Language Differences and Metadata Features on Twitter

Emre Kıcıman
Microsoft Research
Redmond, WA USA
emrek@microsoft.com

ABSTRACT

In the past several years, microblogging services like Twitter and Facebook have become a popular method of communication, allowing users to disseminate and gather information to and from hundreds or thousands (or even millions) of people, often in real-time. As much of the content on microblogging services is publicly accessible, we have recently seen many secondary services being built atop them, including services that perform significant content analysis, such as real-time search engines and trend analysis services. With the eventual goal of building more accurate and less expensive models of microblog streams, this paper investigates the degree to which language variance is related to the *metadata* of microblog content. We hypothesize that if a strong relationship exists between metadata features and language then we will be able to use this metadata as a trivial classifier to match individual messages with specialized, more accurate language models. To investigate the validity of this hypothesis, we analyze a corpus of over 72M Twitter messages, building language models conditioned on a variety of available message metadata.

1. INTRODUCTION

In the past several years, microblogging services like Twitter and Facebook have become a popular method of communication, allowing users to disseminate and gather information to and from hundreds or thousands (or even millions) of people, often in real-time. People use these microblogging services for a variety of purposes, from discussing news and opinions to chatting with friends and coordinating events. Users of these services span from individuals with a few friends to celebrities with millions of fans, and even corporations who are using these services to better communicate with their customers and reach out to their users.

As much of the content on microblogging services is publicly accessible, we have recently seen many secondary services being built atop them, including services that perform significant content analysis, such as real-time search engines and trend analysis services. Researchers have also begun to study microblog content to gain insight into user, group and community behaviors and communication patterns, as well personalization and information finding opportunities. While much of the academic research analyzing microblogging systems has focused on the social graph structures (e.g., studying information dissemination patterns or influence relationships), more recent work has begun to examine microblog textual content. As part of this trend, we are also beginning to see more use of statistical natural language processing techniques, such as n-gram models and latent variable topic models, applied to microblog content.

With the eventual goal of building more accurate and less expensive models of microblog streams, this paper investigates the degree to which language variance is related to the *metadata* of microblog content. While microblog content itself is quite short (Twitter, for example, limits message lengths to 140 characters) there is rich metadata associated with every message, including author metadata such as the name, location, and social details of a user; and easily inferred content metadata such as whether the message is a forward, a reply, contains a web link, or whether other users or topics are explicitly referenced. We hypothesize that if a strong relationship exists between metadata features and language then we can use this metadata as a trivial classifier to match individual messages with specialized, more accurate language models.

To investigate the relationship between metadata features and language, we collected a corpus of over 72M Twitter messages and their metadata, using 64.5M messages for training and reserving the rest for testing. For each metadata feature we studied, we divided the English portion of the corpus into subsets based on its feature value, and used each subset to learn an n-gram model. To quantify language differences, we measured the perplexities among these models, as well as to an n-gram model learned from the entire English portion of the corpus.

In our results we see that some metadata is correlated with language style. For example, as might be expected, the coarse location of users (e.g., their timezone) seems to have a strong relation to their aggregate observed language. As another example, we see language differences in messages based on the number of people who follow (i.e., subscribe to) the author: authors with more than 1000 followers write most differently, whereas the authors with less than a 1000, less than 100 or less than 10 are more similar.

In the rest of this paper, we describe our corpus and analysis techniques in more detail and then present the results

SIGIR 2010 WEB N-GRAM Workshop.
July 23, 2010, Geneva, Switzerland.

of our initial investigation. We conclude by discussing the implications of the work and possible future directions.

2. RELATED WORK

Previous studies have presented detailed characterizations of the Twitter service, and identified distinct classes of users based on their messaging, social and other behaviors [3, 2, 4]. A few studies have studied content on Twitter through systematic analysis of messages [1, 5]. For example, Naaman et al. [5] use human coding of 3379 messages from 350 users to infer two primary classes of users: "'informers"' who post messages that are primarily informational in nature (20% of users); and "'meformers"' who post messages that relate to themselves and their thoughts (80% of users). The recognition of such distinct classes of users and messaging activities leads us to suspect distinct language styles are in use on Twitter. Following [9] that improves information retrieval of web documents through the use of multiple distinct language models for the different fields of a document (e.g., title, abstract, body), we have begun to study whether different classes of messages in Twitter may also be better modeled separately.

