WORD GINI: 語の使用の偏りを捉える指標の提案とその応用

村山 太一　　若宮 翔子　　荒牧 英治
奈良先端科学技術大学院大学
{murayama.taichi.mk1, wakamiya, aramaki}@is.naist.jp

1 はじめに

言葉における「標準」性を定めるのは難しい．「標準」といっても， 様々な解釈が可能である．素朴な解釈としては，使用頻度が高ければ標準的であろうという考え方がある．しかし，大阪弁などの方言一つとっても，使用者が多くても，使用される場所が関西に偏っていることから，標準的な言語とはいえない．同様な偏りは，位置だけでなく，使用者や時間的にも起こりうる．「オレ」といった言葉は使用者が男性に偏り，「おはよう」といった単語も朝の時間帯にしか利用されないという意味で偏りが存在する．このように，「標準」の定義は単なる頻度と乖離する場合がある．

では，これまで，どのように語の標準性が定義されてきたのであろうか？有名な国語辞典の一つであり，標準的な日本語が掲載される岩波国語辞典[*1]には，65,000語が収録されている．また，同義語などの語同士の関係性を取得可能な意味辞書である日本語 WordNet[*2]などを用いて，日本語の基本語彙を抽出する研究 [10,12,14]も盛んである．しかし，これら既存の方法では，流行語や新語に追随することが困難である．

本研究のアイデアは，頻度でなく，偏りという概念を導入し標準を定義する点である．ここでいう偏りは，人，時間，位置の様々な観点に基づくものであり，これを行うためには，位置情報，ユーザ情報，時間情報が付与されているソーシャルメディアデータは最適な材料である．本研究ではソーシャルメディアデータを利用することで，ユーザごとの大量の書き言葉データの取得を行う．このデータを利用し，偏りに着目した新しい指標を提案し，この指標を応用した語彙の標準性の判定の可能性を提示する．

語の偏りを数値化した指標として「偏り値」を提案することの利点は以下の 2 点である．

1. 内省や主観に頼らない標準性の定量評価が可能である点
2. 新語に対しても更新が可能である点

[*1] https://www.iwanami.co.jp/book/b256574.html
[*2] http://compling.hss.ntu.edu.sg/wnja/

2 語の偏りに着目した新たな指標：偏り値

本章では，ソーシャルメディアにおけるユーザの発言を用いて，偏りに着目した基準値の作成方法とその概要について述べる．

2.1 データ

偏りに着目した基準値の作成用のデータとして, 2011年 7 月〜2012 年 7 月に Twitter 上に投稿された位置情報付きツイート 24,817,903 件を利用する．このデータから取得可能な人，時間，場所の 3 つの観点について，それぞれの基準値の作成を行う．

2.2 ジニ係数による偏り値の算出

2.2.1 ジニ係数とは

偏りを表す指標としてジニ係数を利用する．ジニ係数とは，社会における経済格差を表す指標である．算出方法は，n 個の標本 $x_1 \leq x_2 \leq ... \leq x_n$ が存在し，μ_x を標本平均とするとき，以下のように定義される．

$$Gini = \frac{1}{2\mu_x n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| \qquad (1)$$

ジニ係数は 1 に近いほど経済格差が大きいことを示しており，ジニ係数が 1 のときは一人に財が集中している状態を表す．一方，ジニ係数が 0 のときは均等に財が分配されている状態を表す．

2.2.2 偏り値の算出方法

2.1 節のデータのジニ係数を利用して語の偏り値を算出する．今回はデータから読み取ることができる，人，時間，場所の 3 つの偏りに着目する．具体的には，人を投稿ユーザ単位，時間は投稿された月単位，場所は投稿された地点 ($4km^2$ 区分) を一つの単位とみなし，出現した語にそれぞれどの程度の偏りが存在するかを算出する．

以下では，具体的な算出方法を投稿ユーザに着目した偏り値を例に説明する．

1. 与えられたツイートデータを投稿ユーザごとに分類し，全てのツイートデータを形態素解析ソフト

表 1　3 つの偏りにおいて特徴的な語

PER-GINI			TIME-GINI			LOC-GINI		
Rank	Word	偏り値	Rank	Word	偏り値	Rank	Word	偏り値
1	intensité	12.186	1	foursquare	1.882	1	ukb	9.614
2	estimée	12.175	2	loctouch	2.003	2	koj	9.730
3	rapport	11.828	3	忘年会	2.007	3	日本橋駅	9.573
4	alerte	11.931	4	ロケタッチ	1.750	4	箱根湯本駅	9.713
5	tekuteku	12.054	5	初詣	1.955	5	rjss	9.799

