SVM FOR *INDONESIAN* COREFERENCE RESOLUTION WHICH CONTAINS PLURAL ENTITIES

Syifa Muhammad Husni[1], Ken Kinanti Purnamasari[2]

[1,2] Informatics Engineering– University Of Indonesia Computer
Dipatiukur Street 112-114 Bandung City
Email : syifamh94@gmail.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRACT

Coreference Resolution function to find two natural language expressions refer to the same entity. In previous research, Coreference Resolution is used to identify a singular pronoun and name of a person entity. The input data used is the text of the novel speak Indonesia. In a novel will contain a series of the story of one's life and the people around him. So a pronoun for a person not just a singular pronoun, however there are also many people who are being told or a plural pronoun. So on this research will be made Coreference Resolution with plural pronouns and pronoun in the Indonesia language text by using the method of Support Vector Machine (SVM) in order to be able to measure the accuracy of the results that are generated when a plural word identified. As for the input data will be processed, going through the stage of feature extraction preprocessing to to get the format to be used on the SVM method of SVM, on research, there are 7 feature that is used to detect the Coreference Resolution including i-pronoun, a pronoun, i-propernoun, j-propernoun, distance of a string, a distance of word and string match. Based on the test results, the average value of the accuracy obtained in this research is of 61.41%.

Keywords *: Coreference Resolution, Suppport Vector Machine, Natural Language Processing*, POSTAG, *Text Mining.*

1. INTRODUCTION

Coreference Resolution is the process of identifying the equality between the entities and serves to find two natural language expressions that refer to the same entity [1]. Coreference Resolution has been explored in various languages including bahasa Indonesia. In previous research, Coreference Resolution for Indonesia have been built, namely to identify the people who are being told on Indonesia, said the language of the text that could be identified only for pronouns and singular entity name someone. In his research, which is the data input was novel language of Indonesia. The novel is a long prose essay and contains a series of stories surrounding someone with that accentuates the character and nature of each behavior [2].

In previous research, the method of Support Vector Machine (SVM) is used for the novel the language of Indonesia. In the study used 2 trainer data in the form of the novel with the total number of pairs a number 2788 couples coreference and the test data in the form of a novel excerpt 5 contains a pronoun or a named entity, resulting in an average accuracy of 50.14% with highest accuracy and lowest accuracy 63% 38.8% [3].

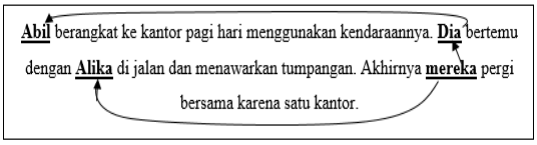
Based on the above description, the value of accuracy on Coreference Resolution using plural pronouns are unknown, so this research will do the addition of detection on the plural pronoun with the training and test data from Indonesia-language novel. The method to be used for the classification is a method of Support Vector Machine (SVM).

This study intends to develop an appropriate koreferensi detector system in a sentence. As for the goal to be achieved from this research is to know the value of accuracy on Coreference Resolution by providing a plural noun for detection using SVM method.

2. RELATED WORK

2.1 Coreference Resolution

Coreference resolution is a subtask of Natural Language Processing (NLP), and also the liquid process to identify two natural language expression that refers to the same entity. Coreference resolution is to identify the purpose of the equality between the entities or between pronouns and the entities referred to in the introduction phase in named entity [1]. Here are illustrations of Coreference resolution in this research.



Pictures 1. Examples Of Coreference Resolution

2.2 Preprocessing

In the development of a system of Coreference Resolution, needs to be done needs to be done so that the stages of data input can be processed by the system. The stage needs to be done is the stage for process data that has not been structured so that be structured according to the needs of the data before it is processed at a later stage or commonly called by preprocessing. In this study the preprocessing performed is Case Folding, Filtering, Tokenizing, and POS Tagging.

