https://doi.org/10.17576/apjitm-2022-1102-09
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 11 No. 2 December 2022: 112 - 122
e-ISSN: 2289-2192

Received: 31 March 2022
Accepted: 25 October 2022
Published: 1 December 2022

NAMED ENTITY RECOGNITION FOR QURANIC TEXT USING RULE BASED APPROACHES

SHASHA ARZILA TARMIZI
SAIDAH SAAD

ABSTRACT

The variety and difference between domains for textual data require customization in the Natural Language Processing component especially in Named Entity Recognition where different domains contain several types of entities. The current NER model is deemed not fit to accurately extract entities from Quranic text due to its unique content. This paper describes the building of a rule-based Named Entity Recognition method to extract the entities that exist in the English translation to the meaning of the Quranic text and its performance evaluation. Named entity tagging, a common task in-text annotation, in which entities (nouns) in the unstructured text are identified and assigned a class. A few rules are built to extract several types of entities such as the name of prophets and people, creation, location, time, and the various names of God. The rules are built mainly using regular expressions and gazetteers. The rules that have been built result in high precision and recall as well as a satisfactory F-score of over 90%. The results from this experiment can be used as annotation in building a machine learning model to extract entities from the same type of domain specifically on the Quranic text or generally in the Islamic domain text.

Keywords: Named Entity Recognition (NER); rule-based method; Quranic text

INTRODUCTION

The Quran contains 114 surahs and 6236 verses in total. It is a holy book in the religion of Islam as it contains various information and stories that serve as guidance for Muslims. These guidance and reminders are often told in the form of narratives thus there are multiple entities that can be found in the Quran. However, it requires a lot of time and energy to manually extract these entities due to the huge number of verses in the Quran. Natural Language Processing (NLP) is a field of data science that helps solve these problems by inventing a method to quickly extract these entities. The method is widely known as Named Entity Recognition (NER). Due to the uniqueness and sanctity of this Holy book and since the original language of this book is in Arabic, the usual NER method cannot be used to accurately extract the entities.

NER is a critical component in automated text processing where it tries to identify phrases referring to concepts of interests in unstructured text. The NER technology in the general domain, which focuses on detecting concept mentions in the text. NER is used to extract and mine the predefined semantic of information from textual data and responsible for recognizing concepts relevant to interpreting the meaning of texts and classifying them according to a set of categories. The main purpose of extracting entities from textual data is to single out the main topics of the data. It can also help summaries the content of a certain text or article. The entities found in the data can give the main idea of what the texts or articles are about. However, NER need to be built and customized to meet the requirement of the text-domain especially in Islamic domain which have the special concept and have their own meaning.

NER as mentioned is an essential component of NLP and functions by carrying out two major tasks. The first task is identifying entities that exist in text and later classifying them into the respective type of entities (Mansouri, A et al., 2018). The common type of entities extracted includes the name of a person, the name of the organization, and the name of places in English. These types of entities are too generic and work well only for general articles in English. Therefore, many studys and research have been carried out to extend and modify the usual NER method to extract even more types of entities and in texts of different languages. For example, modifications are made to the basic NER method for texts in different languages such as Arabic (Antoun W. et al, 2020) (Helwe, C., & Elbassuoni, S., 2019), Chinese (Li X. et al, 2020) (Jia C. et al, 2020) and even Malay (Nadia, U. & Omar, N., 2019). Modifications are also made for texts in various domains. Domain that widely uses NER includes biomedical (Maan Tareq Abd, 2017) and medical domain, for example to extract names of diseases (Zhang T. et al, 2020).

This paper focuses on the extraction from the English translation of the Quranic text using rule-based NER method. This NER method requires manually crafted rules to carry out the task. Rules used for NER task are usually built from grammatical patterns, syntactic features, orthographic features along with hand-made gazetteers (Budi, I., & Bressan, S., 2003). The rules used in this study are mainly made by patterns using regular expression or RegEx, and gazetteers. The NER model is built to extract eight types of entities that will be elaborated in the methodology section. The dataset that is used for the documentation of this paper is the 112 verses of Surah Al-Anbiya. The English Quranic translation that is used for this study is from a well-known Quran translation by Dr. Muhammad Taqiuddin Al-Hilali and Dr Muhammad Muhsin Khan, also known as the Hilali-Khan Quran translation. The results of the extraction are then sent to domain experts in the Quranic field to be verified. The verified results are then analysed for future use of annotation in the building of a machine learning NER model and as a gold standard to compare with, for evaluating the performance of the approaches being done.

