Analisis Entitas Nama pada Teks dengan Menggunakan Metode Robust Disambiguation

Analysis of Name Entities in Text Using Robust Disambiguation Method

Muthia Virliani[1], Moch. Arif Bijaksana[2], Arie Ardiyanti Suryani[3]
[1,2,3] Telkom University; Jl. Telekomunikasi Terusan Buah Batu, Bandung 40257, Telp. (022) 7566456
[1,2,3]Informatics Study Program, Faculty of Informatics, Telkom University, Bandung
e-mail: [1]muthiavirliani15@gmail.com, [2]arifbijaksana@telkomuniversity.ac.id ,
[3]ardiyanti@telkomuniversity.ac.id

Abstrak

Entitas nama merupakan kata benda yang spesifik yang terdapat dalam teks, seperti nama orang, nama negara dan sebagainya. Nama orang dalam teks seringkali sama, yang menyebabkan orang kesulitan untuk membedakan atau mengetahui nama-nama yang sama tersebut adalah orang yang sama atau tidak. Keambiguasian nama juga terdapat pada hadis, seperti nama Abdullah pada nomor hadis 2404 dan 2411, yang menyebabkan keambiguasian antara nama, karena dua nama ini sama dan belum terbukti adalah orang yang sama. Berdasarkan masalah ini, maka penelitian ini berfokus pada disambiguasi entitas nama dengan mempertimbangkan hubungan konteks dan koherensi antara entitas nama. Kedepannya diharapkan, hal ini dapat membantu orang untuk memahami keambigusian nama atau meminimalisir kesalahan tafsir nama. Metode yang digunakan pada penelitian ini adalah Robust Disambiguation, karena dalam metode ini konteks dari entitas nama dipertimbangkan. Output yang dihasilkan berupa pengelompokkan entitas nama berdasarkan orang yang sama atau berbeda, yang diproses dengan Density-based Spatial Clustering of Applications with noise. Evaluasi pada penelitian ini menghasilkan nilai accuracy sebesar 90%, precision sebesar 97% dan recall sebesar 89% yang diperoleh dari nilai aktual dan nilai prediksi.

Kata kunci*—Density-based Clustering, Disambiguasi, Hadis Shahih Bukhari, Jaccard Similarity, Robust Disambiguation.*

Abstract

Named entities are proper nouns or objects contained in a text, such as a person's name, country name, and others. Names of persons in some text are often the same or similar, which makes it difficult for people to distinguish or find out these same names are the same person or not. An ambiguity of names also found in hadith, like the name Abdullah in hadith number 86 and 2411, this causes ambiguity between names because both names are the same and not yet proved to be the same person. Based on this problem, then this study focuses on named entity disambiguation, which considered further context and coherent relation between a named entity. Expected in the future, it would help people to understand the ambiguity of the name or minimize the misinterpretation of names. The method used in this research was Robust Disambiguation because, in this method, the context of the named entity considered. The resulted output obtained was in the form of named entity that grouped based on the same person or different person processed with Density-based Spatial Clustering of Applications with Noise. This research resulted in an accuracy value of 90%, a precision value of 97%, and a recall value of 89% obtained from actual value and predicted value.

Keywords*—Density-based Clustering, Disambiguation, Hadith Sahih Bukhari, Jaccard Similarity, Robust Disambiguation.*

1. INTRODUCTION

Text processing is one of the crucial things in text mining. According to the Great Indonesian Dictionary, text defined as a manuscript in the form of original words from authors or quotations from the holy book for the roots of teachings or reasons. Other than that, there is also a book of hadith which consists of text. The hadith are all the words of the Prophet Muhammad SAW, his actions and subject matter. In hadith studies, there is a term called rawi. Rawi is the person who receives and conveys a hadith. One of the most Sahih or authentic hadith books is the Sahih Bukhari compiled by Imam al-Bukhari because it requires *liqa'* or meeting with two narrators, between rawi and his teacher [1]. In the hadith of Sahih Bukhari, some rawi often have the same name, therefore causing ambiguity between the rawi names. It is hard for ordinary people to understand the ambiguity between names and find connections or differences between one name and another because these two names might be the same, or they might be different people. Also, misinterpreting the name of the hadith should not be done as much as possible, because rawi is related to whether or not a hadith is sahih. For example, the name Abdullah in the hadith number 2404 and 86. These two names are similar but not yet proven to be the same person, so the analysis of name entities and named entity disambiguation in the text of the hadith is required. Named entity disambiguation is a process identifying the meaning of a name to eliminate the ambiguity. If there are several similar names in a context, then the identification process is carried out. The identification process is carried out based on its meaning or semantic relation [2].

