TweetCaT: a tool for building Twitter corpora of smaller languages

Nikola Ljubešić,[*] Darja Fišer,[†] Tomaž Erjavec[‡]

* Dept. of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
nikola.ljubesic@ffzg.hr

† Dept. of Translation, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

‡ Dept. of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

This paper presents TweetCaT, an open-source Python tool for building Twitter corpora that was designed for smaller languages. Using the Twitter search API and a set of seed terms, the tool identifies users tweeting in the language of interest together with their friends and followers. By running the tool for 235 days we tested it on the task of collecting two monitor corpora, one for Croatian and Serbian and the other for Slovene, thus also creating new and valuable resources for these languages. A post-processing step on the collected corpus is also described, which filters out users that tweet predominantly in a foreign language thus further cleans the collected corpora. Finally, an experiment on discriminating between Croatian and Serbian Twitter users is reported.

Keywords: Twitter corpora, open source, less-resourced languages, Croatian, Serbian, Slovene

1. Introduction

Twitter is a microblogging platform that was created in 2006 and offers the users the ability to interact with other members in the community in real time over the Internet or on their mobile phones. The users send short, 140-character messages, called "tweets" to their "followers" (other users who subscribe to those messages). Today, there are more than 550 million Twitter users with over 100,000 new users joining every day. On average, almost 60 million tweets are published every day (Ponzetto and Zielinski, 2013). Tweets are becoming an important data source for natural language processing and corpus linguistics and with its growing popularity Twitter an increasing number of tweets are written in languages other than English (Jacobs, 2011), which is why it is becoming increasingly important to be able to process tweets in other languages as well.

One of the first and best known Twitter corpora is the Edinburgh Twitter Corpus (Petrović et al., 2010) that contains almost 100 million tweets which were collected over a period of two months using the Twitter streaming API. Another well-known Twitter collection is the Stanford Twitter Corpus (Yang and Leskovec, 2011), which contains 467 million posts from 20 million users covering a 7 month period in 2009. However, due to Twitter's terms of service,[1] they are no longer available as a dataset.

McCreadie et al. (2012) have overcome this problem by developing Tweets2011, a methodology to distribute a set of tweet identifiers and a separate tweet crawling tool for downloading the identified tweets. Using the Twitter API to collect tweets has become the standard way of compiling Twitter corpora and has been described and discussed in detail by Kumar et al. (2013).

However, these approaches assume English as the language of interest and cannot be directly used for other languages, especially smaller ones, such as Croatian (4 million speakers), Serbian (7 million) and Slovene (2 million), for which standard techniques would return results of which only a small fraction would be useful as a source of data for the language in questions.

We present an alternative method that uses seed terms and a simple language identification module to find new users as well as new tweets from already known users that tweet in the target language. The tool is named TweetCat, as the basic idea follows the well known BootCat[2] (Baroni and Bernardini, 2004) tool, which collects URLs of Web pages from seed terms in order to build Web-based corpora of particular languages and domains. We build two Twitter corpora, one for Croatian and Serbian, and one for Slovene. In a final experiment we attempt to discriminate between Croatian and Serbian Twitter users with partial success, but obtain interesting insights in the problem and Twitter popularity among the speakers of the two languages. The collection tool as well as the compiled corpora are available under permissive licenses. The tool can easily be adapted for other languages.

The rest of this paper is structured as follows: Section 2 presents the TweetCat tool, Section 3 gives an analysis of the collection procedure, Section 4 details several post-processing steps for corpus clean-up and Section 5 gives conclusions and directions for future work.

[1] https://twitter.com/tos

[2] http://bootcat.sslmit.unibo.it/

2. Description of TweetCaT

Since only a small fraction of the streamed tweets are in the desired language, we use the Twitter searching API to identify the tweets containing user-specified seed terms specific for the language. The seed terms are manually selected to be fairly high-frequency content words that are, however, not in the vocabulary of other languages. The number of such terms can be quite small: we have defined 40 seed terms for Croatian and Serbian and 20 for Slovene.