RELATED WORKS

There are a lot of past works in the field of NLP specifically in using NER to mine information from textual data since these data come from different domain that requires a more specific and specialized NER method. However, there are only ample amount of NER works for the Quranic domain and even for the Islamic domain in general.

Most of the NER related works in the Quran domain are done in languages other than English. A past study for NER focused in extracting the entity person has been done but it is carried out using the Quranic translation in Indonesian language (M.A.K et al., 2019). The NER method used in this study is a machine learning model using the Hidden Markov Model (HMM) resulting to an F-score of 0.76. Another similar NER work is done but using the Conditional Random Field machine learning model (Arvianto et al., 2019) results to an F-score of 0.77, a slight difference compared to the previous work mentioned. However, these two works are focused in extracting only one type of entity which is the person entity.

There is also a past work on the English Quranic translation using a supervised machine learning method named Support Vector Machine (SVM) (Maulana M. et al, 2020). The work focuses on different type of kernels for the SVM model which are linear, sigmoid, Radial Basis Function (RBF) and polynomial. According to the result of the experiment, the best score among all the different type of kernels is using the linear kernel with an F-score of 0.75. However, this work also focuses in extracting one type of entity which is the person entity. This study also uses a different translation of Quran by Saheeh International. This translation

does not include the Arabic names of the prophets and person and does not use the Arabic terms for the name of group of people like 'Zalimun' and 'Muttaqun'. Thus, the dataset used may not reflect the entities used in other Islamic literary text.

Another related NER work that is used as reference in this study is a work in extracting the entity name of Allah from the Quran using rule-based method (Hassan M.M. et al, 2018). Three different features are used in building the rules which are orthography, N-grams, and affixes. This work is done using the original Quranic text in Arabic but the building of the rules in this work helps serve as guidance for this study because of the extraction of the same type of entity. This work does not publish any statistical result of the performance of NER model, but an analysis of the names extracted are made through this study.

A hybrid NER model comprising of supervised machine learning and rule-based method is built to create a classification system on Islamic hadith text (Hasan & Rassem, 2018). This hybrid model is built using a Support Vector Machine (SVM) and pattern matching. The pattern matching is done on the type of questions since the model is used to build a Question Answering System (QAS). The NER model is used to extract person and location entities from Islamic text. The NER model gives a satisfactory result in which the precision is 88.39%, a recall score of 87.66% and F-score of 87.93%

Another related NER work is research on extracting entities using different classifiers from an English translation of hadith documents (Jaber & Saad, 2016). This work uses different type of classifiers, namely Support Vector Machine (SVM), Naïve Bayes (NB) and Max Entropy, to extract six type of entities which are person, location, organization, money, date, and time. This study also built an NER model using the combination of the classifiers mentioned and the performances of all the classifiers are later compared to determine which classifier works best for an Islamic domain. The result of the research shows that the best NER model is the combination of all the classifiers which results in a precision of 96.9%, a recall percentage of 93.6 and F-score of 95.3% which indicates an exceptional performance in extracting the entities from hadith documents.

The study on the past works related work of NER in the Quranic domain shows that a lot of improvements can be made to achieve a more satisfactory method in extracting entities from the Quran. Most of the NER method that has been developed does not pass an F-score of over 0.8. Besides, there are not many past studies that work on extracting different types of entities when the Quran is filled with various types of interesting entities. The extraction of different types of entities can open more opportunity for future works as well as providing a more comprehensive materials for the building of ontology of Quran. The past works also helped in identifying the problems that might occur in the building of the NER model as well as providing the direction for this study.