Mecab とシステム辞書 neologd を用いて単語ごとに分割し，ユーザごとの出現単語回数を値とした行列 (ユーザ×総単語) を作成する．

2. ユーザごと (行ごと) に，すべての単語の出現回数の合計が 1 になるように正規化する．その後，単語ごと (列ごと) に，出現回数の合計が 1 になるように正規化する．
3. 単語ごと (列ごと) に値を特徴として取り出し，2.2.1 節で述べたジニ係数の算出を行う．単語ごとに算出されたジニ係数をもとに偏り値 (*Deviation*) を以下のように定義する．

$$Deviation = -\log(1 - Gini) \qquad (2)$$

この値が高いと語の偏りが著しいことを意味し，値が低いと語が満遍なく用いられていることを意味する．

人，時間，場所の観点に着目して作成した偏り値をそれぞれ，「PER-GINI」「TIME-GINI」「LOC-GINI」と呼ぶ．

2.3　結果

単語を 3 つの偏り値順にそれぞれランキングし，他の偏り値のランクとの差が大きい順に並び替えたものを表 1 に示す．ここで示したのは，他の指標では偏りなく用いられているにも関わらず，特定の指標では偏りが大きいとされている語である．

PER-GINI の上位にはフランス語の「intensité」「estimée」などがあり，特定の外国人ユーザが発言していることが伺える．TIME-GINI の上位には「忘年会」や「初詣」の他に，位置情報サービスの「foursquare」や「loctouch」などがあるが，これは 2012 年初期に流行り出したサービスであることから，発言が全くされていない月と発言が多くされている月があるのが要因である．LOC-GINI の上位には国際空港を示す単語「ukb(神戸空港)」「koj(那覇空港)」「rjss(仙台空港)」や駅名が上位に来ている．

この結果は，「PER-GINI」「TIME-GINI」「LOC-GINI」それぞれにおいて，高い偏り値を示すであろう語が高い値を持っており，期待した結果を十分に示したものだと考える．

次章以降では，偏り値の中でも，特に PER-GINI に着目し，応用に即した利用可能性を検討する．3 章では，PER-GINI を用いた語彙平易化を行う．4 章では，PER-GINI を用いた語の標準性の把握を試みる．

3　応用例 1：語彙の平易化

3.1　語彙平易化の概要

平易化は，難解なテキストを平易なテキストに書き換える手法のことで，言語学習者や子どもといった多くの読者の文章読解の手助けをする手段の 1 つである．近年では特に，大規模パラレルコーパスを用いた統計的機械翻訳の手法 [1,8,9] や，特定の難解な語彙を平易な語彙に置きかえる語彙ベース [3,6,7] の手法が一般的である．

本章では, 後者の語彙ベースの平易化に着目し, PER-GINI が平易化に応用可能か検証する．語彙ベースの平易化は複雑な語彙の識別，語彙代替候補の抽出，質の良い語彙の選択，ランキングという大きく分けて 4 つのタスクからなる．本章では，特に候補のランキングのタスクに着目し，PER-GINI により精度が向上するかどうかを検証する．

3.2　データ・モデル

語彙の平易化の検証のため，英単語の偏り値を英語ツイートデータを用いて作成する． 2016 年 1 月〜6 月に米国で投稿された位置情報付きツイート 29,258,422 件を用い，PER-GINI を算出する．

PER-GINI を用いる語彙平易化のモデルとして，拡張の容易性の観点から Glavas(2015) [2] の候補ランキングにおけるモデルを利用する．これは教師なしの手法で，ターゲット単語と候補単語との単語ベクトル類似度や，ターゲット単語の周りの文脈類似度，情報量，言語モデルを用いた n-gram 頻度などを特徴として利用し，候補単語のランキングを行うモデルである．

Word Embedding Model の実装には Glavas モデル

表 2　5 つのモデルごとの TRank-at-n

モデル	n=1	n=2	n=3
単語ベクトル類似度	0.386	**0.619**	**0.720**
周囲 1 語彙の言語モデル	**0.478**	0.618	0.671
PER-GINI	0.417	0.592	0.649
Glavas+PER-GINI	**0.525**	**0.698**	**0.750**
Glavas モデル [2]	0.520	0.691	0.740