Recently, studies have begun to apply statistical natural language processing techniques to microblog content. For example, Ramage et al. [6] use a form of Latent Dirichlet Allocation (LDA) to discover over 200 models of topic and style within a corpus of 8M twitter messages; and show how this can be used to provide personalized information consumption applications. Similarly, Ritter et al. [7] propose an approach to unsupervised modeling of conversations on Twitter, using LDA combined with a Hidden-Markov Modeling of dialog acts. We believe that such statistical modeling of message content will become an important tool in future research, as well as a foundation for information retrieval and other applications built atop microblogging services.

3. OUR DATASET

We collect our corpus of microblogging content from the Twitter service. Twitter is a social media service that allows users to broadcast short text messages to their "'followers"'. These messages are limited to 140 characters each. Most of these messages are also made publicly visible for searching and discovery. Begun in October 2006, Twitter has grown rapidly, and current reports state that Twitter users are posting 50M messages per day [8].

For our study, we gathered a sample of over 72M messages from Twitter's public stream, collected over the course of 3 days in May, 2010. We set aside a random 10% sample of these messages to use for testing. From the remaining 64M messages, we filtered for English content. We apply the rest of our analysis to these English messages. Table 1 shows basic statistics about the number of tokens, unique unigrams, bigrams and trigrams in our training dataset.

Pre-processing of Twitter messages consists of lower-casing the message and performing word-breaking on white-space and all non-alphanumeric characters. Since most URLs are auto-generated, pseudo-random short URLs, we also replace all URLs with a special "<URL>" token. We do not canonicalize user names or topic tags, annotated by Twitter convention with the '@' and '#' tags, respectively.

	All	English
Num Messages	64,565,242	27,677,009
Num Tokens	772,369,630	394,406,783
Unique 1-grams	34,150,462	5,423,111
Unique 2-grams	–	55,172,026
Unique 3-grams	–	187,217,338

Table 1: Basic characteristics of our Twitter training data set

4. ANALYSIS TECHNIQUE

For each metadata feature we studied, we divided the English portion of the corpus into subsets based on its feature value. For example, to study the relationship between geography and language, we divided our Twitter corpus into subsets based on the value of the message's time zone feature (more details in the next session). For discrete valued features, such as time zone, we only analyzed subsets with a substantial number of messages, e.g., we analyzed the London time zone, but not the Midway Island time zone. For other features, such as the number of followers of a user, we chose to subset boundaries such that all subsets had substantial numbers of messages.

Once we divided our corpus into subsets, we learned a separate n-gram language model for each subset. For each subset, we built a uni-gram, bi-gram and tri-gram language model, using a closed vocabulary based on the vocabulary of our entire training set of English messages. The language models are smoothed using Modified Kneser-Ney smoothing. For each feature we studied, we quantified language differences by measuring the perplexity of each of our learned n-gram models against each subset of data. As a further comparison point, we also measured the perplexity of a language model trained on the entire training set of English messages.

It is important to note that our analysis only serves to identify correlations between individual features and language differences. In particular, our algorithm does not account for potential correlations among features themselves. However, as a first investigation, our analysis does provide a simple but scalable technique that can identify features that can be used as a trivial classifier. In our future work, we plan to investigate and compare additional analysis techniques, such as message clustering and applying mechanisms of latent variable topic models to at least a subsample of our corpus.

5. RESULTS

In this section, we describe our results of investigating metadata features that relate to geography and number of followers of a user.

5.1 Geography and Language Style

As our first analysis, we present the correlation between geography and language style. Of course, it is natural to expect that geography have an impact on language style due to language dialects as well as geographic-specific topics, events, place names, etc.

