Person Named Entity Generation and Recognition for Arabic Language

Ibrahim A. Alkharashi

Computer and Electronic Research Insitiute
King Abdulaziz City for Science and Technology
Riyadh – Saudi Arabia
kharashi@kacst.edu.sa

Abstract

Arabic person named entity has unique characteristics that govern the generation and the analyzing processes. This work presents an infrastructure that has been developed to assist and simplify the processes related to Arabic person named entities such as generation, recognition, translation and transliteration and correction. The infrastructure consists of a single two-dimensional person name map-table that maps the root of a given Arabic person name to its valid pattern. Each valid Arabic person name or Arabic name fraction is projected onto that table. Each projected entry is assigned some properties that identify some affixation and gender characteristics. The paper will focus on two main parts regarding the Arabic person name map-table. The first part will be devoted to discuss the Arabic person named entity production behaviour of Arabic language with emphasis on some statistical findings on the capabilities of roots in producing person names and patterns usability as a model for person name. The second part will discuss the usability of the Arabic person name map-table as a training set to assist the process of person named entity recognition, correction and transliteration.

Introduction

Named Entity Recognition (NER) is the process of identifying proper names including people, locations, and organizations, in a given text. It is a basic process developed for many languages and used to assist many automated applications such as translation, named entity transliteration, text POS Tagging, information retrieval and data mining. Person Named entity is a subclass of Named Entity that deals with person names.

Most of the Arabic person names are generated in the same manner as other Arabic nouns and hence follow the same morphological and syntactical rules. In general Arabic words are derived from set of about six thousand roots using few hundreds patterns. Most likely, if an Arabic person name is taken out of its context, then it will not be recognized as so due to lake special identification marks such as capitalization. As the case of normal text, Arabic person names within a text will be written with no short vowels. After the spread of Islam, some new Person names and Person naming convention have been adapted by Arabic language. Some of those names do not easily follow the grammatical and syntactical rules of Arabic language, and it adds complexity to the processes needed to treat them (Akhtar 2007).

Single Arabic person name comes in different formats including simple, affixed and compound. The simple form representing a simple person named entity such as عمر "Omar" that can be mapped to a root-pattern. Affixed names such as العمر "Alomar" is a simple person name prefixed with one or more name-prefixes. Finally, a compound name such as شمس الدين "Samsuldain" is a name that is composed of two or more simple names with one or more name-prefixes. These variations complicated the process of handling Arabic person named entities. Furthermore, absence of vowels from Arabic text added more complexity to the processes. It is very common to interpret an non-diacritized Arabic word as being a verb, a noun with many meanings or a person name with one or more spilling. For instant, the word "حسن" as a person name, can be interpreted as "حَسَن" to name a male or as "حُسْن" to name a female.

Arabic names are based on a long naming system generated by chaining list of names. This system is in use more or less throughout the Arab world. Affixation and inserting name connectors is a common phenomenon in Arabic person naming system. A given Arabic name can be decomposed into simple names, name affixes and name connectors. For example the name of this Umayyad caliphate "عبدالملك بن مروان" has "ملك" and "مروان" as simple names, "عبد" and "ال" as name prefixes and "بن" as a name connector.

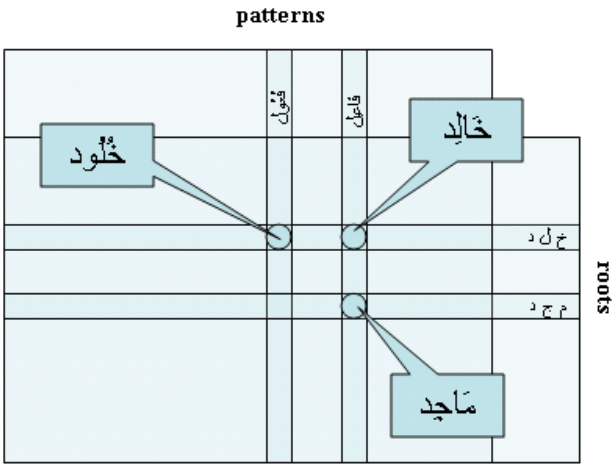
Dataset

Corpus of unique Arabic person names is used to build the dataset. The list consists of more than 100,000 unique given names and surname of Saudi nationalities.

