Collection and Indexing of Tweets with a Geographical Focus

Adrien Barbaresi

Collection and Indexing of Tweets with a Geographical Focus

Adrien Barbaresi
Institute for Corpus Linguistics and Text Technology
Austrian Academy of Sciences
Sonnenfelsgasse 19 – 1010 Vienna
adrien.barbaresi@oeaw.ac.at

Abstract

This paper introduces a Twitter corpus currently focused geographically in order to (1) test selection and collection processes for a given region and (2) find a suitable database to query, filter, and visualize the tweets. Due to access restrictions, it is not possible to retrieve all available tweets, which is why corpus construction implies a series of decisions described below. The corpus focuses on Austrian users, as data collection grounds on a two-tier detection process addressing corpus construction and user location issues. The emphasis lies on short messages whose sender mentions a place in Austria as his/her hometown or tweets from places located in Austria. The resulting user base is then queried and enlarged using focused crawling and random sampling, so that the corpus is refined and completed in the way of a monitor corpus. Its current volume is 21.7 million tweets from approximately 125,000 users. The tweets are indexed using Elasticsearch and queried via the Kibana frontend, which allows for queries on metadata as well as for the visualization of geolocalized tweets (currently about 3.3% of the collection).

Keywords: Computer-Mediated Communication, Web Corpus Construction, Database Solutions, Visualization

1. Introduction

The availability and ease of use has made the online social networking service Twitter one of the most popular data sources for studying social communication (Leetaru et al., 2013). Generally, the interest in Twitter is considered to reside in the immediacy of the information presented, the volume and variability of the data contained, and the presence of geolocated messages (Krishnamurthy et al., 2008). Other social networks do not deliver the same amount of text, especially for German (Barbaresi, 2015b), and more importantly, cannot be deemed as stable in time in terms of popularity and API access (Barbaresi, 2013).

Short messages published on social networks constitute a "frontier" area due to their dissimilarity with existing corpora (Lui and Baldwin, 2014), most notably with reference corpora. Since August 2009, Twitter has allowed tweets to include geographic metadata (Stone, 2009), which are considered to be a valuable source for performing linguistic studies with a high level of granularity, e.g. on language variation (Ruiz Tinoco, 2013). Thus, from the point of view of corpus and computational linguistics, Twitter data are both highly relevant and difficult to process.

Due to access restrictions, mostly mechanical constraints on the API, it is not possible to retrieve all tweets one would need. For example, when using the so-called "gardenhose" streaming API, it is necessary to enter search terms or a geographic window, and a fraction of corresponding data is returned, which may greatly affect results (Morstatter et al., 2013), especially for highly frequent keywords as used by the TweetCat approach (Ljubešic et al., 2014) or for the *German Twitter Snapshot* (Scheffler, 2014). In that sense, focusing on a given geographical region can be a way to provide enough relevant linguistic evidence. However, there are structural characteristics which complicate the collection of tweets from German-speaking countries, and especially Austria, which makes it a interesting test case.

First, even without considering the market penetration of Twitter, the population of the country is comparatively small, so that Austrian users cannot be expected to be easily found at random, all the more since users preferentially connect to other users from their own country (Kulshrestha et al., 2012). Second, geolocated tweets are a small minority, with estimates as low as 2% of all tweets (Leetaru et al., 2013). Third, because of privacy concerns Austrian users can be expected to be very cautious about geolocation services: German twitterers for example are very reluctant to include geographic coordinates in their tweets (Scheffler et al., 2014). Finally, the success at being able to place users within a geographic region varies with the peculiarities of the region (Graham et al., 2014).

2. Design decisions

Following the characteristics stated above, and because corpus construction in the linguistic tradition implies a number of decisions which have to be made explicit (Barbaresi, 2015a), salient methodological issues will be dealt with in detail in this section.

First, while most studies ground on a collection process which is limited in time, the corpus described in this article is a monitor corpus in the sense that it grows constantly with time. Since metadata include the time of posting, it is possible to split the corpus in units of time. More generally, the purpose is to be opportunistic enough during corpus creation in order to enable researchers to tailor subcorpora which match particular interests.

Second, geolocated tweets (*place* element in the JSON response) may be casually sent from Austria, but not really by Austrian users: they can merely be an indication that the user has spent some time in Austria. Furthermore, it is technically possible to spoof one's location either by editing by hand the location field of a given tweet, or by tampering with the GPS device used for geolocation. On the other hand, the field which is sent with each tweet along with the user profile (*user/location* field), if given, refers to the subjective point of view of the users as regards their lo-

cation. It may not seem as objective as mere coordinates, and even when both the profile and the device location are valid, they do not always correspond (Graham et al., 2014) but it is a strong assertion regarding the place users feel at home or related to at least. Here lies the difference between a mere "posted from Austria" predicate and the corpus construction process which leads to tweets hopefully "made in Austria".