と同様の Glove[*3]を利用する．言語モデルは，語彙平易化のタスクに適しており，映画などの字幕から作成されたコーパス SubiMDB [5][*4]を利用する．

3.3　評価方法

上述した Glavas モデルと，そのモデルの特徴に PER-GINI を加えた Glavas+PER-GINI モデルを比較する．テストデータとして文章とその中で最も難しいとされる 1 つの語彙とその語彙に置き換えられる複数の語彙候補ランキングがセットとなった VICOR フォーマットの BenchLS [4] を利用する．

評価指標には，TRank-at-n を用いる．これは，正解データのランキング $r \leq n$ の語彙がモデルの出力するランキングのトップに来る割合を求めたものである．

3.4　結果

Glavas モデルの特徴ごとの比較として，PER-GINI 単体，モデルの特徴の 1 つであるターゲット語彙の周囲 1 語彙の言語モデル確率，単語ベクトル類似度，モデル全体の比較として，Glavas モデル，Glavas+PER-GINI モデルの計 5 つの TRank-at-n を表 2 に示す．結果が示すように，PER-GINI 単体は他の特徴例と比較し最も高い精度は出なかったものの，十分に語彙平易化のランキングを反映しており利用できる特徴の 1 つと考えられる．同時に，Glavas モデルの特徴の 1 つに PER-GINI を用いることで，n の値に関わらず Glavas モデル単体より高い TRank-at-n を示した．これらの結果から，PER-GINI が語彙平易化に対して有効な指標の 1 つであることが示された．

4　応用例 2：語の標準性の判定

4.1　標準的な語について

3 章で PER-GINI を語彙の平易化に用いたが，PER-GINI 自体はその語がどれだけ偏りなく用いられているかを示したものであり，3.4 節の結果が示すように語彙の平易さと一定の相関は見られるものの完全に一致する指標ではないと考える．すなわち，PER-GINI は SNS 上で使用者の偏りがないことを示しており，一般的な語の標準性を示していると考える．

本章ではこの仮説に基づいて，PER-GINI と一般的な語の標準性の関係について検証する．そのために，クラウドソーシングを利用したアンケートにより，PER-GINI が語の平易さと標準さ，どちらをよく説明しているかを調査する．

表 3　クラウドソーシングで調査した語の組み合わせ例

	PER-GINI 値	
	高い	低い
PPDB	式当日に 楠木正成の末裔 日常用語	式の日 楠木正成の子孫 日常の言葉
Simple PPDB	農水省 生活必需品 近ごろ	農林水産省 日用品 最近

4.2　データ・問題設定

正確な PER-GINI を作成するため，2 章で用いた位置情報付きツイートではなく無作為に抽出した日本の 98,775 ユーザが 2009 年 11 月 3 日〜2010 年 3 月 25 日に発言した 241,138,909 ツイートを用いて PER-GINI を求めた．

本調査は，Yahoo!クラウドソーシングサービスにより行った．具体的には，20 歳以上の 300 人を調査協力者とし，意味が同じ 270 組の語を提示し，どちらがより簡単か (平易さ)，もしくはどちらがより一般的か (標準さ) を答えてもらうタスクを行う．各組の語のうち，より偏り値が低い語と調査協力者が選択した語の一致率を評価する．

提示する 2 つの語の組み合わせは，日英対訳コーパスから学習された日本語の言い換えコーパス PPDB [13] と難解な語句と平易な語句の言い換えコーパス Simple PPDB [11] から，無作為に語句の組み合わせを抽出する．その中でも意味が等しい組み合わせを各コーパスから目視で 135 組ずつ選択する．表 3 で用いた語の組の例を示す．

4.3　結果・考察

クラウドソーシングの結果を表 4 に示す．この表は PPDB，Simple PPDB からそれぞれ選択した 135 組の語のうち，「平易=偏り低」は「平易な語を選択する」タスクを課した場合に偏り値が低い語句を選択した人が多かった件数を，「標準=偏り低」は「標準な語を選択する」タスクを課した場合に偏り値が低い語句を選択した人が多かった件数を示したものである．

