Benchmarking Mi-NER: Malay Entity Recognition Engine

Thenmalar Ulanganathan[1], Ali Ebrahimi[1], Benjamin Chu Min Xian[3], Khalil Bouzekri[1],
Rohana Mahmud[2], Ong Hong Hoe[1]
[1]Artificial Intelligence Lab
MIMOS Berhad
Malaysia
email: thenmalar.nathan | ali.ebrahimi | khalil.ben | hh.ong@mimos.my
[2]Department of Artificial Intelligence
Faculty of Computer Science & Information Technology
University of Malaya
rohanamahmud@um.edu.my
[3]benjamin-chu@live.com

***Abstract*— Named entity recognition (NER) is a process of recognizing, identifying, and extracting useful entities, like person, location and organization for information mining from unstructured texts. This paper presents (Mi-NER), a Malay language Named Entity Recognition engine that is developed using a probabilistic approach. The results of benchmarking Mi-NER against existing systems are presented in this paper. In addition, the details of the experimental work are highlighted and discussed. Precision, Recall and F-Measure have been used to measure the results for this evaluation.**

Keywords-Benchmarking; Malay Language; Natural Language Processing; Named Entity Recognition.

I. INTRODUCTION

In recent years, development in semantic analysis of unstructured text has triggered many applications in Text Mining, Summarization, Text Understanding, Information Retrieval and Extraction [1][2].

Named Entity Recognition (NER) is a subtask of information extraction which involves identification of proper nouns in texts, called named entities, and classification of these named entities into a set of pre-defined categories of interest (e.g., Person, Location, Organization) [3]. The main goal of NER is to reduce the manual annotation of named entities in texts by human annotator which is a time-consuming and laborious process. However, in order to automate this process, machines have to be trained, as they need to analyze and understand the content of the text before being able to recognize the named entities. Machine learning techniques including statistical and probabilistic methods have been used to successfully build automated NER engines [4]-[7].

Building a machine learning model necessitates an NER-annotated corpus to be able to detect the correct entity types for new words/phrases based on the context. Such corpora are available for the major languages, such as English, Chinese, Spanish, and Hindi [8]. However, due to the lack of linguistic resources for Malay language, training corpora have to be built from scratch to train NER models. In this paper, a Malay NER engine called Mi-NER is presented and compared with existing Malay NER engines. A manually-built corpus is used to train the NER model. Another two manually-built corpora are used to test the models.

This paper is structured as follows: Section 2 describes the related work on existing NER systems; Section 3 highlights the proposed Machine Learning model of Mi-NER; Section 4 shows the experimental results; Section 5 discusses the results of Mi-NER compared to the existing systems. Finally, Section 6 concludes this paper.

II. RELATED WORK

NER systems exist for various languages, such as English, Dutch, Arabic, Chinese, etc. but only few can be found for Malay language.

In their work, Fong, Y. S. et al. [9] use several text processing modules from A Nearly-New Information Extraction (ANNIE) system. They proposed a method for creating rules and gazetteers for Iban language which is one of the 63 indigenous languages of Sarawak, Malaysia, according to the Dewan Bahasa dan Pustaka (Institute of Language and Literature DBP) [10]. The system includes a tokenizer, a manually-built gazetteer (entity dictionary), a sentence splitter and a part-of-speech (POS) tagger and the use of several rules written with Java Annotation Pattern Engine (JAPE). In this work, rules are designed to detect several named entities including Person, Organization, and Location, as well as other types of entities, such as Time, Monetary, Date and Percentage.