For our analysis, we use the user provided "timezone" as our geographic location indicator. The timezone is a convenient geographic feature for our use because, while the

	Hawaii	Alaska	Pacific	Mtn.	Central	Eastern	Quito	Brasilia	Greenld.	London	Jakarta	Osaka	Tokyo
Hawaii	1573	3078	3623	2795	3018	3294	3094	5591	4228	2027	6623	5051	3529
Alaska	3506	1500	3238	2641	2866	3182	2892	11005	6496	2907	6004	12610	6477
Pacific	2775	1894	1303	1825	2040	2222	2226	11676	6501	2493	2769	11591	5611
Mtn.	4619	4379	5263	1360	2362	2742	2824	13384	7465	2874	17897	13453	7023
Central	4941	4655	5969	1774	1185	2009	1838	13244	7368	2695	24610	14107	6740
Eastern	5586	5208	7244	2053	1943	1216	1767	15560	8475	2648	31850	14535	6953
Quito	5042	4689	6539	2324	2200	2241	1153	8234	6061	2810	26049	13806	7197
Brasilia	8063	8279	10229	5674	6230	6528	6666	724	5810	4909	28775	11331	7465
Greenld.	4437	4776	5966	3642	4006	4170	4030	1932	1536	2868	14817	11179	5962
London	5013	5573	7160	3478	4065	4115	4266	10621	6561	917	21472	15561	7342
Jakarta	5631	4896	5494	5298	5761	6200	6138	17000	9690	4461	1338	12107	7407
Osaka	8276	8086	9359	6599	6944	7340	7252	16236	10461	5444	19994	1598	4495
Tokyo	5682	5546	6589	4521	5006	5043	5222	8904	6811	3635	13864	2386	1265

Figure 2: Perplexity of bi-gram models learned for each time zone (rows) with respect to test data from each time zone (columns)

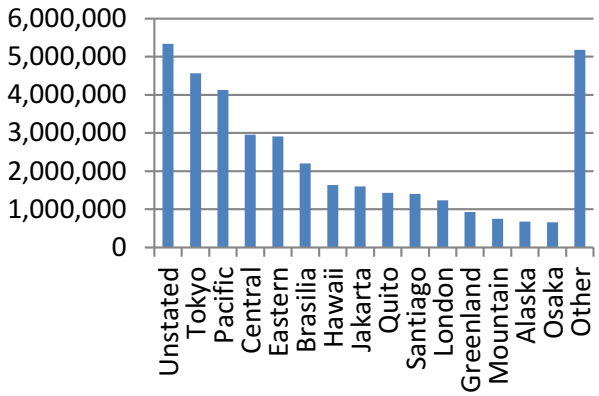

Figure 1: The number of messages collected from the top 13 timezones.

user chooses their own timezone, Twitter explicitly limits the users' choices to 30 different time zones. We build models for the languages of the top 13 time zones, ranked by the number of English language messages collected in each time zone (Figure 1).

In future work, we expect to utilize other geographic information also available in the Twitter metadata such as user-specified city, state, country information, or available GPS coordinates provides richer and more granular information, But, the use of user-specified location metadata requires the recognition and canonicalization of various names for locations (e.g., users refer to New York City as any one of 'nyc', "new york city", "new york, ny", "manhattan", etc). The use of automated or semi-automated location annotation (such as with GPS) can help, but is not yet widespread.

Figure 2 shows the perplexity of each language model with respect to each of the time zones. We see that the primarily English-speaking time zones (North America and London) use similar, though distinctive languages. The Quito time zone is also very similar to North America.

In Table 2, we show a selection of the most frequently used words in each time zone, and their relative likelihood α in comparison to the global English language model, calculated simply as $\alpha = P_{timezone}(token)/P_{global}(token)$. A qualitative inspection uncovers that the language differences among timezones seem to be due to several causes. First, the likelihood of *location names* is strongly associated with timezone. We see place names such as Denver, California, Miami, London, and Shibuya being much likely words in their respective time zones than globally. Secondly, we see variance in the probability of words associated with topics of a geographic nature. The topic "tweetmyjobs" is popular on the Eastern time zone, but not elsewhere, and is related to a company of the same name based in North Carolina. In the London time zone, we see that the topics '#bgt' (Britain's