Simple PPDB に関する 2 つのタスク結果を比較対象とし，マクネマー検定を行った結果，$p = 0.0098$ と

[*3] http://www-nlp.stanford.edu/data/glove.6B.200d.txt.gz
[*4] http://ghpaetzold.github.io/subimdb/

表 4　クラウドソーシングによる一致率調査の結果

	偏り値低	
	平易	標準
PPDB	0.756	0.733
Simple PPDB	0.793	**0.874**

なり有意差がみられた．これらが示すように，Simple PPDB に掲載されている語では偏り値が標準さを表しているが，PPDB では標準さを平易さと同程度の基準でしか表していない．Simple PPDB が主に単語であるのに対し，PPDB は句も扱っているという特徴を持つ．この特徴から，表 3 の「式の日」と「式当日に」のように「日」と「当日」では前者の語が偏り値が低いものの，「式当日に」という語句がより一般的であると判断する人が多くなったと考えられる．

これらの結果から，Simple PPDB に掲載されている１つの語ならば PER-GINI によって語の標準さを表現することができるが，句になると PER-GINI によって標準さを表せない文章も現れるという結果を得た．

5　おわりに

本研究の知見は以下である．

- 語の使用の偏りを使用者，時間，位置という３つの観点から計量した（2 章）．
- 使用者の偏りは，平易化の精度を向上させることから，平易性とも関連した概念である（３章）．
- 使用者の偏りは，平易さよりも標準性をよく説明する概念である（４章）．

今後は，英語の偏り値も標準性を表したものなのかの検証や，今回提示した利用可能性以外に基本語彙選定やリーダビリティ指標などへの活用の検討を行う予定である．また，句を扱う場合の偏り値の算出などの課題を検証していく必要があると考える．

なお，本研究で開発した偏り値が付与されたワードリストはサイト[*5] にて入手可能である．

謝辞

本研究の一部は，JSPS 科研費 JP16H06395, JP16H06399, JP16K16057 および JST ACT-I の支援を受けたものです．

参考文献

[1] William Coster and David Kauchak. Simple english wikipedia: a new text simplification task. In *Proc. of ACL*, pp. 665–669, 2011.

[2] Goran Glavaš and Sanja Štajner. Simplifying lexical simplification: Do we need simplified corpora. In *Proc. of ACL*, 2015.

[3] Colby Horn, Cathryn Manduca, and David Kauchak. Learning a lexical simplifier using wikipedia. In *Proc. of ACL*, pp. 458–463, 2014.

[4] Gustavo Paetzold and Lucia Specia. Benchmarking lexical simplification systems. In *Proc. of LREC*, 2016.

[5] Gustavo Paetzold and Lucia Specia. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proc. of COLING*, pp. 1669–1679, 2016.

[6] Gustavo Paetzold and Lucia Specia. Unsupervised lexical simplification for non-native speakers. In *Proc. of AAAI*, pp. 3761–3767, 2016.

[7] Gustavo Paetzold and Lucia Specia. Lexical simplification with neural ranking. In *Proc. of EACL*, pp. 34–40, 2017.

[8] Lucia Specia. Translating from complex to simplified sentences. *Computational Processing of the Portuguese Language*, pp. 30–39, 2010.

[9] Sanja Štajner, Iacer Calixto, and Horacio Saggion. Automatic text simplification for spanish: Comparative evaluation of various simplification strategies. In *Proc. of RANLP*, pp. 618–626, 2015.

[10] Yuriko Sunakawa, Jae-ho Lee, and Mari Takahara. The construction of a database to support the compilation of japanese learners' dictionaries. *Acta Linguistica Asiatica*, Vol. 2, No. 2, pp. 97–115, 2012.

[11] 梶原智之, 小町守. Simple ppdb: Japanese. 言語処理学会第 23 回年次大会, pp. 529–532, 2017.

[12] 国立国語研究所. 日本語教育のための基本語彙調査. 秀英出版, 1984.

[13] 水上雅博, Graham Neubig, Sakriani Sakti, 戸田智基, 中村哲. 日本語言い換えデータベースの構築と言語的個人性変換への応用. 言語処理学会第 20 回年次大会, 2014.

[14] 島村直己. 多文化共生社会における日本語教育研究　サブプロジェクト：日本語の基本語彙に関する研究. 国語研プロジェクトレビュー, Vol. 3, No. 3, pp. 133–141, 2013.

[*5] http://sociocom.jp/data/gini/