MATERIAL AND METHODS

This study is done in several phases, starting from the pre-processing phase, rules building phase, extraction phase and ends with the evaluation phase. The dataset chosen for this work is all 112 verses of Surah Al-Anbiya, the 21st chapter of the Quran, using the English translation of the Quran by Hilali-Khan. This study is done using the NLP library, SpaCy and is fully written in the Python language. There are a few steps involved in the completion of this study. Figure 2 below shows the flow diagram of the study, showing the general framework of the NER model. The detail of each phase that corresponds to the diagram will be explained later in this section.

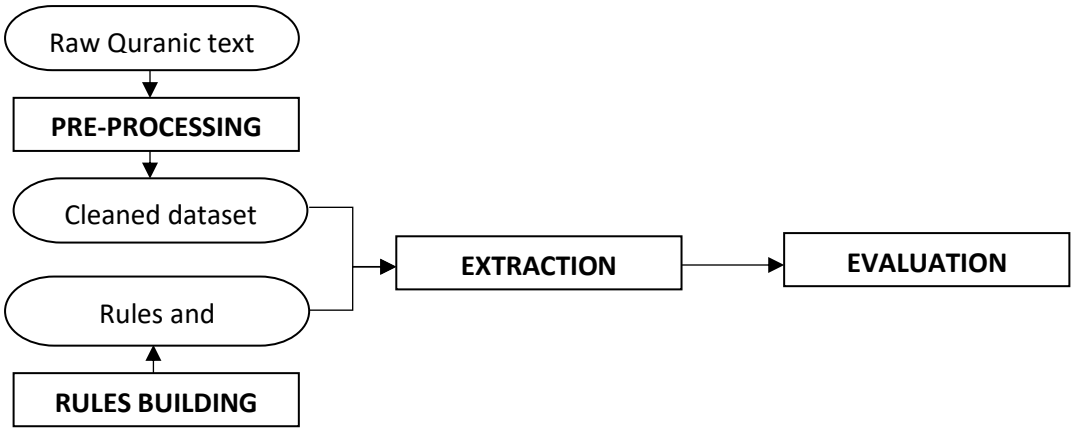

FIGURE 1. Diagram of the flow of NER process

PRE-PROCESSING PHASE

The first step in every NLP study is the pre-processing step. This step is important to remove noisy data thus ensuring the dataset is efficient enough to be used in the study. A good dataset will result to a better performance since irrelevant and unnecessary data are discarded. There are a few steps done in this pre-processing to output a clean dataset. Figure 3 shows how the dataset look like before it is cleaned.

11. How many a town (community) given to wrong-doing, have We destroyed, and raised up after them another people!

12. Then, when they perceived (saw) Our Torment (coming), behold, they (tried to) flee from it.

13. Flee not, but return to that wherein you lived a luxurious life, and to your homes, in order that you may be questioned.

14. They cried: "Woe to us! Certainly we have been Zâlimûn (polytheists, wrong-doers and disbelievers in the Oneness of Allâh)."

15. And that cry of theirs ceased not, till We made them as a field that is reaped, extinct (dead).

16. We created not the heavens and the earth and all that is between them for a (mere) play.[1]

17. Had We intended to take a pastime (i.e. a wife or a son), We could surely have taken it from Us, if We were going to do (that).

FIGURE 2. Example of Raw Dataset

Firstly, the digits that appeared in the dataset are removed since the NER model would not extract digits or numerals. Furthermore, the removal of digits would not alter the context and meaning of the Quranic text because the translation chosen for this study uses words instead of numerals for cardinality as shown in Figure 4. The digits that appeared in the dataset are only used to denote the number of verses.

43. And the king (of Egypt) said: "Verily, I saw (in a dream) seven fat cows,
whom seven lean ones were devouring,
and seven green ears of corn, and (seven) others dry. O notables! Explain to me my dream,
if it be that you can interpret dreams."

FIGURE 3. Example of Cardinals in The Dataset

After that, unnecessary characters, or symbols such as punctuations are also removed from the dataset. Some examples of the punctuations are semi colon, exclamation mark and question mark. Characters like parentheses are reserved for another pre-processing step that is carried out during the last step of this pre-processing phase. Next, co-referencing is done on

terms like ‘He’ and ‘We’ that refers to the entity Allah by replacing those terms with ‘Allah’. This step is done to annotate that the word ‘Allah’ is meant in the verse.

