How Much is Said in a Tweet? A Multilingual, Information-theoretic Perspective

Graham Neubig and Kevin Duh
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japan

Abstract

This paper describes a multilingual study on how much information is contained in a single post of microblog text from Twitter in 26 different languages. In order to answer this question in a quantitative fashion, we take an information-theoretic approach, using entropy as our criterion for quantifying "how much is said" in a tweet. Our results find that, as expected, languages with larger character sets such as Chinese and Japanese contain more information per character than other languages. However, we also find that, somewhat surprisingly, information per character does not have a strong correlation with information per microblog post, as authors of microblog posts in languages with more information per character do not necessarily use all of the space alloted to them. Finally, we examine the relative importance of a number of factors that contribute to whether a language has more or less information content in each character or post, and also compare the information content of microblog text with more traditional text from Wikipedia.

Introduction

One of the characteristics of Twitter and other microblog services is that they impose a hard limit on the number of characters that users may use in a single message. This, and a number of other factors, results in a unique form of writing and interacting that is far removed from that observed in more canonical text. For example, in order to evade the character limit, it is known that people will often use unique abbreviations to properly express what they want to say (Pennell and Liu 2011).

One factor that contributes to how much this sort of abbreviating is necessary is the fact that characters carry different punch in different languages. In particular, a single character in languages with logographic characters such as Chinese or Japanese contains more content than one in English or other languages with a relatively small number of phonetic characters. It has previously been noted in the popular media that this has effects on the way people communicate on Twitter (Rosen 2011).[1] This impels the question: how much is really being said on Twitter? How much information is contained in a single tweet? Does this information content vary from language to language, and if so, why?

In this paper, we perform a multilingual study spanning 26 languages that attempts to quantify this difference exactly from an information theoretic perspective. Specifically, we measure the information content of each tweet using the Shannon entropy according to a statistical language model. This gives us a quantitative measure of the amount of information included in one character or one tweet for each of these languages, and an approximate answer to our question of how much is said in a tweet.

With these results, we further perform an analysis on how communication tendencies on Twitter affect, or are affected by, the information content in several of the concerned languages. For example, it is well known that text on Twitter is rife with unique phenomena such as hash tags, mentions, and retweets (Honey and Herring 2009; Boyd, Golder, and Lotan 2010). We attempt to measure the effect of these features on the information content as measured by entropy. Finally, we compare the information content on Twitter with an identical amount of text from the more canonical domain of Wikipedia, exploring the implications of the limited number of characters and unique writing style of Twitter on the way that users communicate.

Experimental Setup

In preparation for our experiments, we collected 120M tweets from the Twitter public stream over the course of six weeks in June and July of 2012.

Next, we used langid.py (Lui and Baldwin 2012) to identify the language of each tweet. In order to increase the language identification accuracy, before running language identification we lowercased the text and removed hash tags, URLs, user names starting with "@," and characters that were not contained in the character sets of any of the languages in which we were interested. These measures were taken only for the language identification, and the remainder of the analysis uses the tweet text as-is. Given the language identification results, we kept all tweets that had a language identification confidence of over 0.95, which resulted in a

[1] (Rosen 2011) anecdotally reports that using Google Translate, on average Japanese and Thai tweets translate into 260 and 184 English characters, respectively, larger than the 140 limit.

Table 1: The number of tweets and characters in each language.

