Crawling microblogging services to gather language-classified URLs. Workflow and case study

Adrien Barbaresi

Crawling microblogging services to gather language-classified URLs Workflow and case study

Adrien Barbaresi
ICAR Lab
ENS Lyon & University of Lyon
15 parvis René Descartes, 69007 Lyon, France
adrien.barbaresi@ens-lyon.fr

Abstract

This paper presents a way to extract links from messages published on microblogging platforms and their classification according to the language and possible relevance of their target in order to build a text corpus. Three platforms are taken into consideration: FriendFeed, identi.ca and Reddit, as they provide a relative diversity of user profiles and, more importantly, user languages. In order to explore them, I introduce a traversal algorithm based on user pages. As I target lesser-known languages, I tried to focus on non-English posts by filtering out English text. Using mature open-source software from the NLP research field, a spell checker (aspell), and a language identification system (`langid.py`), my case study and benchmarks give an insight into the linguistic structure of the considered services.

1 Introduction

1.1 The 'Web as Corpus' paradigm

The state of the art tools of the 'Web as Corpus' framework rely heavily on URLs obtained from search engines. As a matter of fact, the approach followed by the most researchers of this field consists of querying search engines (e.g. by tuples) to gather links that are crawled in order to build a corpus (Baroni et al., 2009).

Until recently, this method could be used in free corpus building approach. However, increasing limitations on the search engine APIs, which make the gathering process on a low budget very slow or impossible, have rendered its use impossible. All in all, the APIs may be too expensive and/or too unstable in time to support large-scale corpus building projects.

Moreover, the question whether the method used so far, i.e. randomizing keywords, provides a good overview of a language is still open. Other technical difficulties include diverse and partly unknown search biases, partially due to search engine optimization tricks as well as undocumented PageRank adjustments. Using diverse sources of seed URLs could at least ensure that there is not a single bias, but several ones.

The crawling method using these seeds for corpus building may then yield better results, e.g. ensure better randomness in a population of web documents as described by Henzinger et al. (2000).

1.2 User-based URL gathering

My hypothesis states that microblogging services are a good alternative to overcome the limitations of seed URL collections and the biases implied by search engine optimization techniques, PageRank and link classification.

My URL gathering uses a user-based language approach. Its obvious limits are the amount of spam and advertisement. The main cause for bias is the technology-prone users who are familiar with these platforms and account for numerous short messages which in turn over-represent their own interests and hobbies.

However, user-related biases also have advantages, most notably the fact that documents that are most likely to be important are being shared, which has benefits when it comes to gathering links in lesser-known languages, below the English-speaking spammer's radar.

1.3 Interest

The main goal is to provide well-documented, feature-rich software and databases relevant for linguistic studies. More specifically, I would like to be able to cover languages which are more rarely seen on the Internet, which implies the gath-

ering of higher proportions of URLs leading to lesser-known languages. We think that social networks and microblogging services may be of great help when it comes to focusing on them.

In fact, it can be argued that the most engaged social networking nations do not use English as a first communicating language.[1] In addition, crawling these services gives an opportunity to perform a case study of existing tools and platforms.

Finally, the method presented here could be used in other contexts : microtext collections, user lists and relations could prove useful for microtext corpus building, network visualization or social network sampling purposes (Gjoka et al., 2011).

2 Data Sources

FriendFeed, identi.ca and Reddit are taken into consideration for this study. These services provide a good overview of the peculiarities of social networks. A crawl appears to be manageable by at least the last two of them, in terms of both API accessibility and corpus size, which is not the case for Twitter, for example.

2.1 identi.ca

identi.ca is a social microblogging service built on open source tools and open standards, which is the reason why I chose to crawl it at first.

The advantages compared to Twitter include the Creative Commons license of the content, the absence of limitations on the total number of pages seen (to my knowledge), and the relatively small amount of messages, which can also be a problem. A full coverage of the network where all the information may be publicly available is theoretically possible. Thus, all interesting information is collected and no language filtering is used for this website.

