてすべての特徴値が記載されているわけではなく，特徴値が未記載であることの方がむしろ多い．特徴値を決定することは，上述のようにフィールドワークを含む調査と分析が必要になることが多く，一般に容易ではない．

本稿では，機械学習による分類手法を用いて，WALS に登録された各言語の各特徴値が，それ以外の特徴値からどの程度正確に推定することができるかを調べる．また，分類器の適用が効果的な特徴はどれかについても調査する．このような調査をする動機付けとしては，(i) 未記載の特徴値に対して推定値が与えられることで，WALS を利用した統計処理などが可能になる場合がある [1]，(ii) 現状の WALS の冗長性について示唆が得られる，(iii) 特徴値推定のための分類モデルを分析することにより，特徴間における傾向や分類の道標となる言語に関する知見が得られる，などが挙げられる．

[1] もちろん，得られるのは推定値であり分類誤りを含むので，扱いには注意を要する．

2 WALS に関する統計値

WALS には，2014 年 7 月の時点で，2,679 という多くの数の言語が記述されており [2]， また特徴数は 192 である [2]．特徴値が記載されている割合は 17%であり，残りの 83%は未記載である．これを直観的に示すために，図 1 に特徴–言語行列を図示する．この行列は各行が特徴に，各列が言語に対応し，特徴値が記載されている場合は黒い点，そうでない場合は白い点で表されている．この図では白い点が大半を占めており，WALS が疎であることを視覚的に示している．

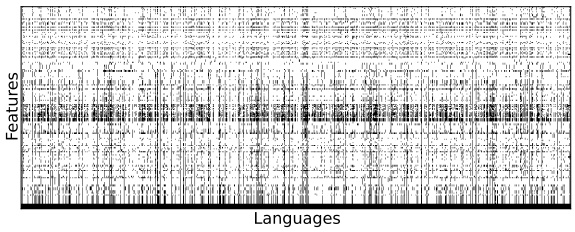

図 1: 特徴–言語行列 (各行が特徴に，各列が言語に対応する．黒い点：特徴値が WALS に記載されている，白い点：記載されていない)

3 関連研究

言語類型論については膨大な文献があるが，ここでは WALS を用いた数理的な研究のいくつかについて触れる．

Daumé and Campbell [4] は，WALS の各特徴を確率変数とみなし，ある特徴が別の特徴に対して影響を与えるかどうかの変数を新たに加えた確率モデルを考えた．影響を与えやすい特徴対には言語普遍性が存在することを示唆しているとしている．彼らはさらにノ

[2] 現在地球上には約 7,000 の言語があるといわれているので，半数以上の言語は未掲載である．

イズ項を加える，言語系統の情報を追加するなどのモデル拡張を行っている．Daumé [3] は，ノンパラメトリックベイズ手法を用いて，地理的な近さと系統的な近さの両方をモデルに組み込み，各特徴が地理的な要因で決定しやすいか否かを示す指標を算出した．

Roy et al. [5] は接置詞に着目し，コーパスを用いた教師無し手法により，各言語が前置詞を用いるか後置詞を用いるかを推定する手法を提案している．

本稿で試みているような特徴値の識別についての包括的な研究はこれまでない．

4 分類実験

機械学習による分類手法を用いて，WALS に登録された各言語の各特徴値が，それ以外の特徴値からどの程度正確に推定することができるかを調べる．そのために，各特徴に対し，その値が既知であるような言語集合を用いた所謂 leave-one-out 法により分類正解率を算出する．すなわち，そのような言語集合内のある言語を評価事例，それ以外の言語を訓練事例として，評価事例の言語の特徴値が正しく推測できるかを調べる．このような実験を値が既知であるすべての言語に対して行い，分類正解率を算出する．分類器構築のための素性ベクトルは，対象としている特徴以外の特徴を二値化して作成する．つまり，特徴は一般に多値であるので，それぞれの値について，その値であれば 1 となり，そうでなければ 0 となるような素性を考える．

4.1 従属特徴，および特殊な特徴

WALS には，様々な角度や粒度で言語特徴が記述されている．完全に等価な特徴対は存在しないものの，特徴間の関連性については，体系的に整理されていないのが実情である．例えば，特徴 81A は，S(ubject) と V(erb) と O(bject) が，SOV や SVO などのいずれの順序で出現するかに関するものであるのに対し，特徴 82A は S と V の順序に関するものである．両者の違いは単純に粒度であり，例えば前者の SVO ならば，後者は VO に決定してしまう．本稿では，この例のように，ある特徴の値が別の特徴が取りうる値を制限してしまうとき，前者は後者の **従属特徴** であるということにする．ただし，ここでいう従属とは，粒度の違いのような本質でない依存性のことであり，言語普遍性のような学問的に興味深い依存性 (例：OV なら後置詞を用い，形容詞-名詞，および関係節-主要部名詞の語順である) とは異なる．

