ツイートの情報量について －情報理論に基づく多言語調査－

Graham Neubig　　Kevin Duh
奈良先端科学技術大学院大学 情報科学研究科

1　はじめに

ツイッター[1]に代表されるマイクロブログサービスは世界中で急速に普及しつつあり、その特徴的なコミュニケーション法と言語現象は、社会学、自然言語処理などの分野で広く研究の対象となっている。その中で、ユーザが 1 つの投稿（ツイッター上では「ツイート」）に利用できる文字数に制限を設けるのがマイクロブログの最大の特徴と言えよう[2]。この文字数の制限がマイクロブログ上のコミュニケーションに大きく影響し、他のテキスト媒体で見られないような特徴的な省略や記述法が用いられることも多い [6]。

しかし、実際には各言語において 1 文字に内在する情報量が異なるため、この文字数の制限は決して世界中の言語で同一な効果があるとは限らない。特に、種類数が多く、1 文字が明確な概念を表す漢字と、種類数が少なく、1 文字が音韻的な特徴しか表さないローマ字を比較すると、前者の方が 1 文字に含まれている情報量が多いことは明らかである。このため、英語を含むローマ字圏の言語と、日本語や中国語を含む漢字圏の言語では 140 文字に記述できる情報の量は必然的に異なり、これがマイクロブログ上のコミュニケーションに影響を及ぼすと考えられる。本研究では、ツイッター上のマイクロブログデータに対して 26 言語に渡る多言語調査を行い、各言語のツイートの情報量を情報理論の観点から分析する。

2　実験設定

実験の準備として、2012 年 6〜7 月の 6 週間に渡りツイッターの公開ストリームから 1.20 億件のツイートを収集し、`langid.py`[4] を用いて言語判定を行った。最後に、言語判定の確率が 0.95 以上のツイートを言語ごとに分類し、確率が 0.95 未満のツイートを調査の対象から外した。その結果、9240 万件のツイートが残り、ツイート 1 件当たりの文字数は 70.8 文字となった。調査対象をさらに 50,000 件以上のツイートが残った言語に絞った結果、26 言語が残った。その諸元を表 1 に示す。この 26 言語は様々な言語族に属し、様々な文字種を利用する。

3　情報量の計測

以降の節では、情報理論の観点から調査対象の言語の様々な傾向を分析する。この分析を行うためには、まずツイートに含まれる情報量を数量的に計測する尺度が必要である。この尺度として、ある情報を確率モデルで表現した時に、その符号化に必要なビット数を表すエントロピーを採用する [7, 1]。今回の調査に関して、様々な言語に渡ってノイズを多く含むマイクロブログのテキストに適切な確率を与えるモデルが必要となる。このような頑健な推定を実現するために、単純かつ言語非依存な文字ベースの n-gram モデルを採用する。n-gram モデルは、文字列 $W = w_1, \ldots, w_i, \ldots, w_I$ の確率を、各要素 w_i の $w_{i-n+1}, \ldots, w_{i-1}$ が与えられた時の条件付き確率の積で近似する [3]。

$$P(W) \approx \prod_{i=1}^{I} P(w_i | w_{i-n+1}, \ldots, w_{i-1}) \qquad (1)$$

n-gram モデルは一般的には単語列に対して利用することが多いが、単語列に対する n-gram モデルは表記の揺れに敏感であり、モデル推定に比較的大きな学習データを要する。さらに、日本語や中国語のような分かち書きされない言語では単語分割が必要となる。このため、本調査では各要素 w_i を単語ではなく文字とし、これらの問題に対処する。

言語モデルの文脈の長さを $n = 7$ とし、平滑化に、文字言語モデルのように低頻度 n-gram が少ないデータにおいても頑健に推定可能な Witten-Bell 法を利用した [8]。言語モデルの学習及びエントロピーの計測の際には 10 分割の交差検定を行った。エントロピーが言語モデルの学習データ量に依存するため、全ての言語に対してデータ量を全ツイートから無作為にサンプルした 50,000 ツイートに固定した。