Sharum, M. Y. et al. [11] use a name index and regular expressions to recognize entities only limited to people's names from Malay texts. Therefore, by focusing on minimal techniques of NER to recognize people's names, they showed that the approach of recognizing people's names can be performed and returns precise results. On the other hand, Rayner Alfred et al. in [12] use rule-based approach to identify named entities in Malay texts. The approach uses Rule-Based Part of Speech (RPOS) tagger, which is Malay

 ISBN: 978-1-61208-542-5

Rule-Based POS tagger that applies a POS tag dictionary and affixing rules in order to identify the word definition [12]. The work revolves around designing rules based on the POS tags. For instance, when the POS tag for a particular word is referring to a proper noun, then a specific rule will be applied to this word in order to determine whether it is a named entity or not. In this work, these rules are designed to detect three major types of named entities, which are Person, Organization and Location.

Semantria [13] is able to process Malay texts and its NER feature is able to automatically extract proper nouns like persons, places, or companies from texts. It comprises of a POS module to tag each of the word tokens with a corresponding POS tags. Subsequently, it performs a series of algorithms to extract relevant named entities from texts [13].

In this paper, we compare the results generated from Mi-NER against the results from Rule-Based NER system in [12] and Semantria [13] for our experimental evaluation.

III. ENTITY RECOGNITION MODEL

Natural language Processing (NLP) uses Linear-Chain Conditional Random Fields (Linear-Chain CRF) in many sub sequential text processing task including NER, POS, and word segmentation [14]. Mi-NER uses Linear-Chain CRF machine learning technique to train its NER model [14]. CRF is a popular probabilistic method for structured prediction. It is a technique which has been applied to several domains including bioinformatics, computer vision and text processing. Sha and Pereira [15] created one of the first large-scale applications of CRFs by matching state-of-the-art performance on segmenting noun phrases in text.

After that, linear chain CRFs has been applied on variety of problems in NLP including NER. In NER models, all of the named entity labels are independent. However, the named entity labels of neighboring words are dependent (e.g., Los Angeles: location, Los Angeles Times: organization). One way to relax this independence assumption is to arrange the output variables in linear chains which are CRF.

A. Training

In this work, we extracted data from news and non-news sources, respectively Bernama [16] news archive and social media (including tweets, blogs and wikis). Those two sources are used in order to build a training corpus which contains a total of 275,322 tokens. 70% of the training set is collected from news and the remaining part is reserved for non-news. The data is annotated by native speakers and verified by two linguistic experts.

The process of building Malay NER model is further described in Fig. 1. There are three main steps including Preprocessing, Annotation and Generation.

