Proceedings of the 13th Workshop on Building and Using Comparable Corpora, pages 14–21
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Constructing a Bilingual Corpus of Parallel Tweets

Hamdy Mubarak, Sabit Hassan, Ahmed Abdelali
Qatar Computing Research Institute
Doha, Qatar
{hmubarak, sahassan2, aabdelali}@hbku.edu.qa

Abstract

In a bid to reach a larger and more diverse audience, Twitter users often post *parallel tweets* —tweets that contain the same content but are written in different languages. Parallel tweets can be an important resource for developing machine translation (MT) systems among other natural language processing (NLP) tasks. In this paper, we introduce a generic method to collect parallel tweets. Using this method, we collect a bilingual corpus of Arabic-English parallel tweets and a list of Twitter accounts who post Arabic-English tweets regularly. Since our method is generic, it can also be used for collecting parallel tweets that cover less-resourced languages such as Urdu or Serbian. Additionally, we annotate a subset of Twitter accounts with their countries of origin and topic of interest, which provides insights about the population who post parallel tweets. This latter information can also be useful for author profiling tasks.

Keywords: Corpus Creation, Machine Translation, Arabic-English, Parallel Tweets, Comparable Corpora

1. Introduction

Extensive usage of social media in recent years has flooded the web with a massive amount of user-generated content. This has the potential to be a very valuable resource for Natural Language Processing (NLP) tasks such as Machine Translation (MT). However, in social media platforms such as Twitter, users typically write content in a very informal way. The users extensively use emoticons, short forms of phrases such as "idk (I don't know)" and follow traits that are far from traits of traditionally written content that follow language rules and grammar closely. Because of the unpredictable and inconsistent nature of content in social media, it is quite difficult to exploit this type of data. In recent years, this issue has gained significant interest among researchers and motivated many of them to work on harvesting useful data from this ever-growing pool of user-generated content. To facilitate this process, we identify and focus on an interesting trait among Twitter users: some Twitter users post tweets with the same message written in different languages —that we will call *parallel tweets*.

Organizations, celebrities and public figures on social media platforms, such as Twitter, try to reach out to as large of an audience as possible. Often the audience consists of individuals who use different languages. To build a connection with this diverse audience, organizations, celebrities, and public figures post tweets in multiple languages to ensure max reach out. Twitter, with traditionally 140 (Now, 280) character limit on the tweets, prompts the users to reach out to their audiences across multiple tweets containing the same message in different languages. In our paper, we propose a method to collect such tweets. These parallel tweets can be a great resource for machine translation. Ling et al., (2013) show that parallel texts from Twitter can significantly improve MT systems. As opposed to crowdsourcing translations that cost money or complex mechanisms of cross-language information retrieval, we provide a free and generic method of obtaining a large amount of translations that cover highly sought after new vocabulary and terminology. For example, in Table 1, we can see that, خدمة إلكترونية is translated to "e-Service" by the user. Google Translate on the other hand, would translate it as "electronic service".

In our proposed method, we first crawl Twitter to collect a large number of tweets and find unique Twitter accounts from these tweets. Then, we filter the accounts to only include those who are likely to post parallel tweets —accounts with high popularity. Then, for each account, we identify candidates for parallel tweets and lastly, we filter the candidate parallel tweets to only include tweets that have a high possibility of being parallel. For filtering candidate parallel tweets, we use a simple dictionary based method along with some heuristics. We also eliminate parallel tweets with repetitive content as we want our collection to capture the diversity of user-generated content on social media without redundancy in the collection.

In this paper, we focus on collecting pairs of Arabic-English parallel tweets using the proposed method. We release 166K pairs of Arabic-English parallel tweets. We also report 1389 accounts that post such parallel tweets regularly. This collection of accounts is valuable as we expect these accounts to continue posting parallel tweets in the future. To demonstrate this effect, we collect parallel tweets from the same users in two different time frames, separated by 16 months, and observe a remarkable growth in the number of parallel tweets collected. This suggests that our resource will grow significantly in the future. We publicly share the parallel tweets by their IDs as well as the usernames of Twitter accounts who post parallel tweets regularly.