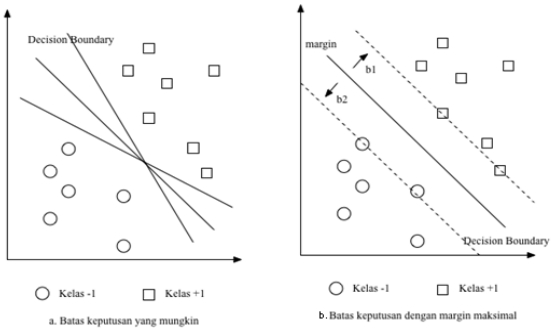
2.3 Feature Extraction

On the research data that has been modified in accordance with the needs of the data system will be in ektrak using some of the features [1]. Where extraction is used to retrieve important information from a data to be used in the classification of SVM. This system uses several features that are useful to determine whether or not two expressions of coreference, feature used that is.

a. Feature to describe words i (i-Pronoun and i-Propernoun).
b. Feature to describe words j (j-Prononun and j-Propernoun).
c. Features for the relation between the word *(Distance Of Word, Distance Of Sentence, dan String Match).*

2.4 Support Vector Machine

Support Vector Machine (SVM) algorithm is one of the most popular Machine Learning for classification and regression. The concept of a simple SVM can be explained as an effort to find the best hyperplane that serves as a dividing two classes in the input space, a dividing hyperplane can be found by measuring the magin, margin is the closest distance between hyperplane with data from each class. The subset of data called the closest training with support vector [4]. The following is an illustration of the SVM.



Pictures 2. The Decision Limits The Data Set

SVM is used to find the function of the separator (klasifier) or optimal hyperplane that can separate the two sets of data. To locate the separator function is to maximize the margin or the distance between two sets of objects from different classes and two sets of such objects is support vector. Linear functions is a separator in the linear SVM. Training data will be declared by $(x_i y_i)$, the feature set for the data class trainer to-i stated by $x_i = \{x_1, x_2, \ldots, x_{iq}\}$, and label classes declared by $y_i \in \{-1,1\}$. *Hyperplane* the separator is defined by the following equation.

$$wx_i + b = 0 \quad (1)$$

Data x_i divided into two classes. The data include class -1 (negative samples) defined as the vector that satisfies the inequality below.

$$wx_i + b < 0 \; untuk \; y_i = -1 \quad (2)$$

While the data included classroom +1 (positive sample) is defined as the fektor that satisfies the inequality below.

$$wx_i + b > 0 \; untuk \; y_i = +1 \quad (3)$$

Where:
x_i = data input
y_i = given label
w = the value of the normal field
b = relative position of the field against the center coordinates

The parameters w and b are parameters to search for the value. If the label data yi =-1, then the following equation will be delimiter.

$$wx_i + b \leq -1 \quad (4)$$

Whereas if the data label yi = + 1, then dividing into the equation below.

$$wx_i + b \geq +1 \quad (5)$$

The values xi and yi can be searched when the value of each pair from the results of the extraction of features and class initialization has been obtained. The result of the extraction of the features altered into a form data format SVM (x), whereas the data classes into the label data SVM (y).

The first step that is how to get the value of αi is changing any input data into vector values (*support vector*) = (y x). Kemudian nilai

Then the vector values of each input data entered into the equation the kernel trick phi $(\emptyset)$.

$$\phi\begin{bmatrix}p_i\\v_i\end{bmatrix} = \begin{cases}\sqrt{{p_n}^2+{v_n}^2} > 2, maka \begin{bmatrix}\sqrt{{p_n}^2+{v_n}^2}-p_i+|p_i-v_i|\\ \sqrt{{p_n}^2+{v_n}^2}-v_i+|p_i-v_i|\end{bmatrix}\\ \sqrt{{p_n}^2+{v_n}^2} \leq 2, maka \begin{bmatrix}p_i\\v\end{bmatrix}\end{cases} \quad (6)$$

The value of p can be obtained from the equations of linear kernel for the value of x below.

$$p = \sum_{i=1}^{n} {x_i}^T x_j \ (i,j = 1,\dots,n) \quad (7)$$

The value of v can be obtained from linear kernel equation for y value here.