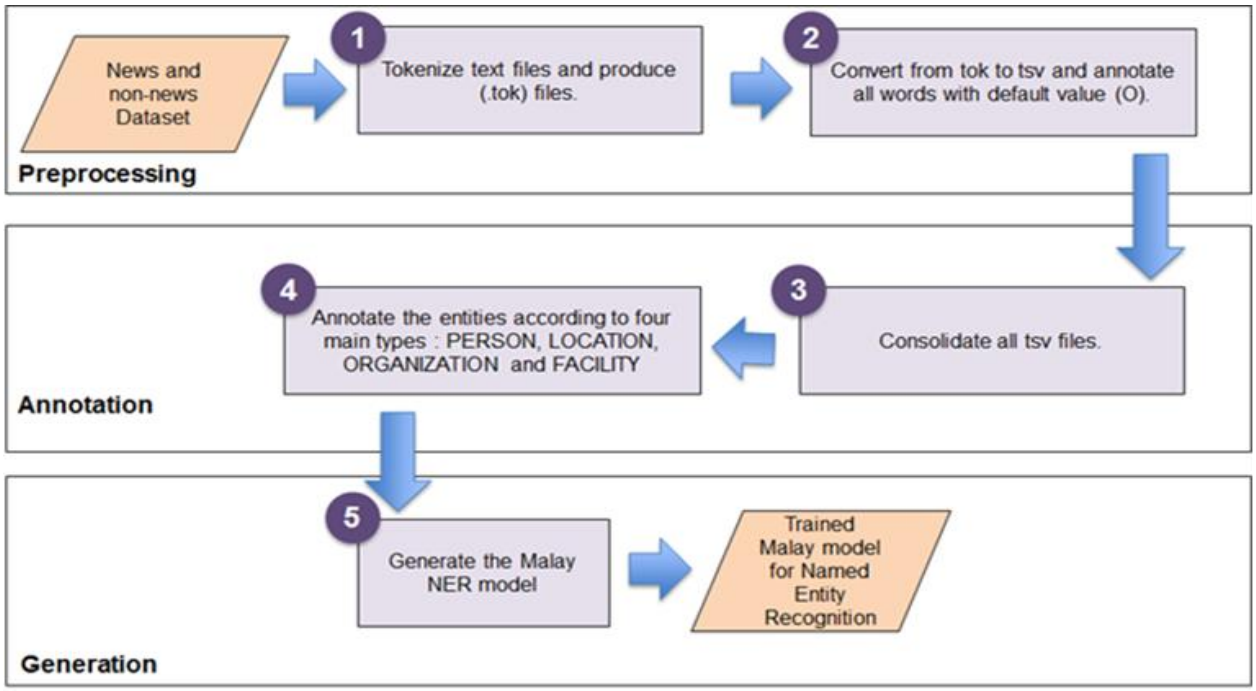


Figure 1. Malay Named Entity Recognition Model Training Process

 ISBN: 978-1-61208-542-5

In the Preprocessing phase, we run the tokenizer to generate a ".tok" file with a token per line. Subsequently, another script is executed to generate a ".tsv" file (Tab Separated Values) where all of these tokens are initialized with a default value "O" (refer Fig. 2). The ".tsv" file is used by Malay native speakers to annotate which are the relevant named entities based on the sentence context. However, ".tsv" file format allows us to save data in which tokens and their entity types are separated by Tab. Four named entity types are considered: PERSON, LOCATION, ORGANIZATION and FACILITY. The output is shown in Fig. 3.

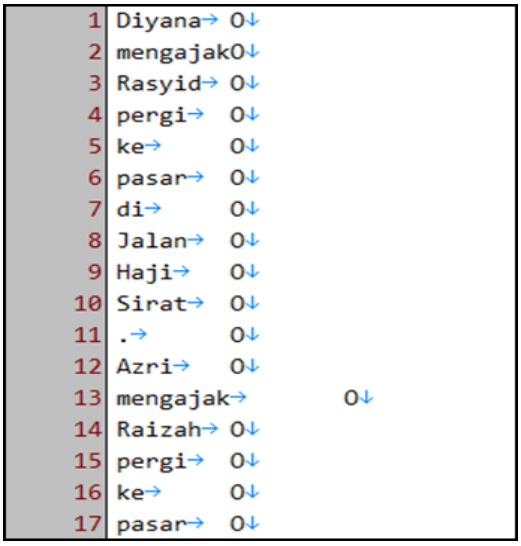


Figure 2. TSV file with initialized value

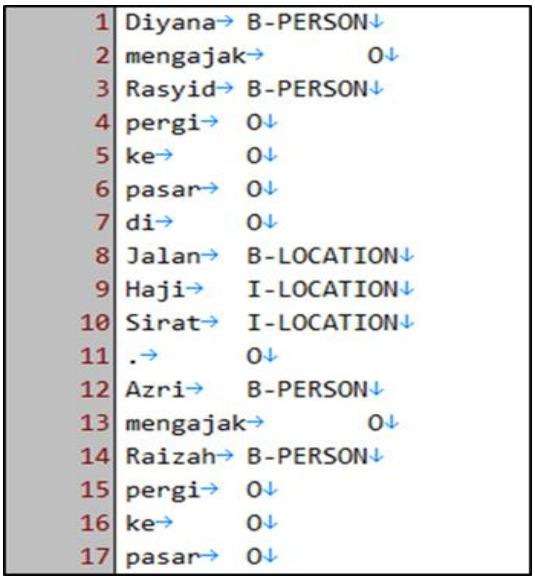


Figure 3. Annotated TSV file by Malay native speakers

The "B" notation is used before any named entity type to represent the beginning element for the named entity. If the name entity contains only one element, "B" notation will represent the beginning as well as ending of the element. Otherwise, if there is more than one element (token) in that named entity, an "I" notation will be assigned to the subsequent token(s) as shown in Fig. 3 (for example, Jalan Haji Sirat).