2.2 FriendFeed

To my knowledge, FriendFeed is the most active of the three microblogging services considered here. It is also the one which seems to have been studied the most by the research community. The service works as an aggregator (Gupta et al., 2009) that offers a broader spectrum of retrieved information. Technically, FriendFeed and identi.ca can overlap, as the latter is integrated in the former. However, the size difference between the two platforms makes this hypothesis unlikely.

The API of FriendFeed is somewhat liberal, as no explicit limits are enforced. Nonetheless, my tests showed that after a certain number of successful requests with little or no sleep, the servers start dropping most of the inbound connections. All in all, the relative tolerance of this website makes it a good candidate to gather a lot of text in a short period of time.

2.3 Reddit

Reddit is a social bookmarking and a microblogging platform, which ranks to the 7th place worldwide in the news category according to Alexa.[2] The entries are organized into areas of interest called 'reddits' or 'subreddits'. The users account for the linguistic relevance of their channel, the moderation processes are mature, and since the channels (or subreddits) have to be hand-picked, they ensure a certain stability.

There are 16 target languages so far, which can be accessed via so-called 'multi-reddit expressions', i.e. compilations of subreddits: Croatian, Czech, Danish, Finnish, French, German, Hindi, Italian, Norwegian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, and Turkish.[3]

Sadly, it is currently not possible to go back in time further than the 500th oldest post due to API limitations, which severely restricts the number of links one may crawl.

3 Methodology

The following workflow describes how the results below are obtained:

1. URL harvesting: social network traversal, obvious spam and non-text documents filtering, optional spell check of the short message to see if it could be English text, optional record of user IDs for later crawls.

2. Operations on the URL queue: redirection checks, sampling by domain name.

3. Download of the web documents and analysis: HTML code stripping, document validity check, language identification.

[1] http://www.comscore.com/Press_Events/Press_Releases/2011/12/Social_Networking_Leads_as_Top_Online_Activity_Globally

[2] http://www.alexa.com/topsites/category/Top/News

[3] Here is a possible expression to target Norwegian users: http://www.reddit.com/r/norge+oslo+norskenyheter

The only difference between FriendFeed and Reddit on one hand and identi.ca on the other hand is the spell check performed on the short messages in order to target the non-English ones. Indeed, all new messages on the latter can be taken into consideration, making a selection unnecessary.

Links pointing to media documents, which represent a high volume of links shared on microblogging services, are excluded from this study, as its final purpose is to be able to build a text corpus. As a page is downloaded or a query is executed, links are filtered on the fly using a series of heuristics described below, and finally the rest of the links is stored.

3.1 TRUC: an algorithm for TRaversal and User-based Crawls

Starting from a publicly available homepage, the crawl engine selects users according to their linguistic relevance based on a language filter (see below), and then retrieves their messages, eventually discovering friends of friends and expanding its scope and the size of the network it traverses. As this is a breadth-first approach, its applicability depends greatly on the size of the network.

In this study, the goal is to concentrate on non-English speaking messages in the hope of finding non-English links. The main 'timeline' fosters a users discovery approach, which then becomes user-centered as the spider focuses on a list of users who are expected not to post messages in English and/or spam. The messages are filtered at each step to ensure relevant URLs are collected. This implies that a lot of subtrees are pruned, so that the chances of completing the traversal increase. In fact, experience shows that a relatively small fraction of users and URLs is selected.

This approach is 'static', as it does not rely on any long poll requests (which are, for instance, used to capture a fraction of Twitter's messages as they are made public); it actively fetches the required pages.

3.2 Check for redirection and sampling

Further work on the URL queue before the language identification task ensures an even smaller fraction of URLs really goes through the resource-expensive process of fetching and analyzing web documents.

The first step of preprocessing consists of finding those URLs that lead to a redirect, which is done using a list comprising all the major URL shortening services and adding all intriguingly short URLs, i.e. less than 26 characters in length, which according to my FriendFeed data occurs at a frequency of about 3%. To deal with shortened URLs, one can perform HTTP HEAD requests for each member of the list in order to determine and store the final URL.

The second step is sampling that reduces both the size of the list and the probable impact of an overrepresented domain names in the result set. If several URLs contain the same domain name, the group is reduced to a randomly chosen URL.

Due to the overlaps of domain names and the amount of spam and advertisement on social networks such an approach is very useful when it comes to analyzing a large list of URLs.