$$v = \sum_{i=1}^{n} {y_i}^T y_j \ (i,j = 1,\dots,n) \quad (8)$$

The calculation result of the substitution of the value of the x and y values into equation 6 rated bias = 1 to get the optimal perpendicular distance considering the positive vector. Then search parameters *αi*, by first looking for the value of the function of each abstract using equation 9. With that in mind, i = 1, .., n, *αi* is searchable on linear equations using equation 10 here.

$$\sum_{i=1,j=1}^{n} \alpha_i {S_i}^T S_j \quad (9)$$

$$\sum_{i=1,j=1}^{n} \alpha_i {S_i}^T S_j = v_i \quad (10)$$

After the parameters *αi*, then enter into the equation here.

$$\widehat{W} = \sum_{i=1,j=1}^{n} \alpha_i S_i \quad (11)$$

The results obtained from equation 11, the next 12 into the equation is used to get the value of w and b.

$$y = wx + b \quad (12)$$

Such that w and b value didapatkanlah (value hyperplane) to classify the second class.

If a function satisfies Mercer's theorem, which States that the resulting kernel matrix must be semi semi of the definite positive, then the function ¬ can be said as the kernel functions [5]. In this study the kernel used is the Linear Kernel.

$$K(x_i, x) = {x_i}^T x \quad (13)$$

The process of classification method of machine used in this system are the LIBSVM, library created by Garbriel Couto.

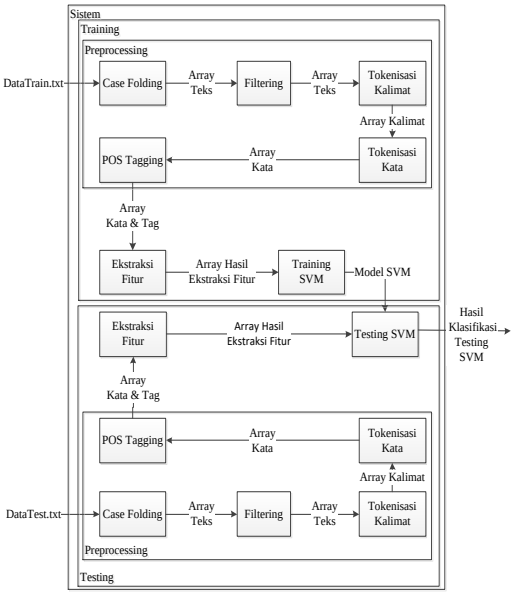
2.5 Analysis of the Problem

Based on observations of the back against the novel was used in the previous studies, many found the plural pronoun can not be detected by the system. While the plural word itself is included in these types of pronouns (personal pronouns). The training data used in previous studies only conduct detection against a single pronoun and name of the person, so that the text that has the plural pronouns are not used when testing. It makes the Coreference Resolution before being limited in detecting, as a plural pronoun to be detected as a kind of Personal pronouns are Pronouns (PRP). Therefore, this research will utilize features of i-pronoun, i-propernoun, j-pronoun and j-propernoun to find out the type of words and features distance of word (DOW), distance of sentence (DOS), and string match for relationships between words that will in comparison.

2.6 Analysis Of The Solution

Based on the explanation of the issues already described earlier, with Coreference Resolution system, he had made in building the system, there will be several stages to be performed i.e. the stage of training for training data, the data for testing, testing stage classification using SVM and calculation accuracy of Coreference Resolution.

The following is a general overview of the system Corefence Resolution built.



Pictures 3. Overview Of The System

2.6 Analysis Of Data Input

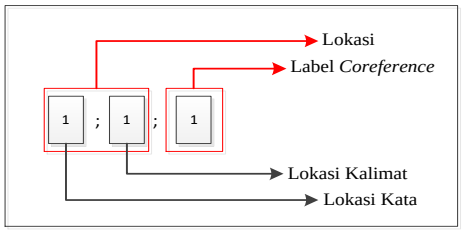
The input data will be processed by the system is data training and testing, data on previous research, labeling the data conducted training to certify coreference to be used in the classification. Here is an example snippet of text from the novel "Heaven Breech" which will be used as training data.