There are some terms/words which may be used for different purposes. A case in point would be the name entity "Tun Razak" which is a name used to call a person (commonly known as former prime minister of Malaysia as well as his son who is current prime minister), location entity (Jalan Tun Razak) as well as a facility entity(Universiti Tun Abdul Razak).

B. Testing

To benchmark the accuracy of the proposed Mi-NER, two collections of tokenized datasets have been created. These two annotated Malay datasets consist of 500 articles from news and non-news sources with 250 articles in each source, for a total of 8649 tokens for the first dataset and 9077 tokens for the second dataset. The articles for the first dataset are extracted from Harian Metro [17] and Utusan Malaysia news archive [18], whereas the second dataset contains selected articles from the Malay dataset developed in [19] by Su'ad Awab, which consists of different categories including art, economics, education, health, information technology, law, literature, sport, and science.

Both datasets are built in the same way as the training dataset (refer Fig. 1). After that, the results are checked and verified by another two linguistic experts. The final results are used as our gold standard to evaluate the results of the proposed Mi-NER engine, rule-based NER engine [12] and Semantria [13].

IV. EXPERIMENTS AND RESULTS

The results of Mi-NER are compared against the rule-based Malay ER proposed in [12] and Semantria [13].

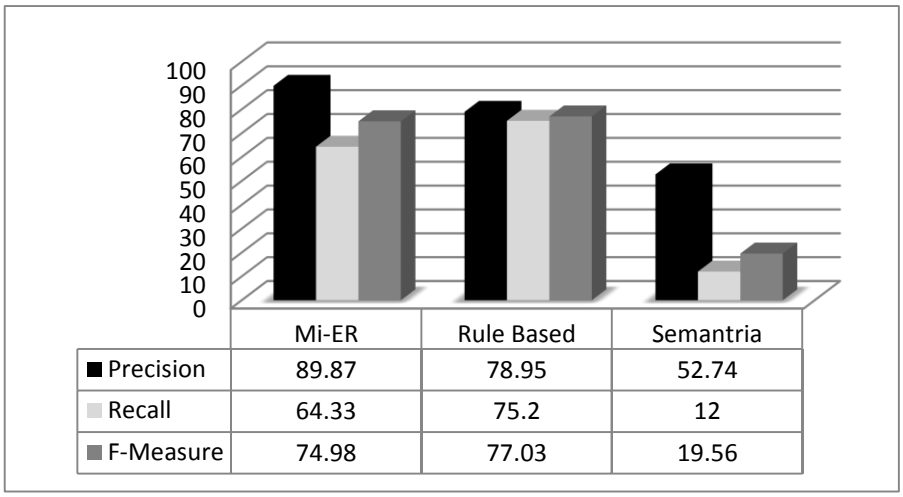
Precision, Recall and F-Measure

CoNLL-2002 [20] shared task is the established approach of evaluating NER systems by using the following measures: Precision, Recall and F-measure. The precision measures the percentage of entities found by the algorithm that are correct. Recall is based on the percentage of named entities defined in the corpus that were found by the evaluation program. F-measure is used to measure the accuracy of precision and recall measures. F-measure can be interpreted as a weighted average of the precision and recall.

Fig. 4 shows the Precision, Recall and F-measure scores resulting from our evaluation of these systems for news dataset. Mi-NER demonstrated highest Precision with the value of 89.87% followed by Rule-Based ER with 78.95% and Semantria with 52.74%. Mi-NER and Rule-Based ER have almost similar F-measure scores, but generally Rule-Based ER performs better in Recall. Semantria has the lowest scores compared to the rest for Precision, Recall and F-measure.

