e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 4 No. 2, Desember, 2020, **Hal. 68-74**
DOI: 10.29408/edumatic.v4i2.2551

Name Disambiguation Analysis Using the Word Sense Disambiguation Method in Hadith

Ageng Prasetio[1], Mochammad Arif Bijaksana[2], Arie Ardiyanti Suryani[3]
[1,2,3]Department of Informatics Engineering, Universitas Telkom
email: agengprs@student.telkomuniversity.ac.id[1], arifbijaksana@telkomunivesity.ac.id[2], ardiyanti@telkomunivesity.ac.id[3]

(Received: 8 September 2020/ Accepted: 18 November 2020 / Published Online: 20 Desember 2020)

Abstrak

Name disambiguation is the problem solving process to find similar names in sentences. The ambiguity of names can be found in hadith of Sahih Bukhari, names "Abdullah bin Amru" in hadiths no 27 and “Abdullah bin Amru” in hadith no 58, These names are the same, but there is no proof they are the same person. This problem is the early indication of ambiguity of name in the hadith. Based in this problem, this research aims to find name disambiguation of hadith narrators with classification by considering the perawi chain. To solved this problem the authors used Word Sense Disambiguation (WSD), WSD is a process to assign the same meaning from the sentences, based on the context in which the word appears. To classify several names in the hadith, the authors used KNN algorithm, by combining the WSD and KNN method can reduce the ambiguity of names in hadith. The data used in this study came from the hadith of Sahih Bukhori through the pre-processing stage. After conducting the research showed a collection of hadith numbers with the same name prediction with an accuracy of 99% at k = 1. Thus, this method can be used for name disambiguation.

Keywords: Disambiguation, Ambiguity, Hadith, WSD

INTRODUCTION

Each word certainly has its meaning, but what happens if there are words that have more than one meaning. This condition then called ambiguous (Agrawal et al., 2019). Based on the Big Indonesian Dictionary, ambiguous means that it has more than one meaning. This ambiguity can raise doubts in the written or spoken sentence. To eliminate ambiguity, a disambiguation process is needed so that the ambiguous word becomes a clear word. Disambiguation is the process of removing an ambiguous word by making it clear words (Zhang & Hasan, 2017). Disambiguation process is to distinguish between the meaning of the same word to be different (Moro et al., 2014).

Hadith is one of the sources of Islamic religious law after the holy book of the Al-Qur'an which is a collection of sayings, deeds, agreements and decrees from Rasulullah Salallahu Alaihi Wasalam (Faizal et al., 2013). By the middle of the second century Hijriah, a large number of tabi'in compiled their works into book form. People who collect and record hadith are called Perawi. From the hadith that have been written into the book, there are the names of the companions of the Rasullullah as perawi of hadith. In hadith Sahih Bukhari's, some of these perawi found the same name in other hadith numbers, therefore causing ambiguity between the perawi names. It is hard for ordinary people to understand the ambiguity between names and find connection or differences between one name and another because these two names may be the same or they may also be different people. This condition is called ambiguity in the perawi names of hadith. For example, as follows in hadith number 27 there is the name "Abdullah bin Amru" and in hadith number 58 there is also the name "Abdullah bin Amru". As another example with a different case it is not the same name but a similar name, in the hadith number 132 there is the name "Abdurrazaq" and at number 40 there is also "Abdurrazzaq" with a difference in the letter "z" wherein hadith number 40 has two letters "zz" in the name" Abdurazaq ", of course, this is a confusion in the name. So that it is necessary to analyze the entity names and the disambiguation of the entity names in the hadith.

The method used in this research is Word Sense Disambiguation which consists of two techniques, corpus-based and knowledge-based (Ali & Rahman, 2018). Corpus-based methods using a supervised approach to induce classifiers from the training data set consisting of a series of sentences that have been labeled, where the label indicates the meaning of the words used. In contrast, knowledge-based methods rely on external resources such as dictionaries, thesaurus, or lexical knowledge based such as WordNet (Saedi et al., 2018). Meanwhile, the corpus-based method is generally found to have a more accurate performance than the knowledge based method (Ali & Rahman, 2018). Meanwhile, Entity Linking is the task of extracting information by linking the entities mentioned in the text collection with the appropriate Knowledge Based (KB) entries (Pan et al., 2015). Entity linking makes it very easy to utilize the large amount of information publicly available in KB about real-world object entities and their relationships to obtain semantic information that can be used for disambiguation about the entities in question.

This research was inherited from paper *Applying Weighted KNN to Word Sense Disambiguation* conducted by Rezapour. This paper used machine learning process with the supervised learning method for word sense disambiguation, based on the k-nearest neighbor algorithm with an accuracy of 76.1%, have proven the effectiveness of using a method that considers the classification of named entities with entries from the most appropriate knowledge base (Rezapour et al., 2011). However, this study did not use dataset from hadith and it is still difficult to find studies that use hadith as a dataset.