表 1:　各言語における投稿の数

言語	文字	言語	文字
英語 (en)	2.71B	イタリア語 (it)	40.0M
日本語 (ja)	625M	ジャワ語 (jv)	36.0M
スペイン語 (es)	723M	中国語 (zh)	16.8M
ポルトガル語 (pt)	348M	ドイツ語 (de)	22.8M
アラビア語 (ar)	275M	タガログ語 (tl)	24.9M
インドネシア語 (id)	142M	スワヒリ語 (sw)	16.8M
韓国語 (ko)	69.4M	ペルシア語 (fa)	9.85M
オランダ語 (nl)	87.5M	ウルドゥー語 (ur)	10.7M
フランス語 (fr)	84.5M	ガリシア語 (gl)	7.83M
トルコ語 (tr)	79.5M	スウェーデン語 (sv)	7.17M
タイ語 (th)	57.6M	ギリシャ語 (el)	6.27M
ロシア語 (ru)	54.8M	ラテン語 (la)	6.91M
マレーシア語 (ms)	60.8M	カタルーニャ語 (ca)	5.76M

[1] `http://www.twitter.com`
[2] ツイッターでは投稿当たり最大 140 文字が利用可能である。
本研究の成果は [5] に詳しい。

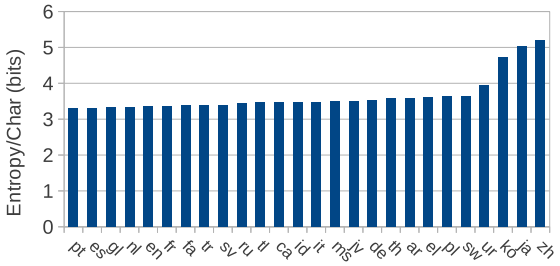

図 1: 文字当たりのエントロピー

4 文字当たりの情報量

まず、各言語における 1 文字当たりの情報量を計測する。先行研究 [1, 2, 9] の調べによると、日本語や中国語など、利用する文字の種類が多い言語では 1 文字当たりの情報量が多いという仮説を立てることができる。これらの言語では 1 文字自体が明確な意味を持ち、ローマ字のような音韻的な特徴しか持たない文字よりも含まれている情報が多いことは直感的にも理解できる。

ツイートコーパスに対して計測した文字ごとのエントロピーを図 1 に示す。この図より、想定通り、中国語、日本語、韓国語のエントロピーが高い結果となったことが分かる。この中で、表意文字である漢字しか用いない中国語が最も高いエントロピーとなり、表意文字と表音文字を両方用いる日本語が続き、主に表音文字を用いる韓国語が 3 番に続いた。全体的に、ローマ字を使用する言語では比較的低いエントロピーが見られ、ローマ字を使用しない言語で最も低くなったペルシア語は下から 7 番目であった。

しかし、使用する文字の種類数以外にも、ツイートの文字ごとのエントロピーに影響し得る要因は様々である。ここでは、情報量に影響し得る要因を 5 つ考慮し、各要因による影響をより詳細に調査する。

使用する文字の種類数： 上述のように、言語で用いられる文字の種類数は大きな要因であると考えられる。ここで、50,000 ツイートのコーパスの中に現れる文字の異なり数を使用する文字数の目安とする。

平均単語長： 空白で明示的に単語の境目を示す言語の中でも比較的短い単語を利用する言語（英語やスペイン語）と比較的長い単語を利用する言語（ドイツ語やスウェーデン語）が存在する。この中で、同一の内容でも長い単語を用いる方が空白の数が少なく、同じものでも少ない文字数で表せるため、長い単語の言語の文字当たりの情報量が多いという仮説も考えられる。

ツイート当たりのツイッター特有の表現： ツイッター上には様々な独特な言語現象が存在する。その例として、「@」から始まるユーザ名、「#」から始まるハッシュタグ、「http」から始まるリンクなどがある。これらは通常の書き言葉と情報量が異なることが考えられ、使用頻度により文字ごとの情報量が異なることも考えられる。

表 2: 26 言語間、ローマ字圏の言語間で測った各要素と文字当たりのエントロピーの R^2 値と相関の比例（「+」）、反比例（「-」）関係。太字はスチューデントの t 検定により $p < 0.01$ で有意な差である。

要素	26 言語		ローマ字圏	
異なり文字数	**75.3%**	(+)	**53.1%**	(+)
平均単語長	**41.3%**	(+)	26.1%	
ツイッター用語	4.4%		35.5%	
リツイート率	18.9%		**44.0%**	(-)
引用率	0.3%		26.3%	
全要因	82.9%		72.1%	