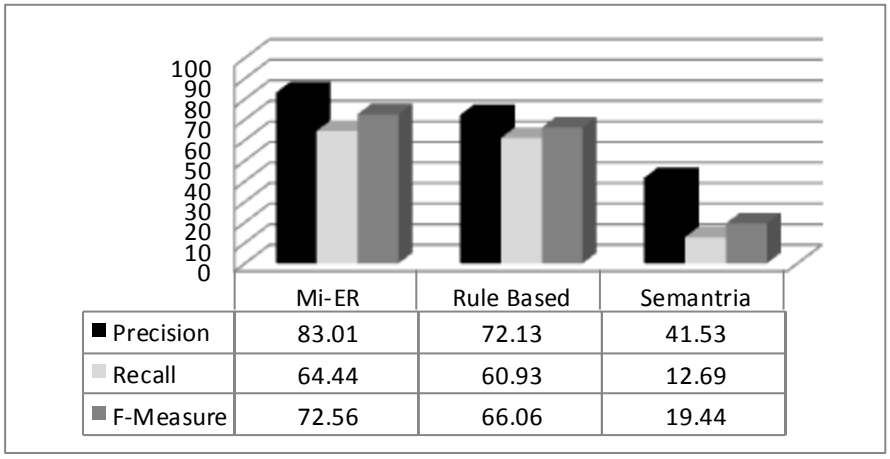
 ISBN: 978-1-61208-542-5

Fig. 5 displays the Precision, Recall and F-measure scores resulting from our evaluation of the three systems for non-news dataset. As can be seen in Fig. 5, Mi-NER demonstrated highest Precision, Recall and F-measure score with the value of 83.01%, 64.44% and 72.56% respectively. For this dataset, Mi-NER performs better than Rule-Based ER for all the scores. Semantria has the lowest scores compared to the rest for Precision 41.53%, Recall 12.69% and F-measure 19.44%.For this dataset, Mi-NER performs better than Rule-Based ER for all the scores. Semantria has the lowest scores compared to the rest for Precision 41.53%, Recall 12.69% and F-measure 19.44%.



	Mi-ER	Rule Based	Semantria
Precision	89.87	78.95	52.74
Recall	64.33	75.2	12
F-Measure	74.98	77.03	19.56

Figure 4. News Dataset Evaluation Results



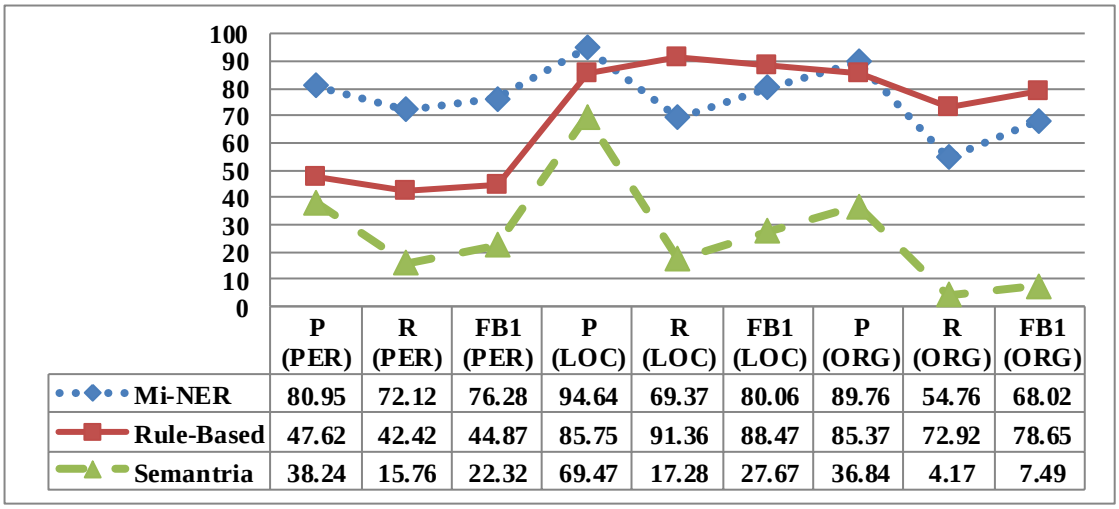
	Mi-ER	Rule Based	Semantria
Precision	83.01	72.13	41.53
Recall	64.44	60.93	12.69
F-Measure	72.56	66.06	19.44