Pacific		Mountain		Central		Eastern		London		Tokyo	
Token	α	Token	α	Token	α	Token	α	Token	α	Token	α
#sfo	5.9	denver	8.8	tx	2.8	#tweetmyjob	5.3	#bgt	8.7	#fm802noa	259.2
#craig	2.6	colorado	7.0	chicago	2.6	nyc	2.4	#eurovision	7.5	#twinglish	195.3
#forsale	5.5	chicago	1.8	texas	2.2	dc	2.2	cumbria	7.2	shibuya	161.2
hella	3.2	ugh	1.5	gon	2.0	miami	2.2	manchester	6.4	#followmejp	148.0
vegas	2.6	gon	1.5	naw	1.9	ny	2.0	uk	6.0	#youtube	76.7
california	2.0	bp	1.4	yall	1.9	boston	1.7	liverpool	5.6	soichi	70.8
favorite	1.2	energy	1.4	favorite	1.3	smh	1.7	london	5.4	#japan	63.5
		favorite	1.3	#tcot	1.8	favorite	1.2	favourite	3.6	youtube	2.1

Table 2: This table shows a selection of time zones and the words that appear more frequently in them as compared to other time zones, together with each word's likelihood ratio relative to the global English language model learned across all time zones.

	Alaska	Hawaii	Pacific	Mountain	Central	Eastern	London	Global
kobe	-3.9 (1.1)	-4.2 (0.5)	-3.9 (1.3)	-3.8 (1.1)	-3.7 (1.5)	-3.7 (1.6)	-5.1 (0.1)	-3.9
obama	-4.1 (0.7)	-4.1 (0.6)	-4.0 (0.9)	-3.8 (1.4)	-3.8 (1.2)	-3.8 (1.4)	-4.4 (0.3)	-3.9
lol	-2.3 (1.1)	-2.5 (0.8)	-2.3 (1.1)	-2.3 (1.2)	-2.3 (1.4)	-2.3 (1.3)	-2.6 (0.7)	-2.4
lool	-4.7 (0.6)	-4.0 (3.2)	-4.9 (0.3)	-5.1 (0.3)	-4.9 (0.3)	-5.0 (0.3)	-3.8 (5.3)	-4.5
haha	-2.7 (1.9)	-2.8 (1.3)	-2.8 (1.3)	-3.0 (1.0)	-3.0 (0.8)	-3.2 (0.6)	-2.9 (1.0)	-2.9
hahaha	-3.1 (2.1)	-3.4 (1.2)	-3.3 (1.4)	-3.6 (0.8)	-3.6 (0.7)	-3.7 (0.6)	-3.7 (1.0)	-3.4
ha	-3.5 (1.0)	-3.5 (1.1)	-3.5 (1.1)	-3.5 (1.3)	-3.5 (1.1)	-3.6 (0.8)	-3.4 (1.4)	-3.6
hahah	-3.8 (1.9)	-3.9 (1.4)	-3.9 (1.5)	-4.1 (1.0)	-4.2 (0.7)	-4.4 (0.5)	-4.2 (0.7)	-4.1

Table 3: This table shows the log-probabilities and relative likelihoods of selected words in the language models built upon different time zones.

Got Talent) and '#eurovision' are mentioned much more frequently than in the global language model. Other examples include the '#craig' (Craigslist) topic in the Pacific time zone, and the '#fm802noa' topic in Japan. Finally, we see dialect variance, such as the spelling variations ("favourite" vs. "favorite"), and colloquialisms such as the use of 'gon' instead of 'going to'.

In Table **??**, we drill into the topic variance across time zones by inspecting the relative popularity of two topic words, 'obama' referring to Barack Obama, and 'kobe' referring to Kobe Bryant, a basketball player. As might be expected, we see that both are more popular within the United States' time zones than in the London or Tokyo time zones. In this table, we also highlight dialect variance by showing the log-probabilities and relative likelihoods 'lol' ("laugh out load") and 'haha' (laughter) and their popular variants. Despite ease of global communication in Twitter, we see a strong correlation between such colloquial usage and geography as represented by time zone.

While it is clear from these results that geographic location and language are correlated, it remains for future work to investigate the most appropriate granularity and representation of location for the purposes of language modeling. While we used time zone in our analysis because of its clarity and convenience (i.e., we do not need inferrance or canonicalization machinery atop time zone metadata as we do for user-specified location information), other location information is also included in Twitter metadata and provides an opportunity to further investigate this question.

5.2 User features

Our next analysis is of a user-specific feature, the number

	<10	<100	<1000	>1000
<10	922	2413	4528	7831
<100	1166	1071	2477	4811
<1000	1682	1341	1216	2317
>1000	3345	2421	2804	1544

Figure 3: This figure shows the perplexity of models learned from each of our follower-groupings, with respect to the other followers

of followers of a user. Unlike bi-directional friend relationships on other social networking services, the one-way follower relationship on Twitter is an indication that a user is interested in reading what another user is broadcasting.