Table 1. Text Input Data Training

Dan yang lebih penting **ia** tak ingin ikan piaraannya kepanasan dan **Zaenab** hangus terbakar. "**Kita** hanya butuh hujan. **Kita** harus memohon Allah agar memberikan sihir hujan!" **Zaenab** mendesis lagi. "Hujan? Untuk apa?" tanya **Kufah** tak mengerti maksud **Zaenab**.

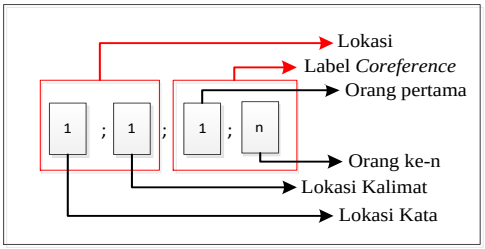
The text of the novel at the top, there are seven pronouns consisting of four person pronoun (Zaenab, Zaenab, Kufa, Zaenab) and three single pronoun (ia, Kita, Kita).

To be able to detect the words, then it should be labeled as key for classification. Given label consists of three numbers, separated by semicolon (;) to pronoun singular and four digits or more for a plural pronoun are separated by a semicolon (;). The example in the following figure.



Pictures 4. Single Label Data [3]

Format data label is used to provide single sign on to a single name or pronouns such as I, she, he, we, Zaenab, Kufah.



Pictures 5. Plural Label Data

Format data label plural is used to give a sign to the plural pronoun as they say, we, us, y'all. To get the location of words and sentences then it can be seen in the following table.

Table 2. The location of the sentences and Words

No	Teks Novel
1	Dan yang lebih penting **ia** tak ingin ikan 1 2 3 4 **5** 6 7 8 piaraannya kepanasan dan **Zaenab** hangus 9 10 11 12 13 terbakar. 14
2	"**Kita** hanya butuh hujan. **1** 2 3 4
3	**Kita** harus memohon Allah agar **1** 2 3 4 5 memberikan sihir hujan!" **Zaenab** 6 7 8 **9** mendesis lagi. 10 11
4	"Hujan? Untuk apa?" tanya **Kufah** tak 1 2 3 4 **5** 6 mengerti maksud **Zaenab**. 7 8 **9**

Coreference labels for the grant that is based on the first person pronoun appears without reference have a coreference. So if the first person pronoun appears to be given the label of coreference-¬ is 1, when the next pronoun appears with the same people then keep 1, and if it is different then being 2 and onwards depending on how many that will appear. the following input data for coreference labels above.

Table 3. Coreference Label

No	Words	Coreference Label
1	Ia	1
2	Zaenab	2
3	Kufah	3
4	Kita	1, 2

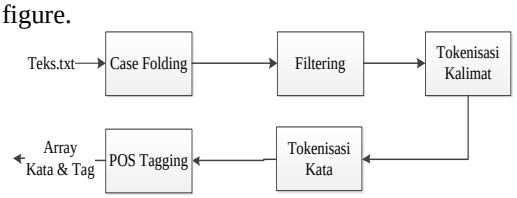
Format data label singular and plural can be seen in the following table, where for the location of words and sentences can be seen in table 2.

Table 4. Data Label

No	Words	Location		Coreference Label	Data Label
		Words	Sentence		
1.	ia	5	1	1	5;1;1
2.	Zaenab	12	1	2	12;1;2
3.	Kita	1	2	1, 2	1;2;1;2
4.	Kita	1	3	1, 2	1;3;1;2
5.	Zaenab	9	3	2	9;3;2
6.	Kufah	5	4	2	5;4;3
7.	Zaenab	9	4	3	9;4;2

2.7 Preprocessing

Preprocessing the input data must be done that will be processed, in order to get the maximum results from each stage to the next stage. Preprocessing phase to be performed i.e. case folding, filtering, tokenizing sentences, tekenizing said, and the POST tagging. Preprocessing stage can be lilhat in the following figure.



Pictures 6. Preprocessing

2.8 Feature Extraction

Feature extraction is a stage to get the data needed information from previous data. Keywords used for the extraction of this feature is the pronouns people such as name, singular, and plural. On the research of this plural word added to add keyword information. Because the word in plural pronouns, here are 7 feature that will be used for the process of classification using SVM method is as follows.