A phenomenon similar to parallel tweets is *comparable tweets*. When a pair of tweets have significant overlap in content and theme but are not exact translations of each other, we call them *comparable tweets*. Since our method is automatic, it is prone to some errors. In our error analysis (section 4), we notice that although some pairs of tweets that were tagged as parallel by our system may not be exact translations of each other, they are actually comparable tweets. Since these pairs of tweets have significant overlap, they can also be useful for many tasks in cross-language information retrieval.

In addition to collecting parallel tweets and Twitter accounts, we also annotate a subset of Twitter accounts for their countries and topics the accounts typically post about. This allows us to understand the demographics of Twitter users who post parallel tweets. This information will be useful in future collections of parallel tweets as we will know in which countries posting parallel tweets is a popular trend and which topics are likely to have many parallel tweets. Moreover, this information can be useful for tasks such as author profiling.

Although in our paper, we present a bilingual corpus of Arabic-English parallel tweets, our generic method can also be adapted for other language pairs and has the potential to be particularly useful for less-resourced languages such as Urdu or Serbian.

In section 2, we survey related work from relevant literature, and in section 3, we present our method and data collected using this method. In section 4, we provide some preliminary assessments for the data quality, and in section 5, we discuss the annotation of accounts for their countries of origin and topics of tweets. Lastly, we conclude with a summary and future work.

2. Related Work

Although the amount of data on social media is growing at an incredible speed and can be a valuable resource for NLP tasks, the utilization of data on social media has been underwhelming. Efforts to use these platforms as a resource for translation are still relatively small.

Sluyter Gäthje et al. (2018) built a parallel resource for English-German using 4000 English tweets that were manually translated into German with a special focus on the informal nature of the tweets. The objective was to provide a resource tailored for translating user generated-content.

Jehl et al. (2012) and Abidi and Smaili (2017) extract parallel phrases by using CLIR techniques. The major difference is that these methods are extracting comparable data, whereas, we want to extract parallel tweets, which we can expect to be closer to true translation. Jehl et al. use a probabilistic translation-based retrieval (Xu et al., 2001) in the context of Twitter for the purpose of training Statistical Machine Translation (SMT) pipeline. For evaluation purposes, Jehl et al. (2012) use crowdsourcing to create a parallel corpus of 1000 Arabic tweets and 3 manual English translations for each Arabic tweet and reports improvement for SMT pipeline. Abidi and Smaili (2017) used topics related to Syria to crawl Twitter and collect 58,000 Arabic tweets and 60,000 English tweets. The tweets are then preprocessed heavily, which requires knowledge of Arabic. Then, the tweets are aligned to produce a corpus of comparable Arabic-English tweets aimed at improving MT systems.

Vicente et al. (2016) present a parallel corpus that covers 5 languages from the Iberian Peninsula, created by automatic collection and crowdsourcing. To align parallel content, Vicente et al. (2016) use measures such as publication date, string length similarity, hashtag and user mention overlap, and Longest Common Subsequence ratio (LCSR). LCSR exploits the similarity of the languages within the Iberian peninsula. The aim of the corpus is to aid in the development of microtext translation systems. Vicente et al. (2016) used the corpus in a shared task to evaluate it.

In comparison to the above methods, our method is more generic, which does not require specific knowledge of the language and can be used for different language pairs. Our method is also relatively simple that uses minimal external resources. The generic and simple nature of our method makes it easily adaptable for less-resourced languages.

Ling et al. (2013) collect parallel content of different languages from single tweets (compare Table 1 and Table 2 for difference). They reported a significant improvement in MT systems. In this work, we will not focus on extracting parallel content from single tweets. However, our methods can be adapted to do so in the future.

Our work also augments existing work in Twitter account annotation. Specifically for Arabic Twitter users, there is a scarcity of resources. Inspired by Mubarak and Darwish (2014), who annotate tweets for their dialects, Bouamor et al. (2019) presented a dataset of 3000 Twitter accounts annotated with their countries of origin. Alhozaimi and Almishari (2018) categorize 80 Twitter accounts into 4 categories of topics the accounts are interested in. It suffices to say that there is a need for such resources and our annotation of Twitter accounts for country and topic, although not our primary goal, is a step forward.