Language	Tweets	Characters	Writing
English (en)	34.2M	2.71B	Latin
Japanese (ja)	13.8M	625M	Chinese/Kana
Spanish (es)	8.39M	723M	Latin
Portuguese (pt)	4.31M	348M	Latin
Arabic (ar)	2.82M	275M	Arabic
Indonesian (id)	1.37M	142M	Latin
Korean (ko)	1.29M	69.4M	Hangul
Dutch (nl)	1.23M	87.5M	Latin
French (fr)	1.03M	84.5M	Latin
Turkish (tr)	921k	79.5M	Latin
Thai (th)	916k	57.6M	Thai
Russian (ru)	660k	54.8M	Cyrillic
Malay (ms)	566k	60.8M	Latin
Italian (it)	513k	40.0M	Latin
Javanese (jv)	342k	36.0M	Latin
Chinese (zh)	323k	16.8M	Chinese
German (de)	296k	22.8M	Latin
Tagalog (tl)	287k	24.9M	Latin
Swahili (sw)	184k	16.8M	Latin
Persian (fa)	131k	9.85M	Arabic
Urdu (ur)	110k	10.7M	Arabic
Galician (gl)	107k	7.83M	Latin
Swedish (sv)	90.7k	7.17M	Latin
Greek (el)	81.9k	6.27M	Greek
Latin (la)	74.4k	6.91M	Latin
Catalan (ca)	71.2k	5.76M	Latin
Polish (pl)	68.1k	4.46M	Latin
Finnish (fi)	60.3k	3.12M	Latin

total of 92.4M tweets and an average of 70.8 characters per tweet.

Keeping all languages that had at least 50,000 tweets in the collection after this processing gave us a total of 26 languages, the statistics for which are shown in Table 1. As can be seen in the Table, these languages cover a variety of language families and writing systems.

Measuring Information Content

In the following sections, we take a look at tweets from an information-theoretic perspective for all of the languages in our collection. In order to perform this study, we first must have a way to measure the amount of information contained in a tweet quantitatively. To this end, we use entropy, a classic criterion based on Shannon's information theory (Shannon 1948; Brown et al. 1992) that tells us how many bits of information are required to encode a message when the message is described with a probabilistic model. In particular, if we have a corpus of text $\mathcal{W}$, the entropy H of the corpus is defined as its negative $\log_2$ probability

$$H(\mathcal{W}) = -\log_2 P(\mathcal{W})$$

If we have a good probabilistic model of language, the entropy of each character in the tweet should give us an approximate idea of how much information is contained therein.

For this examination, we require a probabilistic model that can be used reliably and robustly over a wide variety of languages on microblog text, which is notoriously difficult to analyze and full of orthographic variation. In order to do so, we adopt perhaps the simplest model of language possible, the n-gram model (Chen and Goodman 1996). n-gram models simply approximate the probability of a string $W = w_1, \ldots, w_i, \ldots, w_I$ by predicting each element incrementally, referencing the previous $n-1$ elements:

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

While n-gram models are traditionally calculated over strings of words, word-based models are sensitive to orthographic variation, require a relatively large amount of data to calculate robustly, and require word segmentation or morphological analysis for languages with no word boundaries or rich morphology. Instead, we simply calculate a language model where each element w_i represents a single character, which greatly simplifies most of these problems.

In our experiments, we use the SRILM toolkit (Stolcke 2002) to calculate a character 7-gram model[2] using interpolated Witten-Bell smoothing (Witten and Bell 1991). To measure the entropy over any particular data set, we used 10-fold cross validation, training the model on 90% of the data, and measuring it on 10%, and repeating the process until each of the tweets has been used in the test set exactly once. As entropy of language models tends to fluctuate based on the training data size, we hold the amount of data constant to 50,000 tweets across all languages, sampled at random from the greater body of training data.[3]

Information Content per Character

First, we would like to measure the information content per character in tweets for each of our languages of interest. Given previous discussion in scientific literature (Brown et al. 1992; Chang and Lin 1994; Mori and Yamaji 1997), a reasonable hypothesis is that languages that use a relatively large number of characters in their writing system (such as Chinese, Japanese, and Korean) should have a higher information content per character than other languages. This is due to the fact that these writing systems have larger character sets, with characters generally used to represent full syllables, as opposed to the other writing systems in our study, which approximately have one letter per phoneme.

We demonstrate experimental results measuring the entropy by character for each of the languages in Figure 1. It

[2] We also tried raising the language model order to 10 in preliminary experiments, which reduced overall entropy somewhat, but did not greatly affect our relative results.

[3] 50,000 tweets is enough to give us a relatively stable estimate of per-character entropy. Even when we increased the number of tweets to 500,000 for the languages that had sufficient data, this did not effect the relative rank of any of the languages when they were sorted by entropy.

