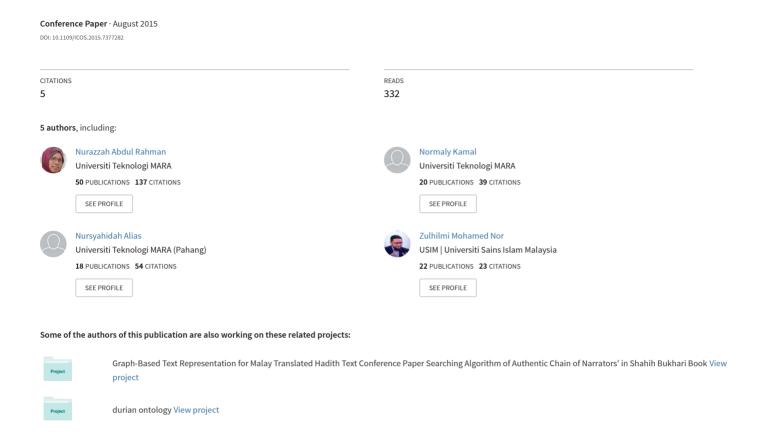
An Identification of Authentic Narrator's Name Features in Malay Hadith Texts



An Identification of Authentic Narrator's Name Features in Malay Hadith Texts

Nurazzah Abd Rahman¹, Nursyahidah Alias¹,
Normaly Kamal Ismail¹

¹Fakulti Sains Komputer dan Matematik
Universiti Teknologi MARA
Shah Alam, Malaysia
{nurazzah, normaly}@tmsk.uitm.edu.my,
syahidah@pahang.uitm.edu.my

Zulhilmi bin Mohamed Nor², Muhammad Nazir b Alias³
²Fakulti Pengajian Quran dan Sunnah Universiti Sains Islam Malaysia Bandar Baru Nilai, Malaysia ³Fakulti Pengajian Islam Universiti Kebangsaan Malaysia Bangi, Malaysia

Abstract - Name is regarded as important to human. A person's name can give impact to him/her either positively or negatively. It is important to recognize or match a person's name with the right person. To do this on the computer, we need to develop a rule so that the computer will be able to recognize the right name with the right person. Before developing the rule, the features about the person's name itself need to be identified. So far, there has been very little work on recognizing a person's name in the Malay texts. This research identifies features of authentic narrator's name in the Malay hadith texts. We extracted the authentic narrators' names manually from hadith texts in the purpose of identifying the features of the names. We extracted manually 455 narrator's name from 150 Malay hadith texts. From the extracted name, exist same person of narrator's name with different types of spelling. There are also many forms of the narrator's names which refers to one person. We also identify the name manner in the Malay hadith texts and added 3 more name manners. Then, we developed the rule to recognize the narrators' names in the Malay hadith texts based on the identified features.

Keywords - features; Malay text; rule-based; identification

I. INTRODUCTION

The Malay hadith texts are the translated version of the Arabic hadith texts. The Arabic hadith texts use the classical Arabic language which is the standard language of Islam from 1500 years ago [1]. The narrators' names in the Malay hadith texts involve the process of transliteration [2].

A person's name in the Arabic hadith texts could be translated into one or many forms [1]. For example, a narrator's name such as a such as a can be translated into the Malay texts as Aisyah, A'isyah or 'Aishah. As mentioned in [1], standard sets or corpora can be obtained if the research community always uses the corpora as it will produce standard data. For example, [13] used Sunan Tarmizi's book for his research. Therefore, it is difficult to get the standard spelling of a narrator's name in the Malay hadith texts because there is no standard Malay hadith corpus in the research community. In addition, identifying a narrator's name can also pose a problem because of the many forms of name and their similarities that exist in the corpora. Thus, it

will take a long time for researchers to determine the narrator's name accurately [4].

The purpose of identifying the characteristics of narrators' name is to develop rule-based to recognize the narrators' names in the Malay hadith texts. Moreover, the verification of a hadith narrator's name is very important because it is the first step in the process of checking the authentication of a hadith [1]. Furthermore, digitized Malay hadith texts came from the printed version and were converted into texts. Therefore, the verification of the hadith narrators' names will become important in the digitized version. Existing related works that recognize persons' names in texts include [8] which designs the algorithm for English-based personal names and [9] which is a design to handle the standard and Egyption colloquial Arabic texts. In addition, [10] explores the mapping relation between English person names and Chinese translated names.