As a solution to this problem, this research was conducted to eliminate the ambiguous names in the hadith of Shahih Bukhari with clustering techniques, using a supervised approach (Rezapour et al., 2011) with a classification based on k-nearest neighbor (Angreni et al., 2018). Where the results of query instances are classified based on the majority of categories in KNN (Zhang et al., 2018). After the classification results are determined, then the information is extracted by connecting the entities mentioned in the context of the text by connecting to the Knowledge-Based (KB) that corresponds to similar words. This study aims to build a dataset containing a set of named entities (Shen et al., 2018), determine the results of disambiguation testing using Word Sense Disambiguation and Entity Linking and measure system performance based on precision, recall and f1-score. In addition, this study has limitations, that is the language used is limited to Indonesian.

METHOD

In this study, the method used is described in stages in the sub-sections according to the workflow of this research, starting from the extraction of names in the hadith of Sahih Bukhari at the pre-processing stage, classifying the same names using Word Sense Disambiguation, and connecting name entities to the knowledge base. The end of this workflow ends with performance measurement. The process carried out on the system can be seen in Figure 1.

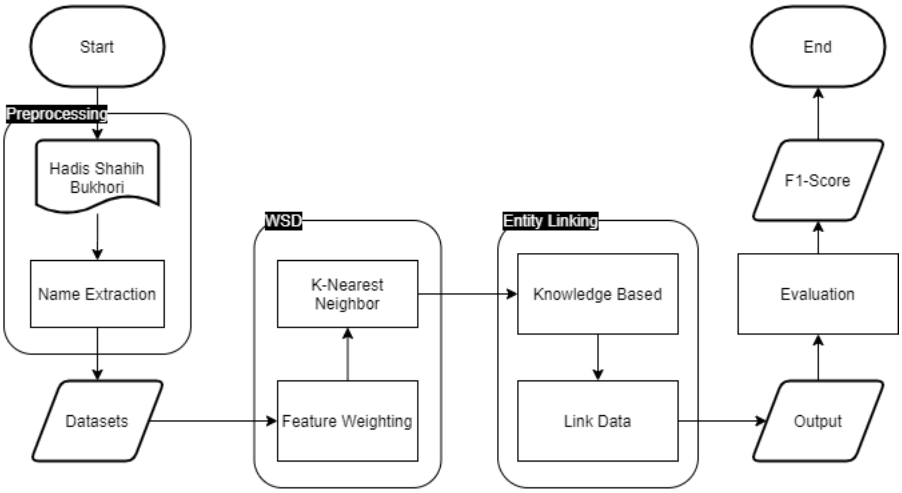

Figure 1. System Overview

Pre-processing

Pre-processing is the first step for preparing data before further process. In general, data pre-processing is done by eliminating incompatible data or transforming data into a form that is easier for the system to process (Oscar, 2019). Not all data from Sahih Bukhari's hadith are used. Therefore, data pre-processing was carried out to sort out the important data that will be used in this study. Before implementing pre-processing, the researcher then manually labels the Bahasa Indonesia translation of the hadith data to separate the entity names, hadith numbers and perawi chains.

Feature Weighting

Feature weighting is carried out to give weight to the value of each named entity. This process divided in two stages, first the countvectorizer is a process to calculate how often the actual name appears, so that the possibility of the same name appears in each hadith number. For example, Abbad bin Tamim in the hadith number 134 gets the value of 1 which indicates that there is one name that appears in the hadith data. Stage two is TFIDF (Upendraa & Sudheer, 2016) the process is to calculate how important name of the entity appears, if a score between 1 means high, then the word is relevant, whereas if a score between 0 means low, then the word is not important. With TFIDF, we can find out the importance of words in the data and we can find out the appearance of each word in other data based on the calculated IDF value from the formula.

$$Tfidf = tf(t,d) \ x \ idf(t) \qquad (1)$$

Where *tf (t, d)* or Term Frequency is a number that is calculated how often terms (words) appear in data. Meanwhile, *idf (t)* or Inverse Document Frequency is the weight value of terms or words that appear with the formula in equation (2).

$$Idf(t) = log \frac{N}{df+1} \qquad (2)$$

Data that has been feature weighted will be processed using K-Nearest Neighbor (KNN). The best *k* value for KNN depends on data (Zhang et al., 2018). The classified data will be processed to validate the real name data which will be done semantically and linked to the Knowledge-Based DBpedia (Pan et al., 2015). Where the data obtained is taken from the DBpedia database (Parravicini et al., 2019), making DBpedia a source of knowledge that is available on the Web using Semantic and Linked data technology (Zapilko et al., 2016).