3. Methodology and Corpus Construction

Before diving further into the methodology, it's important to have a good understanding of the phenomenon of parallel tweets. In this section, we will provide details of the phenomenon on Twitter and the various options used by the platform users, followed by our methodology and details of collected corpus.

3.1. Parallel Tweets

If a pair of tweets are translations of each other, we call them parallel tweets. It's important to distinguish between parallel tweets and tweets that contain parallel data. Table 1 and Table 2 contain examples of parallel tweets and tweets containing parallel content respectively. Our focus is on the scenario of Table 1. We can identify several characteristics of parallel tweets that are important for developing the methodology. We observe that the tweets are usually consecutive or within a short period of time. The presence of certain words in both tweets can indicate that they are parallel tweets. It suffices to check if there is a significant overlap between the two tweets.

3.2. Methodology

Our methodology follows a three-step procedure. First, we collect candidate parallel tweets from Twitter users who are likely to post parallel tweets. In the second step, we filter candidate parallel tweets to obtain our collection of parallel tweets. In order to improve the quality of the corpus, in the third step, we remove duplicate tweets and exclude accounts who post repetitive tweets.

3.2.1. Collecting Candidate Parallel Tweets

Step 1: search Twitter for a large number of tweets using commonly appearing words in the targeted language pair, alternatively, we can use language filter if available; e.g

Account	Country	Language	Tweet
HukoomiQatar	Qatar	English	e-Service | The Ministry of Economy and Commerce provides a number of services to the Qatari nationals
		Arabic	خدمة إلكترونية | تقدم وزارة الاقتصاد والتجارة مجموعة من الخدمات للمواطنين القطريين
ArifAlvi	Pakistan	English	I pray for the quick recovery of Mr Nawaz Sharif. May Allah restore him to full health. I am sure the government will ensure all medical facilities.
		Urdu	مں نواز شریف صاحب کی جلد صحت ىابى کىلئي الله کی بارگاہ مں دعا گو ہوں اور امد کرتا ہوں کہ حکومت تمام طبی سہولات کی فراہمی ىقىنى بنائي گی
SerbianPM	Serbia	English	Sam Parker, Congratulations to @vonderleyen and the new Commission team. We look forward to working with you over the next five years as we prepare Serbia for EU Membership.
		Serbian	Честитке @vonderleyen и новом тиму Европске комисије. Радујемо се што ћемо сарађивати са вама у наредних пет година док припремамо Србију за чланство у ЕУ.

Table 1: Examples of parallel tweets

Account	Country	Language	Tweet
SerbianPM	Serbia	Serbian English	Поносна сам на представљање најбољих српских производа у економском Павиљону на другом кинеском међународном сајму увоза ЕКСПО у Шангају #CIIE #Србија Proud to see the best of #Serbia on display at the Economic Pavilion of the China International Import Expo in Shanghai #CIIE
KuwaitAirways	Kuwait	Arabic English	احجز مع العطلات إلى المدينه المنورة على عروض درجه رجال الأعمال،، للمزيد من المعلومات اتصل على 1806060 Book your trip to Madinah with our Business Class offers For more information call 1806060

Table 2: Examples of tweets with parallel content (inside same tweet)

"lang:ar" in case of Arabic. **Step 2:** Collect all the unique accounts from these tweets. **Step 3:** At this point, it's important to understand who is likely to post parallel tweets. Our assumption is that most likely the Twitter user will have a large number of followers. In this step, we shortlist the accounts based on number of followers. **Step 4:** We collect all available tweets from the shortlisted accounts but exclude tweets that are too short as they would compromise the richness of the corpus. **Step 5:** For each tweet, we check language of the tweet along with language of previous and next tweet as we expect the user to post parallel tweets within a short period of time. If the languages form our target pair of languages, we consider the corresponding tweets to be candidate parallel tweets.

3.2.2. Filtering Candidate Parallel Tweets

Once we have the candidate tweets, we need to identify which ones are indeed parallel tweets. In our language pair, let us call the first language L1, and second language L2. We assume availability of a dictionary that maps words from L1 to L2. In our candidate pair of parallel tweets, let us call the tweet from L1 to be T1 and the tweet from L2 to be T2.