リツイート率： ツイッター上で、他のユーザのツイートを共有する「リツイート」と呼ばれる仕組みがあり、ツイートの冒頭に「RT」と元の発信源のユーザ名を追加することでリツイートであることを表す。リツイートされた投稿と通常の投稿の内容に差があるとも考えられるため、各言語におけるリツイート率は言語のツイートの情報量に影響する可能性がある。

引用率： リツイートではあるユーザがそのまま他のユーザのツイートを共有するが、他のユーザの投稿を共有しながら自分独自のコメントを追加する「引用」という形もある。前者はツイートの冒頭に「RT」が現れ、後者はツイートの内部に「RT」が現れるためにこの 2 つの現象を区別できる。

これらの要因がツイートの情報量に与える影響を調べるために、入力を上記の 5 つの要因、出力を言語の文字ごとのエントロピーとする重回帰を行い、各要因の重みについての考察を行う。ほとんどの値をそのまま利用するが、異なり文字数が極端に大きい中国語や日本語の影響を緩和するために、異なり文字数の対数を取ってから回帰に利用する。これらの要因とエントロピーの相関を測る尺度として、回帰に用いる要因が説明する分散の割合を表す R^2 値を利用する。各要因の影響を個別に調べるために、要因を 1 つずつ回帰に用いた場合の R^2 を調べ、全ての要因を組み合わせた場合の回帰の結果も調べる。更に、調査対象の 26 言語を全て考慮した場合の結果以外にも、言語が使用する文字の数の影響を取り除くためにローマ字圏の言語のみを用いた結果も調査する。

表 2 の結果から、全ての要因を 26 言語の回帰に用いた際、R^2 が 82.9%となり、ローマ字圏の言語に限った場合は 72.1%となった。この結果から、考慮した要因は各言語の文字当たりのエントロピーを比較的正確に推定できていると言える。想定通り、最も文字当たりの情報量と相関の強い要因は使用する異なり文字数であった。この傾向が様々な文字体系の言語を考慮した場合でも、ローマ字圏の言語に限った場合でも見られた理由として、英語のように通常のローマ字のみを使用する言語と、ス

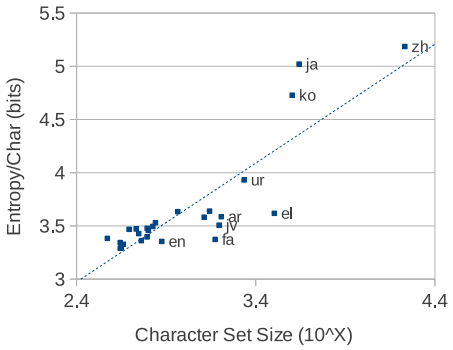

図 2: 異なり文字数と文字当たりのエントロピーの相関

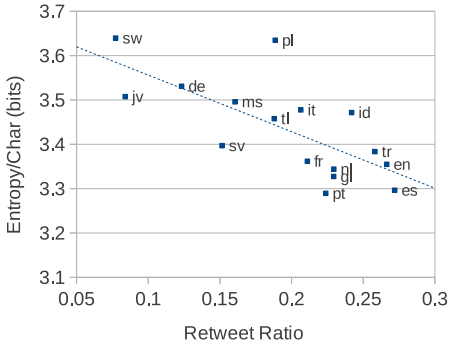

図 3: ローマ字圏の言語におけるリツイート率と文字当たりの情報量の相関

ウェーデン語のようにアクセントを表す付加記号を利用する言語が両方存在し、同じローマ字でも事実上利用する文字の数が異なるためである。異なり文字数とエントロピーの相関の詳細を図 2 に示す。

全ての言語に対する回帰では、平均単語長もエントロピーと有意な相関が見られたが、これは日本語や中国語といった特異点の影響であると考えられる。平均単語長を計算するために、単語を「空白で区切られた文字列」と定義したが、日本語や中国語は通常の書き言葉に空白による区切りを用いないため、空白で区切られた文字列は非常に長いものとなる。これらの言語は異なり文字数も多いため、単語長と異なり文字数を個別に考慮する時に両方の要因は文字エントロピーとの有意な相関があるように見られる。