Named Entity Recognition implementation depends on domain of the reserach and type of language used. The NER developed to specific domain perform not well when it is used to other domain [14]. For example NER developed for recognition the political text only cannot be used for recognition in other domain [19]. Very little research work has been found on recognizing person names in Malay texts. Research on Malay NER done by [17], [14] used news artikel in Bernama and Mstar website, [7] used news artikel in Utusan Online website, [18] used old Malay manuscript Corpus with Jawi character. But, research on Malay NER used Malay hadith text not yet explore.

This paper has 4 sections. Section 1 is the introduction. Section 2 discusses the methodology of the research. Section 3 puts forward the results and discussions while Section 4 presents the conclusion of the paper.

II. METHODOLOGY

The process of selecting narrator's name features was done manually because it is the most widely used method [1]. This research will use [4] as reference in identifying the name manner of narrators in the Malay hadith texts as shown in Table I. Table I shows the name manner of narrators in the Arabic texts.

Usually, the authentic narrator's name in Malay hadith text is structured as in Table II.

From Table II, P refers to the narrator's name. The word "bin" refers to son-of someone (in this case, Q), while "binti" refers to daughter-of somebody (in this case, Q). Q is the father, R is the grandfather and S is the fourth grandfather of P. It is important to note that, P, Q, R and S may consist of more than one word. Structure 3 in Table II indicates that a narrator's name may be written until the fourth grandfather's name. Why is this so? According to the expert, it is very common for a Malay person to have the same name. Thus, to differentiate those with similar name is by adding the forefathers' names to their own name. Thus, by having their names written until the fourth grandfather would determine the genealogy of the family. This, in turn, will also determine the name with the right narrator. The Family and Luqob name manner is also used for the same reason. It is because there are 200 authentic narrator names only in Shahih Bukahri book.

A person's name in unstructured Malay texts has a sequence of capital letters [7][16] that we used as the first step to identify a name. By referring the name manner in Table I, Table III explains the methodology that we used to identify every component in the name manner in the Malay hadith texts.

TABLE I. NAME MANNER OF NARRATOR IN ARABIC TEXT

No.	Name Manners	Example
NM1	First Name	Safina
NM2	First Name + Father	Ismail Ibn Ibrahim
NM3	First name + Father +	Bakr Ibn Yahya Ibn Zaban
	Grandfather	
NM4	First Name + Father +	Abdul Rahman Ibn Ibrahim
	Family	Damashcky
NM5	First Name + Family	Khallad al-Saffar
NM6	First Name + Father nickname	Yazed Ibn Abu Zaid
NM7	First Name + Father nickname + Family	Said Ibn Abi Said Maqbari
NM8	First name + Father +	Abdullah Ibn Mohammed
11110	Grandfather + Fourth	Ibn
	Grandfather Nickname	Aqeel Ibn Abi Talib
NM9	Surname	Abu Zubair
NM10	Surname + Father	Umm Habiba bint Jahsh
NM11	Surname + Family	Abu Malik al-Ash'ari
NM12	Surname + Father	Abu Bakr Ibn Abi Shaybah
	Surname	
NM13	Ibn + Father Name	Ibn Umar
NM14	Ibn + Father Nickname	Ibn Abi Mary
NM15	Family Relation	His father, his grandfather,
		his
		master, his uncle, his brother,
		my
		father etc.
NM16	Luqab	Al Amash, Nubayshah al
		Khair
NM17	Family	Al Ouzaii

The Malay hadith texts as the data source were extracted from the Lidwa Pusaka website [3] manually. We copied every hadith in the website to individual text file for each of the hadiths. The website provides 62,000 hadiths that are

divided into 9 most popular hadith books. The software comes from Indonesia and, therefore, uses Bahasa Indonesia in the text. The reason we used the Lidwa Pusaka website was because there is no digitized Malay hadith texts available in Bahasa Malaysia with complete authentic narrator name chain (sanad). So far, there exists only the summary of authentic narrator name chain in digitized Malay hadith texts in Bahasa Malaysia [6], [5]. We only extracted the data from Shahih Bukhari's book only as it is the most authentic book in hadith [12].