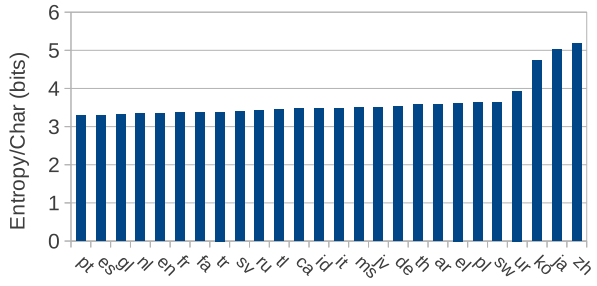

Figure 1: Entropy per character in a tweet for each of the languages.

can be seen that in general, this hypothesis is correct. Chinese, with it's fully logographic alphabet, has the highest information content with 5.19 bits of entropy per character, and Japanese is next, followed by Korean. On the other hand, the languages that have the lowest information content per character tend to be those originating from the Iberian peninsula: Spanish, Portuguese, and Galician. In general, languages written in Latin text tend to have a lower information content per character, with the first non-Latin language being Persian, with the 7th lowest entropy.

However, there are a number of factors other than the number of characters used to write a language that could possibly have an effect on the amount of information contained per character in a tweet. For example, the occurrence of phenomena not directly related to the language itself, but specifically linked to the unique writing style on Twitter may also have an effect on the information content as measured by entropy. Here, we focus on five different factors that could possibly effect information content:

Character Set Size: We can expect the character set size to be the largest factor here. We empirically determine the character set size of a particular language to be the number of unique characters used in the said language within our 50,000 tweet corpus.

Characters/Word: Languages with longer words will use a single word where other languages may use two, which can be expected to result in slightly more efficient use of characters. For example, the term "pop music" in English can naturally be shortened to "Popmusik" in German, reducing the number of characters used by one.

Twitter Terms/Tweet: On Twitter, there are a number of unique linguistic phenomena such as user mentions starting with "@," hash tags starting with "#," and external links starting with "http." It is likely that these terms will have a different information content than more traditional text in the language, so the frequency of their usage may also influence information content.

Retweet Ratio: A large number of tweets are "retweets," where a user simply shares another user's post, preceded by the letters "RT." It is conceivable that retweeted tweets will either have more or less information content than other tweets, so we include this as a factor in the analysis.

Quote Ratio: Finally, in contrast to retweets, where the user simply reposts another user's post as-is, in quotes the user adds an additional comment on top of the original post. We make the distinction between retweets and quotes by whether the letters "RT" occur at the beginning of the post or after original content respectively.

Table 2: R^2 values and direction of correlation between each factor and per-character entropy over either all languages, or only languages written in Latin script. Correlations significant to $p < 0.01$ according to Student's t-test are written in bold, with "+" indicating a direct correlation and "-" indicating an inverse correlation.

Factors	All		Latin	
Char Set Size	**75.3%**	(+)	**53.1%**	(+)
Chars/Word	**41.3%**	(+)	26.1%	
Twitter Terms/Tweet	4.4%		35.5%	
Retweet Ratio	18.9%		**44.0%**	(-)
Quote Ratio	0.3%		26.3%	
All Factors	82.9%		72.1%	

In order to measure the relative contribution of each of these factors we perform a linear regression using each of the factor values as input, and attempt to predict the entropy value of the language. We use most of the values as-is, but instead of using the raw character set size, we use the logarithm of the character set size, which helps to achieve a more linear correlation with entropy and mitigate the impact of Korean, Japanese, and Chinese on the overall results.[4] As a measure of how well correlated the factors are with the entropy, we use the standard measure of the fraction of variance described by the factors considered in the regression, the R^2 value. R^2 is defined as

$$R^2 = 1 - Res/Var, \tag{2}$$

where Res is the sum squared residual between the linear regressor predictions and targets, and Var is the variance of the targets.

To isolate the contribution of each of the factors to entropy, we show results with each of the five factors used individually, as well as a regression with all of the factors combined. In addition, we show results for when all of the languages are considered, and also when we only consider languages that mainly use the Latin character set to neutralize some of the impact that different writing systems will have on the results.