ローマ字圏の言語に焦点を絞った場合、リツイート率と文字当たりのエントロピーの間に有意な負の相関が見られた。その詳細を図 3 に示す。この負の相関の原因は自明ではないが、リツイートされた投稿とリツイートされなかった投稿を人手で調べた結果、リツイートされた投稿はリツイートされなかった投稿に比べて通常の書き言葉に近く、マイクロブログに頻繁に現れる略語や絵文字などが比較的少なかった。これらの要因によって、同一の内容でもリツイートされたテキストの文字数が比較的多く、情報量の計算に影響したと考えられる。

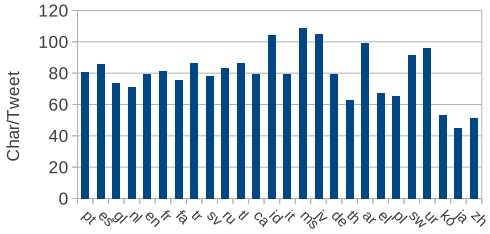

図 4: 投稿当たりの文字数

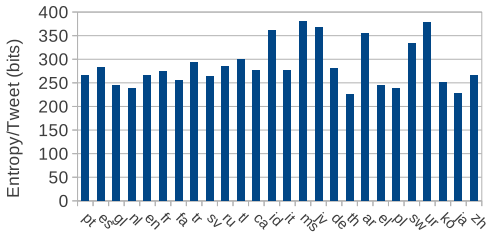

図 5: 投稿当たりのエントロピー

5 ツイート当たりの情報量

漢字圏の日本語と中国語の 140 文字に含まれる情報がローマ字圏の英語の 140 文字に含まれる情報に比べて明らかに多い。この事実をマイクロブログに対して当てはめると、一見、日本語や中国語のマイクロブログ投稿は英語のマイクロブログ投稿より情報が多いと言えると思われる。しかし、この考察はツイートの情報量の全てを語っているわけではない。実際には、ツイートの中で 140 文字に渡る詳細な現状報告がある一方、「食べた！」のような極端に短いものもある。ツイートの情報を語るには、文字に含まれる情報量だけではなく、ツイートの内容を表現するために用いられる文字数も考慮する必要がある。

まず、各言語におけるツイート当たりの平均文字数を図 4 に示す。文字当たりのエントロピーとツイート当たりの文字数の関係を強調するために、言語の順を図 1 と同じ文字エントロピーの降順とする。この図の結果の中で最も印象的なのは日本語、中国語、韓国語の平均文字数である。これらの言語は文字当たりの情報量が多いが、平均文字数がほかの言語に比べて明らかに少ない。実際には、その他の言語の平均文字数が約 70 文字である中、この 3 つの言語は約 50 文字にとどまる。同一の意味をより少ない文字数で伝えられる言語で実際に使用されている文字数も少ないことは直感的な結果ではあるが、ツイートの情報量が文字の情報量に単純比例するわけではないことを証明する結果であるとも言える。

平均文字数の多い言語に目を向けても興味深い結果が得られる。マレー語、ジャワ語、インドネシア語は全てインドネシア列島で広く利用される言語であり、この地域における独特なツイッターの使い方を示唆する。前節で述べた要因を調査した結果、この 3 つの言語では引用

表 3: 26 言語間、ローマ字圏の言語間で測った各要素と投稿当たりのエントロピーの R^2 値と相関の比例（「+」）、反比例（「-」）関係。太字はスチューデントの t 検定により $p < 0.01$ で有意な差である。

要素	26 言語		ローマ字圏	
異なり文字数	0.5%		19.3%	
平均単語長	15.0%		14.6%	
ツイッター用語	18.5%		**71.8%**	(+)
リツイート率	0.5%		19.2%	
引用率	**43.8%**	(+)	**80.2%**	(+)
全要因	71.7%		91.0%	

率はそれぞれ 44.8%、53.9%、33.0%に達し、全言語の中央値である約 3%の引用率を大幅に上回っていることが分かる。引用では、元の投稿の著者の発言に加えて自分のコメントを更に付け加えるため、実質的に 1 投稿の中に 2 投稿分の情報が含まれているため、平均文字数が長くなることは不思議ではない。これ以外の外れ値であるスウェーデン語、アラビア語、ウルドゥー語についても全て同様の傾向を示した。