Each compound person named entry in the collection is decomposed into smaller name fractions. Further more, name affixation articles, such as ال and أبو, were identified, removed, and listed in another table. Finally, each simple name or name fraction is tagged with left and right affixability (ability to accept affixes) and gender properties and then projected onto the proper cell in root-pattern map-table. At the end of this tedious manual process, a total of sixty thousands cells were created with all properties assigned. As shown in Figure 1 the map table is virtually a two dimensional table that maps root with pattern for every valid Arabic person name fraction.

Two more tables were created as result of this process, namely the pattern and affix/connector tables. The Pattern table lists all patterns used in the map-table along with usage frequencies. Table 1 list the most frequently used patterns. The other table is the name affixes/connector. Table 2 lists the most common name affixes and name connectors used in Arabic. A third table was created and added to the dataset to assist the process of person name entity recognition. The table is the trigger token list. The table lists set of tokens that are used to determine whether the proceeding and/or the succeeding word may trigger a certain person named entity. Table 3 shows partial entries from trigger token list.

Figure 1: Arabic name map-table for the dataset



pattern	names		pattern	names	
	count	%		count	%
فَعلَة	3603	5.74	مَفعَلِيّ	603	0.96
فَعَل	3364	5.36	فُعَيلِل	599	0.95
فَعَّال	3096	4.93	فَاعُلِي	517	0.82
فُعَيل	2890	4.60	مُفَاعِل	506	0.81
فَعلَان	2406	3.83	مَفعُول	425	0.68
فُعَيلَة	2261	3.60	فَعلِيل	407	0.65
فَعلِي	2235	3.56	فَعُولِيّ	374	0.60
فُعَيلِيّ	1852	2.95	أفعَال	363	0.58
فَاعَل	1809	2.88	أفعَل	346	0.55
مَفعَل	1680	2.68	مِفعَال	346	0.55
فَعَالِي	1552	2.47	فُوَيعِل	337	0.54
فَعلَاء	1327	2.11	فُعَيلِى	330	0.53
فَعُّول	1283	2.04	فَعُولَة	326	0.52
فَعَّالَة	1253	2.00	فَعَّة	325	0.52
فَعلل	1179	1.88	فُعلُولَة	325	0.52
فَعلَا	1139	1.81	مُفَيعِل	323	0.51
فَاعِلَة	1136	1.81	فَعِي	315	0.50
فَعلِيَّة	997	1.59	فَعلِلَة	292	0.47
فَعلَى	964	1.54	فَعلَاوِيّ	291	0.46
فُعَيلَان	933	1.49	فَعِيلِيَّة	284	0.45
فُعلُول	837	1.33	فَعلَين	274	0.44
فَعلَانِيّ	749	1.19	فُعَيلِيل	264	0.42
مُفعِلَة	687	1.09	فُعَيلَاء	256	0.41
فِعلَال	644	1.03	فَعِيلَا	252	0.40
فَعَلِي	641	1.02	فَاعُول	250	0.40

Table 1: Most frequently used patterns

Affixes / connectors					
آل	آل	أمة	بنو	ذي	عبدرب
أب	أم	أمة	بني	رب	عبيد
أبا	أمة	أو	بو	سي	عز
إبن	أهل	با	بي	سيد	لال
أبو	أبى	باي	ذا	سيدة	لالو
أبى	أخو	ب	ذات	سيدة	للا
أخو	ال	بن	ذو	عبد	ل
ال	أم	بنت	ذوي	عبدة	ولد

Table 2: Common name affixes and connectors.