Figure 5. Non-News Dataset Evaluation Results

Fig. 6 illustrates the Precision, Recall and F-measure scores specifically in terms of Person, Location and Organization entities comparing each system in our evaluation for news dataset. Referring to the data in Fig. 6, generally Mi-NER performs better than Rule-Based ER in terms of recognizing Person entities but Rule-Based ER performs better than Mi-NER in recognizing Location and Organization entities for both Recall and F-measure scores. On the other hand, Semantria has the lowest scores for all Precision, Recall and F-measure for recognizing Person, Location and Organization entities.

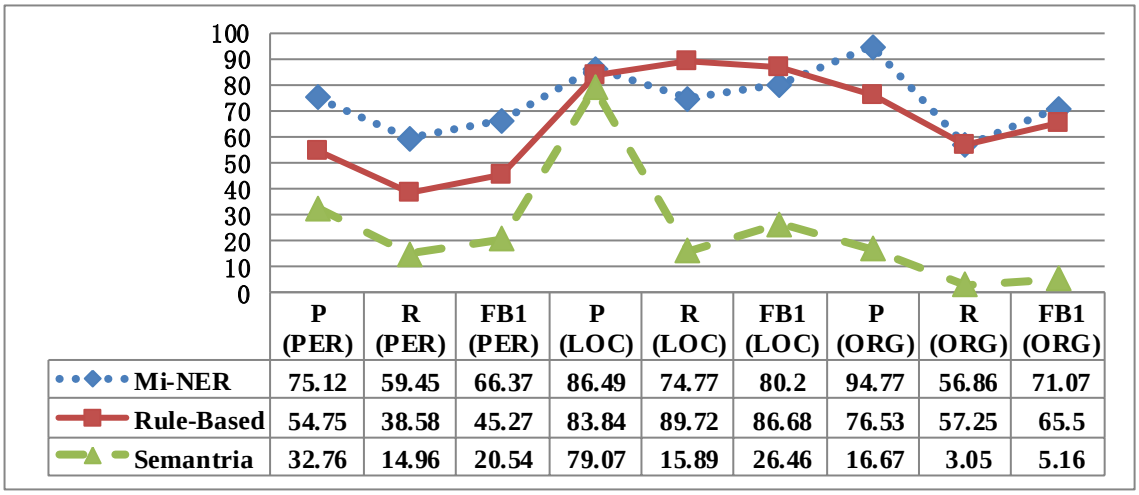
Fig. 7 represents the Precision, Recall and F-measure scores specifically in terms of Person, Location and Organization entities comparing each system in our evaluation for non-news dataset. Based on the results, generally Mi-NER performs better than Rule-Based ER in terms of recognizing Person and Organization entities but Rule-Based ER performs better than Mi-NER in recognizing Location entities for Recall and F-measure scores. On the other hand, Semantria has the lowest scores for all Precision, Recall and F-measure for recognizing Person, Location and Organization entities.