3.3 Language identification

Microtext has characteristics that make it hard for 'classical' NLP approaches like web page language identification based on URLs (Baykan et al., 2008) to predict with certainty the languages of the links. That is why mature NLP tools have to be used to filter the incoming messages.

A similar work on language identification and FriendFeed is described by Celli (2009), who uses a dictionary-based approach: the software tries to guess the language of microtext by identifying very frequent words.

However, the fast-paced evolution of the vocabulary used on social networks makes it hard to rely only on lists of frequent terms, therefore my approach seems more complete.

A first dictionary-based filter First, a quick test is used in order to guess whether a microtext is English or not. Indeed, this operation cuts the amount of microtexts in half and enables to select the users or the friends which feature the desired response, thus directing the traversal in a more fruitful direction.

The library used, enchant, allows the use of a variety of spell-checking backends, like aspell, hunspell or ispell, with one or several locales.[4] Basically, this approach can be used with other languages as well, even if they are not used as discriminating factors in this study. We consider this option to be a well-balanced solution between processing speed on one hand and coverage on

[4] http://www.abisource.com/projects/enchant/
All software mentioned here is open-source.

the other. Spell checking algorithms benefit from years of optimization in both areas.

This first filter uses a threshold to discriminate between short messages, expressed as a percentage of tokens which do not pass the spell check. The filter also relies on software biases, like Unicode errors, which make it nearly certain that the given input microtext is not English.

`langid.py` A language identification tool is used to classify the web documents and to benchmark the efficiency of the test mentioned above. `langid.py` (Lui and Baldwin, 2011; Lui and Baldwin, 2012) is open-source, incorporates a pretrained model, and covers 97 languages, which is ideal to tackle the diversity of the web. Its use as a web service makes it a fast solution, enabling distant or distributed work.

The server version of `langid.py` was used, the texts were downloaded, all the HTML markup was stripped, and the resulting text was discarded if it was less than 1,000 characters long. According to its authors, `langid.py` could be used directly on microtexts. However, this feature was discarded because it did not prove as efficient as the approach used here when it came to a substantial amounts of short messages.

4 Results

The surface crawl dealing with the main timeline and one level of depth has been performed on the three platforms.[5] In the case of identi.ca, a deep miner was launched to explore the network. FriendFeed proved too large to start such a breadth-first crawler so that other strategies ought to be used (Gjoka et al., 2011), whereas the multireddit expressions used did not yield enough users.

FriendFeed is the biggest link provider on a regular basis (about 10,000 or 15,000 messages per hour can easily be collected), whereas Reddit is the weakest, as the total figures show.

The total number of English websites may be a relevant indication when it comes to establishing a baseline for finding possibly non-English documents. Accordingly, English accounts for about 55% of the websites, with the second most-used content-language, German, only representing about 6% of the web pages.[6] So, there is a gap between English and the other languages, and there is also a discrepancy between the number of Internet users and the content languages.

4.1 FriendFeed

To test whether the first language filter was efficient, a testing sample of URLs and users was collected randomly. The first filter was emulated by selecting about 8% of messages (based on a random function) in the spam and media-filtered posts of the public timeline. Indeed, the messages selected by the algorithm approximately amount to this fraction of the total. At the same time, the corresponding users were retrieved, exactly as described above, and then the user-based step was run. One half of the user's messages was kept, which, according to the real-world data, is realistic.

The datasets compared here were both of an order of magnitude of at least 10^5 unique URLs before the redirection checks. At the end of the toolchain, the randomly selected benchmark set comprised 7,047 URLs and the regular set 19,573 URLs.[7] The first was collected in about 30 hours and the second one in several weeks. According to the methodology used, this phenomenon may be explained by the fact that the domain names in the URLs tend to be mentioned repeatedly.

Language	URLs	%
English	4,978	70.6
German	491	7.0
Japanese	297	4.2
Spanish	258	3.7
French	247	3.5

Table 1: 5 most frequent languages of URLs taken at random on FriendFeed

According to the language identification system (`langid.py`), the first language filter beats the random function by nearly 30 points (see Table 2). The other top languages are accordingly better represented. Other noteworthy languages are to be found in the top 20, e.g. Indonesian and Persian (Farsi).

