Arabic to English Person Name Transliteration using Twitter

Hamdy Mubarak, Ahmed Abdelali
Hamad Bin Khalifa University (HBKU), Doha, Qatar
{hmubarak, aabdelali}@qf.org.qa

Abstract

Social media outlets are providing new opportunities for harvesting valuable resources. We present a novel approach for mining data from Twitter for the purpose of building transliteration resources and systems. Such resources are crucial in translation and retrieval tasks. We demonstrate the benefits of the approach on Arabic to English transliteration. The contribution of this approach includes the size of data that can be collected and exploited within the span of a limited time; the approach is very generic and can be adopted to other languages and the ability of the approach to cope with new transliteration phenomena and trends. A statistical transliteration system built using this data improved a comparable system built from Wikipedia wikilinks data.

Keywords: Transliteration, Named Entities, Social Media, Tweet Normalization, Arabic Language Variations

1. Introduction

With the emergence of social media outlets, millions of users exchange messages daily. This rapid expansion raises new challenges related to retrieval and extraction in a multilingual scope. Named Entities processing has been recognized as a key technique that supports a number of Natural Language Processing fields (Callan and Mitamura, 2002) and (Khalid et al., 2008). Using traditional approaches for building transliteration resources (Kirschenbaum and Wintner, 2010; Hálek et al., 2011) or mining them from text and news (Darwish et al., 2012; Kumaran et al., 2010; Sajjad et al., 2011) might not keep the pace with rapid expansion of information form such outlets. The social media outlets are providing large volume, high-value, content that is being sought by researchers, both in business and academia. Opinion mining (Lukasik et al., 2015; Manoochehr et al., 2013; Agarwal et al., 2011), customer relation, eBusiness, eHealth (Paul and Mark, 2011; Luis et al., 2011) are examples for disciplines that are exploiting these resources.

The amount of data generated from the tweets only surpasses 500 millions tweets per day\(^1\), as such, it presents a unprecedented type of versatile resource that can be utilized namely for transliteration. Unlike similar resources, Twitter data includes explicit data about user, location, language, social network...etc.

In our paper, we present results of experiments for harnessing large number of tweeps\(^2\) information to build a transliteration module that can be used to support translation as well as cross-language information retrieval. The advantage of using tweets versus other methods is the accuracy as well as the freshness. While linguistic resources such as Encyclopedia, Onomasticons might require time to maintain and update. Social media are becoming a faster way to get large amount of information. The occurrence frequency of a given item reflect well the accuracy and its standard use. For our case-study language “Arabic”, we were able to collect over 880,000 unique Arabic users with their transliteration to English in a period of few months. This is 500% more than all the data extracted from Wikipedia (WK) (see Table 1). Even though, data from Twitter might not totally substitute high-quality, consistent and collaboratively edited data from WK.

It is common to note variations within a language. Researchers have studied and documented such phenomena in corpora (Abdelali, 2004; Abdelali and Cowie, 2005). The large amount of data from Twitter persistently disclose current trends and methods used to transcribe names. Given the Arabic name “أحمد” (AHmd)\(^3\) Wikipedia accounts for 56% of the times the name is transliterated as “Ahmed”, 40% ”Ahmad”, 4% to ”Ahmet, Akhmad, Akhmet, Achmad”. For the name “أشرف” (A$rf)” 93.5% ”Asr” 7% ”Achraf”. Twitter data proved to be far more richer and new phenomena and trends were observed and learned from these data. We note that the former names were transliterated in further more ways. “محمد” (AHmd) was transliterated into “ahmed, ahmad, ahmd, a7mad, a7med, a7md, and ahmmd” and “شرف” (A$rf) transliterated into “asr, ashrf, arsh, shrf, achr, ashrf””. The study provides details for collecting, processing and validation for the usability of this resource which is being made publicly \(^4\). We built a transliteration model using character-based model and we were able to achieve higher scores in BLEU comparing to an equivalent set from WK data (Kirschenbaum and Wintner, 2010).

The remainder of this paper is organized into the following sections: Review for the state-of-the-art and related research, Twitter data collection and pre-processing, followed by experiments and lastly results and a conclusion.