Step 1: We remove stopwords from both tweets[1]. **Step 2:** We remove commonly known suffixes and prefixes from words of T1 and T2 and assume the remaining parts are stems.[2] Such surface-level (and light) stemming yields reasonably good result while being easily applicable to less-resourced languages. We anticipate that using complex stemmer/lemmatizer or a high-coverage lookup table when available would yield better accuracy of the collected tweets, but we opted to examine the accuracy of our approach in low-resourced scenario where these resources are typically unavailable. **Step 3:** We look up stems of T1 in the dictionary and check if the stem appears in T2 after mapping from L1 to L2. If it does, we count it as a "match". **Step 4:** If the number of matches exceeds a threshold, we tag the pair as parallel tweets.

The matching threshold in step 4 can be changed to obtain corpus of different quality. Higher threshold will result in higher quality corpus, but lower number of parallel tweets. To decide this threshold, we take a subset of the data and annotate it manually, identifying if they are indeed parallel. Then, we plot number of parallel tweets retained for

[1] https://sites.google.com/site/kevinbouge/stopwords-lists

[2] Example: in our English surface stemming, we just removed 's', 'ed' and 'ing' from the end of words.

Correctness	English tweet	Arabic tweet
Correct	GOAL! Scored by Chang Jin Moon (Shabab Al Ahli Dubai) 35 min. Shabab Al Ahli Dubai 1 Emirates 0 #SAHvEMR	هدف! سجله شانج جن مون (شباب الأهلي دبي) دقيقة 35.شباب الأهلي دبي 1 الإمارات 0 #SAHvEMR
Wrong	GOAL! Scored by Chang Jin Moon (Shabab Al Ahli Dubai) 35 min. Shabab Al Ahli Dubai 1 Emirates 0 #SAHvEMR	نهاية الشوط الأول: شباب الأهلي دبي 1 الإمارات 0 #SAHvEMR (Translation: The end of the first half: Shabab Al Ahli Dubai 1 Emirates 0 #SAHvEMR)

Table 3: Example of duplicate tweets

Account	English tweet	Arabic tweet
QatarPrayer	It's now **Fajer** athan time **4:05am** according to Doha city local time and its suburbs. #Qatar	حان الآن موعد أذان الفجر 4:05ص حسب التوقيت المحلي لمدينة الدوحة وضواحيها. #قطر
	It's now **Asr** athan time **3:06pm** according to Doha city local time and its suburbs. #Qatar	حان الآن موعد أذان العصر 3:06م حسب التوقيت المحلي لمدينة الدوحة وضواحيها. #قطر

Table 4: Example of account posting repetitive tweets. Differences between English tweets (templates) are written in bold.

different thresholds and the corresponding errors.

3.2.3. Improving Quality of Corpus

At this point, we noticed that, since each tweet is compared with its preceding and succeeding tweet, it's possible that the tweet has matching words exceeding the threshold for both the previous and next tweet. Table 3 illustrates this issue[3]. This is an uncommon occurrence but to address this issue, we pick the pair that has a higher number of matches. We also noticed that some accounts posted repetitive tweets that are extremely similar to each other. These accounts mostly follow a template for posting tweets and are likely to be bots. Table 4 shows an example of such accounts. These accounts are not very useful for the purpose of creating a corpus for machine translation. To identify these accounts, we plot number of words in all the tweets posted by the account against the number of unique words among them. If the ratio of unique words versus total words is below a threshold, we exclude the account.

To increase the quality of the collected Arabic-English tweets, we can use complex Arabic word segmenter to split prefixes and suffixes, for example Farasa word segmenter (Darwish and Mubarak, 2016; Abdelali et al., 2016), or lemmatizer (Mubarak, 2018), and for English we can use Porter stemmer (Porter, 1980). We leave this for future work.