The tool is written in Python and iterates a user-specified number of times over the two basic steps:

1. querying the search API with the goal of identifying new users and
2. retrieving new tweets of already known users

In the first step we query the search API with every seed term and check if the retrieved users are already in our user index. If they are not, we retrieve their timeline of recent tweets and perform basic language identification over this collection. We determine the language of the tweets by comparing their vocabulary against a user-defined list of very frequent words. Again, this list can be quite small, for Croatian and Serbian we used 87 words and 40 words for Slovene. We temporarily discard users with timelines of less than 100 tweets because this turns out to be a minimum for a reliable language estimate.

Once a user tweeting in the language is identified, (s)he is added to the user index and checked for followers and friends (users (s)he follows) assuming that they will tweet in the same language. Each of the followers and friends is subjected to the same language identification procedure as the initial user.

The second step consists of retrieving new tweets from the timelines of all known users. No additional language identification is performed once a user has passed the initial check since this would significantly complicate our straight-forward procedure which main goal is to produce high-recall data collections that can be filtered later on, when the maximum amount of information on users is available.

The whole procedure presented in pseudocode is as follows:

```
function lang_id(author):
  timeline=tokenize(timeline_api(author))
  return coverage(timeline,function_words)>=threshold

for seed_term in seed_terms:
  for hit in search_api(seed_term):
    if hit.author not in user_index:
      if lang_id(hit.author):
        user_index.add(hit.author)
        for follower in follower_api(hit.author):
          if follower not in user_index:
            if lang_id(follower):
              user_index.add(follower)
        for friend in friend_api(hit.author):
          if friend not in user_index:
            if lang_id(friend):
              user_index.add(friend)

for user in user_index:
  for tweet in new_tweets_api(user):
    output(tweet)
```

The tool outputs the tweets in a simple XML format, where each tweet is given its metadata along with the name of the author and the text of the tweet, as shown below:

```
<tweet id="429333550584721413"
    created_at="2014-01-31T19:21:09"
    retrieved_at="2014-02-01T17:29:55.98763"
    favorite_count="19" retweet_count="1">
<screen_name>dfiser3</screen_name>
<text>Kurc pa petkov večer, v katerem je
    beseda izpiti samostalnik,
    ne glagol.</text>
</tweet>
```

The information about the number of users that favorited or retweeted the tweet is likely to change during time. The fact that this collection will be distributed as a script for downloading the collection via the API has the benefit of keeping these data as up-to-date as possible.

3. Analysis of the collection procedure

While the tool can be run for a specified period of time or to gather a specified amount of text, it is our goal to run the tool continuously in order to compile monitor corpora for all three languages, especially because historical Twitter data is not available through the API. To gain insight into the dynamics of collecting the tweetosphere for smaller languages we present a timeline of identifying new tweets and users in Figure 1 for the first 235 days of running the tool.

The figure shows that both curves drop rapidly in the first ten days and stabilize from that point on. For Croatian and Serbian an average 320k tweets are collected daily in the first ten days, which drops by half in the next 10 days. After that period the number of new tweets stays constant on around 120k tweets daily. The number of new users behaves in a similar fashion although with much smaller numbers, identifying on average 2.3k users daily in the first ten days, going down to constant 110 users a day for the remainder of the timeframe.

For Slovene we observe a similar phenomenon with a day average of 59k tweets in the first ten days and constant 20k new tweets in the remainder of the timeframe. In the first ten days we identify just below 400 users daily which drops off to average 13 new users a day.

As can be seen, this approach yields a reasonably sized Twitter collection in a matter of days by collecting 2.6 million of Croatian and Serbian tweets and half a million of Slovene tweets in the first week. Running the tool for a longer period produces a smaller, but constant and still significant stream of new data.