TABLE II. AUTHENTIC NARRATOR'S NAME STRUCTURE IN MALAY HADITH TEXTS

Structure	Malay name
1	P bin/binti Q
2	P bin/binti Q bin/binti R
3	P bin/binti Q bin/binti R bin/binti S

TABLE III. METHODOLOGY TO IDENTIFY THE NAME MANNER IN MALAY HADITH TEXTS

Component of name manner	Explaination
First Name	Could be more than one word
Father	Appears after the first word "bin" in the name
Grandfather	Appears after the second word "bin" in the name
Fourth Grandfather Nickname	Appears after the third word "bin" in the name
Surname	 Starts with the word "Abu / Ummu / Abi". "Abu / Abi" has the same meaning which is 'the father-of' somebody. The different spelling indicates the
	different use of the word according to Arabic grammar. "Abu" is used to show the subject in a sentence. "Abi" is used as a word that requires that another word in the sentence. • "Ummu" means the 'mother-of' somebody.
Father Surname	 Appears after the first word "bin" in the name Starts with the word "Abu / Abi".
Ibn	Starts with the word "Ibnu"."Ibnu" means the 'son-of' somebody
Family Relation	Starts with the word "Bapaknya / Kakeknya / pamannya / bapakku / "
Family	Starts with the word "Al"Needed as family name list for Arabic
Luqab / Appelation	Starts with the word "Al"Needs to refer to the meaning of word

However, this research used only 150 of hadith texts to extract the narrators' names from the texts. The narrators' names of the hadith were also extracted manually. We copied each of the narrator names in the Malay hadith texts into a list of narrator's names. This research will identify the name manner of narrators in the Malay hadith texts. Then,

we developed the rule-based using Python programming [11] based on the name manner identifed for a single Malay hadith text. The overall step in this research is show in Figure 1.

III. RESULTS AND DISCUSSION

The extraction of narrators' names from the texts produced 455 names with different spelling. However, the data show that one person's name can be spelt differently. Table IV shows the different ways of spelling a person's name.

For the narrator's names 1 and 2, the difference is marked at the beginning of the name. Name 2 has apostrophe mark (') at the beginning of both first name and family name, unlike Name 1 which only has the apostrophe mark at the first name. However, both names refer to the same person. Such difference came about during the transliteration process. As a rule (linguistically), the character begins with apostrophe mark in 'Abdullah is used with the ξ (ain) character in Arabic, while Abdullah with the (alif) character in Arabic carries no apostrophe mark. However, sometimes, transliterators tend to ignore this linguistic rule. Names 3 and 4 also have different spelling in which Name 4 contains double 'r' character in the spelling of the name. The extraction was done from only one hadith book, the Shahih Bukhari's book, available in the same website. However, the data produced different spelling names for the same meaning [1]. This is because there is no standard transliteration for narrators' names in the Malay hadith texts.

TABLE IV. DIFFERENT SPELLING OF THE SAME PERSON'S NAME

Name No.	Extract form Malay
	hadith text
1	'Ubaidullah bin
	Abdullah
2	'Ubaidullah bin
	'Abdullah
3	Abdurrazaq
4	Abdurrazzaq

Table V shows the ambiguity in a person's name; Names 5 and 6 are spelt in the same name until the third word. Name 6 has another word, followed by his father's family name. This is similar to Names 7 and 8, where the latter also added the family name after the father's name. According to the expert, family or surname name manner is used to distinguish people with similar first or second name.

TABLE V. AMBIGUITY IN PERSON'S NAME WITH MANY FORMS

Name No.	Extract form Malay hadith text
5	Abu An Nu'man
6	Abu An Nu'man 'Arim bin Al Fadlal
7	Alqamah bin Waqash
8	Alqamah bin Waqash Al Laitsi

Based on [4], Table VI shows the name manner of narrators in the Malay hadith texts.

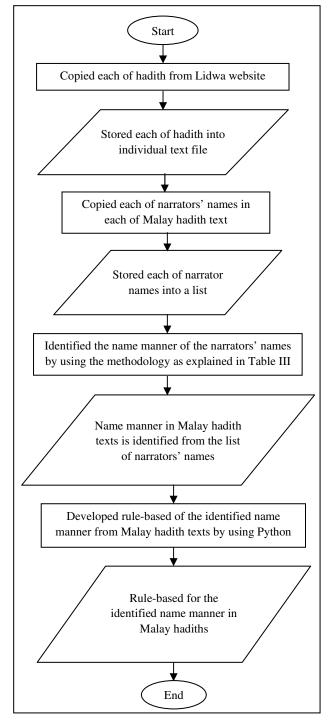


Fig. 1. Overall steps in this research

As shown in Table VI, in the Malay hadith texts, the word "Ibn" in Arabic is translated as "bin" or "Ibnu" to describe the son of somebody in Bahasa Malaysia. "Ibn" is translated to "bin" when it is used in the middle of the name and is translated to "Ibnu" when it is used at the beginning of the name. Ambiguity is also present between the name