3.3. Arabic-English Parallel Tweets Corpus

Using the method described in Section 3.2., we collect a corpus of 166K Arabic-English parallel tweets and 1,389 accounts who regularly post them. For our collection of Arabic-English parallel tweets, first, we collect 175M Arabic tweets in March 2014 using Twitter API with language filter assigned to Arabic; "lang:ar". From these tweets, we identify 15,000 unique accounts who have more than 5,000 followers and collect available tweets from these accounts. Since very short tweets (less than or equal to 5 words) are not that useful for many NLP tasks such as MT, we exclude them from our collection. Once we have a large number of tweets, we carry out the procedure in Section 3.2. in two stages, separated by 16 months. During the first stage, we collect 120K parallel tweets from these accounts in July 2018. We expect these accounts to continue to post parallel tweets. Therefore, in November 2019, we collect parallel tweets from the same accounts again. During this stage, we collect more than 83K additional pairs of tweets. At this point, we have 203K parallel tweets. We can see that our collection grew significantly in the span of 16 months. Therefore, we can expect the collection to grow further in the future. To illustrate possible growth in the future, Table 5 shows the top 5 accounts (according to the number of parallel tweets collected) and their posting rate of parallel tweets. To reduce the margin of error, we removed duplicates from the collection as described in Section 3.2. During the whole procedure, we use Buckwalter Lexicon (Buckwalter, 2004) as a dictionary to calculate degree of matching between two tweets. If the degree of matching exceeds threshold of *3*, we consider the tweets to be parallel. The matching threshold of 3 is found experimentally and justified in section 4.

Then, we calculate ratio of unique words and total number of words in tweets posted by each account. If this ratio falls below the threshold of 0.1, we exclude the account and all the tweets posted by the account. This threshold is also decided on experimentally, which is described in section 4. Finally, we end up with 166K tweets posted by 1,389 accounts.

[3] In all tables, in case of wrong English translation, the correct translation is given inside parentheses.

4. Quality of Corpus

In order to determine the quality of our collected corpus and identify the thresholds described in section 3, we select a subset of candidate parallel tweets and annotate them manually. To select this subset of tweets, we notice that, after removal of short tweets, the average number of words in tweets is 23. We randomly select 1,000 pairs of tweets who match on at least 10% of the mean number of words (rounded up, 10% of 23 is 3). We categorize these 1,000 tweets as "parallel" (translations of each other), "comparable" (they have significant overlap in content) or "unrelated" (errors) manually. Table 6 shows examples of the different categories.

Figure 1 depicts experimentation on degree of matching used as threshold to decide whether a pair is indeed parallel. In Figure 1, we group tweets that are parallel and comparable together and consider unrelated tweets as errors. We can see that at threshold of 3, we achieve less than 10% error rate. Going from threshold of 3 to 4, we lose 22.3% (from 1,000 to 777) of the tweets while reducing the error by only 2% (from 95 out of 1,000, which is 9.5%, to 58 out of 777, which is 7.5%). We can see the trend that when the threshold is increased, we lose a significant portion of tweets, while reducing error by only a small fraction. Since with threshold of 3, we retain large number of tweets while having less than 10% error rate, we decide that 3 is an appropriate threshold for our corpus.

Account	Number of Parallel tweets	Rate of Posting (Per Day)
HukoomiQatar	2,615	3.18
culturebah	2,311	1.69
AshghalQatar	2,202	2.16
DMunicipality	1,974	2.23
QF	1,944	2.11

Table 5: Accounts with highest posting rate of parallel tweets

To identify accounts who post repetitive tweets, we calculate the ratio of unique words and total words posted by accounts. If the ratio falls below a threshold, we consider the account to post repetitive content. In order to find an appropriate threshold, we plot the ratio of number of unique words and total words for each account against number of tweets posted by that account. We can see from Figure 2 that there are few accounts who have a high number of tweets and fall below the ratio of 0.1. KuwaitMet is one such account (posted ∼7,000 tweets, with ratio less than 0.01). KuwaitMet is the official account of Kuwait Meteorological Department. They post many tweets every day using a template-like format that differ only in certain values such as wind speed or rain amount, while rest of tweet content is the same. Parallel tweets from such accounts are not desirable as they do not contribute to the richness of corpus and therefore, we exclude them from our corpus.