المغني	العقيد	الحاكم	المحترم
المغنية	الفريق	الحاكمة	الموقر
المقاول	الفنان	الدكتور	حفظه الله
المقاولة	الفنانة	الدكتورة	رحمه الله
المقدم	القبطان	الرئيس	رضي الله عنه
الملاح	القسيس	الرائد	يرحمه الله
الملاحة	الكاتب	الرسول	الأستاذ
الملازم	الكاتبة	الرقيب	الأستاذة
المهندس	اللواء	الشيخ	الملك
المهندسة	المرحوم	الشيخة	الملكة
النبي	المساعد	الطبيب	الأمير
النقيب	المساعدة	الطبيبة	الأميرة
الوزير	المشرف	الطيار	الإيدميرال
الوزيرة	المشرفة	الصحابي	الإستشاري
الوكيل	المعلم	التابعي	الإستشارية
الوكيلة	المعلمة	العريف	الجنرال

Table 3. Trigger token list

Arabic Person Name Generation

Three interesting observations can be made about personal name generation of Arabic language. First observation is about roots capability of producing personal named entities. Roots vary in their generation capability, some roots such as "ح م د" are very generous in producing person names and some do not contribute at all to the name list. Table 4 shows the most productive Arabic roots. The second observation is that only 16 Arabic patterns contribute to the production of more than 50% of Arabic person names. The last observation is regarding number of patterns per pattern length. Figure 2 shows that patterns of six characters in length contribute more to the pattern map. That does not mean that patterns of six characters produce more names. As matter of fact Figure 3 shows that patterns of length of five letters are the most productive patterns for Arabic person names.

root	names count	root	names count
حمد	146	رود	60
سلم	112	نصر	57
سعد	107	صلح	56
حسن	90	صبح	56
عبد	89	شرف	55
عمر	80	عقل	55
رشد	71	جهر	53
سمر	71	خلد	52
برك	68	حول	52
خضر	66	عرف	52
زهر	66	نعم	52
قبل	63	سرح	51
نور	63	جمل	50
جبر	61	حضر	50
ربع	61	خلف	50
روح	61	رزق	50

Table 4: Most productive roots

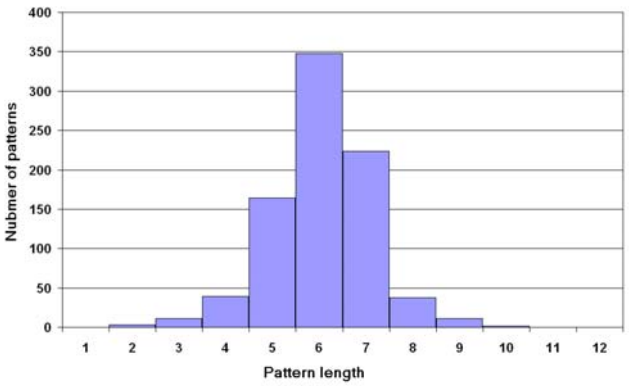


Figure 2: Number of patterns per pattern length

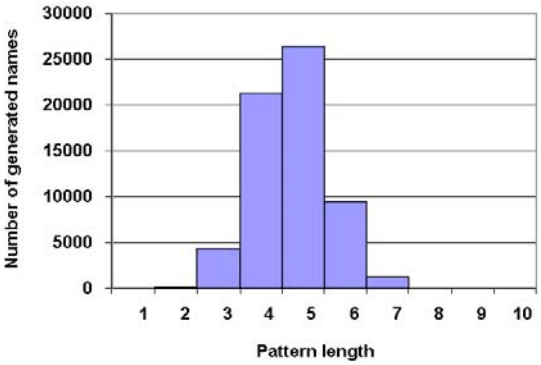


Figure 3: Number of generated names per pattern length

The Dataset as a Transliteration Tool

Transliteration is a fundamental task in any machine translation systems (Kashani, 2007). It also has potential needs in many information system applications (Al-anzi, 2004; Larkey, et al 2003).