[5] Several techniques are used to keep the number of requests as low as possible, most notably user profiling according to the tweeting frequency. In the case of identi.ca this results into approximately 300 page views every hour.

[6] http://w3techs.com/technologies/overview/content_language/all

[7] The figures given describe the situation at the end, after the sampling by domain name and after the selection of documents based on a minimum length. The word URL is used as a shortcut for the web documents they are linked to.

Language	URLs	%
English	8,031	41.0
Russian	2,475	12.6
Japanese	1,757	9.0
Turkish	1,415	7.2
German	1,289	6.6
Spanish	954	4.9
French	703	3.6
Italian	658	3.4
Portuguese	357	1.8
Arabic	263	1.3

Table 2: 10 most frequent languages of spell-check-filtered URLs gathered on FriendFeed

4.2 identi.ca

The results of the two strategies followed on identi.ca led to a total of 1,113,783 URLs checked for redirection, which were collected in about a week (the deep crawler reached 37,485 user IDs). A large majority of the 192,327 total URLs apparently lead to English texts (64.9%), since only a spam filter was used.

Language	URLs	%
English	124,740	64.9
German	15,484	8.1
Spanish	15,295	8.0
French	12,550	6.5
Portuguese	5,485	2.9
Italian	3,384	1.8
Japanese	1,758	0.9
Dutch	1,610	0.8
Indonesian	1,229	0.6
Polish	1,151	0.6

Table 3: 10 most frequent languages of URLs gathered on identi.ca

4.3 Reddit

The figures presented here are the results of a single crawl of all available languages altogether, but regular crawls are needed to compensate for the 500 posts limit. English accounted for 18.1% of the links found on channel pages (for a total of 4,769 URLs) and 55.9% of the sum of the links found on channel and on user pages (for a total of 20,173 URLs).

The results in Table 5 show that the first filter was nearly sufficient to discriminate between the links. Indeed, the microtexts that were under the threshold led to a total of 204,170 URLs. 28,605 URLs remained at the end of the toolchain and English accounted for 76.7% of the documents they were linked to.

Language	URLs	%	Comb. %
English	863	18.1	55.9
Spanish	798	16.7	9.7
German	519	10.9	6.3
French	512	10.7	7.2
Swedish	306	6.4	2.9
Romanian	265	5.6	2.5
Portuguese	225	4.7	2.1
Finnish	213	4.5	1.6
Czech	199	4.2	1.4
Norwegian	194	4.1	2.1

Table 4: 10 most frequent languages of filtered URLs gathered on Reddit channels and on a combination of channels and user pages

Language	URLs	% of total
English	21,926	76.7
Spanish	1,402	4.9
French	1,141	4.0
German	997	3.5
Swedish	445	1.6

Table 5: 5 most frequent languages of links seen on Reddit and rejected by the primary language filter

The threshold was set at 90% of the words for FriendFeed and 33% for Reddit, each time after a special punctuation strip to avoid the influence of special uses of punctuation on social networks. Yet, the lower filter achieved better results, which may be explained by the moderation system of the subreddits as well as by the greater regularity in the posts of this platform.

5 Discussion

Three main technical challenges had to be addressed, which resulted in a separate workflow: the shortened URLs are numerous, yet they ought to be resolved in order to enable the use of heuristics based on the nature of the URLs or a proper sampling of the URLs themselves. The confrontation with the constantly increasing number of URLs to analyze and the necessarily limited re-

sources make website sampling by domain name useful. Finally, the diversity of the web documents put the language recognition tools to a test, so that a few tweaks are necessary to correct the results.

The relatively low number of results for Russian may be explained by weaknesses of `langid.py` with deviations of encoding standards. Indeed, a few tweaks are necessary to correct the biases of the software in its pre-trained version, in particular regarding texts falsely considered as being written in Chinese, although URL-based heuristics indicate that the website is most probably hosted in Russia or in Japan. A few charset encodings found in Asian countries are also a source of classification problems. The low-confidence responses as well as a few well-delimited cases were discarded in this study as they account for no more than 2% of the results. Ideally, a full-fledged comparison with other language identification software may be necessary to identify its areas of expertise.