In this study, to analyze and determine how accurate the system performance is, the F1-Score is used. The Confusion Matrix table is used to measure the performance of system output classification problems with more than two classes. The confusion matrix itself is needed to help obtain calculation accuracy by considering the predicted and actual value (Upendraa & Sudheer, 2016) Table 1 is an overview of the confusion matrix.

Table 1. Confusion Matrix

	Actual Positive	**Actual Negative**
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Based on these components, it will be used to calculate the Precision, Recall and F1-Score. The following is an explanation of precision, recall and F1-score.

Precision is the percentage indicating the accuracy of the system in predicting data into positive classes.

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

Recall shows the sensitivity of the system in predicting positive classes.

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

F1-Score is a value that shows the system's performance in doing prediction.

$$F1\text{-}Score = 2 \times \frac{(Precision \times Recall)}{(Precision+Recall)} \tag{5}$$

RESULT AND DISCUSSION

Result

In this study, the authors used 755 test data, based on the results of pre-processing. The output is the entity name along with the hadith number and perawi chain. From this test data, there are several names which is similar but have different hadith numbers as shown in table 2.

Table 2. Result of Pre-processing

Name	Hadith Number	Perawi Chain
Abdullah	63	Abdullah > Syubah > Qotadah > Anas bin Malik
Abdullah	86	Abdullah > Umar bin Said bin Abu Husain > Abdullah bin Abu Mulaikah > Uqbah bin Al Harits
Abdullah bin Abbas	6	Abdullah bin Abbas > Abu Sufyan bin Harb
Abdullah bin Abbas	49	Abdullah bin Abbas

As seen in the table 2, after the pre-processing stage, each name with a different hadith number has a different perawi chain. Hadith names and numbers are the results of the extraction of names and perawi chains are the order of hadith narrators. Table 3 is an example of the test results from some of the test data.

Table 3. Sample Test Results

Name	Hadith Number	Predicted Name	Same Name in Hadith
Abdullah	46	Abdullah	5, 31, 63, 86, 122
Abdullah	86	Abdullah	5, 31, 46, 63, 122
Abdah	92	Abdushshamad	93
Jarir	68	Jarir	118, 120
Jarir	120	Jarir	68, 118
Khalid	133	Al Laits	3, 11, 27, 61, 80, 101, 113

As seen in the table 3 above are the results after classification using the KNN, the actual names and hadith numbers are obtained according to the original data set and the predictive data for the same name and number prediction. From the test results, a total of 755 names of entities entered into the system, 471 names that are considered ambiguous, and other names for categories will be ignored. For each ambiguous name with a different hadith number is tested against each other, resulting in a predictable name. Determine the same name in the hadith based on the name prediction and classification results with the k-nearest neighbor. Table 4 is the result of DBpedia links based on entity names.

Table 4. DBpedia Link Result

Name	Hadith Number	Link
Abdullah	46	-
Abdullah	86	-
Abdah	92	-
Jarir	68	-
Jarir	120	-
Khalid	133	['Khalid'], 'http://dbpedia.org/resource/Khalid_ibn_al-Walid']

As seen in table above 4 to validate between the real name and the predicted name, by semantically linking the real name to the knowledge-based in DBpedia, a similar name is obtained by matching the real name. There are several links are not available, this is because there is no name that matches the data provided. Another factor that affects this is the limited knowledge-based data available. The evaluation results can be seen in Table 5.

Table 5. Evaluation

K	Precision	Recall	F1-Score
1	99%	99%	99%
2	67%	62%	51%
3	53%	44%	32%
4	44%	33%	20%

As seen in the table 5, the researcher tries to calculate for each *k* value, and the results of the evaluation vary depending on the *k* value tested. The best *k* value used in this study is 1 which is the value of the closest neighbour, when the *k* value is higher between 2,3 and 4, the resulting accuracy will be smaller. This figure was selected based on trials conducted after the feature weighting stage to determine the best *k* value based on the results of evaluation of precision, recall (Suryaningsih, 2020) and f1-score.

Discussion

The pre-processing results in Table 2 are used as input in the classification process by paying attention to their closest neighbors. The pre-process results were obtained from the raw data of Hadith Sahih Bukhari which has been manually labeled on the names and hadith numbers, so that the names and hadith numbers of each criteria were obtained as well as the perawi chain which were the order of the people narrating the hadith. The pre-processing results are used in the next stage, feature weighting was carried out in two stages of Countvectorizer and TFIDF. Feature weighting is used to assign a value to each name by considering how often the names mentioned and how important the names are described to the formulas 1 and 2. After all names have weights then the k-nearest neighbor process is carried out by considering the weight of each name to determine the prediction of the same name as in Table 3.