As shown in Figure 3, we see that while there are noticeable differences in language among the groupings of messages whose authors had less than 10, 100, and 1000 followers, the largest language difference occurs among messages whose authors had more than 1000 followers. Inspecting the differences in word probabilities, we do not find the same kinds of topic or dialect variance as we did with our language models conditioned on geography. Instead, we find differences in ego-centric words, such as 'I', 'me' or 'my', as well as in words that are indicative of how one uses Twitter (e.g., words such as 'RT' indicating a retweeting or forwarding of a message, and our '<URL>' token referencing a web page).

As shown in Table 4, we find that while usage of ego-

	<10	<100	< 1000	> 1000	Global
I	-1.6 (1.0)	-1.5 (1.1)	-1.5 (1.0)	-1.7 (0.7)	-1.6
my	-2.0 (1.1)	-2.0 (1.1)	-2.1 (1.0)	-2.3 (0.7)	-2.0
me	-2.2 (1.0)	-2.1 (1.0)	-2.1 (1.0)	-2.3 (0.8)	-2.2
you	-1.9 (1.0)	-1.9 (1.0)	-1.9 (1.0)	-1.9 (1.0)	-1.9
your	-2.5 (1.0)	-2.5 (1.0)	-2.5 (1.0)	-2.4 (1.3)	-2.5
rt	-2.3 (0.4)	-2.1 (0.8)	-1.8 (1.2)	-1.9 (1.0)	-1.9
<URL>	-1.6 (1.3)	-1.8 (0.9)	-1.8 (0.9)	-1.5 (1.5)	-1.7

Table 4: This table shows the log-probabilities and relative likelihoods of selected words in the language models built upon the messages of users with different numbers of followers.

centric words does not vary significantly for user groups with less than 1000 followers, there is a significant drop in the use of ego-centric words by users with more than 1000 followers. There is mixed evidence of a similar shift in the usage of 2nd-person words such as 'you' or 'your'. Again there is little difference in the likelihood in user groups with less than 1000 followers, but we do see that 'your' is more likely in the language model conditioned on users with more than 1000 followers.

Users with different numbers of followers appear to use Twitter's retweeting and URL referencing functionality at different frequencies as well. The 'RT' token is likely to appear in language models conditioned on users with fewer followers, and the 'URL' token is more probable in the language models built from user groups with either less than 10 or more than 1000 followers than in the other models.

6. DISCUSSION AND FUTURE WORK

In our study, we found that there is a strong correlation between language and metadata features, including both author and message characteristics. We quantified the relationship between the language style as represented by an n-gram model and features such as coarse-grained location and the number of followers of an author.

Our planned future work falls along three directions. First, we plan to strengthen our core results by continuing to gather additional data, experimenting with additional algorithms such as clustering techniques and LDA. Second, in addition to the geographic and user-metadata features we consider in this paper, we are analyzing content-related metadata, such as whether a message includes a URL, or is a retweet, the time-of-day. As part of this, we are also investigating questions of appropriate granularity. For example, assuming sufficient available data, would it be better to represent location at the city, state or country granularity instead of coarse-grained timezones? Finally, we are planning to experiment with techniques to build more accurate language models by taking advantage of the differences in language styles, and to apply these language models to various analyses of social network data, such as information retrieval and information extraction.

7. REFERENCES

[1] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS '10: Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[2] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *WebKDD/SNA-KDD*, 2007.

[3] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.

[4] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[5] M. Naaman, J. Boase, and C.-H. Lai. Is it really about me? message content in social awareness streams. In *CSCW*, 2010.

[6] D. Ramage, S. Dumais, and D. Liebling. Characterizing Microblogs with Topic Models. In *ICWSM*, pages 130–137, 2010.

[7] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *NAACL*, 2010.

[8] Twitter. Measuring tweets. `http://blog.twitter.com/2010/02/measuring-tweets.html`, 2010.

[9] K. Wang, X. Li, and J. Gao. Multi-style language model for web scale information retrieval. In *SIGIR 2010*, 2010.