From the results in Table 2 we can see that when all five factors are used, an R^2 of 82.9% is achieved when all languages are considered, and an R^2 of 72.1% is achieved when only Latin languages are targeted. Overall this indicates that the elements considered are relatively good predictors of

[4]This is also motivated from an information theoretic perspective, as the entropy of a single character under a uniform distribution will be equal to $\log_2(V)$ where V is the character set size.

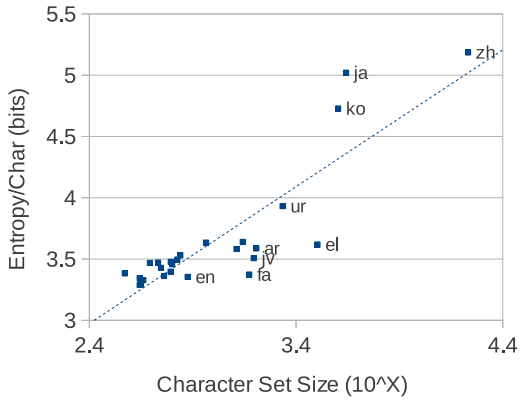

Figure 2: A plot of the correlation between character set size and per-character entropy.

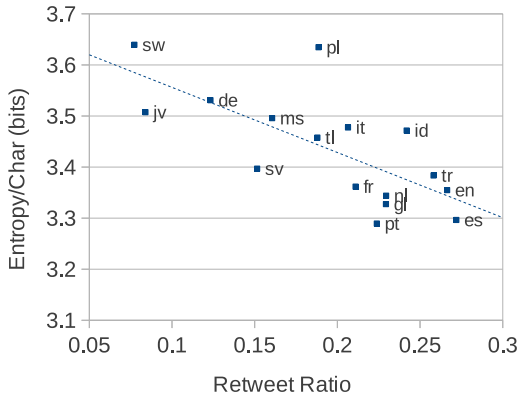

Figure 3: A plot of the correlation between retweet ratio and per-character entropy for languages with Latin character sets.

the per-character information content in each language. As expected, the largest contributing factor was the character set size, which had a strong correlation with per-character entropy. This was true both when languages with different writing systems were considered, and for Latin texts, likely as a result of some languages (like English) using plain, unaccented alphabetical characters, and others (such as Swedish) using a richer set of diacritics. A detailed plot of correlation between character set size and per-character entropy is shown in Figure 2.

For the regression over all languages, characters/word was also found to be significant by the linear regression, although this is likely an artifact of the fact that languages without explicit word boundaries (such as Japanese and Chinese) also happen to have large character sets. When both character set size and characters/word are considered in concert, characters/word does not have a significant effect.

When only languages written in the Latin character set are considered, the ratio of retweeted posts has a significant negative correlation with character-based entropy. We demonstrate this trend with in further detail in Figure 3. While the reason for this negative correlation is not immediately obvious, a manual examination of retweeted and non-retweeted text provides a very clear explanation. Our inspection found that posts that get retweeted tend to be more well-formed language, while regular Twitter posts have a higher percentage of unique abbreviations, smileys, misspellings, and other non-canonical linguistic phenomena. Thus, the less predictable texts tend to be assigned a higher entropy, resulting in the negative correlation between retweet ratio and per-character entropy. We examine the relationship between Twitter text and more canonical text further in the sections below.

Information Content per Tweet

It has been noted before in the popular media (Rosen 2011) that 140 characters in Japanese or Chinese contains signifi-

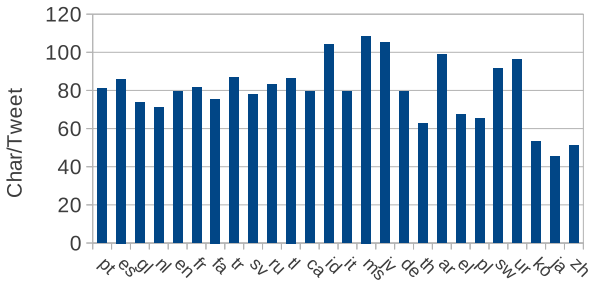

Figure 4: The average number of characters in one tweet.

cantly more information than an equivalent number of characters in English. With respect to Twitter, this would seem to indicate that a Japanese or Chinese tweet is generally saying more than an English tweet. However, this observation only tells half the story. In reality, tweets are of all sizes, from a short four letter exclamation “Yes!” to a full 140 character account of the author’s current situation or opinions. To know how much is said in tweets, we need to consider both the amount of information that is packed into a single character, and the number of characters authors choose to use to say what they want to say.