上記の特徴 81A と 82A の例のように，従属特徴が特徴値を完全に決定してしまう場合，推定しようとしている特徴の値が実際には既にわかっていることを意味する．つまり，実際の応用では起こりえない状況である．また，完全に決定してしまうのでなく，取りうる値を制限するような場合でも，学習される分類規則は言語一般の性質をとらえるという観点からは，全く重要でない．以上のことから，我々は従属特徴を除いた状況でも，実験を行う．より具体的には，各特徴に対する従属特徴集合を記述した従属特徴リスト [3] を作成し，推定対象である特徴の従属特徴は，分類器の素性として用いないことにする．本稿の実験では，従属特徴を素性から除去する際は，その従属特徴と同じ chapter[4] に属する特徴もすべて素性集合から削除している．

また，ごく一部の言語にのみ関連する特徴も存在する．例えば，特徴 139A と 140A は手話に関する特徴であり，それらの値を他の言語について推定しようとすべきでない．我々は，139A と 140A，および同様の理由から 10B と 39B を分類実験に用いないことにする．また，chapter 90，chapter 143，chapter 144 については，chapter 内に関連する特徴が多く存在するので，それぞれ特徴 90A，143A，144A のみを使用し，90B–G，143B–G，144B–Y は分類実験に用いないことにする．

4.2 同系統言語の扱い

同系統の言語は，同じ特徴値を持つことが多い．よって，ある言語の特徴値を推定するときに，それと同系統の言語が訓練データに存在すれば，一般に分類性能は上がるだろう．しかし，応用を考えると，同系統の言語が少ないあるいは存在しないからこそ，特徴値を推定したいような状況も考えられる．また，ある系統の特徴値の分布を学習してしまうことにより，言語的普遍性が捉えられなくなる可能性がある．よって，本稿では同系統の言語を訓練データから除外する設定の実験も行う [5]．具体的には，WALS に登録されている語族 (family) が同一な言語を同系統とみなす．

上述したのは同系統の言語を訓練事例として用いないという旨であるが，語族 (family) および語派 (genus) の情報は素性としても一切用いないことにする．

[3] http://lr.pi.titech.ac.jp/~takamura/typology.html より入手可能．

[4] WALS の特徴は，chapter という単位にグループ化されている．多くの chapter は特徴を一つだけ含む．

[5] 本稿では考慮に入れていないが，より厳密にするためには，地理的な近さの影響も考慮する必要がある [1].

表 1: 分類正解率．表中の Yes は，同系統言語を訓練データとして用いた場合，従属特徴を素性として用いた場合に対応する．macro はマクロ平均を表し，micro はマイクロ平均を表す．

同系統言語	従属特徴	分類正解率 (%)	
		macro	micro
No	No	58.67	72.09
Yes	No	61.92	74.66
No	Yes	65.07	80.31
Yes	Yes	67.91	82.94
ベースライン		55.52	53.77

4.3 実験設定

分類器としては Support Vector Machine (SVM) を用いる．SVM を多値分類に用いるため，one-versus-rest 法を用いた．正則化パラメータ C については，0.01，0.1，1，10，100 の 5 値から最適な値を選んだ．対象にしているデータ全体で一度しか出現しない特徴値は削除した．また，SVM のある分離平面を学習する際に正例もしくは負例の数が 0 になる場合は，学習不可能であるとして，分離平面の構築を中止した．そのような場合は，対応する特徴値は取り得ないとみなす．

4.4 実験結果および考察

まず，同系統言語の利用および従属特徴の利用に関していくつかの実験設定があるので，それぞれについての分類正解率を表 1 にまとめる．ベースラインとしては，それぞれの特徴について最多数の特徴値を出力する分類器を設定した．同系統言語及び従属特徴を使用した場合は約 83%，どちらも使用しない場合は約 72%の (マイクロ平均) 正解率となっている．ベースラインと比較すると，特にマイクロ平均の向上が著しい．マイクロ平均がマクロ平均を大きく上回っていることから，評価データのサイズが大きい，すなわち (leave-one-out 法を用いているため) 訓練データのサイズが大きいほど正解率が高くなっていることわかる．

加えて，上述のベースラインからの分類器の正解率の上昇が最も大きい 10 特徴を，表 2 に示す．高いものでは 30 ポイントを上回る正解率上昇があることがわかる．また，上位に挙げられているものの多くは，主要部と補部の順序に関するものとして解釈できる．

各特徴について，実際に値が未記載であるような言語のその特徴値を推定するため，あらためて値が既知である言語をすべて用いて分類器の構築を行った．これによる特徴値の推定結果などは，上記 leave-one-out 実験の結果と共にリソースとして公開する [6]．