FIGURE 4. Example of Phrase with Co-Referencing

Figure 5 shows the phrase that is involved with co-referencing. The word ‘We’ in the phrase refers to Allah, thus it will be changed to *“…and spend out of what Allah have provided for them…”*. This step is necessary since the entity ‘Allah’ is one of the significant entities that must be extracted.

The final step in the pre-processing is the removal of phrases in brackets and parentheses. These phrases are just extended explanation of the verse, or reference for the verse thus it is not relevant to the actual dataset because it is not mentioned in the original Quranic verse. Figure 6 shows the example of phrase containing parentheses.

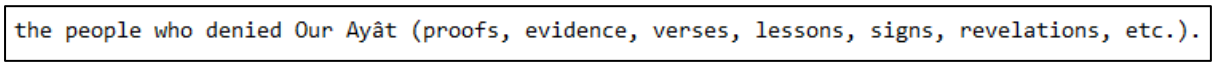

FIGURE 5. Example of Phrase with Parentheses

The phrase in Figure 5 shows that the sentence in parentheses is referring to ‘Our Ayat’. Thus, it is just an explanation to what ‘Our Ayat’ means and the sentence does not exist in the actual Quranic text, so it is discarded from the dataset as it is necessary and may bring inaccurate result in this study. Figure 7 shows the dataset after the pre-processing steps have been carried out.

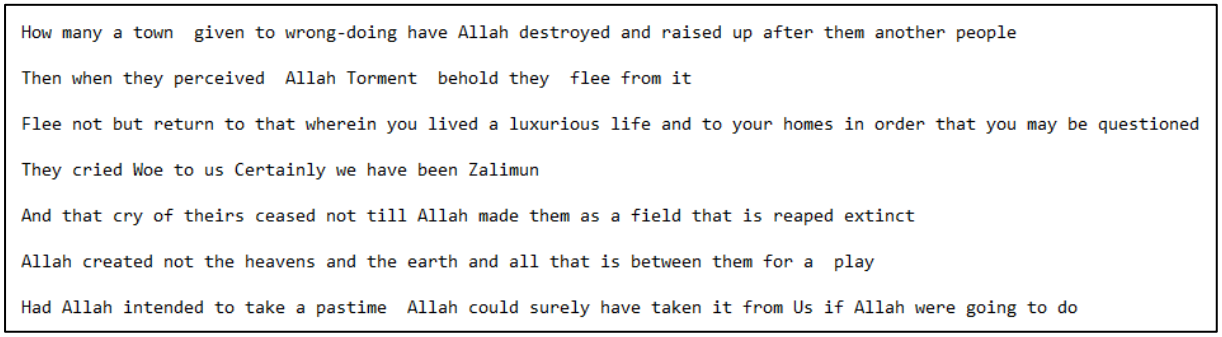

FIGURE 6. Example of Cleaned Dataset

RULES BUILDING PHASE

This phase is crucial as the rules that has been built would determine the performance of the method. The rules for this study are categorized into two types, which are using gazetteers and patterns that are extracted with RegEx (regular expression). This study will extract eight types of entity and different rules are built for each type of entity. The entities extracted are the entities that commonly appeared in the Quran. Table 1 shows the entities extracted along with the rule used for each entity.

TABLE 1. Types of Entities Extracted

Entity Tag	Type of Entity	Example	Type of Rule
PER	Individual, person	Maryam, Qarun	Using gazetteers
PER-PROPHET	Prophets	Adam, Musa	Using gazetteers
PER-GROUP	Name of a group of people	Muhajirun, Children of Adam	Using gazetteers and patterns
LOC-AFTERLIFE	Places in afterlife	Firdaus, Hell	Using gazetteers
LOC	Name of the city, historical places, and mountains	Babylon, Makkah, Egypt, Mount Sinai	Using gazetteers
GOD	Name of God	Allah, All-Knowing	Using patterns
CREATION	Creations such as angels, jinns	Jibril, Iblis	Using gazetteers
STIME	The time that refers to Qiamat	Last Day, Day of Recompense	Using patterns