This research inherited from the study conducted by Hoffart J. [3] and Yang Li [4]. The study Conducted by Hoffart J. about collective disambiguation used data from the knowledge base derived from Wikipedia, where the framework consists of prior probability, the similarity between context, and coherence among candidates. This study builds a weighted graph and computes a dense subgraph with an accuracy of 87.31%. The study conducted by Li Y. about Mining Evidences for Named Entity Disambiguation used knowledge base such as Wikipedia. This study considered the association between entities and context, also the topical coherence from Wikipedia cross-page links. This study obtained an accuracy of 86%. Both studies have proved the effectiveness of using methods that consider the context and coherent. But these two studies did not use a dataset from the hadith and still hard to find the study used hadith as the dataset. Some other related studies are research conducted by Mark Dredze [5], about Entity Disambiguation for Knowledge Base Population using Knowledge Bases derived from Wikipedia pages with 95% accuracy achieved. The framework of this study consists of the selection candidate list, entity disambiguation, and feature set of entity linking with any knowledge base. The study conducted by Maria Pershina [6] about Personalized Page Rank for Named Entity Disambiguation with an accuracy of 89.9% using AIDA harvested from Wikipedia, in this study, a new algorithm contrived for collective disambiguation without requiring parameters [7]. Also, there is a study conducted by Ayman Alhelbawy about Graph Ranking for Collective Named Entity Disambiguation with an accuracy of 87% used AIDA [8]. The framework of this study is for collective disambiguation using a graph model and model coherence by links between nodes.

As a solution to this problem, this study presents an analysis of the named entity to disambiguate ambiguous names in the hadith text by considering semantic and lexical relationships. This study complementary the problem of name disambiguation by trying to develop the scope of the dataset with hadith data, which is still rare. Also, this research can minimize mistakes when interpreting rawi names, where rawi is the aspect of validity in hadith.

The method used in this study is Robust Disambiguation that combines several approaches, consisting of Popularity Prior, Context Similarity, and Coherence among entities. Popularity Prior is a process to calculate how often an object refers to particular name entities. The most common way is to estimate the occurrence of a name that refers to specific name entities or numbers [3]. Context Similarity is a process by considering the context of the mention or named

entity. The chosen approach for Context Similarity is Keyphrase-based Similarity. In this approach, for each mention or name entity, a context is constructed from all words in a text [3]. Each named entity represented as a set of words or phrases. Coherence among entities is a significant process because most writings deal with several semantical topics [3]. In this process, candidate entities considered for a different mention, so it can be determined and calculated as an assumption or notion of coherence. Jaccard Similarity is also used to calculate the value of similarity between the same name at context similarity and coherence approach. These methods used because to eliminate the ambiguity of the same names, need to consider the context between named entities or semantic relations. The output of this study is in the form of named entity that is grouped based on the same person or not and the data linkage, using DBSCAN or Density-Based Spatial Clustering of Applications with Noise. This method used because DBSCAN can determine how many clusters are needed, so there is no need to give some k values. This study aims to build a dataset containing a set of named entities, find out the test results of disambiguation using Robust Disambiguation and Jaccard Similarity, find out the test results of cluster analysis using DBSCAN and measure system performance from accuracy, precision, and recall. Also, this study has limitations that are the language used is only limited to Indonesian or English.

2. RESEARCH METHOD

In this study, the methods used are explained step by step in the subsection according to the workflow of this study, which starts from the selection of the dataset, processed the dataset into structured data with text preprocessing, eliminating the ambiguity of the same name with name disambiguation, and clustering using DBSCAN and PCA. The end of this workflow ended with performance measurement.

2.1 Dataset

The chosen dataset obtained from the book of Sahih Bukhari, this book is a collection of hadith compiled by Imam Bukhari. There is a structure in the hadith consisting of sanad, matan, and rawi. Sanad is the genealogical link of the people that connect to matan, where all behavior, statement, and others are matan hadith. The person who conveys the hadith called rawi. The dataset used in addition to the hadith is the Wikipedia Page and articles on a website as text input. Wikipedia dataset used as a comparison.

2.2 Text Preprocessing

Text preprocessing is a crucial step for selecting and cleaning data before further processing. The chosen dataset, as explained in the previous subsection, is the hadith Sahih Bukhari. Figure 1 represents and illustrates the text processing steps.

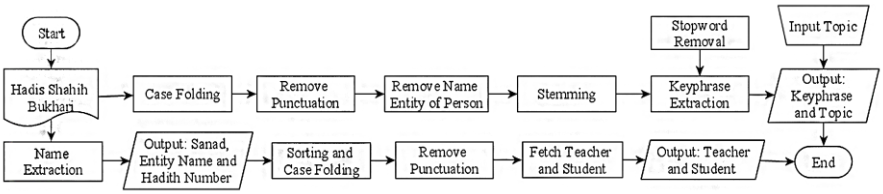

Figure 1. Flowchart of the Text Preprocessing

As shown in the illustration of Figure 1, the chosen dataset then cleaned with text preprocessing. The text preprocessing divided into two processes that are keyphrase extraction and name extraction. These two processes carried out separately. The output of these two processes used for the name disambiguation and the steps of these two processes explained below.