また，例として，WALS に記載されていない日本語の特徴について，学習されたモデルを用い実際に推定した結果を分析する．結果は表 3 の通りである．まず，14A，15A，16A，17A は強弱アクセントに関するものであり，高低アクセントに基づいた言語である日本語にはそもそも定義されないと考えてよい [7]．文法的性の有無 (30A，31A，32A) については，無であると正しく推定された．ここで，これらの特徴値の推定の際には，お互いを訓練事例として用いていない (すなわち，例えば 30A の推定には 31A および 32A は用いていない) ことに注意されたい．これら特徴値の分離平面で重みが大きかった素性として，類別詞 (55A) が選択的あるいは義務的になっていることが挙げられる．宮本 [6] の考察では，文法的性と類別詞は同様の働きを持つので，それらの存在は排他的になりやすいことが示唆されており，これにより実験結果が説明可能である．また特徴 81A は，S，V，O に関して二つの順番が支配的であるような言語についてのものであり，それにあてはまらない日本語には定義できない．表記体系 (141A writing systems) については，音節的体系 (syllabic, 日本語ではひらがな及びカタカナ) が存在すると推定された．この特徴の訓練事例は 6 言語しかなく，学習には十分な量とはいえない．また，この 6 言語はすべて音節的体系とアルファシラビック (alphasyllabic) のいずれかに属すので，実際の正解である表語文字体系と音節的体系の混合 (mixed logographic-syllabic) を推定することは不可能である．特徴 110A 及び 130B については，より専門的な分析を必要とする．

5 おわりに

言語の特徴を集約したデータである WALS に対して，未記載の特徴値を推定する実験を行った．その際，推定に影響を与える要素として，同系統言語及び従属特徴の存在を挙げ，実験結果における影響を調べた．従属特徴のデータ，実際の推定結果のデータは，さらなる研究の発展のために公開する．事例調査として，日本語の未記載特徴値の推定結果について考察を行った．

[6] http://lr.pi.titech.ac.jp/~takamura/typology.html より入手可能．

表 2: 分類器とベースラインの正解率の上昇 (上位 10 特徴)．同系統言語及び従属特徴は使用していない．

特徴 ID	特徴名	上昇 (%)
85A	Order of Adposition and Noun Phrase	36.43
83A	Order of Object and Verb	33.51
95A	Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase	31.52
88A	Order of Demonstrative and Noun	29.74
97A	Relationship between the Order of Object and Verb and the Order of Adjective and Noun	28.42
86A	Order of Genitive and Noun	26.58
89A	Order of Numeral and Noun	25.85
144A	Position of Negative Word With Respect to Subject Object and Verb	25.06
29A	Syncretism in Verbal Person/Number Marking	24.24
51A	Position of Case Affixes	23.96

表 3: 日本語の未記載特徴の推定値．推定された特徴値のカラムの先頭に書かれた数字は，WALS により与えられている特徴値の ID である．また，ここでは正則化パラメータは $C = 1$，従属特徴は使用しない場合の結果を示している (日本語は孤立言語であるため同系統言語の使用/未使用は結果に影響を与えない)．スコアは SVM の識別関数の出力である．

特徴 ID	特徴名	推定された特徴値	スコア
14A	Fixed Stress Locations	7 Ultimate	1.680
15A	Weight-Sensitive Stress	5 Unbounded: Stress can be anywhere	1.445
16A	Weight Factors in Weight-Sensitive Stress Systems	1 No weight	0.356
17A	Rhythm Types	2 Iambic	2.061
30A	Number of Genders	1 None	6.591
31A	Sex-based and Non-sex-based Gender Systems	1 No gender	6.609
32A	Systems of Gender Assignment	1 No gender	6.599
81B	Languages with two Dominant Orders of Subject, Object, and Verb	1 SOV or SVO	2.201
110A	Periphrastic Causative Constructions	2 Purposive but no sequential	0.906
130B	Cultural Categories of Languages with Identity of 'Finger' and 'Hand'	1 Hunter-gatherers	-0.124
141A	Writing Systems	4 Syllabic	0.144

参考文献

[1] Bernard Comrie. *Language Universals and Linguistic Typology*. University of Chicago Press, 1981.

[2] Matthew S. Dryer and Martin Haspelmath (eds.). The world atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2014-07-03.).

[3] Hal Daumé III. Non-parametric Bayesian areal linguistics. In *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pp. 593–601, 2009.

[4] Hal Daumé III and Lyle Campbell. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 65–72, 2007.

[5] Rishiraj Saha Roy, Rahul Katare, Niloy Ganguly, and Monojit Choudhury. Automatic discovery of adposition typology. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pp. 1037–1046, 2014.

[6] 宮本正興. 名詞のクラス. 『言語』セレクション第 1 巻, pp. 115–122, 2012.

[7] 斎藤純男. 日本語音声学入門. 三省堂, 2013.