The gazetteers contain a specific list of names of the entities. The gazetteers were built by using references from the Quran corpus built by the Language Research Group from the University of Leeds. The gazetteers are also modified to fit the dataset in terms of spelling. The building of rules using pattern method takes into account the similarities between the possible entities. There are three entities that use the pattern method which are PER-GROUP, NAME_OF_GOD and STIME. The pattern that is used for the entity 'PER-GROUP' is the suffix '-un' because the translation uses Arabic words like 'Zalimun' and 'Musrifun' according to its original Arabic word structure. The entity 'NAME_OF_GOD' also uses the pattern method because the Quranic translation uses prefixes such as 'All-' in 'All-Knowing' and 'Oft-' in 'Oft-Forgiving'. Another pattern method used is to extract 'STIME' that relies on the use of the word 'Day' in noun phrases that is used to refer to the term Qiamat.

EXTRACTION PHASE

After the completion of the pre-processing of dataset and the building of the rules, the crucial step which is the execution of NER. The extraction is carried out using the specific rules for each entity created in the first step with the help of the SpaCy library. The execution outputs the type of the entity along with the name of the entity itself. Figure 5 shows the example of the output from the execution of rule-based NER method.

```
And Allah bestowed upon him Ishaq  and  Yaqub  Each one Allah made righteous
(GOD,  Allah )
(GOD,  Allah )
(PER-PROPHET, Ishaq )
(PER-PROPHET, Yaqub )
```

FIGURE 7. Example of NER Output

EVALUATION PHASE

Before the NER model is evaluated, the result is sent to domain expert who is knowledgeable in the Quran field for validation. This is because the Quran is a Holy Book for Muslims, and the purity of the content should be preserved and monitored by experts. The verified result is then analysed to evaluate the NER model using the precision, recall, and F-

score method that derives from the confusion table (Goutte, C., & Gaussier, E., 2005). The extracted entities are categorised into true positive, true negative, false positive and false negative to calculate the scores. Precision is defined by the performance of the method in extracting and labelling the entities correctly among all the extracted entities. The precision is calculated using the following equation:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

The recall score calculates the correctly extracted entities among all the entities that should be extracted. This score can be represented using the following equation:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The true positive from this experiment refers to the correctly labelled entities, the false positive refers to the incorrectly labelled entities meanwhile the false negative refers to the entities that are not extracted by the NER method. The overall performance for each entity is then calculated using F-score equation as the following:

$$F\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

RESULTS AND DISCUSSION

Before the study is carried out, a preliminary experiment is done using the pre-existing NER library named SpaCy. This experiment is done on the exact same dataset that is used for the rule-based NER method. An overall F-score of only 0.61 is obtained from this preliminary experiment which is not satisfactory. Table 2 shows the preliminary experiment result obtained by SpaCy library focusing on person, nationalities, date, time, location, and cardinals.

TABLE 2. Preliminary result of NER on Quranic text using SpaCy NER

Type of Entity	Precision	Recall	F-Score
PERSON	0.43	0.84	0.57
NORP	0.20	1.00	0.33
DATE	1.00	1.00	1.00
TIME	1.00	1.00	1.00
LOC	1.00	0.45	0.62
CARDINAL	1.00	1.00	1.00

This SpaCy NER model does not accurately recognize crucial entity such as Allah. This existing model also has a low performance in extracting the person entity. This is a huge issue since the Quran contains a lot of person entities considering its history content and these entities must be correctly classified to accurately analyse the content of a specific chapter. This model is also unable to recognize the locations mentioned in the Quran. The SpaCy NER model is also unable to extract the NORP or the nationality and religious group entity correctly. This is mainly because the NORP entities that exist in the Quran are not similar to the entities that can be found in the general text. The result of this preliminary experiment serves as strong evidence to show that some customization should be made to cater to the different entities that exist in this Quranic text.

The result for the rule-based NER model is verified by the domain expert and has been classified into True Positive, False Positive, and False Negative in order to calculate the precision, recall, and F-score values as explained in the evaluation phase section. Table 3 shows the performance of the NER model in extracting each entity according to its precision, recall and F-scores.