Third, since language cannot reliably be used as a proxy for location (Graham et al., 2014), no language selection is undertaken. For the same reason, retweets are included, even if the original messages may have been posted from other locations and in another context, because they are still considered to be meaningful. They can be removed for further studies by using the metadata as well as the "RT" mentions in the messages (Ruiz Tinoco, 2013). Furthermore, the use of typical Austrian-German words do not seem to lead to a substantial amount of users, due to the mobility of users and due to the difficulty to define a "national variety" (Ebner, 2008), which separates this case from languages like Croatian or Slovene (Ljubešic et al., 2014).

Fourth, geocoding algorithms can be used to help recreate absent geolocation metadata, using textual mentions of place (Leetaru et al., 2013) or linguistic cues (Scheffler et al., 2014) based on the identification of "local words" (Cheng et al., 2010). On the one hand, there are potential ambiguities in place names that have to be resolved to establish a reliable list of Austrian places, which implies a significant amount of work with an unknown outcome. On the other hand, the tweets are not exclusively in German and I do not agree with the segmentation of Austria in one bloc as used by (Scheffler et al., 2014). That is why no attempt is undertaken to recreate location metadata.

Finally, so-called "heavy tweeters" (Krishnamurthy et al., 2008) as well as peculiarities of the API (Morstatter et al., 2013) raise the question of sampling processes. Although human users usually entertain a stable amount of stable relationships (Gonçalves et al., 2011), it is conceivable that heavy users as well as machine-generated tweets account for distortions in the corpus. Additionally, the random sampling methodology used by Twitter to generate the streams of tweets is rarely put into question (Zafar et al., 2015). This means that steps have to be taken in order to minimize the impact of differences in user activity as well as potentially unknown sampling biases.

3. Implementation

To sum up the methodological concerns, what is needed is a method allowing to find and collect tweets from Austrian users with a reasonable precision. My method uses different modules as presented in figure 1. The first component can be considered to be a "lurker" module in the sense that it merely listens to the Twitter's streaming API[1] to collect geolocated tweets whose coordinates are in or close to Austria. Tweets featuring geolocation in Austria or with a user profile location field linked to Austria are singled out. The corresponding user names are then passed to a second module which fetches user streams in order to analyze them. Additionally, the social networks (friends and followers) are crawled (Kumar et al., 2014) in order to find other potentially interesting users, which makes the operation comparable to an API-side focused or scoped crawling (Olston and Najork, 2010). The communication with the API relies on the Python wrapper *twython*.[2]

The constant filtering is meant to optimize the collection. In fact, there are mechanical constraints on both ends: access to the API on user level is limited to 180 requests per slot of 15 minutes, and on the other side unneeded content may clutter up storage devices. Additionally, I found that potentially interesting users are geographically and linguistically very mobile; they may use several languages and be tied to several home places. Finally, even among users who use geolocation services, the proportion of tweets with actual location data may greatly vary, so that users are unequally productive in this respect.

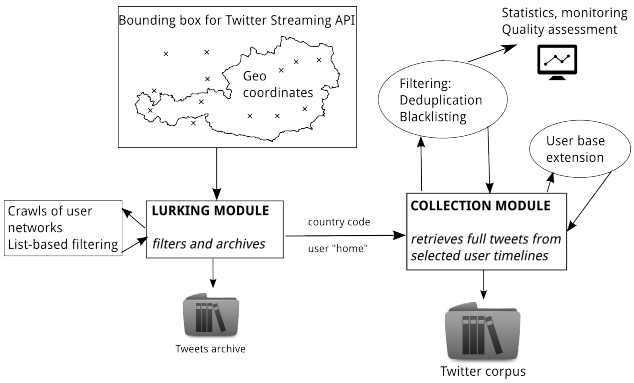

Figure 1: Schema of the implementation

Studies have shown that it is desirable to gather a set of users which is both large and diverse (Zafar et al., 2015), so that the collection process is opportunistic despite a rather conservative setting concerning location: at least 50% of geolocated tweets per user have to be in Austria. Positives in the user location field are found on token level using a fixed list of case-insensitive cues: nationwide mentions (e.g. Austria), all regions (*Bundesländer*), well-known landscapes (e.g. *Waldviertel*), and top-20 cities. For the sake of completeness, the main quarters of the major cities (e.g. *Josefstadt* in Vienna) as well as major geographical features (e.g. valleys and rivers) have been added, however they seem to be rarely used. Quantitatively speaking, the number of users found that way (around 125,000) is concordant with results from market studies, with an estimated number of 140,800 Austrian users in September 2015.[3]

The corpus is constantly growing, and so is the user base. Filtering steps include the deduplication of tweets and the blacklisting of unwanted users, which both yield statistical information for quality assessment. At the same time, remaining tweets are scanned for other user names in replies or retweets, whose timelines are retrieved and stored if they match the location criteria. In order to avoid bias by heavy twitterers, the timelines are fetched at random intervals among the range of valid users.