Tabel 5. Ekstraksi Fitur

The feature to describe words *-i* .	
i-Pronoun	describe the value of i-pron detected the POST tags as PRP. If true then the value 1, otherwise the value is 0.
i-Propernoun Pharse	describe the value of i-prop being detected the POST tags as the NNP. If true then the value 1, otherwise the value is 0.
The feature to describe words *-j*	
j-Pronoun	describe the value of j-pron detected the POST tags as PRP. If true then the value 1, otherwise the value is 0.
j-Propernoun Pharse	describe the value of j-prop being detected the POST tags as the NNP. If true then the value 1, otherwise the value is 0.
The feature to describe words *-i* and words*-j*	
Distance Of Words (DOW)	Describes the distance between the words i and j, this value will be derived from the word distance j reduced the distance of said i then divided by the number of words in the text document.
Distance Of Sentence (DOS)	medeskripsikan the distance between the sentences i and j, this value will be obtained from a range of sentences reduced sentence distance i j and then divided by the number of words in the text document.
String Match	describe the string on i and j, if string on i and j are the same then the value is 1, if it is different then the value is 0.

Extraction of feature is divided into two, extracting features for data training and testing data, the following are examples of data input for training. Here is a sample input data for training.

Table 6. Example Input Data Extraction Features

Data Input
dan yang lebih penting **ia** tak ingin ikan piaraannya kepanasan dan **zaenab** hangus terbakar. **kita** hanya butuh hujan. **kita** harus memohon allah agar memberikan sihir hujan **zaenab** mendesis lagi. hujan untuk apa tanya **kufah** tak mengerti maksud **zaenab**.

At the stage of extraction feature for training, data labels. txt is utilized to indicate words that will be detected, so it can make a pair of words-i and j. In defining the word spouse-i and j, the location of the words i>-j (i > j). Where is the location of the word in question is the total number of words in the text of the novel can be seen in the table above.

For feature i-pronoun, i-propernoun, j-pronoun, and j-propernoun, pair i and j will be in checking based on the tags that have been given by the process of POS Tagging. For example the word **Ia**$_1$ (j) having tag **"PRP",** so i-pronoun given the value 1 and i-propernoun given a value of 0. For the word **Zaenab**$_1$ (i) has the tag "NNP", then a pronoun are given a value of 0 and j-propernoun is given a value of 1.

The feature Distance Of Words (DOW) is used to find the value of word spacing, word-spacing i and j start from the beginning of the word in the text. For example, namely the words i **zaenab** with location = 12 words and words he said location j = 5. As for the formula that is used as an example here.

$$DOW = \frac{\text{Lokasi Kata } i - \text{Lokasi Kata } j}{\text{Jumlah kata dalam dokumen teks}} = \frac{12-5}{38} = 0.18$$

The feature Distance Of Sentence (DOS) is used to find the value of distance between sentences, sentences i and j starts from the beginning of the sentence in the text,. For example, namely the words i zaenab location sentence = 1 and j = sentence he location 1. As for the formula used is as follows.

$$DOS = \frac{\text{Lokasi Kalimat } i - \text{Lokasi Kalimat } j}{\text{Jumlah kalimat dalam dokumen teks}} = \frac{1-1}{4} = 0$$

The following example from the data feature extraction stage training, who've gone through Preprocessing stage.