2.2.1 Keyphrase Extraction

The results or output of the keyphrase extraction in the form of Keyphrases that accompanying each entity name. The following steps carried out in this process:

1. Case folding and remove punctuation
   The dataset or hadith text converted to lowercase and every punctuation in the text deleted to ease the next step.
2. Remove the named entities of persons.
   The next step is to removed named entities of persons. This step conducted to obtained the contents of the hadith or can be called matan. The method used is by labeling the text to separate the names of persons and the contents of the hadith or matan text.
3. Stemming
   Matan hadith that has no names entities of persons, then processed by Stemming. Stemming is the process of reduced words to their stem.
4. Stopword removal and Keyphrases extraction
   The result of the stemming process then processed with stopword removal to eliminated or deleted common words that have a high frequency of occurrence. Data that already cleaned from the previous steps then extracted to obtained the keyphrase.

2.2.2 Name Extraction and Fetch Teacher-Student

The results or output of the name extraction, in the form of named entity or rawi, sanad, and hadith number, including the resulting output in the form of mention teacher and student. The following are step by step carried out in this process:

1. Name extraction
   Dataset or hadith text are manually labeled to separate the entity names of persons, hadith number, and sanad from its contents or matan.
2. Sorting and case folding
   The output obtained from the name extraction process then sorted by the same name and converted to lowercase letters.
3. Remove punctuation
   In this step, all punctuation marks then deleted to make the next process easier.
4. Fetch teacher and student
   The first step of getting the teacher and student is searching for a rawi or name in sanad and find the position. After the location of the name or rawi found, then take a name on the right and left side of the rawi. Based on the knowledge of Rijal al-Hadith[9], the left side of a rawi on a sanad is the teacher, and the right side of a rawi on a sanad is the student.

2. 3 Name Disambiguation

The results of the text preprocessing above then processed with the name disambiguation. The Keyphrase results from the text preprocessing used for the Context Similarity process. The output in the form of named entity or rawi, sanad, hadith number, and mention of teacher-student used in the Coherence process. Figure 2 represents and illustrates the steps of name disambiguation.

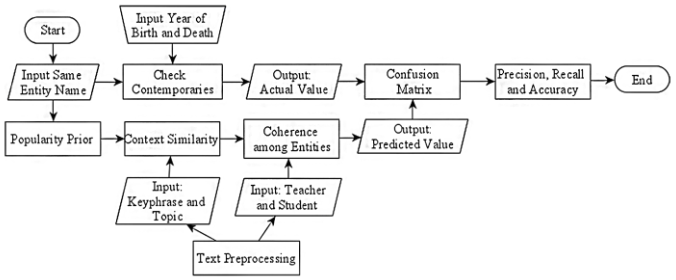

Figure 2. Flowchart of the Name Disambiguation Process

As shown in the illustration of Figure 2, name disambiguation divided into two processes, that are Contemporaries and Robust Disambiguation that consist of Popularity Prior, Context Similarity, and Coherence among Entity. These two processes carried out separately. The output of these two processes used for clustering and the steps of these two processes explained below.

2. 3.1 Robust Disambiguation

The process of robust disambiguation consists of three steps which carried out separately. Below is an explanation of these three steps.

1. Popularity Prior

   In Popularity Prior, the process carried out is to estimate the frequency of a rawi occurrence in a text of the hadith. As an example, the name Abdullah in hadith number 2404 get an occurrence value of 0.0109.

2. Context Similarity

   In the Context Similarity, there are two stages. First, the extracted Keyphrase then processed into a single keyword. Second, the topics and keywords that accompany the two ambiguous name entities then proceeded with each other to get similarity value with Jaccard Similarity. Jaccard Index or Jaccard Similarity Coefficient is a method for calculating the similarity between two objects. Equation 1 is a calculation between two sets of objects, A and B.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

   Where intersection cardinalities of A and B then divided by the union cardinalities of the two objects, and if A and B are empty, then $J(A, B) = 1$. The range of similarity values is $0 \leq J(A, B) \leq 1$.

3. Coherence among Entities

   In Coherence among Entities, there are two stages. First, on the two named entities that same, the teacher and student that accompany the two ambiguous names, then processed with each other to get similarity value with Jaccard Similarity. Second, the results obtained from the first stages then determined based on its similarity value with assumptions. If the teacher and student of the two named entities are different, then it can be said that the two names are different people, and the given output is 0. If the teacher and student of the two named entities are the same or one of them is different, then it can be said that the two names are the same people, and the given output is 1.

2. 3.2 Contemporaries

Contemporaries as an adjective, according to the Great Indonesian Dictionary has

meaning as one period, at the same time, and existing at the same time. In this process, there are two stages. First, the two entities with the same name are then examined from the year of birth and death, to find out whether these people are contemporaries or not. An examination of the year of birth and death is by sequence matcher. Second, the results obtained from the first stages then determined based on its value with conditions. If the year of birth and death are the same, then the given output value is 1. If the year of birth and death are different, then the given output value is 0. The year of death obtained manually from the hadith encyclopedia [10].

2. 4 Clustering Named Entity

DBSCAN is a clustering algorithm with an estimated minimum data density level based on the minimum number of points or minpts within a radius eps [11]. Radius eps is the maximum distance between two points. The core point is the amount of data in radius eps that more than minpts and the amount of data in radius eps that smaller than minpts referred to as a border. Noise is a point whose density is out of range [12]. Equation 2 is a formulation of Euclidean Distance to calculate the distance between the core point and other points.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

Where x is the data center of the cluster, y is the data attribute, n is the amount of data, and the range is $0 \leq d(x, y) \leq 1$. Figure 3 is a clustering process using DBSCAN and PCA [13].