[1] https://dev.twitter.com/streaming/overview

[2] https://github.com/ryanmcgrath/twython

[3] http://de.statista.com/statistik/daten/studie/296135/umfrage/twitter-nutzer-in-oesterreich/

4. Indexing and results

To keep up with the growing amount of tweets, a specific search engine has been chosen. The interest of NoSQL databases to deal with the feature-rich content return by the Twitter API is known (Kumar et al., 2014). Two main components of the open-source *ELK* stack (Elasticsearch, Logstash, Kibana) are used, namely Elasticsearch[4] to index the tweets and Kibana[5] to provide a user-friendly interface to queries, results, and visualizations. The main drawbacks result at the time being from the lack of linguistic processing: a rather unprecise lemmatization of queries and results by the search engine as well as a lack of linguistic annotation. These tasks will require a substantial amount of testing due to the multiple languages and the difficulty of twitter messages.

Although it is not primarily a search engine for linguists, Elasticsearch takes advantage of the native JSON format of the tweets as well as of a number of relevant field types after a subsequent mapping, which allows for refined queries on text and metadata, for instance "the *-erl* diminutive form in tweets from users with more than 10 followers and with the city of Klagenfurt mentioned in the home location field". In the current implementation, using Kibana's syntax, this query translates to *text:*erl AND user.followers_count:[10 TO *] AND user.location:Klagenfurt*. In order to give a user-friendly access to the results, dashboards can be configurated out of a series of indicators (see figure 2).

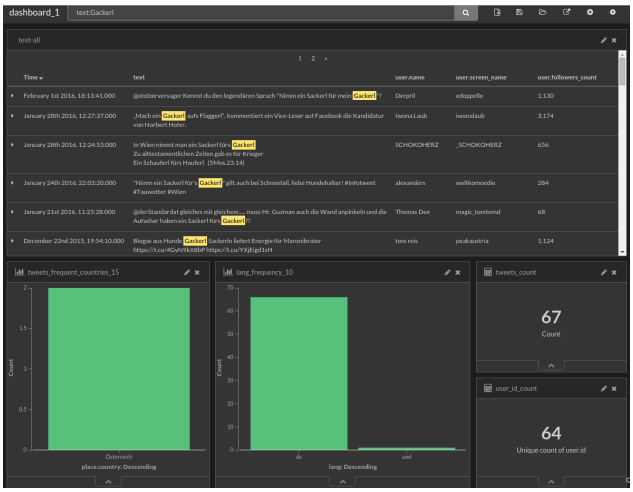

Figure 2: Example of dashboard view

The most frequent languages according to the metadata delivered by Twitter are English (42.2% of all tweets) and German (40.5%), with a number of less frequently represented languages such as Turkish (2.8%), Spanish (1.3%), and Japanese (0.9%). The amount of tweets whose language could not be determined by Twitter is relatively low (6.5%), which indirectly yields insights on the quality of the corpus. This information is confirmed by the mean length of the tweets (100.4 characters and 12.7 tokens).

[4]https://www.elastic.co/products/elasticsearch

[5]https://www.elastic.co/products/kibana

The proportion of geolocated tweets (3.3%) is better than in the comparable *German Snapshot* (Scheffler, 2014), where it amounts to 1.1%. Their distribution by country is largely in favor of Austria (75.0% of geolocated tweets), with a number of other less prominent countries such as the USA (6.2%), Germany (4.1%), and Turkey (1.6%). These figures show that it is necessary to target Austria in comparison to a general approach targeting German. Visualizations of geographical data can be constructed "out of the box" as soon as coordinates have been mapped as geographical data in the database, which allows for the projection of geolocated tweets on a map.

A heat map centered on Austria is shown in figure 3. The distribution of tweets is mostly in line with population distribution, with the exception of Klagenfurt. It highlights the prominence of Vienna and its airport as well as the importance of commuters and travellers, with train tracks partially visible. Holiday resorts such as ski stations are also depicted on the map, which altogether prompts for geographical and sociological analyses of mobility.