Table 7 Sample Data Feature Extraction Training

No	I		J		feature i		feature j		Relation		
	word-i	tag-i	word-j	tag-j	i *pron*	i *prop*	j *pron*	j *prop*	DOW	DOS	*String Match*
1	zaenab$_1$	NNP	ia$_1$	PRP	0	1	1	0	0.18	0	0
2	kita$_2$	PRP	ia$_1$	PRP	1	0	1	0	0.26	0.25	0
3	kita$_2$	PRP	zaenab$_1$	NNP	1	0	0	1	0.08	0.25	0
4	kita$_3$	PRP	ia$_1$	PRP	1	0	1	0	0.37	0.5	0
5	kita$_3$	PRP	zaenab$_1$	NNP	1	0	0	1	0.18	0.5	0
6	kita$_3$	PRP	kita$_2$	PRP	1	0	1	0	0.11	0.25	1
7	zaenab$_3$	NNP	ia$_1$	PRP	0	1	1	0	0.58	0.5	0
8	zaenab$_3$	NNP	zaenab$_1$	NNP	0	1	0	1	0.39	0.5	1
9	zaenab$_3$	NNP	kita$_2$	PRP	0	1	1	0	0.32	0.25	0
10	zaenab$_3$	NNP	kita$_3$	PRP	0	1	1	0	0.21	0	0
11	kufah$_4$	NNP	ia$_1$	PRP	0	1	1	0	0.76	0.75	0
12	kufah$_4$	NNP	zaenab$_1$	NNP	0	1	0	1	0.58	0.75	0
13	kufah$_4$	NNP	kita$_2$	PRP	0	1	1	0	0.5	0.5	0
14	kufah$_4$	NNP	kita$_3$	PRP	0	1	1	0	0.39	0.25	0
15	kufah$_4$	NNP	zaenab$_3$	NNP	0	1	0	1	0.18	0.25	0
16	zaenab$_4$	NNP	ia$_1$	PRP	0	1	1	0	0.87	0.75	0
17	zaenab$_4$	NNP	zaenab$_1$	NNP	0	1	0	1	0.68	0.75	1
18	zaenab$_4$	NNP	kita$_2$	PRP	0	1	1	0	0.61	0.5	0
19	zaenab$_4$	NNP	kita$_3$	PRP	0	1	1	0	0.5	0.25	0
20	zaenab$_4$	NNP	zaenab$_3$	NNP	0	1	0	1	0.29	0.25	1
21	zaenab$_4$	NNP	Kufah$_4$	NNP	0	1	0	1	0.11	0	0

2.9 SVM Clasification

At the stage classification, each pair of words i and j are given positive values 1 or negative 1 to label (yi). The value of the label used for the positive, and coreference if not coreference if negative. At the stage of classification using SVM method, the stage will be done is to rehearse first then the data will be performed or testing

2.9.1 *Training* SVM

At this stage, the results of the ektraksi features is converted to the format of the SVM, where data are used as training data i.e. .txt and also novel lebel .txt, where the label is used to provide data for pembelajara training, or to know the relation words have the same equality, so the data will yield a positive classroom training in a coreference labels have the same class, and will be negative if different coreference labels. And the results of the training will be stored as a model for use as svm prediction data for testing

2.9.2 *Testing* SVM

At this stage of Testing, will be made a prediction of where the models resulting from the training data will be used. So data testing by classifying using the model from training data. in this study, the classification is done using LIBSVM, i.e. library SVM for PHP from Gabriel Coutco which can dikunjugi on GITHUB.

2.10 Testing The System

System testing is performed to ensure the system is running in accordance with the design of the system have been made before. System testing includes testing the functionality and testing accuracy.

2.10.1 Testing Functionality

Black box testing is used to test specific functions of the software being designed. The following is a table of plan testing black box.

Table 8 Plan Test *Black Box*

No	Actifity	The expected output	Types of tests
1	Case Folding	The text of the novel with lowercase	*Black Box*
2	Filtering	The text of the novel without any characters other than "a-z","."," -", and spasi	*Black Box*
3	Tokenizing Sentence	A collection of sentences with without delimeter point	*Black Box*
4	Tokenizing Words	The Words	*Black Box*
5	POS Tagging	The Words And Tag	*Black Box*
6	Determination Of Word Pairs *Training*	Couple Words *i* and *j*	*Black Box*
7	Determination Of Word Pairs Testing	Couple *i* and *j*	*Black Box*
8	The Determination Of The Value Of The Feature Extraction	The group *coreference* and the result is a number feature extraction	*Black Box*
9	Training SVM	Model SVM	*Black Box*
10	Testing SVM	The result of the classification of SVM coreference a pair that has the class.	*Black Box*

2.10.2 Testing Of The Accuracy

Testing accuracy on this research was conducted to find out the results the accuracy of Coreference Resolution system by adding detection for plural pronouns and using SVM method on the text speak Indonesia.