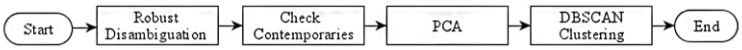

Figure 3. Flowchart of the Clustering Process

Below is a step by step explanation of Figure 3. The dataset used in the clustering process is the values obtained from the results of name disambiguation.

1. Robust Disambiguation and Contemporaries
   In this step, the values generated from the Robust Disambiguation and Contemporaries taken and used as a dataset.
2. PCA or Principal Component Analysis
   PCA is a method to reduce the dimensions of the dataset. In this step, the data obtained from the first step then reduced by PCA. The components selection on PCA is to increase the results of clustering using DBSCAN with an explained variance of 74% and three components.
3. DBSCAN Clustering
   In this step, data results from dimension reduction then processed with the DBSCAN clustering algorithm. From several experiments with DBSCAN to get good clustering results are using parameters with a core point of 2 and neighbor distance or optimal value for eps is 2.11.

2. 5 Performance Measurement

In this study, performance measurements [14] are using precision, recall, and accuracy based on a combination of values in the confusion matrix, that obtained from the predicted and actual values. Below is the step by step explanation of performance measurement to obtained precision, recall, and accuracy values, that calculated separately.

1. Predicted and actual value
   As shown in the illustration of Figure 2, the result of name disambiguation with Contemporaries generated the actual value, while Robust Disambiguation generated the predicted value. In this step, there are two stages. First, for the same two names, the results

of the value obtained from Robust Disambiguation, if more than the threshold, then it can be said that the two names are the same people, and the given output is Y. If the opposite, then the given output N. Second, for the two ambiguous names, the result of the value obtained from Contemporaries, if Contemporaries value equal to 1, then the given output Y because regarding the two same persons. If the opposite, then the given output N. Threshold is the parameter limit of similarity value, the threshold value is 0.5 that obtained from several testing.

2. Confusion matrix
The predicted and actual value obtained from the first step then combined to get the value of the confusion matrix. Table 1 is a representation of the confusion matrix, there are four combinations of predicted values and actual values [15], which consist of True Positive (TP), False Positive (FP), True negative (TN), and False Negative (FN).

Table 1. Representation of confusion matrix

	Class	Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

TP is positive data that predicted true, and FP is negative but predicted positive. TN is data negative that predicted true, and FN is positive but predicted negative.

3. Precision
The combined value of the confusion matrix then calculated using a precision formula. Equation 3 is a precision formula, where precision is a ratio prediction of true positive compared to overall positive predicted [16].

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

4. Recall
The combined value of the confusion matrix then calculated using a recall formula. Equation 4 is a recall formula, where the recall is a ratio prediction of true positive compared to all data that correctly positive (true positive or false negative).

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{4}$$

5. Accuracy
The combined value of the confusion matrix then calculated using a accuracy formula. Accuracy is the ratio prediction of TP and TN compared to all combination value of predicted and actual, to find out the performance of a model. Equation 5 is an accuracy formula.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

3. RESULT AND DISCUSSION

In this study, testing carried out with 100 test data from text preprocessing. The test data consists of 33 named entities, where every name has three or four same names with different

hadith number that compared to each other. The results and discussion for each process explained in the subsection below.

3. 1 Text Preprocessing

The results of the text preprocessing used as the test data for the testing in the name disambiguation process. Table 2 was an exemplification of several test data as a result of text processing. The contents of the table in Indonesian because of the data hadith in Indonesian.

Table 2. An Examplification of the test data

Name	Hadith	Teacher	Student	Topics	Keyword	Death	Born
Abdullah	2404	hibban bin musa	yunus	hibah	['nama', 'ridho', 'undi', 'hibah', 'jatah', 'jalan', 'gilir', 'kecuali', 'cari', 'zamah', 'laku', 'malam', 'isteri', 'saudah', 'tuju']	181	118
Abdullah	2411	abdan	yunus	hibah	['seraya', 'syahid', 'bebas', 'sisa', 'orang', 'kebun', 'beban', 'bicara', 'uhud', 'bunuh', 'buah', 'kurma', 'peristiwa', 'lunas', 'desak', 'hutang', 'perang', 'petik', 'hak', 'keliling', 'temu', 'pagi', 'piutang', 'pohon', 'besok', 'bukti', 'duduk', 'doa', 'berkah', 'terima', 'allah']	181	118
Abdullah	86	muhammad bin muqatil abu al hasan	umar bin said bin abu husain	ilmu	['putri', 'kendaraan', 'madinah', 'cerai', 'perempuan', 'nikah', 'orang', 'temu', 'wanita']	117	40

As shown in the table above, for each name with a different hadith number has its mention or data attributes or features, which are the results of text preprocessing. Attribute name and number of hadith are the results of the name extraction. As explained in section 2, the teacher and student attribute obtained from fetch teacher and student by searching for a name in the sanad and looking for its position, then take the names on the right and left. Figure 4 is one example of the results of the name extraction of sanad.

hibban bin musa → abdullah → yunus → az zuhriy → urwah → aisyah radliallahuanha

Figure 4. Example of sanad

The keyword attribute is the result of the Keyphrase extraction process, where the obtained keyphrase output converted into a single keyword to facilitate the process of name disambiguation. Figure 5 is an illustration of the Keyphrases to the keyword conversion process.