Diacritic restoration for non-diacritized Arabic names is an essential step in transliteration process. Researchers have taken different approaches to diacritized Arabic name prior to transliteration. Alghmdi (2005) approach is achieved by using a lookup table to find one or more diacritized names that match a given input name. Al-anzi (2004) proposed stochastic models for automatic diacritics restoration of Arabic names.

Many other researchers based their transliteration approaches on utilizing parallel corpora. Hermjakob et al (2008) developed a trained Transliteration system based on a bitext of seven million sentences and Google's English terabyte ngrams. Hassan et al, (2007) proposed a language independent approach to extract two lists of named entities to be used for transliteration, from two pair of aligned documents. Fei (2005) developed several language-independent features to capture phonetic and semantic similarity measures between source and target named entity pairs to solve various named entity translation problems presented in different language pairs including Arabic to English. Tan et. al (2005) utilize comparable corpora to develop an unsupervised named entity transliteration approach using temporal and phonetic correlation.

To illustrate the usability of the dataset for transliteration, we use a pattern match score formula along with pattern usage frequency as suggestive tool for the process of diacritic restoration.

Pattern match score formula is a very simple selecting tool to suggest the most appropriate patterns that can be used to diacritize a given Arabic person name. It will give a single point if a letter from the person name matches an original pattern letter (i.e. "ف", "ع", "ل"). And it will give two points if a letter from the person name matches a non original pattern letter (e.g. "م", "أ", "ت", "ة"). For example, for the name "أبـي", all three letters patterns with no weak letters (e.g. "فَعَل", "فَعل", "فُعُل", "فِعِل") will give the same match score of three. However list of patterns with a weak LAM letter ("فَعِـي", "فَعِـيّ", "فَعِّـي", "فُعَـي") will score four points. As a result, pattern selection should be applied on the second list rather than the first. Within a list of patterns with the same high score, a pattern or more with the highest frequencies will be suggested. Those patterns then will be used to restore diacritic prior to transliteration. Table 5 shows sample scoring results and frequency based selection for some supplied person names.

match score	pattern	frequency
أثلاوي		
8	أفعَلِلي	7
9	أفعَالِيّ	23
9	فَعلَاوِيّ	247
9	فَعَلَاوي	4
9	فُعلَاوي	2
9	فِعلَاوِيّ	32
بطيحاء		
4	أفعَلِلي	7
4	أنفَعلَل	1
4	أفعَلِلي	2
5	أفعَلَان	8
7	فَاعِلَاء	15
8	فَعلَلَاء	15
8	فَعِيلَال	7
9	فُعَيلَاء	253
باطل		
3	أفعُل	2
3	أفعَل	337
3	تَفَعُّل	21
5	فَاعَل	36
5	فَاعُل	20
5	فَاعِل	1730
بعقان		
5	أفعَال	363
5	إفعَال	24
5	إفعَال	53
5	فَعلَاء	1265
5	فَعلَاة	26
5	فَعلَات	128
6	فَعلَال	218
7	فَعلَان	1943
7	فَعَلَان	53
7	فَعِلَان	6
7	فُعلَان	174
7	فُعَّلَان	7
7	فِعلَان	210

Table5: Sample scoring results

To further investigate the capability of the dataset to suggest a correct pattern for a given Arabic personal name, we run series of experiments. From the dataset we generated a rough data of more than 60000 entries of person names. Each entry has two values; the name (which can be a name fraction) and its correct pattern. Same names might be listed more than once, and hence associated with different patterns. The generated data is divided into to sets; randomly selected training set (80%) and testing set (20%). We run the same experiment ten times, each time with different training and testing sets.
For each run we generated a frequency list out of the training set. For each entry in the testing set, we run the selecting tool to suggest the most appropriate pattern list. From the suggested pattern list we recorded the position of the correct pattern, if any. Table 6 counts the number of times the correct pattern appears in a given position in the suggested pattern list. Figure 4 shows the probability that the correct pattern for a given name appear in a corresponding sequence. In summery, results show that for set of totally new names, the probability that the correct pattern is appear in the first three suggested patterns is 0.94, the probability that the correct pattern is appear in the first two suggested patterns is 0.86 and probability that the correct pattern is appear as the first suggested patterns is 0.69.