manners for family and luqab/appelation because it begins with the word "Al". As shown in Table VI, the word "Al Aswad" is categorized as luqab/appelation while "Al Humaidi" is categorized as family name manner. To differentiate these words, we refer to the experts to determine the meaning between them. It is described that "Al-Aswad" means "Black" while "Al Humaidi" refers to family name.

Family relation name manner also poses another issue in terms of vocabulary used since the Malay hadith texts are written in Bahasa Indonesia. For example, the word "Kakek" in Bahasa Indonesia, which means "grandfather", should be translated to "datuk" in Bahasa Malaysia. Similarly, the word "paman" in Bahasa Indonesia, meaning "uncle", should be translated to "bapa saudara" in Bahasa Malaysia.

TABLE VI. NAME MANNER OF NARRATORS IN THE MALAY HADITH TEXTS

Name Manner	Malay Text
First Name	Aisyah
First Name + Father	Bisyir bin Muhammad
First name + Father +	Ishaq bin Abdullah bin Abu
Grandfather	Thalhah
First Name + Father +	Alqamah bin Waqash Al Laitsi
Family	
First Name + Family	Washil Al Ahdab
First Name + Father	Utsman bin Abu Syaibah
nickname	
First Name + Father	Sa'id bin Abu Sa'id Al Maqburi
nickname + Family	
Surname	Abu Az Zanad
Surname + Father	Abu Suhail bin Malik
Surname + Family	Abu Said Al Khudri
Ibn + Father Name	Ibnu Sirin
Ibn + Father Nickname	Ibnu Abu Bakrah
Family Relation	Bapaknya, Kakeknya, pamannya,
	Bapakku
Luqab / Appelation	Al Aswad
Family	Al Humaidi

In this study, we found additional three name manners in the Malay hadith texts as shown in Table VII.

TABLE VII. ADDITIONAL NAME MANNERS IN THE MALAY HADITH TEXTS

Name Manner	Malay Text
First Name + Father +	'Amru bin Yahya bin Sa'id bin
Grandfather + Fourth	'Amru Al Makki
Grandfather + Family	
Surname + Relation + Lord	Abu Murrah -mantan budak
Name + Father Lord Name	Uqail bin Abu Thalib
First Name + Relation + Lord	Nafi' mantan budak 'Abdullah
Name + Father Lord Name	bin 'Umar bin Al Khaththab

Table VII shows the presence of a name, "Abu Murrah - mantan budak Uqail bin Abu Thalib", as authentic narrator name in a Malay hadith text. The question is, which one of the narrator's names should we store for the hadith? Should it be "Abu Murrah" and "Uqail bin Abu Thalib" or "Abu Murrah -mantan budak Uqail bin Abu Thalib"? According

to the experts, the name "Abu Murrah" only is sufficient to be stored as narrator name for the hadith.

Then, we developed the rule-based in Python programming to recognize the narrators' names for a single Malay hadith text. In this paper, we are not able to show the full hadith text because of the space limitation. The full hadith text can be accessed from the Lidwa Pusaka website [3] in Sahih Bukahri Book based on the number of hadith shown in Table VIII. Table VIII also shows the rule-based and outputs.

Rules No. 1 and No. 2 in Table VIII used a part of the following Malay hadith text as an input:

Telah menceritakan kepada kami 'Abdullah bin Yusuf berkata, telah mengabarkan kepada kami Malik dari Ishaq bin 'Abdullah bin Abu Thalhah dari Anas bin Malik berkata,...