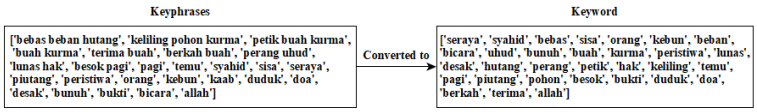

Figure 5. Converted Keyword

The topics, year of death, and year of born attribute inputted manually, these attributes obtained from hadith encyclopedia of sahih Bukhari, where the year of born and death is in Hijri.

3. 2 Name Disambiguation

The results of the text processing explained above, then used as test data in the name disambiguation process. The teacher and student attribute used for the Coherence process, attribute of topics, and keywords used for the Context Similarity process. Also, attribute death and born used for the Contemporaries process. Table 3 was an exemplification of the testing result from some test data.

Table 3. An Examplification of the testing result

Name	Hadith Number (1)	Hadith Number (2)	Prior	Context	Coherence		Score	Pred.	Contemporaries		Act.
					Teacher	Student			Born	Death	
Abdullah	2404	2411	0.0109	0.53	0	1	0.77	Y	1	1	Y
Abdullah	2411	86	0.0099	0.03	0	0	0.02	N	0	0	N
Abdullah	86	2404	0.0168	0.045	0.12	0	0.03	N	0	0	N
Abdurrahman	2249	5205	0.0068	0.035	1	1	0.53	Y	1	1	Y
Abdurrahman	5205	5232	0.0185	0.055	1	1	0.54	Y	1	1	Y
Abdurrahman	5232	2249	0.0185	0.03	1	1	0.53	Y	1	1	Y
Abdurrazzaq	40	132	0.0149	0.05	0.33	1	0.53	Y	1	1	Y
Abdurrazzaq	132	5237	0.0147	0.03	0.14	1	0.52	Y	1	1	Y
Abdurrazzaq	5237	40	0.005	0.03	0.2	1	0.52	Y	1	1	Y
Abu 'Awanah	5204	5217	0.0104	0.03	0.2	0	0.02	N	1	1	Y
Abu 'Awanah	5217	5243	0.0238	0.54	0.2	1	0.78	Y	1	1	Y
Abu 'Awanah	5243	5204	0.078	0.03	1	0	0.52	Y	1	1	Y
Amru	110	119	0.0294	0.51	1	0.2	0.77	Y	1	1	Y
Amru	119	21	0.013	0.01	0	0.17	0.01	N	0	0	N
Amru	21	110	0.0177	0.04	0	0	0.03	N	0	0	N

As shown in the table above, for each ambiguous name with a different hadith number tested against each other, and produce predictive and actual values. Predictive value is the combined value of Prior, Context, and Coherence obtained from the test results of each process. Prior values obtained from the calculation of the frequency of occurrence name in the text of hadith numbers in the Popularity Prior process, these values are small because of the occurrence name is less frequent compared to the sum of all words. In the process of Context Similarity, the resulting context values from testing is small because the keywords that accompany named entities are unique. And for the results of the coherence process after being tested, a new assumption is obtained that is for the same two names might be the same person, although the teachers and students different. This assumption based on the predicted results compared to the actual value. Actual values are the result of the value obtained from the process of Contemporaries based on the results of the sequence matcher year of birth and death of the same two names.

From the table, the predicted and actual values consist of output with Y or N, where Y is the same person, and N is different. As explained in section 2, the predictive value output determined by the threshold. The threshold value used is 0.5, where this value obtained from several testing, and this value gives good results. For example, in the attribute of the score, the name of Abdullah in hadith number 2404 and 2411, the score obtained is 0.77. This score more than the threshold, then the given output is Y. While for the actual value, if the results of the sequence matcher both born and death equal to 1, then the given output is Y other than that, then N. For example, the results of born and death of Abdullah in hadith number 2404 and 2411 both are equal to 1, then the given output is Y.

3. 3 Clustering Named Entity

From the testing results with name disambiguation, for name clustering, the value used is Prior, Context, Coherence, and Contemporaries. These values are then reduced by PCA to facilitate the clustering process with DBSCAN. Figure 6 is a visualization of dimensions reduction with PCA, the green line represents the cumulative variance, from the percentage of variance calculated for the first n components. Meanwhile, the component variance is a percentage of variance calculated for each principal component.