4. Post-processing

The amount of data collected in the first 235 days (hrsrTweets and slTweets) is given in the first half of Table 1. We decided to post-process this data collection with the idea of producing Twitter corpora suitable for a wide range of applications.

We focused on further tuning language identification by removing users who tweet more in languages similar to the languages of interest and users who tweet significantly more in any other language than in the languages of interest. It is important to notice that we were not able to make

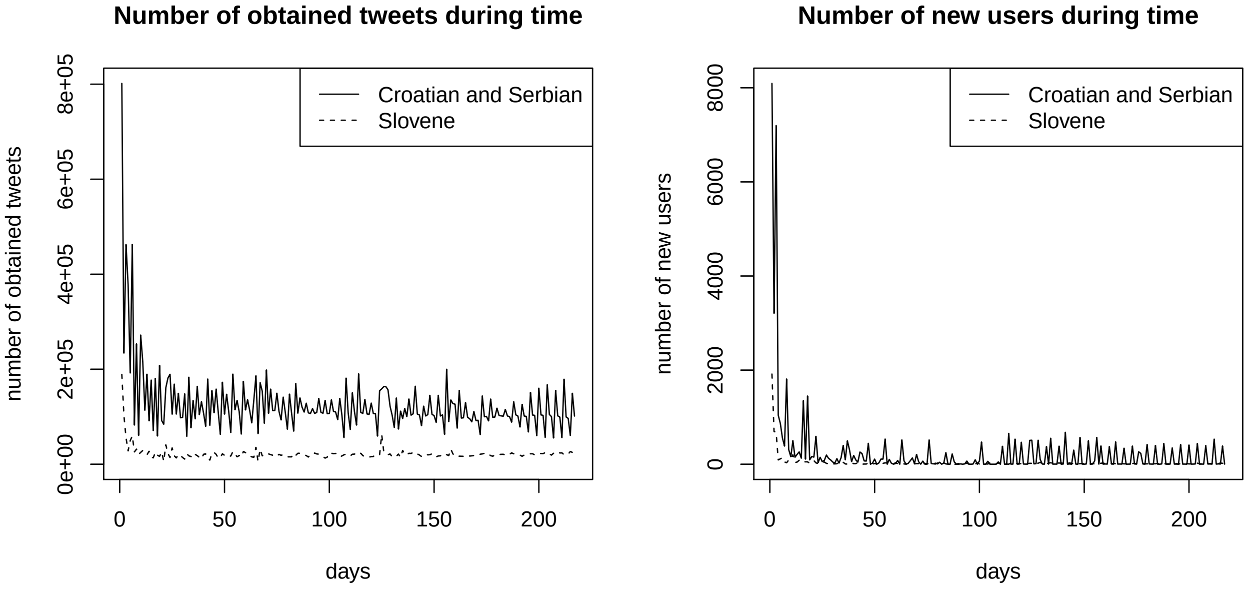

Figure 1: The number of new tweets and users obtained during time on both corpora

a decision of this quality during the data collection process because of a smaller amount of information available. It could be possible to extend the TweetCaT tool with an additional filter that would be applied after enough data was collected and we consider this to be one of the most likely improvements for the next version of the tool.

We first used the Python module langid.py (Lui and Baldwin, 2012) with off-the-shelf models to identify the language of each tweet with mentions, retweets, URLs and hashtags removed. In both corpora langid.py identified tweets written in 97 different languages (all the languages langid.py is trained on) showing that per-tweet language identification is a very difficult task. In each corpus around 50% of the tweets are identified as being in the sought language(s), followed by English with around 15-20% of tweets, similar languages having around 10%, with each of all other languages being identified in less than 1% of the tweets.