---

\(^1\)See http://www.internetlivestats.com/twitter-statistics/

\(^2\)Tweep: A person who uses the Twitter online message service to send and receive tweets.

\(^3\)Buckwalter Transliteration

\(^4\)http://alt.qcri.org/resources/
### Table 1: Statistics from WK using interwiki links for Named Entities translation/transliteration.

|   | en(k) | fr(k) | de(k) | es(k) | ar(k) |
|---|-------|-------|-------|-------|-------|
| en | 5967.8 |       |       |       |       |
| fr | 599.6  | 907.2 |       |       |       |
| de | 578.3  | 469.6 | 857.4 |       |       |
| es | 439.8  | 397.1 | 340.6 | 699.1 |       |
| ar | 154.1  | 133.1 | 120.9 | 136.8 | 233.2 |

2. Related Work

WK as a free multilingual encyclopedia, provides a valuable resource for parallel information that can be easily processed and deployed in cross-language Named Entity (NE) disambiguation, resolution and translation.

Wentland et al. (2008) used WK to build Heidelberg NE Resource (HeiNER), a large multilingual resource that is used for NE disambiguation, translation and transliteration. The resource contains lists of NEs with various sizes in 15 languages. They used triangulation cross languages to expand the initial lists. The size of the English list was 1.74 million entries. The numbers decrease sharply for non-Western languages.

Similarly, Halek et al. (2011) built a bilingual lexicons for English-Czech that was used to improve transliteration in a Statistical Machine Translation (SMT) task. Using the new mined resource improved the score with about 0.5 BLEU points.

Sajjad et al. (2011; 2012) mined transliteration from parallel corpora to improve SMT system. Their unsupervised transliteration mining system uses a parallel corpus to generate a list of word pairs and filters transliteration pairs from that. The system will be retrained on the filtered dataset and this process is iterated several times until all transliteration word pairs were detected. The approach proved fruitful with a BLEU improvement of up to 0.4 points.

Yoon et al. (2007) proposed a phonetic method for multilingual transliteration. The approach exploits the string alignment and linear classifiers that were trained using the Winnow algorithm to learn transliteration characteristics. The results achieved were improved over earlier results reported by Tao et al. (2006); methods built using pure linguistic knowledge.

Yoon et al. (2007) used Mean Reciprocal Rank (MRR) to measure the performance of the transliteration system tested on Arabic, Chinese, Hindi and Korean. The main challenges with former approaches is both unrobustness or dependability on scarce resources that are not easy to find. Data collected from Twitter can extend rapidly and complement the resources in WK.

3. Collecting Names from Twitter

When creating a new account on Twitter, user fills full name (in any characters; less than 20 characters), and an email. Twitter might suggest some user names (unique account names) based on the combinations of the user’s full name and email. User may select from the suggested names or write a new one (in alphanumeric characters only) as shown in Figure 1. This restriction compels the user to transliterate his/her name. Hence, for our case-study, we proceed to collect full names written in Arabic with their transliterations using Twitter user ID (username field).

Figure 1: Creating a new account on Twitter; user is required to provide an alphanumeric username.

Figure 2 shows some of the name-pairs that can be collected using the above approach. In profile, a user can also provide a location which can be a country name, city name, or a landmark name. To map user locations to Arab countries, we used a list which contains the top unique 10K user locations with their mapping to Arab countries by the aid of GeoNames geographical database (Mubarak and Darwish, 2014).

In our experiment, we collected Arabic tweets by issuing the query “lang:ar” against Twitter API. We extracted user’s full name, username, and user location. The language filter can be changed to collect names in other languages along with their transliterations. Between Mar. 2014 to Aug. 2014, we collected approximately 7.3M tweets written by 936K unique users, and 557K (or 60%) of their names have Arabic characters in the full name field. We cleaned the data as it will be detailed further and extracted full name written in Arabic (Name<sub>arb</sub>) that has an overlap above a certain threshold with username written in Latin characters (Name<sub>trans</sub>), along with user location (loc). Sample results are shown in Table 2, where we can note that the transliteration uses standard mapping such as UNGEGN romanization standard (UNGEGN, 2003); additionally, other non-standard transliterations are used such as the case of using numbers “7” and “3” instead of letters “ح” and “ع” respectively, and also transliterating the Arabic letter “ق” to “c” which is not very common.