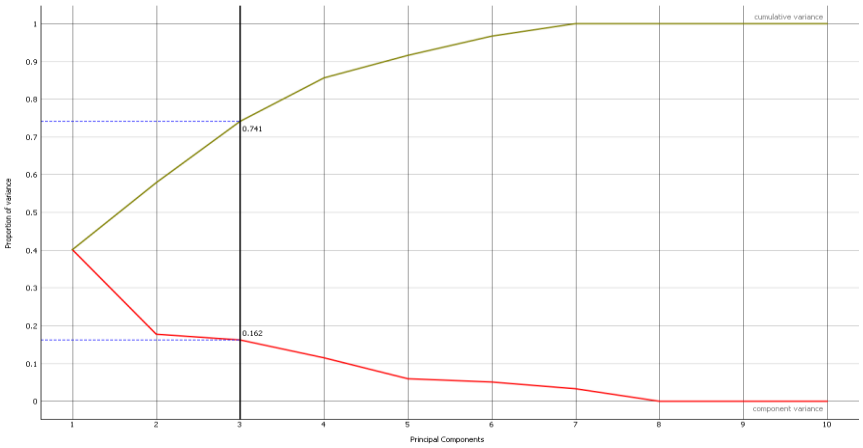

Figure 6. Visualization of PCA

As shown in the figure above, all of the proportion variances of the cumulative value with three components can explain around 74% variance of test data, or in other words, explain how dispersed the data is. The exact coefficient values from the first nine rows of the three Principal Component showed in Table 4, where this coefficient is a correlation between rows and columns of data. On PC2, the higher the Principal Component coefficient, then the data linkage is high. If the negative and positive coefficient value in PC3 or PC1 close to 0, then the data linkage is low.

Table 4. Principal Component coefficient

Name	Hadith Number (1)	Hadith Number (2)	PC1	PC2	PC3
Abdullah	2404	2411	-1.888	0.329	1.356
Abdullah	2411	86	3.498	-0.483	1.179
Abdullah	86	2404	3.474	0.459	0.690
Abdurrahman	2249	5205	-0.721	-1.331	-1.045
Abdurrahman	5205	5232	-0.820	0.328	-2.403
Abdurrahman	5232	2249	-0.760	-0.739	-1.301
Abdurrazzaq	40	132	-0.647	-1.151	0.279
Abdurrazzaq	132	5237	-0.705	-1.145	-0.753
Abdurrazzaq	5237	40	-0.690	-1.904	-0.294
Abu 'Awanah	5204	5217	1.219	-0.408	-1.399
Abu 'Awanah	5217	5243	-2.020	2.089	-0.581
Abu 'Awanah	5243	5204	-0.767	-1.262	-1.632
Amru	110	119	-1.856	0.507	2.140
Amru	119	21	3.577	-1.873	2.291
Amru	21	110	3.512	-0.041	1.399

All the result of dimension reduction above, then clustered using DBSCAN. The parameters used in this clustering process are core points neighbors of 2, and neighbor distance or optimal value for eps is 2.11. The distance between the core to other points calculated by Euclidean. The parameters used based on several experiments with these parameters obtained good clustering results. For more details of these parameters shown in Figure 7. This figure

illustrates the density distribution of the data and gives the reasoning of an ideal selection for Neighborhood distance setting.

As shown in the figure above, all points with the highest k-distance considered as noise,

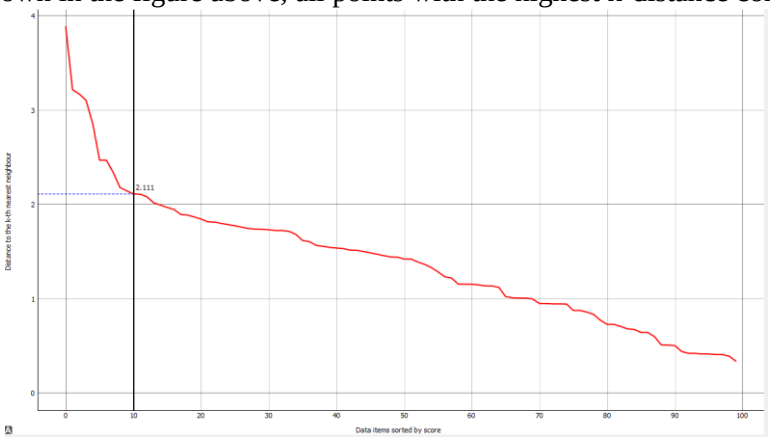

Figure 7. Sorted graph with distance

and other points, added to the cluster. Figure 8 is a visualization of distribution points or names from clustering results with a scatters plot. Informative projection in this figure from Axis x that represents the results of PC1, while Axis y represents the results of PC3, also for the label based on the named entities. From this figure, showing that points with cross shapes represent contemporaries and circles represent non-contemporaries, and each point divided into five clusters, that are C1 or blue, C2 or red, C3 or green, C4 or orange, and C5 or yellow. The gray point with the shape of a cross is the core point. These clusters represent the data linkages.

