Understanding Language Diversity in Local Twitter Communities

Amr Magdy[1], Thanaa M. Ghanem[2], Mashaal Musleh[3], Mohamed F. Mokbel[4]

[1,4]Dept. of Computer Science and Engineering, University of Minnesota, Minneapolis, MN
[2]Dept. of Information and Computer Sciences, Metropolitan State University, Saint Paul, MN
[3]KACST GIS Technology Innovation Center, Umm Al-Qura University, Makkah, KSA
{[1]amr,[4]mokbel}@cs.umn.edu, [2]thanaa.ghanem@metrostate.edu,
[3]mmusleh@gistic.org

ABSTRACT

Twitter is one of the top-growing online communities in the last years. In this poster, we study the language usage and diversity in Twitter local communities. We identify local communities in Twitter on a country-level. For each community, we examine: (1) the language diversity, (2) the language dominance and how it differs from local to global views, (3) demographic representativeness of tweets, and (4) the spatial distribution of different cultural groups within the community. We show fruitful insights about language usage on Twitter which can be exploited in language-based applications on top of tweets, e.g., lingual analysis and disaster management. In addition, we provide an interactive tool to explore the spatial distribution of cultural groups, which provides a low-effort and high-precision localization of different cultural groups.

1. INTRODUCTION

Twitter is one of the most popular social media where people used to post opinions, news items, updates on on-going activities,...etc. Everyday, 500+ million tweets are posted by 320 millions users. With such popularity, many techniques have exploited tweets for language-based analysis. This includes disaster management [3], multi-lingual usage [2], and language identification [4]. In most of these tasks, an implicit assumption has been made that English language is dominating other languages on Twitter to the extent that it could work as a language proxy for other languages [5], so that analyzing English tweets is enough to deduce conclusions about Twitter community. However, some crucial applications, like disaster management, are highly dependent on *local* language usage. For example, during China floods in 2012, propagating information about victims' locations on the Chiense Twitter (Sina Weibo) saved more than two hundred souls [1]. This imposes an important question weather language usage in popular social media is different on the local and global scales.

In this poster, we conduct a study to analyze and understand different aspects of spatial-language interaction in Twitter data, using a half billion of worldwide geo-tagged tweets. The whole dataset is used as a single global community and we identify country-level local communities based on tweets' locations. Then, we study four aspects of language usage within global and local communities. First, the diversity of language usage within both global and local communities using different measures. Second, the language dominance in the local communities and how this compares to the global Twitter community. This deduces fruitful insights on the overall language usage in Twitter data and clearly shows the prime importance of local languages in Twitter-based applications. Third, the representativeness of Twitter global and local communities for demographics of the real population. This relies on comparing language diversity measures with the data collected by international organizations, e.g., UNESCO. Fourth, the spatial distribution of different languages, and so cultural groups, within the country. This is provided as an interactive web-based tool that provides a low-effort and high-precision localization for different cultural groups inside the country. Such localization is of interest for several users, e.g., administrative authorities to localize Syrian refugees.

2. BACKGROUND AND DEFINITIONS

2.1 Def nitions

Our study mainly work on two concepts:

(1) **Twitter local community** of a certain country is defined by the set of all tweets posted within the spatial extent of this country.

(2) **Cultural group** is defined as the group of tweets that are posted in the same language. Throughout the study, we use Greenberg's language diversity index (LDI) as one of the measures to assess cultural diversity. LDI is used in UNESCO World Report on Cultural Diversity. LDI gives the probability of randomly selecting two persons with different native languages from a certain group of people. The higher LDI value, the higher cultural diversity.

2.2 Datasets

In our study, we use 445+ millions geo-tagged tweets that are collected through Twitter public streaming APIs during the period of October 12, 2013 to March 6, 2014. Each tweet

This research is capitally supported by NSF grants IIS-0952977, IIS-1218168, IIS-1525953, CNS-1512877, and the University of Minnesota Doctoral Dissertation Fellowship.

HT '16 July 10-13, 2016, Halifax, NS, Canada

ACM ISBN 978-1-4503-4247-6/16/07.

DOI: http://dx.doi.org/10.1145/2914586.2914612

Table 1: Diversity by LDI

Country	LDI
Macedonia	0.884
AAT	0.865
NA	0.857
Austria, Armenia	0.832
Morocco	0.821

is associated with a country using its geo-tag and public geographic datasets[1]. For language data, we use the language attribute, that is attached to tweets, as it comes from Twitter. To enrich our insights from the measured statistics, we compare our statistics with official organizations and major geographical database providers. Specifically, We use ISO 3166 and GeoNames country information datasets for getting country names and statistics on spoken languages. We also use UNESCO World Report on Cultural Diversity for getting UNESCO values of Greenberg's language diversity index (LDI) for different countries.