TABLE 3. Evaluation Based on Entities

Type of Entity	Precision	Recall	F-Score
PER-PROPHET	1.0	1.0	1.0
PER-GROUP	0.91	1.0	0.95
LOC-AFTERLIFE	1.0	0.25	0.4
GOD	0.93	0.92	0.92
STIME	1.0	1.0	1.0

From this evaluation, the entity PER-PROPHET and STIME have the best performance. This is because the entity PER-PROPHET uses a complete gazetteer containing different spellings of the Prophets' name to make sure that no entity is left out. The entity STIME has a great performance because there are only two entities found in the dataset. Different result may be obtained using different dataset containing more of this entity. The entity PER-GROUP has a better recall score compared to precision because there are a few entities that are wrongly labelled but there is no entity that is unrecognized. For example, the pattern that is used to extract the entity 'PER-GROUP' incorrectly classified the word 'borders' as a name of a group. This issue can cause a bigger problem if assessed in a dataset that contains a lot of 'PER-GROUP' entities and should be improved for future work. The entity LOC-AFTERLIFE has good precision because there is no wrongly labelled entity. However, the NER model did not recognize the entity 'heaven' which belongs to the LOC-AFTERLIFE category and this word appears a few times in the dataset, resulting to a low recall score. For the entity GOD, some errors in co-referencing results in the shown precision and recall score because some entities are left out, meanwhile the others are wrongly labelled. The result also concludes that the building of good gazetteers will improve the recall score of the NER model. Therefore, this entity-based evaluation helps in improving the rules for each entity because the problems that lead to such performance can be analysed.

Table 4 shows the overall scores of the precision, recall, and F-score of this experiment, as well as providing insight of the performance of the model in general.

TABLE 4. Rule-Based NER Method Evaluation Scores

Type of Evaluation	Value
Precision	0.98
Recall	0.91
F-Score	0.94

The scores of the evaluation are over ninety percent which is considered incredibly good. The range of this result is expected because the rules are manually crafted to fit the name of entities with high accuracy. The extraction for the entity type that uses gazetteers shows a remarkably high accuracy except for the entity location because it fails to classify the entity 'heaven' and 'earth' that appear multiple times in the dataset. The entities extracted using RegEx show a lower accuracy compared to the entities extracted using gazetteers since the patterns built are not specific enough and results in wrongly labelled entities. Despite the satisfactory performance of this NER method, based on the verification and comments from the domain expert, there are some issues that occur during the pre-processing part related to the dataset. The co-referencing for the term Allah has changed the meaning and context of the Quranic text which may cause some of the extractions to be invalid. The future work of this study may include the building of a good co-referencing method for this domain.

CONCLUSION

The rule-based NER method described in this paper successfully extracted entities from the 112 verses of Surah Al-Anbiya with a very high F-score of 0.94 thus showing that the rules built are accurate and valid except for some of the issues mentioned in the result section. The analysis from this experiment can serve as a guide for the annotation and building of the future machine learning NER method for Quranic text.

ACKNOWLEDGEMENT

This research study was financially supported by the Fundamental Research Grant Scheme (FRGS/1/2019/ICT02/UKM/02/2) provided by the Ministry of Higher Education, Malaysia.

REFERENCES

Antoun, W., Baly, F., & Hajj, H. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 9-15.

Budi, I., & Bressan, S. 2003. Association rules mining for name entity recognition. In *Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003*, 325-328.

Farhan Dzaky Arvianto, Moch Arif Bijaksana, Arief Fatchul Huda. 2019. People Entity Recognition in Indonesian Quran Translation with Conditional Random Field Approach. 2019 7th International Conference on Information and Technology (ICoICT), 1-5.

Goutte, C., & Gaussier, E. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European conference on information retrieval*, 345-359.

Hasan, A. M., & Rassem, T. H. 2018. Combined support vector machine and pattern matching for arabic islamic hadith question classification system. In *International Conference of Reliable Information and Communication Technology*, 278-290.

Hassan, M.M., Dhamyaa Al-Nasrawi, Redha J. Hassan, Noor T. Mahdi. 2018. Rule Based Method of Named Entity Recognition for Matching Allah's Finest Names in Holy Quran. Journal of Engineering and Applied Sciences 13, no.10, 3618-3623.