TABLE VIII. RULE-BASED AND OUTPUTS

Rule Rule-based Hadith	Output
No. No.	output
	Name2': 'bin',
	Jame3': 'Malik',
	Jame1': 'Anas'}
1 2 3 7 1	nas
Name2>%s)\s bi	in
	Ialik
À-Z]{1}[a-	
z]{2,})" % b	
2. a=""" 164 {"	Name1':
r"(?P <name1 ""<="" th=""><th>Abdullah"}</th></name1>	Abdullah"}
>%s[A- 'A	Abdullah
Z]{1}[a-	
z]{2,})" % a	
3. a=""" 212 {"	Name2': 'bin',
	Jame3':
	sma'il",
>[A-Z]{1}[a- 'N	lame1': 'Musa'}
z]{2,})\s(?P<	Iusa
Name2>%s)\s bi	in
	ma'il
A-Z]{1}[a-	
z]{2,}%s[a-	
z]{2,})" %	
(b,a)	
	Name2': 'Ath',
	Jame3': 'Thawil',
3(/// (lame1':
1 - 1	Iumaid'}
	umaid
3(7)/ (th hawil
Z]{1}[a-	nawii
$\begin{bmatrix} z_{j}(1)[a^2]\\ z_{j}(2,)\end{bmatrix}$ "	
	Name2': 'Al',
	Jame3':
	Iubarak',
	Vame1': 'Ibnu'}
	nu
A-Z]{1}[a- A	
z]{2,})" % a M	Iubarak

Rule No. 1 is to identify narrator name with *First Name* + *Father* in the name manner with only 1 word for *First Name* and *Father*. Table IX explains the details of the rule.

Until this paper is written, we found that character '(aspostrophe) was not recognized as a part of narrator's name. Therefore, Rule No. 2 is developed to recognize the '(aspostrophe) in front of the narrator's name. Rule No. 3 is developed to recognize the '(aspostrophe) in the middle of a narrator's name beginning with index 4 in a word. We still need to explore the '(aspostrophe) to be recognized as a part of narrator's name where it exists.

Rules No. 4 and No. 5 in Table IX used a part of the following Malay hadith text as an input:

Telah menceritakan kepada kami Nu'aim berkata, telah menceritakan kepada kami Ibnu Al Mubarak dari Humaid Ath Thawil dari Anas bin Malik berkata, "Rasulullah shallallahu 'alaihi wasallam bersabda:...

Rule No. 4 is developed to recognize the *First Name* + *Family* name manner. When we began developing Rule No. 4, we wished that the rule would recognize the *Ibnu Al Mubarak* name. However, instead, it recognized *Humaid Ath Thawil* name in the text. The Name2 pattern in Rule No.4 did not recognize a name that only has 2 characters such as *Al*. Therefore, we developed Rule No.5 to solve the problem.

TABLE IX. EXPLANATION OF RULE NO.1

Statement	Explanation
b='bin'	A variable named b with value bin
r"(?P <name1>[A-Z]{1}[a-z]{2,})</name1>	Name1 pattern begins with a capital letter for index 1 and the following letter begins with a small letter with index 2 until the end
\s	White space
(?P <name2>%s)</name2>	Name2 pattern is a string
\s	White space
(?P <name3>[A- Z]{1}[a-z]{2,})"</name3>	Name1 Pattern begins with a capital letter for index 1 and the following letter begins with a small letter with index 2 until the end
% b	Variable b refers to string that declares in the Name2 pattern

IV. CONCLUSION

This research has shown the name manner for authentic narrators' names in the Malay hadith texts. The narrators' names were transliterated from the Arabic texts to Malay texts in Bahasa Indonesia. We extracted manually 455 narrator's name from 150 Malay hadith texts. From the extracted name, exists same person of narrator's name with different types of spelling. There are also many forms of the narrator's names which is refers to one person. We also identified the name manner adapted from [4] and added 3 more name manners that we found in the Malay hadith texts. Then, we developed the rule-based in Phyton regarding the name manner that we had identified. This is the first research contribute to identify person name entity in specific domain which is Malay hadith text. Next, we will combine all the rule-based to recognize name manner for authentic narators in the Malay hadith texts and develop the indexer for authentic narrator name for the hadith texts in the Bahasa Malaysia. For the future work, we will develop an algorithm to recognize the authentic narrator chain in the Malay hadith texts and evaluate the algorithm.