3. RESULTS AND CONCLUSIONS

In this section, we present our study results and conclusions on Twitter local communities of different countries. In our dataset, we have identified 206 Twitter local communities, each is corresponding to one country. Each community is divided into cultural groups. The dataset contains 55 different languages with average of 18 languages used within a single community and standard deviation of 12. Due to space limitations, we include only the most important results. Our poster presented in the conference would include more results. Below a summary of our results on four aspects of language usage.

Language diversity. In our full results, we use three measures for diversity within the community: (i) total number of languages, (ii) number of languages that cover 80% of tweets, and (iii) LDI as defined above. The last two measures have shown robust and consistent results as both of them consider the distribution of language usage within the community. For example, USA is the most diverse based on the first measure and encounters tweets with 44 different languages. However, 85% of USA tweets are posted in English and only 15% in all other languages, which shows much less diversity than other communities. Table 1[2] shows the most diverse local Twitter communities based on LDI. As shown, Macedonia shows the most diversity followed by Australian Antarctic Territory and Netherlands Antilles. In these three territories, the number of languages that cover 80% of their tweets are nine, seven, and six. This shows much more diversity beyond all other communities, that have 80% of the tweets in one to three languages only.

Language domination. Our analysis shows that tweets of 133 countries (∼65% of the countries) are dominated by the first spoken language in the country while the remaining 73 countries are dominated by a non-first language. This clearly shows that language domination in local Twitter communities is mostly for local language rather than international languages like English. In fact, most of countries that are dominated by English although it is not the first language, which are 41 out of 73, encounter low Twitter activity. This shows that English cannot work as a language proxy for other languages when the application is concerned with the spatial extent. Our full results shows that the domination of English in Twitter global community is interpreted by the high Twitter activity from USA and UK. In fact, 81.6% of the whole tweets are posted in only seven languages while 48 languages form only 18.4%. This confirms the observation that global language domination exist in Twitter global community which does not contradict with the domination of local languages in local communities.

[1]https://hiu.state.gov/data/data.aspx

[2]AAT: Australian Antarctic Territory, NA: Netherlands Antilles

Table 2: Number of countries that encounters % difference in LDI, e.g., 16 countries with difference in LDI values $\leq$ 5%

% of LDI Difference	Number of Countries
1	4
3	10
5	16
7	22
10	33

Demographic representativeness. To assess the validity of using Twitter as a representative for actual population, we consider language diversity based on LDI from tweets compared to real LDI values from UNESCO World Report on Cultural Diversity. Worldwide, for 206 countries and territories, we found a weak Pearson correlation of 0.25 between Twitter and real LDI values. However, we identified 33 countries (∼16% of the countries) that having less than or equal to 10% difference in LDI value between Twitter and the real value. This brings the attention again for focusing on local aspects of Twitter data. Although the global Twitter community does not look representative for the human population, certain local Twitter communities may represent their actual population. Table 2 shows the number of countries that encounters a certain difference in LDI values. For example, there are 16 countries with difference in LDI values less than or equal 5%. Our full results show the countries with the least difference in LDI values, which are the most promising candidates for more investigation on demographic representativeness of their tweets.

Spatial distribution of cultural groups. In our poster, we present a tool that enables visual analysis for language spatial distribution within a certain country. Using this tool, one can visually identify the spread of local cultural groups within the country through a web-based interface. This may be of interest for different users, e.g., local authorities to deal with certain situations like Syrian refugees. Our tool facilitates a low-effort and high-precision localization for different cultural groups around the country.

4. REFERENCES

[1] Sina Weibo, China Twitter, comes to rescue amid flooding in Beijing. http://thenextweb.com/asia/2012/07/23/sina-weibo-chinas-twitter-comes-to-rescue-amid-flooding-in-beijing/.

[2] I. Eleta and J. Golbeck. Multilingual Use of Twitter: Social Networks at the Language Frontier. *Computers in Human Behavior*, 41, 2014.

[3] I. V. et. al. Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster. In *ACL*, 2014.

[4] M. Graham, S. Hale, and D. Gaffney. Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer*, 66(4):568–578, 2014.

[5] The Geography of Twitter: Mapping the Global Heartbeat. irevolution.net/2013/06/09/mapping-global-twitter-heartbeat/.