To understand the coverage of our corpus, we count the total number of words (Tokens) and number of unique words (Types) in the set of English and Arabic tweets separately.

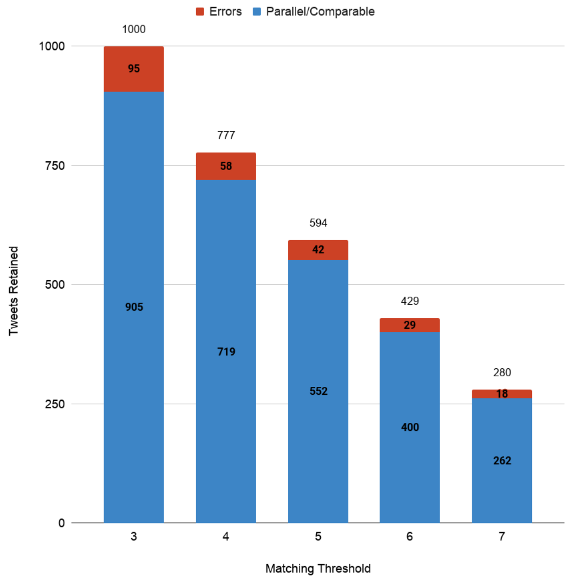

Figure 1: Error comparison of matching threshold

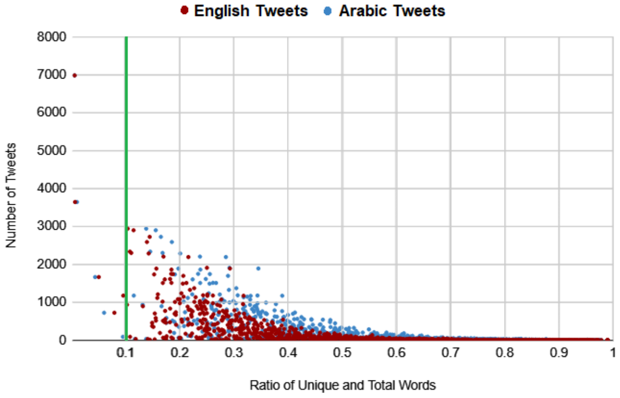

Figure 2: Number of tweets vs. ratio of unique words. Threshold (in Green) for discarded accounts and their respective volume of words.

Table 8 shows this information. The large number of unique words is expected as Twitter users write in different styles and use many words that are not found in the dictionary.

The trade-off in our method for improving accuracy and ratio of unique and total words is the number of tweets. If the thresholds is too high in the above cases, we will lose a significant amount of data.

Table 7 shows evaluation of the final corpus that we present on the 1,000 manually annotated pairs of tweets. We can see that with our current settings, we obtain reasonably good performance as, 68.1% are indeed parallel tweets, 22.4% tweets that are comparable and only 9.5% pairs are errors. If we group parallel and comparable tweets together, we achieved 90.5% accuracy.

Lastly, to address the concern regarding the translation quality as well as the originality of these translations, we evaluate how the parallel tweets compare with Google Translate using MT evaluation metrics such as BLEU score, NIST, Translation Edit Rate (TER) and Word Error Rate (WER). We take a random 100 pairs of parallel tweets.

Category	English tweet	Arabic tweet
Parallel	#LGgram - one of the lightest laptops in the world! Can you guess its weight?	جهاز #LGgram هو أنحف كمبيوتر محمول في العالم ! هل تستطيع أن تحزر وزنه ؟
Comparable	@k_seghir advices freshmen to follow their passion whilst enjoying the educational journey. Learn both inside and outside the classroom.	مدير الجامعة يدعو الطلبة الجدد للاستمتاع في رحلتهم التعليمية داخل وخارج القاعات الدراسية. (Translation: The university president invites new students to enjoy their educational journey inside and outside the classroom)
Error	Live: The press conference begins with a tour through Dilmun Hall.	مباشر: معالي الشيخة مي تؤكد بأن اختيار قاعة دلمون لعقد المؤتمر الصحفي لما له من دلالة على اثار البحرين. (Translation: Live: Her Excellency Sheikha Mai confirms that the choice of Dilmun Hall to hold the press conference...)