3.1 Data Collection and Preprocessing

Using the data collected; a number of steps were used to process this data including:

- Name<sub>arb</sub>, Name<sub>trans</sub>, and loc are normalized as described in DARWISH et al. (2012) (ex: convert letters “&lt; | p, y)” to “&lt; | p, y)” in order, and map non-Arabic decoration characters

http://www.geonames.org/

http://dev.twitter.com

"ISO 3166-1 alpha-2 codes" is used for country codes.
Figure 2: Collecting username information from Twitter in different languages.

Table 2: Samples of extracted names from Twitter Collected data along with their countries.

| Full name          | Username          | Country |
|--------------------|-------------------|---------|
| فارس بن سعود      | farisbinsaud      | SA      |
| حسام جودة         | 7ossamGouda       | EG      |
| عادل الاحطيب      | 3adelalkhteeb     | SY      |
| أمين رفيع          | amineratic        | DZ      |
| السيدа التقليدية   | SAIDA_CHEBBI      | TN      |

Table 3: Name cleaning of characters decoration.

- Titles are removed, ex: “د. الشيخ (d., AlSyx), meaning Dr., Sheikh”, also Mr, Miss, etc.

3.2. Informal Character Writings

Name<sub>trans</sub> sometimes have numbers to represent Arabic letters that have no exact sounds in Latin languages. These numbers are similar in shapes to Arabic characters as shown in Table 4.

3.3. Dialectal Variations in Names

From names that are mapped to Arab countries (using user location), we extracted variations of mapping Arabic characters to Latin equivalents in different countries or regions. Table 5 lists common variations for characters that are affected by the dialects used in Arab countries or regions. These variations are used to classify Arabic names geographically, i.e. inferring a country or a region given only the full username written in Arabic on Latin characters (Mubarak and Darwish, 2015).

3.4. Transliteration Similarity Score

Our hypothesis for name transliteration between Name<sub>arb</sub> and Name<sub>trans</sub> needed a gauge to measure and quantify the similarity between them. Given a Name<sub>arb</sub> is transliterated using elaborate mapping scheme similar to Buckwalter transliteration. We took into consideration removing of name title, informal writings and dialectal variations, some characters are considered equivalent (ex: k=q, 

---

*The list of characters mapping is available at http://alt.qcri.org/resources/TwitterAr2EnTranslit.tgz

*Regions: Gulf (GLF), Egypt (EG), Levant (LEV), and Maghreb (MGR)
names, and 21% of the extracted Arabic names are mapped.

5. Resource Description

The data released from this task includes 881,310 name pairs that can be used for Arabic to English person name transliteration with their respective score. For each name pair, we have the original username, normalized username (Arabic name), user screen name (English transliteration), one of the Arab countries (if possible) according to user location, name tokenization, and similarity score (transliteration accuracy).

The published resource includes also a list of 719 character mapping. The resources are publicly available from http://alt.qcri.org/resources/TwitterAr2EnTranslit.tgz

6. Evaluation and Results

To assess the quality of this resource, we randomly selected 1,000 name pairs from the original names having Arabic characters, and counted how many of these names are extracted as valid transliteration name pairs using our system. The precision (P) was 0.96, the Recall (R) was 0.97, and F1-Measure was 0.965. For example, the system gave the accepted this name pair. To further explore the potentials of using the resource in Machine Translation; We used a statistical phrase-based MT system to build a character-based translation model to experiment with different data processing schemes and evaluate the new data. The system was built with the Moses toolkit default settings. The language model used in the system was implemented as a five-gram model using the SRILM-Toolkit (Stolcke and others, 2002). We compiled three datasets. T100 uses only Twitter data with a threshold of 100. T50 data with threshold greater or equal to 50. in addition to data from WK. We build an additional dataset that was the combination of T50 and WK. For the data used to build the models for evaluation, we randomly extracted two sets of 2000 pairs and used one set for development and the other for evaluation. The remaining data was held for training and building the models. The same approach was applied uniformly on WK data. The results in Table 7 shows that the data collected from twitter cannot be transliterated using model trained on Wikipedia. A Strong indication of
the difference between these two data. On the other hand combining both data proves to be beneficial for processing both datasets. This could be explained by the richness of the twitter data and the consistency of WK data.