The data used as the data-i.e. from 8 Indonesia-language novels contained 126 pronouns consisting of the name of the person, the person pronoun (singular and plural), the number of pairs of words that are formed is 7875 pair. Text or stories that serve as training data has a diverse pattern, this will affect the results of the classification because the result of the extraction of features will have a lot of patterns.
Based on the results of detection of word pairs that have been analyzed and compared with the results of the system, said couples formed

perfectly. Testing system of classification of Coreference Resolution using 10 test data taken from novels and 5 by using the method of SVM, then generate the accuracy as follows.

Tabel 9. The Test Results

The Novel	File	The total number of words couple	Total	
			True	False
5 Cm	cm1.txt	55	30	25
	cm2.txt	36	24	12
Dear Nathan	dear1.txt	45	34	11
	dear2.txt	78	46	32
Garis Waktu	garis1.txt	45	25	20
	garis2.txt	45	23	22
Rindu	rindu1.txt	55	33	22
	rindu2.txt	66	43	23
Teman Hati	teman1.txt	36	27	9
	teman2.txt	91	54	37
Total		552	339	213

Here is the calculation for the value of the overall accuracy of the classification on the novel SVM method.

$$Akurasi = \frac{Jumlah\ yang\ diklasifikasi\ secara\ benar}{Jumlah\ sample\ pasangan\ kata\ yang\ diuji}\ x\ 100\%$$

$$= \frac{339}{552}\ x\ 100\%$$
$$= 61.41\ \%$$

Based on the table above, the test results of Coreference Resolution System using SVM method of generating Accuracy 61.41%

1. The pattern of training data that was generated there is still a shortage for some patterns of the story and some pronouns so classifying is not appropriate as follows..
   a. introduction of the pronoun **kau** in file garis2 .txt.
   b. Exposure story on file, there are .txt cm2 pattern which mentions plural words in adjacent as **kau** guys and **kalian** who have different coreference object.
   c. pronoun who appeared first before the entity named many experienced errors in classification.
2. Yet it does for detection of pronouns as the zealous **budakmu**, because the words that belong to often appear, and when separated, words **budak** will be detected as a pronoun, and the word **mu** can be detected as the PRP candidates coreference..

3 CONCLUSION

This section will explain the conclusions and suggestions of the research.

3.1 Conclusions

Based on the results of the research that has been done, it can be known that Coreference Resolution language Indonesia containing multiple entities can be made and used as a pronoun for the detection of people who serve as the coreference. On this research uses 8 data training in the form of excerpts novels that make up the 7875 pair words and 10 test data for a total of 552 pairs of words that contain multiple entity, pronouns, and named entity. Accuracy resulting from these studies amounted to 61.41%.

3.2 Advice

Based on the results of the testing that has been done, then the expected suggestions on future development for Coreference Resolution system that is.

1. Maximize the performance of the post Tagging to be able to recognize the type of pronoun that has the word suffixes.
2. Maximize the performance of SVM in finding the best for handling hyperlane Coreference Resolution.
3. Add Rule-Based processes to be able to detect the pronoun of ownership can be made as coreference.
4. Add a feature Named Entity Recognation (NER), to be able to recognize named entity such as a person's name, organization name and place name.

BIBLIOGRAPHY

[1] J. Gao, F. Kong, P. Li, and Q. Zhu, “Research of noun phrase coreference resolution,” *Proceedings. 2011 International Conference Asian Language. Processing.* 2011.

[2] Kamus Besar Bahasa Indonesia. 2017. https://www.kbbi.web.id/, 20 Februari 2018 15.30

[3] Ayu Linggar Sari, “Coreference Resolution Menggunakan Metode SVM Pada Novel Berbahasa Indonesia,” Perpustakaan Unikom 2017.

[4] L. Assaffat, “Analisis Akurasi Support Vector Machine Dengan Fungsi Kernel Gaussian Rbf Untuk Prakiraan Beban Listrik Harian Sektor Industri,” *Momentum*, 2015.

[5] A. D. Putri, “Klasifikasi Dokumen Teks Menggunakan Metode Support Vector Machine Dengan Pemilihan Fitur Chi-Square,” *skripsi. Inst. Pertan. Bogor, Bogor*, 2013.