In order to examine this, we first plot the average number of characters in a tweet for each language in Figure 4. To emphasize the relationship between bits of entropy per character and bits of entropy per tweet, we leave the languages ordered by character-wise entropy as in Figure 1.

This graph shows some interesting results. The most striking aspect of this figure is that Japanese, Chinese, and Korean, the languages with the highest information content per character are on the far low end with regards to the number of characters used in a tweet. In fact, all of these three lan-

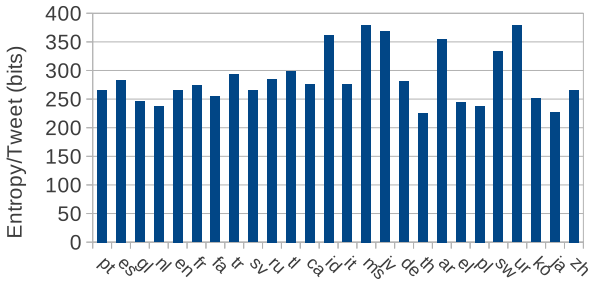

Figure 5: The amount of entropy in bits per tweet.

guages stop around an average of 50 characters per tweet, while the majority of languages written in other scripts use 70 or more characters. While this is somewhat intuitive — the number of characters necessary in a tweet should be inversely proportional to the necessary length to say something meaningful — it does mean that the information content of tweets cannot be simply measured by the amount of information packed into the characters.

On the other side of the spectrum, we also have interesting results. The languages that use the most characters, Malay, Javanese, and Indonesian, are all languages used widely in the Indonesian archipelago, indicating a unique culture of Twitter use in this area. An examination of the factors introduced in the preceding section demonstrated that this was due to an extremely high ratio of quoted re-posts, where the author adds additional comments to another author's post, essentially resulting in content from two or more authors included in a single Tweet. The ratios for Malay, Javanese, and Indonesian were 44.8%, 53.9%, and 33.0% respectively, an order of magnitude over the average of 3% for most languages in the question. The other outliers on the character length graph, Swedish, Arabic, and Urdu, all showed a similar trend.

Next, we measure the average number of bits of information included in a tweet in each language, showing the results in Figure 5, again in the same order as the previous charts. We define information content per tweet as:

$$\text{Entropy/Tweet} = \text{Entropy/Char} \times \text{Char/Tweet}$$

From this graph, we can see that the correlation between the amount of information in a character and the amount of information in a tweet is tenuous at best. This indicates that even though Chinese, Japanese, or Korean speakers could say more in 140 characters if they so chose, in the majority of the cases they do not choose to do so. On the other hand, the languages where the speakers tend to approach the 140 character limit in most of their tweets do show a significant trend of containing more information than other languages.

In order to examine the factors that contribute to the amount of information contained in a tweet in more detail, we perform a regression over the same five factors in the previous section, this time choosing entropy per tweet as our regression target. The results are shown in Table 3. From this

Table 3: R^2 values and direction of correlation between each factor and per-tweet entropy for all or only Latin languages.