Table 6: Examples of corpus evaluation

Parallel Tweets	Comparable Tweets	Unrelated Tweets
68.1%	22.4%	9.5%

Table 7: Evaluation of the corpus

Accts	Tweets	English		Arabic	
		Tokens	Types	Tokens	Types
1,389	166K	3.8M	380K	3.6M	450K

Table 8: Corpus statistics

BLEU	NIST	TER	WER
27.74	4.55	72.47	77.23

Table 9: Comparison of parallel tweets with Google Translate output

The English tweets from these 100 pairs are used as reference. The Arabic tweets from these 100 pairs are used as input to Google Translate and the outputs from Google Translate are compared with the reference tweets using the above metrics. This comparison is summarized in Table 9. The moderately low values of BLEU score and NIST, along with moderately high TER and WER also suggest that these parallel tweets are indeed human translations.

IDs of parallel tweets, list of Twitter accounts and manual annotation can be downloaded from the Qatar Computing Research Institute resources page `http://alt.qcri.org/resources` or the direct link: `http://bit.ly/2xApE8V`

5. Country and Topic Annotation

To understand the demographics of users who post parallel tweets, we annotate the top 200 accounts, who contribute to 80% of total collected parallel tweets, for their countries of origin and topics of interest. This annotation can be useful for other purposes such as author profiling as well.

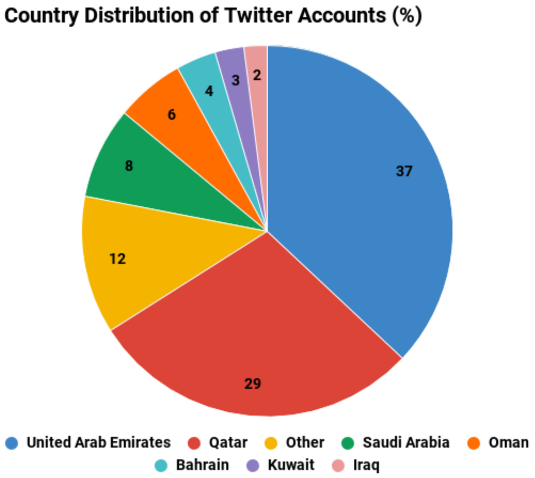

Figure 3: Distribution of accounts according to country

5.1. Country Annotation

We annotate the accounts for their countries of origin. This is not always straightforward as Twitter users may use different kinds of location names on their profiles. We consider city name, country name or flags to get an indication of the country for the account. The distribution of countries is presented in Figure 3. We can see that posting parallel tweets is particularly popular in the Gulf region (UAE, Qatar for example). In the Gulf region, both English and Arabic are used extensively as the population is multilingual. Therefore, we can expect other multilingual communities to be a potential source for parallel tweets as well.

5.2. Topic Annotation

We also annotate the accounts for a topic they are most likely to tweet on. This is done by going through the Twitter profile and identifying the most common topic across tweets. We assign one topic to a profile and categorize