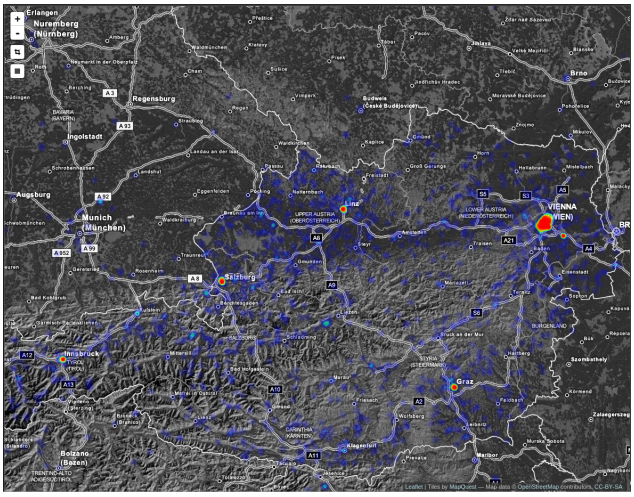

Figure 3: Heat map of all geolocated tweets

5. Conclusion

I introduced a monitor corpus of tweets from Austrian users. The data collection grounds on a two-tier detection process addressing corpus construction and user location issues. The emphasis lies on short messages whose sender (1) mentions a place in Austria as his/her hometown or (2) often tweets from places located in Austria. The resulting user base is then queried and enlarged using random sampling. The current volume of the corpus is 21.7 million tweets from approximately 125,000 users, which is roughly comparable to the *German Snapshot* (Scheffler, 2014) in terms of volume with a number of users one order of magnitude smaller. The tweets are mainly written in English and German. The proportion of geolocated tweets is 3.3%, 75.0% of which come from Austria.

Future work includes work on fine-grained differences in geolocations which could improve the quantitative throughput as well as the qualitative value of the corpus. In the

same perspective, ambiguities of gazetteers have to be reduced to a minimum in order to use them in the user selection process, as the corpus collection will be extended to Germany and Switzerland. Further, user names could be used in order to improve filtering and get insights on distributions of language and gender in the corpus (Jaech and Ostendorf, 2015). Last, tweet identifiers can allow for reuse of the corpus (McCreadie et al., 2012) which could also be done with user identifiers.

6. Bibliographical References

Barbaresi, A. (2013). Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of the 51th Annual Meeting of the ACL, Student Research Workshop*, pages 9–15.

Barbaresi, A. (2015a). *Ad hoc and general-purpose web corpus construction*. Ph.D. thesis, ENS Lyon.

Barbaresi, A. (2015b). Collection, Description, and Visualization of the German Reddit Corpus. In *2nd Workshop on Natural Language Processing for Computer-Mediated Communication, GSCL conference*, pages 7–11.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768. ACM.

Ebner, J. (2008). *Duden: Österreichisches Deutsch*. Dudenverlag.

Gonçalves, B., Perra, N., and Vespignani, A. (2011). Modeling users' activity on Twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656.

Graham, M., Hale, S. A., and Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.

Jaech, A. and Ostendorf, M. (2015). What Your Username Says About You. *arXiv preprint arXiv:1507.02045*.

Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). A Few Chirps about Twitter. In *Proceedings of the First Workshop on Online Social Networks*, pages 19–24. ACM.

Kulshrestha, J., Kooti, F., Nikravesh, A., and Gummadi, P. K. (2012). Geographic Dissection of the Twitter Network. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 202–209.

Kumar, S., Morstatter, F., and Liu, H. (2014). *Twitter Data Analytics*. Springer.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5).

Ljubešic, N., Fišer, D., and Erjavec, T. (2014). TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. *Proceedings of LREC*, pages 2279–2283.

Lui, M. and Baldwin, T. (2014). Accurate Language Identification of Twitter Messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 17–25.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., and McCullough, D. (2012). On Building a Reusable Twitter Corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1113–1114. ACM.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of ICWSM*.

Olston, C. and Najork, M. (2010). Web Crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246.

Ruiz Tinoco, A. (2013). Twitter como Corpus para Estudios de Geolingüística del Español. *Sophia Linguistica: working papers in linguistics*, (60):147–163.

Scheffler, T., Gontrum, J., Wegel, M., and Wendler, S. (2014). Mapping German Tweets to Geographic Regions. In *Workshop Proceedings of the 12th KONVENS conference*.

Scheffler, T. (2014). A German Twitter Snapshot. In *Proceedings of LREC*, pages 2284–2289.

Stone, B. (2009). Twitter Blog: Location, location, location. https://web.archive.org/web/20090823032127/http://blog.twitter.com/2009/08/location-location-location.html.

Zafar, M. B., Bhattacharya, P., Ganguly, N., Gummadi, K. P., and Ghosh, S. (2015). Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream. *ACM Transactions on the Web (TWEB)*, 9(3):12.