次に、各言語における投稿当たりの情報量をエントロピーに基づいて計測し、図 5 に示す。順番は上記の図と同じく文字当たりのエントロピーの昇順とする。図の結果から、文字当たりの情報量と投稿当たりの情報量にほとんど相関がないことが確認できる。つまり、中国語、日本語、韓国語などの著者は文字数制限の 140 文字までに多くの情報を盛り込むことができるにも関わらず、ほとんどの場合では 140 文字を全て使い切ることはない。逆に、制限に近い文字数を利用する傾向にある言語は 1 投稿の情報量が他の言語より多い。

ツイートの情報量に影響を及ぼす要因を詳細に調べるために、前節と同じ 5 つの要素を入力とする重回帰を行い、投稿当たりのエントロピーを回帰の推測対象とする。表 3 の結果から、文字当たりのエントロピーに大きな影響を及ぼした異なり文字数はやはり投稿当たりのエントロピーにほとんど影響を及ぼさないことが見て取れる。その一方、使用する言語の特性ではなく、その言語のユーザの行動特性が投稿当たりの情報量に影響を及ぼすことも分かる。このことは、引用率、ツイッター特有の表現の使用率と投稿当たりの情報量の間には統計的に有意な相関があることから分かる。

最後に、言語ごとの平均情報量に加えて、各言語におけるツイートの情報量の分布を調べた。図 6 に、日本語、英語、インドネシア語という 3 つの異なった性質を持つ言語の分布を示す。この結果から、英語と日本語では、投稿当たりのエントロピーは 2 つの頂点を持つ分布となっていることが分かる。最初の頂点は 100〜300 ビットの情報を含むツイートからなり、文字数の制限を無視してマイクロブログを書いた場合に自然とこの程度の情報が含まれることが多いと考えられる。2 つ目の頂点は最大の 140 文字に近づいた時に現れ、著者が文字数の制限に内容が収まるように内容を調整した影響が見られる。インドネシア語では最初の頂点が見られない理由は前述した引用率の高さに起因するのであろう。最後に、英語の 140 文字は約 480 ビットの情報を記述できることは図 6 より分かる。これに比べて、日本語の平均的なツイートは英語のツイートとほぼ同等の情報を有するが、約 1 割の日本語ツイートは 480 ビットを超えており、英語の 140 文字を使い切っても表現できなかったことも図の結果から見て取れる。

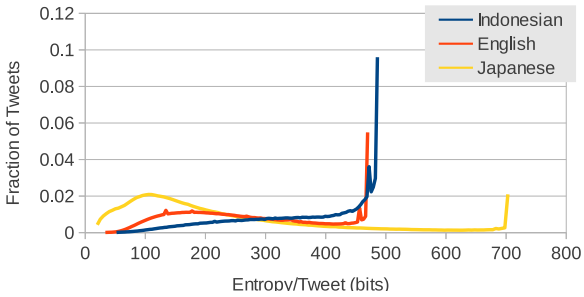

図 6: 3 言語におけるツイートの情報量分布

6 まとめ

本研究は多言語に渡り、ツイートの情報量について考察を行った。その結果、多くの文字を使用する言語は文字当たりの情報量は多いが、使用する文字の数は投稿当たりの平均情報量にほとんど影響はないことが明らかになった。その一方、引用率、リツイート率などの行動的要因は投稿当たりの情報量に影響を及ぼす。今後の課題として、ユーザの影響力とツイートの情報量の関係、文字の制限に直面した際、ユーザはいかに伝えたい内容を表現するかなどを調べていきたい。

参考文献

[1] P. E. Brown, V. J. D. Pietra, R. L. Mercer, S. A. D. Pietra, and J. C. Lai. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18, 1992.

[2] J. Chang and Y.-J. Lin. An estimation of the entropy of Chinese – a new approach to constructing class-based n-gram models. In *Proc. Rocling Computational Linguistics Conference VII*, 1994.

[3] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proc. ACL*, 1996.

[4] M. Lui and T. Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pp. 25–30, July 2012.

[5] G. Neubig and K. Duh. How much is said in a tweet? a multilingual, information-theoretic perspective. In *Proceedings of the AAAI Spring Symposium on Analyzing Microtext*, Stanford, California, March 2013.

[6] D. Pennell and Y. Liu. Toward text message normalization: Modeling abbreviation generation. In *Proc. ICASSP*, 2011.

[7] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[8] I. Witten and T. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *Information Theory, IEEE Transactions on*, 37(4):1085–1094, 1991.

[9] 森, 山地. 日本語の情報量の上限の推定. 情報処理学会論文誌, 38(11), 11 1997.