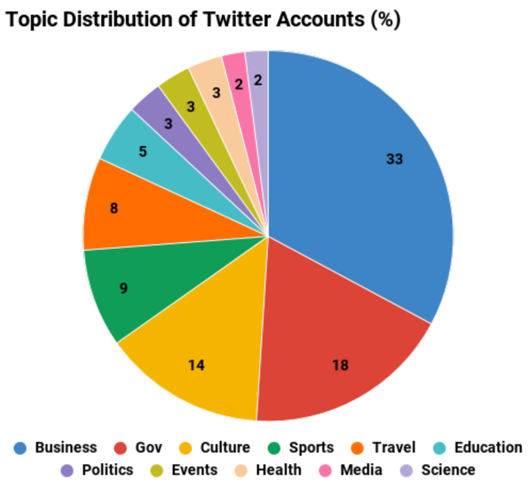

Figure 4: Distribution of accounts according to topic

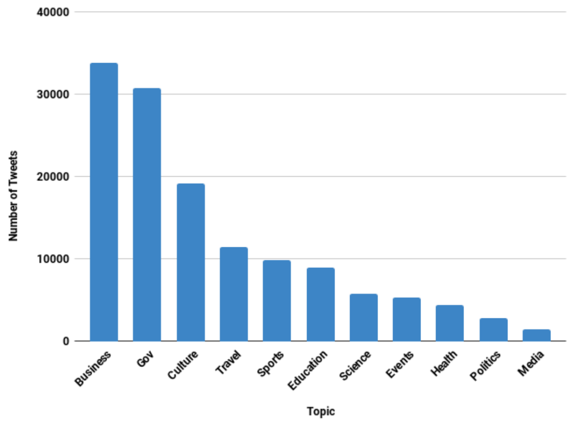

Figure 5: Distribution of tweets according to topic

tweets by that profile to be of that topic. Although the accounts may post tweets related to different topics, for our purposes, a broad understanding of the distribution at the tweet level suffices. Figure 4 shows us the distribution of topics across profiles and Figure 5 shows us the tweet distribution. We can see that majority of the parallel tweets are posted by business (corporations, banks, companies, etc.) or government entities (embassies, ministries, municipalities, etc.) This information can help us in the future to refine our search for accounts who post parallel tweets. During the annotation process, we noticed an interesting phenomenon. Some government or business entities do not post parallel tweets from the same account but use different accounts to post tweets that are translations of each other. For example, the accounts MoI_Qatar and MoI_Qatar_En are two accounts maintained by the same government entity (Ministry of Interior). While MoI_Qatar posts tweets in Arabic, MoI_Qatar_En posts same content translated into English. This has the potential to be an additional resource for parallel tweets and our method can be adapted in future to get those accounts and obtain more parallel tweets.

6. Conclusion and Future Work

In this paper, we have presented a method for collecting parallel tweets of different languages. Using this method, we have collected a bilingual corpus of Arabic-English tweets with over 166K parallel tweets. Although our method has a margin of error, we evaluated how different thresholds can be adjusted to increase accuracy or improve quality of corpus. In addition to the listing of accounts who post such tweets, we have also annotated these accounts with their respective countries of origin and topic that they are likely to tweet on. In the future, we plan to assess the impact of adding such resource to MT systems and use complex stemmer/lemmatizer to improve corpus quality and study its effect on MT performance. We also plan to replicate the same efforts and method to collect data for less-resourced languages.

7. Bibliographical References

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. In *Proceedings of NAACL-HLT 2016 (Demonstrations)*, pages 11–16. Association for Computational Linguistics.

Abidi, K. and Smaili, K. (2017). How to match bilingual tweets ? In *6th NLP 2017 - Computer Science Conference Proceedings in Computer Science & Information Technology (CS & IT)* , Computer Science Conference Proceedings in Computer Science & Information Technology (CS & IT), Sydney, Australia, February.

Alhozaimi, A. and Almishari, M. (2018). Arabic twitter profiling for arabic-speaking users. *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–6.

Bouamor, H., Hassan, S., and Habash, N. (2019). The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy, August. Association for Computational Linguistics.

Buckwalter, T. (2004). *Buckwalter Arabic Morphological Analyzer Version 2.0 LDC2004L02.Web Download.* Philadelphia: Linguistic Data Consortium.

Darwish, K. and Mubarak, H. (2016). Farasa: A new fast and accurate arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1070–1074.

Jehl, L., Hieber, F., and Riezler, S. (2012). Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 410–421, Montréal, Canada, June. Association for Computational Linguistics.

Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mubarak, H. and Darwish, K. (2014). Using twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Lan-*

guage Processing (ANLP), pages 1–7, Doha, Qatar, October. Association for Computational Linguistics.

Mubarak, H. (2018). Build fast and accurate lemmatization for arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Sluyter-Gäthje, H., Lohar, P., Afli, H., and Way, A. (2018). FooTweets: A bilingual parallel corpus of world cup tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Vicente, I. S., Alegría, I., España-Bonet, C., Gamallo, P., Oliveira, H. G., Garcia, E. M., Toral, A., Zubiaga, A., and Aranberri, N. (2016). TweetMT: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2936–2941, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Xu, J., Weischedel, R., Weischedel, R., and Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 105–110, New York, NY, USA. ACM.