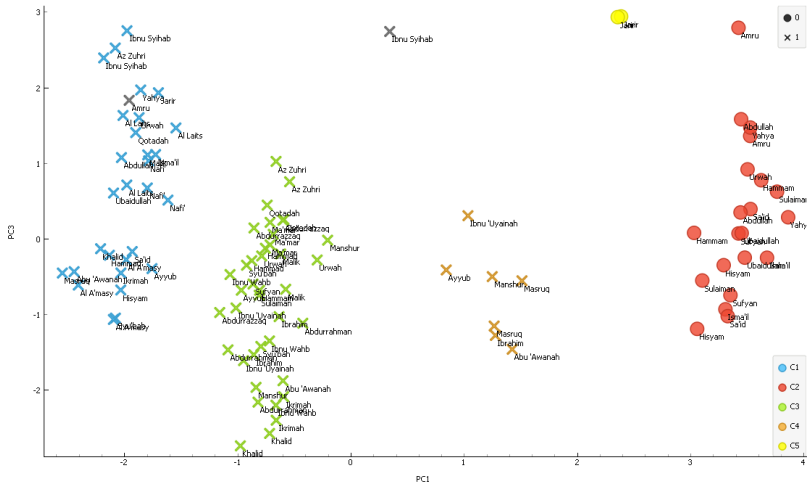

Figure 8. Visualization of Clustering Results with Scatter Plot

There are several interpretations of Figure 8. Table 5 is an interpretation of each cluster with cross shapes and detail about linkages and contemporaries. As an example, there are three names of Nafi' in C1 that can be said are the same person. The names are acknowledged to be the same person because, in the C1, the data linkage between name is very high and also contemporaries.

Table 5. Interpretation of clusters

Cluster	Interpretation	Detail
C1	For the same name with a different hadith number, if all the same names are in C1, then the same names are the same person.	Data linkages are the highest, and the persons are contemporaries.
C3	For the same name with a different hadith number, if all the same names are in C3, then the same names are the same person.	Data linkages are quite high and the persons are contemporaries.
C1, C3, or C4	Every same name with a different hadith number, if there is a name in one of these three clusters and that name also found between these three clusters, then the same names are the same person.	The level of data linkages is different and the persons are contemporaries.

Table 6 is an interpretation of the same names in different clusters, which is cross and circles shape with the detail. As an example, there are two names of Jarir in C5 that can be said are the same person. The name of Jarir also found in C1, but the name in C1 and the name in C5 are different persons. The names are acknowledged to be the different persons because, between C1 and C5, the data linkage between name is low and also not contemporary.

Table 6. Interpretation of diffrent clusters

Cluster	Interpretation	Detail
C1 or C5	If there are names with different hadith numbers in the same cluster, then the same names are the same person, and if that name also found within these two clusters, then the same names are different persons.	Low data linkage and the persons are contemporaries.
C1 or C2	If there are the same names with different hadith numbers in the same cluster, then the same names are the same person. If different, then the same names are different persons.	Low data linkage and the persons are contemporaries.
C3 or C2	The interpretation of these clusters is the same as cluster C1 or C2.	Low data linkage and the persons are contemporaries.
C4 or C2	The interpretation of these clusters is the same as cluster C1 or C2.	Low data linkage and the persons are contemporaries.

Data linkage on each cluster based on the results obtained from the name disambiguation, that consist of Prior, Context, Coherence, and Contemporaries that already reduced by PCA.

3. 4 Performance Measurement

From the results of testing with the name disambiguation above, the predicted and actual values used to measure performance with a confusion matrix. The combination of values obtained from the predicted and actual values are True Positive, True Negative, False Negative, and False positive, where data negative is the different person, and data positive is the same person. True Positive of 67 names that correctly predicted as the same persons. True Negative of 23 names that correctly predicted as different persons. False Positive of 2 names predicted as the different persons, but the truth is the same. False Negative of 8 names predicted as the same persons, but the truth is different. This combination of value used to calculate the precision, recall, and accuracy value. Table 7 is the result of the comparison evaluation.

Table 7. Comparison results of evaluation

Dataset	Combination of Predicted and Actual				Precision	Recall	Accuracy
	TP	TN	FP	FN			
Hadith	67	23	2	8	97%	89%	90%
Wikipedia	74	20	3	4	96%	95%	93%

The data used for comparison are Wikipedia. The name disambiguation process using Wikipedia data is almost the same as the hadith data. In Popularity Prior using Wikipedia data, the most commonly used approach is Wikipedia-based frequencies, by estimating the appearance of a name in the anchor text link, which refers to specific entities, specific numbers, or inlinks. The chosen approach for Context Similarity using Wikipedia is also Keyphrase-based Similarity. In this approach, for each mention or name entity, a context is constructed from all words in a text. Each named entity represented as a set of words or phrases. In the Coherence process, candidate entities considered for a different mention, so it can be determined and calculated as an assumption or notion of coherence. In this process, Wikipedia articles between the two names entities measured by Wikipedia incoming links. As shown in the table above, the accuracy value with Wikipedia data is higher because the value generated in the Context Similarity process with Wikipedia is higher than hadith data, where the Keyword in hadith data more unique. The recall value of the Wikipedia data is also higher, which is due to the smaller predicted error or False Negative value. As explained in Section 1, both of these testing and methods used inherited from research conducted by Hoffart J. [3] and Li Y. [4].