7. Conclusion

In this paper, we presented a methodology for harvesting valuable data from Twitter and used it for person name transliteration from Arabic to English. The collected data, that is being made publicly available, improved transliteration system. Additionally, when compared to collected data from WK, Twitter data has supplementary benefits: 1) Huge amount of parallel data, 2) Dialectal variations coverage, and 3) Informal writings. Our future work will aim to extend this approach to other languages with focus on languages with low presence in WK.

8. Bibliographical References

Abdelali, A. and Cowie, J. (2005). Regional corpus of modern standard arabic. In 2ème Congrès International sur l’Ingénierie de l’Arabe et l’Ingénierie de la langue. Algeria, pages 1–11.

Abdelali, A. (2004). Localization in modern standard arabic. volume 55, pages 23–28. Wiley Subscription Services, Inc., A Wiley Company.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Pasbourg, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media, pages 30–38.

Callan, J. and Mitamura, T. (2002). Knowledge-based extraction of named entities. In Proceedings of the eleventh international conference on Information and knowledge management, pages 532–537. ACM.

Darwish, K., Magdy, W., and Mourad, A. (2012). Language processing for arabic microblog retrieval. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 2427–2430. ACM.

Hálek, O., Rosa, R., Tamchyna, A., and Bojar, O. (2011). Named entities from wikipedia for machine translation. In Conference on Theory and Practice of Information Technologies, pages 23–30, Vrátna dolina, Slovak Republic.

Khalid, M. A., Jijkoun, V., and De Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering. In Advances in Information Retrieval, pages 705–710. Springer.

Kirschenbaum, A. and Wintner, S. (2010). A general method for creating a bilingual transliteration dictionary. In LREC10, pages 273–276, Valletta, Malta.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. (ACL’07), Prague, Czech Republic.

Kumaran, A., Khapra, M. M., and Li, H. (2010). Whitepaper on news 2010 shared task on transliteration mining. In Proceedings of the 2010 Named Entities Workshop: Shared Task on Transliteration Mining. ACL.

Luis, F.-L., Randi, K., and Jason, B. (2011). Review of extracting information from the social web for health personalization. Journal of Medical Internet Research, 13(1).

Lukasik, M., Cohn, T., and Bontcheva, K. (2015). Estimating collective judgement of rumours in social media. arXiv preprint arXiv:1506.00468.

Manoocoehr, G., James, S., and David, Z. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications, 40(16):6266–6282.

Mubarak, H. and Darwish, K. (2014). Using twitter to collect a multidialectal corpus of arabic. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pages 1–7, Doha, Qatar.

Mubarak, H. and Darwish, K. (2015). Classifying arab names geographically. In Proceedings of the ACL 2015 Workshop on Arabic Natural Language Processing (ANLP), pages 1–8, Beijing, China.

Paul, M. J. and Mark, D. (2011). You are what you tweet: Analyzing twitter for public health. In ICWSM, pages 265–272.

Sajjad, H., Fraser, A., and Schmid, H. (2011). An algorithm for unsupervised transliteration mining with an application to word alignment. ACL-HLT’11, Portland, OR, USA.

Sajjad, H., Fraser, A., and Schmid, H. (2012). A statistical model for unsupervised and semi-supervised transliteration mining. (ACL’12), Jeju, Korea.

Stolcke, A. et al. (2002). SRILM – an extensible language modeling toolkit. In Proceedings of the International Speech Communication Association (INTERSPEECH’02), Denver, CO, USA.

Tao, T., Yoon, S.-Y., Fister, A., Sproat, R., and Zhai, C. (2006). Unsupervised named entity transliteration using temporal and phonetic correlation. In EMNLP, pages 250–257. Association for Computational Linguistics.

UNEGGN, W. G. o. R. S. (2003). Report on the current status of united nations romanization systems for geographical names. version 2.2. January.

Wentland, W., Knopp, J., Silberer, C., and Hartung, M. (2008). Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In Proceedings of the 6th LREC, Marrakech, Morocco.

Yoon, S.-Y., Kim, K.-Y., and Sproat, R. (2007). Multilingual transliteration using feature based phonetic method. (ACL’07), pages 112–119, Prague, Czech Republic.