In table 3, several predictions of names that are different from the original name are found, such as "Abdah" in hadith 92 predicting the same name as "Abdushshamad" this is because the weight value between name “Abdah” and “Abdushshamad” has the same weight so that is considered by the system as the same name. As a validation step, Entity Linking is used by utilizing knowledge-based from DBpedia by connecting name entities with the most appropriate knowledge-based data entry in DBpedia.

As can be seen in table 4 which is the result of name validation on DBpedia, if there is a suitable name it will display a link leading to it, but there is a name that does not have an appropriate link, this is because the same name is not found in knowledge-based DBpedia. From the evaluation results in table 5 with the criteria for the *k* value equal to 1 to 4, it can be seen that the tendency of the f1 score value will decrease along with the increasing of the *k* value. F1 score was drop drastically for k = 4 with an f1 score of 20%. This is because the spread of data is increasingly reaching the farthest neighbors. The highest f1 value in this system is found at k = 1, which is the closest neighbor. So that the KNN method can be used to distinguish entities from the names of hadith narrators.

In previous research, the dataset used was not the name entity of hadith narrators instead TWA senses tagged data and focused only on six words (Rezapour et al., 2011). The evaluation results of the previous research obtained an accuracy of 76.1%. These results were obtained from the average calculation of six ambiguous words. Meanwhile in this research, the evaluation results use precision, recall and f1-score calculations. The feature weighting in the previous research was not used countvectorizer and TFIDF like in this research.

CONCLUSION

Based on the results of this research using the k-nearest neighbor method, were obtain 62,4% of total data considered ambiguous by the system. By using the F1 score obtained the best value at k = 1 is equal to 99% as the performance value of the system. This value is obtained from the results of the comparison names predictions and actual names. Based on these results using WSD and KNN can be used for name disambiguation with a system performance of 99%.

REFERENCES

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction. *Journal of Economic Perspectives*, *33*(2), 31–50.

Ali, M. Y., & Rahman, A. U. (2018). Knowledge-Based & Corpus-Based Methods for Evaluation of Semantic Relatedness of Concepts in Knowledge Graphs. *International Journal of IT & Knowledge Management*, *11*(2), 81–86.

Angreni, I. A., Adisasmita, S. A., Ramli, M. I., & Hamid, S. (2018). Pengaruh Nilai K Pada Metode K-Nearest Neighbor (KNN) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan. *Rekayasa Sipil*, *7*(2), 63–70.

Faizal, P. R. M., Ridhwan, A. A. M., & Kalsom, A. W. (2013). The Entrepreneurs Characteristic from al-Quran and al-Hadis. *International Journal of Trade, Economics and Finance*, *4*(4), 191–196.

Moro, A., Raganato, A., & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation. *Transactions of the Association for Computational Linguistics*, *2*, 231–244.

Oscar, H. (2019). *Basics of Data Preprocessing*. https://medium.com/easyread/basics-of-data-preprocessing-71c314bc7188

Pan, X., Cassidy, T., Hermjakob, U., Ji, H., & Knight, K. (2015). Unsupervised entity linking with abstract meaning representation. *In North American Chapter of the Association for Computational Linguistics*, 1130-1139. Denver: Association for Computational Linguistics.

Parravicini, A., Patra, R., Bartolini, D. B., & Santambrogio, M. D. (2019). Fast and accurate entity linking via graph embedding. *Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*. 1-9. Amsterdam : ACM,

Rezapour, A. R., Fakhrahmad, S. M., & Sadreddini, M. H. (2011). Applying Weighted KNN to Word Sense Disambiguation. *Proceedings of the World Congress on Engineering*, London : Imperial College London, 6-8.

Saedi, C., Branco, A., António Rodrigues, J., & Silva, J. (2018). WordNet Embeddings. *Proceedings of The Third Workshop on Representation Learning for NLP*, 122-131. Australia : Association for Computational Linguistics.

Shen, Y., Yun, H., Lipton, Z. C., Kronrod, Y., & Anandkumar, A. (2018). Deep Active Learning for Named Entity Recognition. *International Conference on Learning Representations*, 1-15. Canada: arxiv.org

Suryaningsih, S. (2020). Building Synonym Sets for English WordNet with Robust Clustering using Links Method. *Edumatic : Jurnal Pendidikan Informatika*, *4*(1), 57–62.

Upendraa, B., & Sudheer, B. (2016). KNN TFIDF Based Named Entity Recognition. *International Journal of Scientific and Research*, *1*(12), 35–39.

Zapilko, B., Schaible, J., Wandhöfer, T., & Mutschke, P. (2016). Applying Linked Data Technologies in the Social Sciences. *KI - Kunstliche Intelligenz*, *30*(2), 159–162.

Zhang, B., & Hasan, M. (2017). Name Disambiguation in Anonymized Graphs using Network Embedding. *In: ACM on Conference on Information and Knowledge Management*. Singapore : ACM, 1239-1248.

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2018). Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1774–1785.