We encoded the results of this procedure on the level of each tweet yielding the following final format of each tweet:

```
<tweet id="429333550584721413"
    created_at="2014-01-31T19:21:09"
    retrieved_at="2014-02-01T17:29:55.98763"
    favorite_count="19" retweet_count="1"
    lang="sl" prob="0.999999999423"
    norm_length="72">
<screen_name>dfiser3</screen_name>
<text>Kurc pa petkov večer, v katerem je
    beseda izpiti samostalnik,
    ne glagol.</text>
</tweet>
```

After annotating each tweet with the language attribute, we applied a user-level filter based on two criteria. With the first criterion we remove users that tweet more in a defined set of similar languages than in the language(s) of interest removing thereby errors in our simple language identification procedure. With the second criterion we remove users that tweet twice as often in any other language than the language(s) of interest removing thereby users that tweet probably mostly in English.

The first criterion filters out 16% of users in the Croatian and Serbian collection and 20% of users in the Slovene collection. The second criterion filters out a further 20% of users in the Croatian and Serbian collection and 25% of users in the Slovene one. The intersection between the sets of users that were filtered out with each criterion is quite high, 85%, while12% of users satisfy the second criterion only and 2% of users only the first criterion. This points to the direction that using the second criterion, i.e. removing users who tweet in any other language twice as often, could be sufficient for this task.

After performing post-processing, the size of the current 235 day corpora (called hrsrTwitterCorpus and slTwitterCorpus) by the number of users, number of tweets and number of words is given in the second part of Table 1.

Finally, we applied a recently developed language identifier (Ljubešić, 2014) that showed to be very efficient in discriminating between very closely related languages, in particular Croatian and Serbian web data. It is based on unigram language models trained on 1.9 billion tokens from the hrWaC Croatian top-level-domain web corpus[3] and 900 million tokens of from the srWaC Serbian top-level-domain web corpus[4] with a simple maximum-a-posteriori (MAP) decision rule. On the task of discriminating between those languages on web text, the approach has shown to cut the error of the very efficient Blacklist classifier (Tiedemann and Ljubešić, 2012) four times. It is important to note that

[3] http://nlp.ffzg.hr/resources/corpora/hrwac/

[4] http://nlp.ffzg.hr/resources/corpora/srwac/

	users	tweets	words
hrsrTweets	51,381	26,047,874	290,557,841
slTweets	7,284	4,504,745	55,492,816
hrsrTwitterCorpus	41,807	21,360,940	235,952,967
hrTwitterCorpus	4,465	2,070,381	23,410,410
srTwitterCorpus	26,869	14,072,777	157,531,327
slTwitterCorpus	5,483	3,035,304	38,465,311

Table 1: The number of users, tweets and words in the initial data collections and the final corpora

the Blacklist classifier was already performing at the 100% accuracy level on newspaper texts, but has shown to struggle with everything the Web has to offer.

The language identifier was run on the collection of each user's tweets with URL-s, hashtags and mentions removed. Beside the MAP decision we also calculated a normalized log-probability for each language so that the sum of those normalized probabilities sums to -1.

Manual inspection of the results revealed that Serbian users are correctly classified, but that users identified as borderline Croatian in many cases are actually Montenegrin, who use the ijekavian yat reflex, but partially use Serbian-specific lexis and Serbian-specific syntactic patterns.

By plotting the distribution of the normalized log probability for each user regarding the Croatian model, which is depicted in Figure 2, we realized that the Montenegrin (and possibly Bosnian?) tweet distribution is actually visible on the right side of the left, bigger, Serbian user distribution. This was also our first realization that the Twitter service is obviously very popular in a country of only 600 thousand inhabitants. Additionally, we were surprised by the drastically larger number of users tweeting in Serbian than in Croatian.

Having in mind our current limitations regarding the capability to discriminate between Croatian and Serbian Twitter users, we defined for now two subcorpora, hrTwitterCorpus and srTwitterCorpus, where the srTwitterCorpus contains users that were classified by MAP as Serbian (dotted line in Figure 2) while for the hrTwitterCorpus the normalized log-probability criterion was set to -0.49 (full line in Figure 2). The final sizes of those corpora are given in Table 1. The normalized log-probabilities for each user are distributed as part of the corpora enabling each user to redefine the language criterion.