REFERENCES

- Khaled Shaalan. A Survey of Arabic Named EntityRecognition and Classification. 2014 Association for Computational Linguistics, 2014.
- [2] Lei, G., Zhou, M. L., Yao, J. M., & Zhu, Q. M. A supervised method for transliterated person name identification. In 2nd International Symposium on Electronic Commerce and Security, ISECS 2009 vol. 2, pp. 117–120). doi:10.1109/ISECS.2009.212, 2009.
- [3] Perusahaan Software Indonesia dan SIMRS Saltanera. Software Hadits/Hadis/Hadist Online Indonesia. http://www.lidwa.com/app/?k=abudaud&n=1, 2015.
- [4] Yehya M. Dalloul. An Ontology-Based Approach to Support the Process of Judging Hadith Isnad. Islamic University of Gaza. Master Thesis, 2013.
- [5] Nurazzah Abd Rahman, Zainab Abu Bakar & Tengku Mohd Tengku Sembok. Query expansion using thesaurus in improving Malay Hadith retrieval system. 2010 International Symposium in Information Technology (ITSim), vol. 3, pg: 1404 – 1409, IEEE, 2010.
- [6] Normaly Kamal Ismail, Nurazzah Abdul Rahman & Zainab Abu Bakar. Mutiara Hadis. http://sigir.uitm.edu.my/webhadis/, 2007.
- [7] Mohd Yunus Sharum, Muhamad Taufik Abdullah, Md Nasir Sulaiman, Masrah Azrifah Azmi Murad, Zaitul Azma Zainon Hamzah. Name Extraction for Unstructured Malay Text. IEEE Symposium on Computers & Informatics, IEEE, 2011.
- [8] Cihan Varol & Coskun Bayrak. Hybrid Matching Algorithm for Personal Names. ACM Journal of Data and Information Quality, Vol. 3, No. 4, Article 8, 2012.
- [9] Hesham H. Abdel Ghafour, Ali EI-Bastawissy & Abdel Fattah A. Heggazy. AEDA: Arabic Edit Distance Algorithm, Towards a New Approcah for Arabic Name Matching. IEEE, 2011.
- [10] Xiaoting Zhu & Zhao Lu. A Novel Ontology-based Approach of Chinese Person Name Disambiguation. 2012 Eighth International Conference on Semantics, Knowledge and Grids. IEEE, 2012.
- [11] Zed A. Shaw. Learn Python the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code, Third Edition. Addison-Wesley. United States of America, 2014
- [12] Ahmad Sanusi bin Azmi, Zulhilmi bin Mohamed Nor & Mohd Norzi bin Nasir. Kekeliruan Dan Kritikan Terhadap Sahih Bukhari: Satu Tinjauan Awal. E-Jurnal Penyelidikan dan Inovasi, Pusat Pembangunan Penyelidikan, Kolej Universiti Islam Antarabangsa Selangor, 2013.
- [13] Normaly Kamal Ismail, Nur Hamizah Mat Saad, Sidi Bukhari Sidi Omar Tengku Mohammad Tengku Sembok. 2D Visualization of Terms and Documents in Malay Language. 5th International Conference on Information and Communication Technology for the Muslim World. IEEE, 2013.
- [14] A. Rayner, L.L. Chin, O.C. Kim, and A. Patricia. Malay Named Entity Recognition Based on Rule-Based Approach. International Journal of Machine Learning and Computing, Vol. 4, No.3, June 2014.
- [15] A. Rayner, L.L. Ching, O.C. Kim, A. Patricia, F.T. Soo, N. Mohd, R.B.H. Mohd, and H. Ahmad. A Rule-Based Named-Entity Recognition for Malay Articles. Advanced Data Mining and Applications, Lecture Notes in Computer Science, vol. 8346, pp. 288-299, 2013.
- [16] M. Alireza, A.L. Suriani, and M. Ali. Named Entity Recognition Approaches. International Journal of Computer Science, vol. 8, pp. 339-344, 2008.
- [17] N.M.Z. Ahmad, B.A. Abu, S.S. Mariyam, and G.T. Abd. Using linguistic patterns in FCA-based approach for automatic acquisition of taxonomies from Malay text. Proceedings - International Symposium on Information Technology, ITSim, vol. 2, 2008.
- [18] A.B. Juhaida, O. Khairuddin, N.M. Faidzul, and M.M. Zamri. Part-of-Speech for Old Malay Manuscript Corpus: A Review. Journal of

- Communications in Computer and Information Science, vol. 378 CCIS, pp. 53-66, 2013.
- [19] M. Aboaogo, and M.J.A.Aziz. Arabic Person Names Recognition by Using a Rule Based Approach. Journal of Computer Science 9, vol. 7, pp:922-927, 2013.