Helwe, C., & Elbassuoni, S. 2019. Arabic named entity recognition via deep co-learning. *Artificial Intelligence Review*, *52*(1), 197-215.

Jaber, M. J., & Saad, S. 2016. NER in English translation of hadith documents using classifiers combination. *Journal of Theoretical and Applied Information Technology*, *84*(3), 348.

Jia, C., Shi, Y., Yang, Q., & Zhang, Y. 2020. Entity enhanced BERT pre-training for Chinese NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6384-6396.

Li, X., Yan, H., Qiu, X., & Huang, X. 2020. FLAT: Chinese NER using flat-lattice transformer. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6836–6842.

M.A.K., R.M. & Bijaksana, Moch & Huda, Arief. 2019. Person Entity Recognition for the Indonesian Qur'an Translation with the Approach Hidden Markov Model-Viterbi. Procedia Computer Science. 157, 214-220.

Maan Tareq Abd. 2017. Extended Distributed Prototypical for Biomedical Named Entity Recognition. *Asia-Pacific Journal of Information Technology & Multimedia*, 1-10.

Mansouri, A., Affendey, L. S., & Mamat, A. 2008. Named entity recognition approaches. *International Journal of Computer Science and Network Security*, *8*(2), 339-344.

Maulana M. A., Moch Arif Bijaksana, Arief Fatchul Huda. 2019. Entity Recognition for Quran English Version with Supervised Learning Approach. Indonesia Journal on Computing (Indo-JC) 4.3:77-86.

Nadia, U., & Omar, N. 2019. Malay Named Entity Recognition Using Rule Based Approach. *Asia-Pacific Journal of Information Technology & Multimedia*, 1-10.

Zhang, T., Yu, Y., Mei, J., Tang, Z., Zhang, X., & Li, S. 2020. Unlocking the power of deep pico extraction: Step-wise medical ner identification. *arXiv preprint arXiv:2005.06601*.

Shasha Arzila Tarmizi
Saidah Saad
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
a171183@siswa.ukm.edu.my, saidah@ukm.edu.my

APPENDIX

Gazetteer for name of person:

```
{
    "prophets":
        ["Adam", "Idris", "Nuh", "Hud", "Salih", "Ibrahim", "Lut", "Ismail", "
            Ishaq", "Yaqub", "Yusuf", "Ayyub", "Shuaib", "Musa",
            "Harun", "Dhul-Kifl", "Dawud", "Sulaiman", "Iliyas", "Alyasa", "
                Yunus", "Zakariyya", "Yahya", "Isa", "Muhammad", "Ahmad", "
                Messiah"],

    "angel": ["angel of death", "Harut", "Marut", "Jibril", "Malik", "Mikail", "
        Angel of Death", "Kiraman Katibin", "guards of Hell"],

    "group": ["Children of Adam", "Ad", "Ansar", "Muhajirun", "bedouins", "
        Children of Israel", "people of the Cave",
        "owners of the Elephant", "people of the Ditch", "dwellers of Rass", "
            Dwellers in the Wood", "Yajuj and Majuj", "Madyan",
        "dwellers of Al-Hijr", "Quraish", "Romans", "Thamud", "Tubba"],

    "individual": ["Dhul-Qarnain", "Haman", "Imran", "Jalut", "Talut", "Luqman",
         "Maryam", "Firaun", "Qarun", "As-Samiri", "Uzair",
        "wife of Al-Aziz"],

    "jinn": ["jinn", "Iblis", "Shaitan"]

}
```

Gazetteer for the location entity:

```
{
    "afterlife": ["Firdaus", "Garden of Delights", "Garden of Delight", "Adn", "
        Hell", "Sidrat-ul-Muntaha", "Paradise of Abode",
         "Paradise", "Gardens", "Rivers", "Salsabil", "Sijjin", "Zaqqum"],

    "city": ["Babylon", "Badr", "Bakkah", "Hunain", "Iram", "Makkah", "
        Al-Madinah", "Yathrib"],

    "historic": ["Al-Ahqaf", "Egypt", "Saba"],

    "mountain": ["Judi", "Al-Marwah", "As-Safa", "Arafat", "Mount Sinai", "
        Al-Hijr"]

}
```