Factors	All		Latin	
Char Set Size	0.5%		19.3%	
Char/Word	15.0%		14.6%	
Twitter Terms/Tweet	18.5%		**71.8%**	(+)
Retweet Ratio	0.5%		19.2%	
Quote Ratio	**43.8%**	(+)	**80.2%**	(+)
All Factors	71.7%		91.0%	

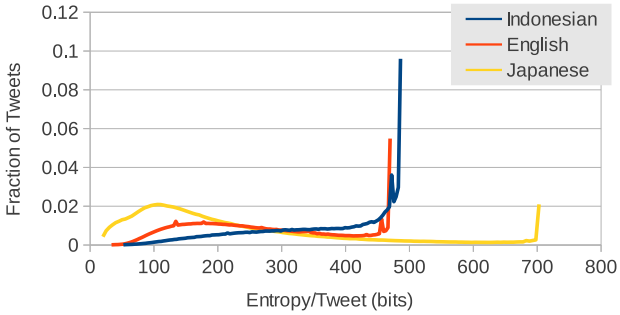

Figure 6: The fraction of tweets with a particular amount of information content.

table, we can see that the size of the character set has almost no effect on the amount of entropy contained in a tweet. This result indicates that while authors with larger character sets at their disposal could possibly write more in 140 characters if they so chose, they rarely exercise this ability, choosing instead to write on average approximately the same amount of content as is written in microblog posts in most other languages. On the other hand, behavioral factors that influence the length of tweets such as the widespread usage of quotes or Twitter-specific terms (such as hash tags, mentions, and links) are a greater indicator of a particular language's average information content per tweet.

Finally, one may wonder about the distribution in the amount of information included in each tweet, in addition to the average statistics presented so far. We calculate the distribution in entropy over tweets in each language by first finding the distribution over the number of characters per tweet, then multiplying the characters by the average per-character entropy. For presentation purposes, we show in Figure 6 the distributions of three representative languages: Japanese, Indonesian, and English. From these results, we can see that for English and Japanese, the amount of information in tweets tends to follow a two-peaked distribution, with the majority of tweets containing 100-300 bits of information, but also with a peak at the 140 character limit where the authors adjust their content to make sure it fits into a single tweet. For Indonesian, the first hump is nonexistant, perhaps due to the propensity for quoting other's tweets more frequently than the other languages. Examining the fraction of tweets exactly at the 140 character limit,

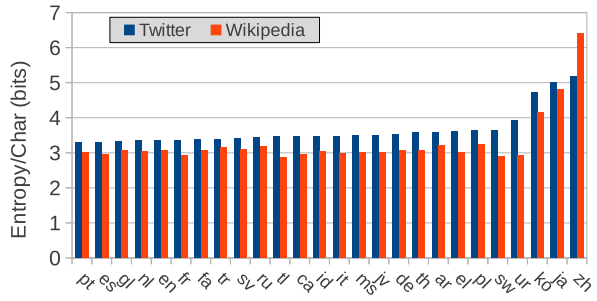

Figure 7: Entropy per character in tweets and Wikipedia text for each of the languages.

we can see that the values are 9.6%, 5.5%, and 2.1% for Indonesian, English, and Japanese respectively. Finally, while the average tweet in Japanese has a similar amount of information to that in English, it should be noted that on the upper end of the spectrum approximately 10% of tweets in Japanese do contain more information than can be packed into 140 characters in English.

Comparison with Canonical Text

While the previous section investigated cross-lingual differences in the amount of information contained in tweets, text on Twitter is only a small, special subset of language. Thus, a natural question is "how does the information content of a tweet compare to that of other genres of text?" This question was also independently inspected by (Pang and Ravi 2012), who found that English news text had a lower entropy per word than social media text. In order to provide a simple comparison towards answering this question in each of the 26 languages examined in the previous section, we gather text from Wikipedia.[5]

The procedure for measuring entropy is generally the same as that described previously, with the exception of the number of sentences. As Wikipedia sentences naturally tend to be longer than tweets, we choose 20,000 sentences for each of the languages, which gives approximately the same amount of text as 50,000 tweets. The sentences were cleaned of markup using the WikiExtractor.py[6] tool, and 20,000 were sampled randomly from the whole Wikipedia corpus, and arranged in random order.