It is important to stress that the produced corpora do not contain tweets in the specific language only, but all tweets of users that mostly tweet in that language. This design criterion was motivated by the fact that cleaning up a Twitter collection on the level of tweets runs the danger of removing also mixed-language tweets, which are very frequent esp. with English, and whether such tweets should actually be removed heavily depends on the intended usage of the collection.

5. Conclusion

We have presented TweetCaT, a freely available tool for collecting Twitter corpora of smaller languages and the Twitter corpora of Croatian, Serbian and Slovene we have built with the tool. The procedure aims to collect the largest corpus with the least strain on the available Twitter APIs. The tool is published on github[5].

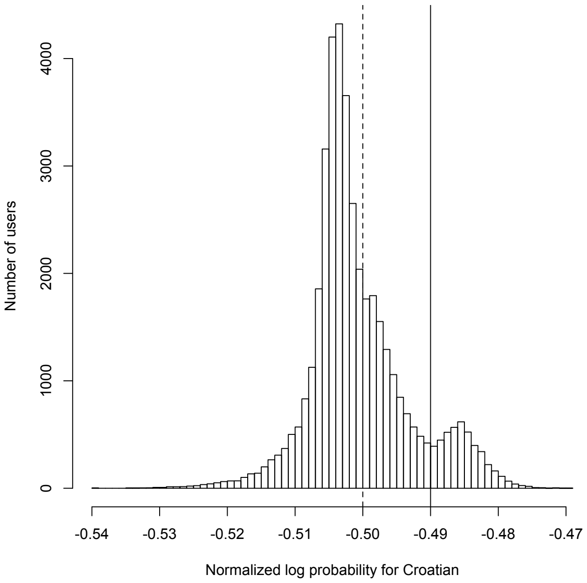

Figure 2: Distribution of the normalized log probability for Croatian regarding the discrimination between Croatian and Serbian on the user level

In our further work we plan to make the corpora available via a concordancer. We are in the process of making the corpora available for download following the Twitter Terms of Service, as a set of tweet identifiers, their annotations and a tweet crawling tool for downloading and recreating the corpus[6]. We are also working on way to normalize the language of Croatian and Slovene tweets, which is often colloquial and informal, in order to be able to process it with tools tranined on standard language; we have currently taken the first steps in this direction, by using character-based statistical machine translation on a test collection of Slovene tweets (Ljubešić et al., 2014). Finaly, it would be interesting to further investigate language identification for very closely related languages, in order to be able to reliably distinguish not only Croatian and Serbian tweets, but also Montenegrin and Bosnian ones.

6. Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement no. PIAP-GA-2012-324414 (project Abu-MaTran) and the Research programme of the Slovenian Research Agency P2–0103 ”Knowledge Technologies”.

7. References

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *In Proceedings of LREC 2004*, pages 1313–1316.

[5] https://github.com/nljubesi/tweetcat

[6] http://nlp.ffzg.hr/resources/corpora/twitter/

Frank Jacobs. 2011. 539 - Vive le tweet! a map of Twitter's languages.

Shamanth Kumar, Fred Morstatter, and Huan Liu. 2013. *Twitter Data Analytics*. Springer, New York, NY, USA.

Nikola Ljubešić. 2014. {bs,hr,sr}WaC: Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the WAC-9 Workshop*.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišser. 2014. Standardizing tweets with character-level machine translation. In *Proceedings of the 15th International Conference, CICLing 2014*, Lecture Notes in Computer Science. Springer.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *ACL (System Demonstrations)*, pages 25–30.

Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. 2012. On building a reusable twitter corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1113–1114, New York, NY, USA. ACM.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 25–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simone Paolo Ponzetto and Andrea Zielinski. 2013. Exploiting social media for natural language processing: Bridging the gap between language-centric and real-world applications. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Tutorials)*, pages 5–6, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, New York, NY, USA. ACM.