run	Number of times the correct pattern appears in a given position in the suggested pattern list for 10200 names						
	1st	2nd	3rd	4th	5th	6th	7th
1	7059	1739	797	340	111	68	45
2	7007	1685	791	347	127	64	58
3	6996	1794	766	334	112	65	58
4	7096	1754	725	342	121	69	64
5	6956	1809	807	359	120	67	58
6	7024	1806	774	338	119	58	48
7	7005	1816	817	349	121	70	45
8	7146	1687	740	336	129	68	48
9	7021	1759	746	342	125	64	48
10	7038	1710	741	329	123	52	69

Table6: Sample scoring results

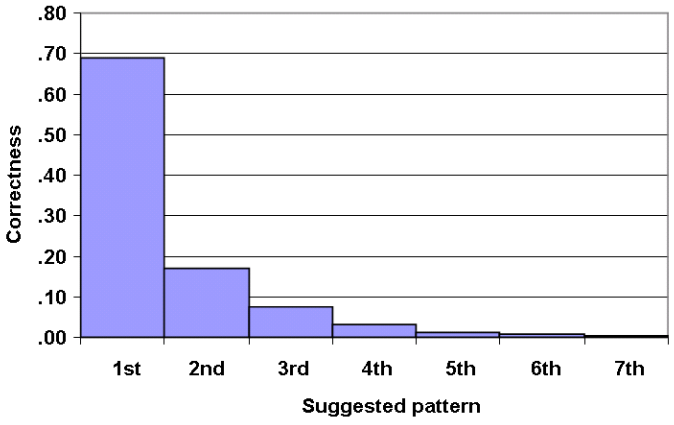


Figure 4: probability of a correct pattern shown in suggested pattern list

Conclusions and Future Work

This work presented a basic dataset that can be used to assist processes related to Arabic person named entity generation and recognition. Some interesting findings about Arabic person named entity generation were introduced. Some of those findings were utilized to automate and assist the Arabic name transliteration. In a future work, the dataset will be utilized and tested in the process of Arabic person named entity recognition.

References

Akhtar, N. (2007). Indexing Asian names. The Indexer (25)4, 12-14.

Al-anzi, F. S (2004). Stochastic Models for Automatic Diacritics Generation of Arabic Names. Computers and the Humanities. 38(4), 469-481. Springer Science, Netherlands.

Alghamdi, M. (2005). Algorithms for Romanizing Arabic names. Journal of King Saud University: Computer Sciences and Information. 17: 1-27.

Hassan H. & Sorensen J. (2005). An Integrated Approach for Arabic-English Named Entity Translation. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 87–93, Ann Arbor.

Hassan A., Fahmy H. & Hassan H. (2007). Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora. Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP, 2007), AMML Workshop.

Hermjakob, U., Knight, K. & Daum´e III, H. (2008). Name translation in statistical machine translation learning when to transliterate. In Proceedings of ACL-08: HLT, pp 389–397, Columbus, Ohio, June. Association for Computational Linguistics.

Kashani M. (2007). Automatic Translitration from Arabic to English and its impact on translation. Master thesis in computer science. Simon Fraser University. Burnaby, BC, Canada.

Larkey, L., AbdulJaleel, N., Connell M. (2003) What's in a Name?: Proper Names in Arabic Cross Language Information Retrieval. CIIR Technical Report IR-278.

Tao, T., Su-Youn, Y., Fister, A., Sproat, R. and Zhai, C. (2006). Unsupervised Named Entity Transliteration Using Temporal and Phonetic Correlation. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006). Association for Computational Linguistics. 250–257.