 ISBN: 978-1-61208-542-5



	P (PER)	R (PER)	FB1 (PER)	P (LOC)	R (LOC)	FB1 (LOC)	P (ORG)	R (ORG)	FB1 (ORG)
Mi-NER	80.95	72.12	76.28	94.64	69.37	80.06	89.76	54.76	68.02
Rule-Based	47.62	42.42	44.87	85.75	91.36	88.47	85.37	72.92	78.65
Semantria	38.24	15.76	22.32	69.47	17.28	27.67	36.84	4.17	7.49

Figure 6. News Corpus Analysis for Person(PER), Location (LOC), and Organization(ORG).
P: Precision, R: Recall, FB1: F-Measure



	P (PER)	R (PER)	FB1 (PER)	P (LOC)	R (LOC)	FB1 (LOC)	P (ORG)	R (ORG)	FB1 (ORG)
Mi-NER	75.12	59.45	66.37	86.49	74.77	80.2	94.77	56.86	71.07
Rule-Based	54.75	38.58	45.27	83.84	89.72	86.68	76.53	57.25	65.5
Semantria	32.76	14.96	20.54	79.07	15.89	26.46	16.67	3.05	5.16

Figure 7. Non-news Corpus Analysis for Person(PER), Location (LOC), and Organization(ORG).
P: Precision, R: Recall, FB1: F-Measure

V. DISCUSSION

Experiments conducted show that Mi-NER has the highest precision for both the datasets and the highest recall and F-measure scores when tested using the non-news dataset. Experiments also show that the recall of Mi-NER is almost the same for both datasets. However, the Rule-Based ER has a noticeably better recall for the news dataset and thus a higher F-measure score.

Based on the results presented in Fig. 6 and Fig. 7, Mi-NER has the best results for detecting Person entities for both datasets. In fact, all types of entities (i.e., Persons, Locations and Organizations) are detected by Mi-NER with roughly close values. This is mainly because Mi-NER is trained equally on the different types of entities unlike the case of other NER systems. For example, the Rule-Based ER has significantly higher scores for locations and organizations. This is because it supports the detection of locations and organizations with additional dictionaries containing large numbers of locations and organizations to be detected. The dictionaries in addition to the rules used to detect locations and organizations significantly boost the scores for these types. Persons are detected in the Rule-Based ER using only the available rules which fail to cover all possible person names. As a result, the Rule-Based ER detects locations better than Mi-NER. However, Mi-NER outperforms the Rule-Based ER in terms of detecting organizations for the non-news dataset as shown in Fig. 7. News dataset is expected to have a large number of organization entities unlike the non-news dataset. This also affects the training of Mi-NER where it has a higher

 ISBN: 978-1-61208-542-5

chance to learn features about fewer numbers of organizations in the non-news dataset.

Semantria locates entities using a Malay POS tagger and a predefined set of entities. Its list of entities can be customized by user's queries to detect the entities that most concern the user. Results show that Semantria has the lowest figures for both datasets. This can be due to problems with the used Malay POS tagger or limitations in the list of entities for Malay language. However, it shows a high precision of detecting locations which indicates that Semantria uses a rich list with sufficient number of location entities unlike other entity types.

Based on our finding, one of the advantages of Mi-NER is to detect the organizations based on the short forms of their suffix like "Sdn" as the short form of Sendirian and "Bhd" as the short form of Berhad and etc. The strategy to add more variation of different types of organization's suffix would enable the system to be more powerful in order to find Person and Organization entities. This strategy applied for Person entities by adding more people's name with different salutations, such as Dato, Datuk, Dato Seri, Datuk Seri, etc. This strategy helps Mi-NER system to differentiate the Location and Person entities, as some persons' name are assigned for a location and by introducing a variety of salutations, Mi-NER system is able to get better results for these cases.

VI. CONCLUSION

In this paper, we have presented a Linear-Chain CRF machine learning technique to train Mi-NER model and benchmark against the rule-based Malay NER proposed in [12] and Semantria [13].

From a qualitative comparison point of view, Mi-NER performs better on the Precision, yet it needs more improvement on the Recall. However, we showed that a statistical approach to develop a NER Engine performs better on some aspects of NER than other rule-based entity recognizers especially when using a large corpus for training. However, there are some challenges with building Malay NER model which caused by lack of online linguistic resources including Malay words have many derivative words that change the syntactic meaning as well as insufficient and limited digital resources. Those factors may restrict applying of machine learning methods and semantic approaches.