A common practice known as cloaking has not been addressed so far. In fact, a substantial fraction of web pages show a different content to crawler engines and to browsers. This Janus-faced behavior tends to alter the language characteristics of the web page in favor of English results.

Regarding topics, a major user bias was not addressed either. Among the most frequently shared links on identi.ca, for example, many are related to technology, IT, or software, and are mostly written in English. The social media analyzed here tend to be dominated by English-speaking users, either native speakers or second-language learners.

In general, there is room for improvement concerning the first filter. The threshold could be tested and adapted to several scenarios. This may involve larger datasets for testing purposes and machine learning techniques relying on feature extraction.

The contrasted results on Reddit shed a different light on the exploration of user pages: in all likelihood, users mainly share links in English when they are not posting them on a language-relevant channel. The results on FriendFeed are better from this point of view, which may suggest that English is not used equally on all platforms by users who speak other languages than English. Nonetheless, there seems to be a strong tendency for the microblogging services discussed here to be mainly English-speaking.

Last but not least, the adequateness of the web documents shared on social networks has yet to be thoroughly assessed. From the output of this toolchain to a full-fledged web corpus, other fine-grained instruments (Schäfer and Bildhauer, 2012) as well as further decisions (Schäfer et al., 2013) are needed along the way.

6 Conclusion

I presented a methodology to gather multilingual URLs on three microblogging platforms. In order to do so, I performed traversals of the platforms and used already available tools to filter the URLs accordingly and identify their language.

I provide open source software to access the APIs (FriendFeed and Reddit) and the HTML version of identi.ca, as an authentication is mandatory for the API. The TRUC algorithm is fully implemented. All the operations described in this paper can be reproduced using the same tools, which are part of repositories currently hosted on GitHub.[8]

The main goal is achieved, as hundreds, if not thousands, of URLs for lesser-known languages such as Romanian or Indonesian can be gathered on social networks and microblogging services. When it comes to filtering out English posts, a first step using an English spell checker gives better results than the baseline established using microtexts selected at random. However, the discrepancy is remarkable between the languages one would expect to find based on demographic indicators on one hand, and based the results of the study on the other hand. English websites stay numerous even when one tries to filter them out.

However, the discrepancy between the languages that one would expect to find (based on demographic indicators and the results of the study) is remarkable.

This proof of concept is usable, but a better filtering process and longer crawls may be necessary to unlock the full potential of this approach. Lastly, a random-walk crawl using these seeds and a state of the art text categorization may provide more information on what is really shared on microblogging platforms.

Future work perspectives include dealing with live tweets (as Twitter and FriendFeed can be queried continuously), exploring the depths of identi.ca and FriendFeed, and making the directory of language-classified URLs collected during this study publicly available.

[8] https://github.com/adbar/microblog-explorer

7 Acknowledgments

This work has been partially funded by an internal grant of the FU Berlin (COW project at the German Grammar Dept.). Many thanks to Roland Schäfer and two anonymous reviewers for their useful comments.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Eda Baykan, Monika Henzinger, and Ingmar Weber. 2008. Web Page Language Identification Based on URLs. *Proceedings of the VLDB Endowment*, 1(1):176–187.

Fabio Celli. 2009. Improving Language identification performance with FriendFeed data. Technical report, CLIC, University of Trento.

Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. 2011. Practical recommendations on crawling online social networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1872–1892.

Trinabh Gupta, Sanchit Garg, Niklas Carlsson, Anirban Mahanti, and Martin Arlitt. 2009. Characterization of FriendFeed – A Web-based Social Aggregation Service. In *Proceedings of the AAAI ICWSM*, volume 9.

Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. 2000. On near-uniform URL sampling. In *Proceedings of the 9th International World Wide Web conference on Computer Networks: The International Journal of Computer and Telecommunications Networking*, pages 295–308. North-Holland Publishing Co.

Marco Lui and Timothy Baldwin. 2011. Cross-domain Feature Selection for Language Identification. In *Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 553–561.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493.

Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. 2013. The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In Stefan Evert, Egon Stemle, and Paul Rayson, editors, *Proceedings of the 8th Web as Corpus Workshop*, pages 7–15.