The per-character entropy of the Wikipedia text in comparison with that of tweets is shown in Figure 7. It can be seen that for almost every language, the per-character entropy is significantly lower for Wikipedia than it is for tweets with the exception of Chinese and Japanese. The fact that per-character entropy is lower for most languages is natural, in that Wikipedia text is significantly more well-formed, and contains fewer abbreviations and lexical variation than Twitter text, as is noted by (Pang and Ravi 2012). On the other hand, for Chinese and Japanese the increase (or lack of change) of entropy can be explained by the fact that Twitter contains a larger fraction of Latin characters in the form of hash-tags, mentions, and links, while Wikipedia uses a greater fraction of characters from the large set of Chinese characters, as well as Japanese Hiragana, and Katakana. The actual fractions of Chinese/Japanese native characters in Twitter and Wikipedia are 35.5% and 63.6% for Chinese, 59.2% and 77.1% for Japanese respectively. As characters from these larger sets contain more information per character, having a larger fraction of these characters will lead to a higher average entropy.

[5]Retrieved Sept. 20, 2012.

[6]http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

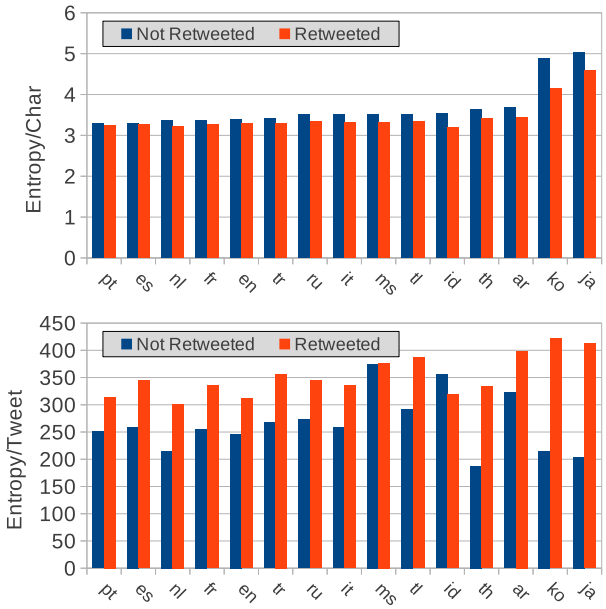

Figure 8: Entropy per character and tweet in non-retweeted and retweeted posts in each of the languages.

Finally, given the fact that the more well-formed text of Wikipedia has a lower entropy per character than standard tweets, we can return to our discussion of the difference between non-retweeted and retweeted posts from the previous section and see if we achieve similar results there as well. In order to measure the entropy, we take 50,000 non-retweeted and 50,000 retweeted posts from all of the 15 languages in the collection for which sufficient data exist, and use the previously described language model training procedure. To mitigate potential problems caused by posts retweeted by multiple people being counted multiple times, we remove all duplicate posts before measuring the entropy. The results of this examination can be found in Figure 8.

The first result that we can notice from the figure is that per-character information content is lower for retweeted posts in all of the 15 languages. Considering our observation that retweeted posts tend to have more consistent, canonical language, this is a natural result given the similar re-

sults on Wikipedia.[7] However, when looking at per-tweet information content, we can see the exact opposite result; in all languages except Indonesian and Malaysian (which have a disproportionally high ratio of non-retweet quotes), the per-tweet entropy is higher for retweeted posts. In other words, while the language of retweeted posts tends to be more consistent, retweeted posts tend be longer and thus contain more information on the whole. This is particularly true for languages with more information per character such as Japanese and Korean, with retweeted posts containing twice the information of their non-retweeted counterparts. This characteristic of retweeted posts, among many other factors (Suh et al. 2010), could also provide hints about which tweets are more likely to be retweeted.

Related Work

Examination of multilingualism on social media in general and Twitter in particular is an emerging field that has seen some research in the previous years. Much of this has focused on network-based approaches to quantify the relationships between people on these networks (Kulshrestha et al. 2012; Quercia, Capra, and Crowcroft 2012). For example, (Kulshrestha et al. 2012) studied the global follow network of Twitter users and measured the percentage of links that cross linguistic or geographical boundaries.