Consequently, some improvement will be made especially to improve the current results of Mi-NER in terms of recall by training the model with different variations of sentences.

ACKNOWLEDGMENT

We would like to acknowledge Miss Amiera Syazreen Mohd Ghazali, research assistant at MIMOS Berhad who has contributed to the process of annotating training and testing datasets used in Mi-NER systems. We are also grateful to Dr. Rayner Alfred, associate professor of Computer Science, University Malaysia Sabah and Leow Chin Leong, University Malaysia Sabah for their help to evaluate their Rule-Based ER system on our testing dataset.

REFERENCES

[1] S.Jusoh and H.M. Alfawareh, "Techniques, applications and challenging issue in text mining," International Journal of Computer Science Issues(IJCSI), vol.9, no.6 , pp. 431-436, 2012.

[2] J.Steinberger and K. Jezek, "Using latent semantic analysis in text summarization and summary evaluation," in Proc. ISIM'04, pp. 93-100, 2004.

[3] S.A. Golder and B.A. Huberman, "Usage patterns of collaborative tagging systems". Journal of Information Science, vol.32, no.2, pp. 198-208, 2006.

[4] K.P. Murphy, Machine learning: a probabilistic perspective. 2012: MIT press.

[5] C.Malarkodi, R. Pattabhi and L.D. Sobha. "Tamil ner–coping with real time challenges," in 24th International Conference on Computational Linguistics, pp. 23, 2012.

[6] R. Vijayakrishna and S.L. Devi, "Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields," in IJCNLP, pp. 59-66, 2008.

[7] J. Piskorski, "Named-entity recognition for Polish with SProUT," in Intelligent Media Technology for Communicative Intelligence, Springer. p. 122-133,2005.

[8] C. Neudecker, "An Open Corpus for Named Entity Recognition in Historic Newspapers," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 2016. Portorož, Slovenia: European Language Resources Association (ELRA).

[9] Y.S. Fong, B. Ranaivo-Malançon and A.Y. Wee. "NERSIL-the Named-Entity Recognition System for Iban Language," in PACLIC, pp. 549-558, 2011.

[10] D.B.d. Pustaka, Malay for The Institute of Language and Literature. 2016; Available from: http://prpm.dbp.gov.my/.

[11] M.Y. Sharum, M.T. Abdullah, M.N. Sulaiman, M.A.A. Murdan and Z.A.A. Hamzah, "Name extraction for unstructured Malay text," in Computers & Informatics (ISCI),IEEE Symposium on. IEEE, pp. 787-791, 2011.

[12] R. Alfred, et al. "A Rule-Based Named-Entity Recognition for Malay Articles," in International Conference on Advanced Data Mining and Applications. Springer, pp.288-299, 2013.

[13] https://www.lexalytics.com/. Semantria., Available from: https://www.lexalytics.com/semantria, Retrieved September, 2016.

[14] A. Culotta, A. McCallum and J. Betz. "Integrating probabilistic extraction models and data mining to discover relations and patterns in text," in Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, pp. 296-303, 2006.

[15] F. Sha and F. Pereira. "Shallow parsing with conditional random fields," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, pp.134-141, 2003.

[16] BERNAMA. BERNAMA Archived News, Available from: http://www.bernama.com/bernama/v8/newsarchive.php, 2015.

[17] H. Metro, My Metro, Available from: http://www.hmetro.com.my/, Retrieved September, 2016.

[18] U, Utusan., Utusan Online, Available from: http://www.utusan.com.my/, Retrieved September, 2016.

[19] T. Baldwin and S. Awab "Open source corpus analysis tools for Malay," in In Proc. of the 5th International Conference on Language Resources and Evaluation. 2006. Citeseer.

 ISBN: 978-1-61208-542-5

[20] E.F. Tjong Kim Sang and F. De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in Proceedings of the seventh conference on Natural language learning at HLT-NAACL, Association for Computational Linguistics, pp.142-147, 2003

 ISBN: 978-1-61208-542-5