For the results of Hadith evaluation, the recall value influenced by the prediction error of the same person predicted as different persons because the predictive value is small, and this recall value measures the quantity of system. The recall value obtained was 89%, where this value is pretty good. Factors that affect precision are contemporaries that caused by the absence of a birth year of a name, and this precision value measures the quality of the system of 97%. And for the accuracy value influenced by all results combined values where the higher the prediction error, the smaller the accuracy. The accuracy value of 90% signifies closeness to the Actual value.

4. CONCLUSIONS

Based on the results of the testing with Robust Disambiguation, the accuracy obtained is equal to 90% of the same name input that correctly identified as the same or different people. Also, based on the results of testing analysis and interpretation of clustering with DBSCAN reduced by PCA, the results correctly displayed these same names grouped according to the same or different people based on clusters or data linkage and their contemporaries. The precision and recall values of the test results were high does not mean that the entire system was great because test data presented only reach 100 data, and results can change depending on the data. The context similarity value obtained based on the test results for hadith was small, so only using this method was less effective. Because compared with the context similarity value for Wikipedia was high.

Then based on the results of the entity name analysis concluded that the same name surely the same person if the teacher, student, year of birth, year of death, and topics between the names were the same and related. Overall based on the testing, the Robust Disambiguation method can be applied to eliminate the ambiguity of named entity in hadith data because the evaluation results obtained pretty good. This research presents a complementary study to the name disambiguation problem from another point of view, where this study using another dataset that is hadith, which is still very rare in this problem. This research also helps to identify ambiguous names in the hadith, to minimize mistakes when interpreting names, because of names or rawi important aspect.

5. SUGGESTIONS FOR THE FUTURE WORKS

Suggestions for further development are to consider the synonyms of the words in the keywords that accompany the named entity before calculating the value of similarity because, based on the results of keyword extraction, most of the keywords produced are unique. When cleaning the data, normalization should be with lemmatization, because this process aims to produce words into their basic form. Also, when testing needs to increase the amount of test data, so the results more accurate, and more names identified.

REFERENCES

[1]. K. Pendidikan, D. A. N. Seni, and B. Islam, "Ulumul hadits," *Ulumul Hadist*, 2017.

[2]. X. Han and J. Zhao, "Structural Semantic Relatedness: A knowledge-based method to named entity disambiguation," in *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2010.

[3]. J. Hoffart *et al.*, "Robust disambiguation of named entities in text," in *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2011.

[4]. Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan, "Mining evidences for named entity disambiguation," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, vol. Part F128815, doi: 10.1145/2487575.2487681.

[5]. M. Dredze, P. Mcnamee, D. Rao, A. Gerber, and T. Finin, "Entity disambiguation for knowledge base population," in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, 2010.

[6]. M. Pershina, Y. He, and R. Grishman, "Personalized page rank for named entity disambiguation," in *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2015, doi: 10.3115/v1/n15-1026.

[7]. E. F. Y. Hom, F. Marchis, T. K. Lee, S. Haase, D. A. Agard, and J. W. Sedat, "AIDA: an adaptive image deconvolution algorithm with application to multi-frame and three-dimensional data," *Journal of the Optical Society of America A*, 2007, doi: 10.1364/josaa.24.001580.

[8]. A. Alhelbawy and R. Gaizauskas, "Graph ranking for collective Named Entity Disambiguation," in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 2014, vol. 2, pp. 75–80, doi: 10.3115/v1/p14-2013.

[9]. Mr. Suryadi, "REKONSTRUKSI KRITIK SANAD DAN MATAN DALAM STUDI HADIS," *ESENSIA: Jurnal Ilmu-Ilmu Ushuluddin*, 2015, doi: 10.14421/esensia.v16i2.996.

[10]. H. H. Batubara, "Pemanfaatan Ensiklopedi Hadis Kitab 9 Imam sebagai Media dan Sumber Belajar Hadis," *Muallimuna : Jurnal Madrasah Ibtidaiyah*, 2017, doi: 10.31602/muallimuna.v2i2.769.

[11]. T. N. Tran, K. Drab, and M. Daszykowski, "Revised DBSCAN algorithm to cluster data with dense adjacent clusters," *Chemometrics and Intelligent Laboratory Systems*, 2013, doi: 10.1016/j.chemolab.2012.11.006.

[12]. E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems*, 2017, doi: 10.1145/3068335.

[13]. H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*. 2010, doi: 10.1002/wics.101.

[14]. N. Japkowicz, "Why question machine learning evaluation methods? (An illustrative review of the shortcomings of current methods)," in *AAAI Workshop - Technical Report*, 2006.

[15]. S. Visa, B. Ramsay, A. Ralescu, and E. van der Knaap, "Confusion matrix-based feature selection," in *CEUR Workshop Proceedings*, 2011.

[16]. M. R. Ghorab, D. Zhou, A. O'Connor, and V. Wade, "Personalised Information Retrieval: Survey and classification," *User Modelling and User-Adapted Interaction*, 2013, doi: 10.1007/s11257-012-9124-1.