There has also been some work on examining how the language one uses affects the usage patterns of Twitter, such as the usage of hash tags, mentions, and retweets, or the number of tweets posted in a certain time period (Weerkamp, Carter, and Tsagkias 2011; Hong, Convertino, and Chi 2011). For example, (Hong, Convertino, and Chi 2011) found that certain languages have a higher percentage of tweets containing hashtags (18% in German: 18%, 14% English), while others have few (5% in Japanese, Indonesian, and Malay). For the percentage of tweets containing retweets, Indonesian and Malay are high (at 39% and 29%, respectively), while Japanese, German, and French are low (at 7%, 8%, and 9%, respectively). An open question is how much of these differences can be attributed to language characteristics, such as the entropy/character measure here, and how much can be explained by social dynamics.

We believe this paper is the first formal study of the effects of multilingualism on the actual amount of information contained in tweets. In addition, while it has been previously noted in the popular press that the amount of information that can be expressed in Chinese or Japanese is larger than that of English (Rosen 2011), this study also shows that this does not necessarily affect the actual amount of information that users choose to post.

Finally, there have been attempts to perform machine translation between languages within the 140-character Twitter limit (Jehl 2011). If we make the reasonable assumption that information content will be approximately identical across tweets, the fraction of tweets in the source language that can be expressed in 140 characters in the target language (as deducible by Figure 6) will help indicate the number of tweets that can be translated without loss.

[7] A small fraction of this difference can also be attributed to the fact that retweeted posts always start with "RT" and are thus easier to predict, but even when these characters are removed retweeted text still has a lower entropy per character.

Conclusion and Perspectives

This paper has examined the amount of content included in tweets from a multilingual, information theoretic perspective. In particular, we found that:

- Languages with large character sets tend to have more information per character, but this has little to no correlation with the average amount of information actually contained in a tweet.
- On the other hand, behavioral factors, such as the propensity to quote other's words, are better predictors of the average amount of information in tweets in a language.
- In comparison with more canonical Wikipedia text, across most languages, tweets generally contain more information per character, likely a result of Twitter-specific abbreviations and a less consistent writing style.
- Retweeted tweets tend to have more information on the whole, but less information per character, a result of the more consistent style of writing and a larger average number of characters per tweet.

On the other hand, this study only scratched the surface of the many possible correlations between language, behavior, and information content. For example: can the writing style of influential opinion leaders be captured by an information-theoretic measure such as entropy? When faced with character limits, how do authors reduce the amount of information in their tweets to fit into the boundaries? In what situations do authors split their posts over multiple tweets, and what connotations does this have from the information-theoretic perspective? All of these questions present interesting research directions for future work that could likely be further pursued with the tools we introduced in this paper.

References

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *Proc. HICSS'10*, 1–10.

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

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

Chen, S. F., and Goodman, J. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Honey, C., and Herring, S. C. 2009. Beyond microblogging: Conversation and collaboration via Twitter. In *Proc. HICSS'09*, 1–10.

Hong, L.; Convertino, G.; and Chi, E. H. 2011. Language matters in Twitter: A large scale study. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Jehl, L. 2011. Machine translation for Twitter. Master's thesis, The University of Edinburgh.

Kulshrestha, J.; Kooti, F.; Nikravesh, A.; and Gummadi, K. 2012. Geographic dissection of the Twitter network. In *International AAAI Conference on Weblogs and Social Media*.

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

Mori, S., and Yamaji, O. 1997. An estimate of an upper bound for the entropy of Japanese (in Japanese). *Journal of the Information Processing Society of Japan* 38(11).

Pang, B., and Ravi, S. 2012. Revisiting the predictability of language: Response completion in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pennell, D., and Liu, Y. 2011. Toward text message normalization: Modeling abbreviation generation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Quercia, D.; Capra, L.; and Crowcroft, J. 2012. The social world of Twitter: Topics, geography, and emotions. In *International AAAI Conference on Weblogs and Social Media*.

Rosen, R. J. 2011. How much can you say in 140 characters? a lot, if you speak Japanese. *The Atlantic*.

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

Stolcke, A. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Speech and Language Processing (ICSLP)*.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proc. of SocialCom*, 177–184.

Weerkamp, W.; Carter, S.; and Tsagkias, M. 2011. How people use Twitter in different languages. In *WebScience*.

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