

Using Stanford NER and Illinois NER to Detect Malay Named Entity Recognition

S. Sulaiman, R. Abdul Wahid, S. Sarkawi, and N. Omar

Abstract—The goal of NER is to detect named entities in an open document. Many techniques are used to solve the NER problem. Most Malay Named Entity Recognition uses rule based and gazette to tag the names for each entity. In this paper, we tested online news articles using Stanford NER and Illinois NER to measure the capability of these NER tools to detect Malay Named Entities. The results are computed using the CoNLL evaluation metric. Stanford NER tends to produce higher results on F1 and Precision compared to Illinois NER. In the future, more NER systems will be evaluated to measure the compatibility of the tools to recognize Malay Named Entities.

Index Terms—Malay named entity recognition, named entity, Malay.

I. INTRODUCTION

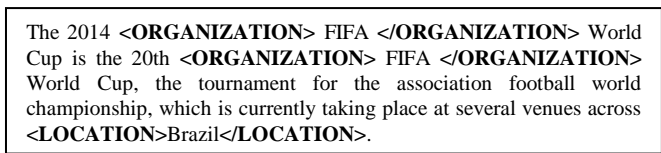
NER is important for NLP, document retrieval, morphological analysis, and information extraction [1]-[3]. NLP tools can be used to improve the processing of clinical records, as performed by [4]. A named entity extractor was used to differentiate between patient's and physician's names. Within Information Retrieval (IR), NER improves the detection of relevant documents [5]. Problems occur when many names need to be stored in the gazette. In some situations, a small gazette is sufficient to give good precision and recall [6].

The aim of NER is to detect named entities in open documents, such as websites and online newspapers. A named entity refers to a phrase representing a specific class.

In speech recognition, NER is usually tasked to detect named entities. This task is considered to be more difficult, since the capitalization of words, and generally the words themselves, is approximated by Automatic Speech Recognition (ASR) technologies. Optical Character Recognition (OCR) faces a similar problem in detecting named entities [7]. Fig. 1 is an example of Stanford Named Entity Recognition; marked with the two entity types; <Organization> and <Location>.

Many techniques and algorithms have been used to solve the NER problem. Previous researchers generally used handcrafted rules to overcome these problems. However, most recently, they used supervised machine learning or a collection of training examples to automatically stimulate

rule-based systems. When training examples were unavailable, a rule based system was preferred [8].



The 2014 <ORGANIZATION> FIFA </ORGANIZATION> World Cup is the 20th <ORGANIZATION> FIFA </ORGANIZATION> World Cup, the tournament for the association football world championship, which is currently taking place at several venues across <LOCATION>Brazil</LOCATION>.

Fig. 1. Stanford named entity recognition example.

Many researches have been done to recognize named entities in other languages, including English [9], Arabic [10], Chinese [11], and Indian [12]. These languages use different techniques to tackle issues regarding NER. However, these languages (including Malay) have their own morphologies. Alfred, R *et al.*, 2014 [13] used a rule based approach to develop Malay Named Entity Recognition. Several dictionaries were used to handle the named entities, like person, location and organization. A reasonable output depends on correct rules being used; and most crucially, all dictionaries must be up to date to achieve correct results. Some NERs use Gazettes to keep data about people's names, places, organizations, and many other forms of information regarding proper nouns. In this article, we conducted an experiment to measure the relevance of Illinois and Stanford NERs in Malay documents.

This article is divided into six sections. Section II presents an overview of related studies. Section III describes the Illinois and Stanford NERs. Section IV clarifies the test collection. Section V presents the experiments and results. Finally, Section VI presents our conclusions and possible directions for future research.

II. RELATED STUDIES

A. NER for English and German Languages

Nothman, J *et al.*, 2013 [3] proposed a Learning Multilingual NER, by exploiting the text and structure of Wikipedia articles. Each Wikipedia article was classified into a Named Entity (NE) type and approximately 7,200 articles were labelled manually. A heuristic approach was used and the results showed an accuracy of approximately 95% [3].

Nadeu, D, 2007 [14] developed a Named Entity Recognition (NER) system to classify rigid designators, such as proper names, biological species, and temporal expressions in text. This NER system was built using a semi-supervised system to recognize four NE types. It was expanded, by improving its key technologies, and applied to 100 NE types. The results showed that limited supervision was required to build a complete NER system [14].

Manuscript received July 24, 2015; revised December 15, 2015.

S. Sulaiman, R. Abdul Wahid, and S. Sarkawi are with the Sultan Idris Education University, 35900 Tg Malim, Perak, Malaysia (e-mail: suliana@fskik.upsi.edu.my, rohaizah@fskik.upsi.edu.my, suliana@fskik.upsi.edu.my).

N.Omar is with Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia (e-mail: no@ukm.edu.my)

B. NER for Arabic Language

Naji F. Mohammed, 2012 [15] designed an NER system based on a neural network approach to recognize named entities of Arabic texts. The use of a machine learning approach to classify NER from Arabic text, based on a neural network technique, was proposed. A neural network was used to automatically learn to detect component patterns and make intelligent decisions based on available data. This could also be used to classify new information within large databases. The accuracy of the system was 92%. The results showed that the neural network approach achieved better results than a decision tree [15].

C. NER for Malay Language

Rayner *et al.*, 2014 [13] proposed a rule-based Named Entity Recognition algorithm for Malay articles. The proposed Malay NER was designed based on Malay part-of-speech (POS) tagging feature and contextual features that were implemented to handle Malay articles. In order to handle three entity types, a number of manually accessed dictionaries were created to handle person, location and organization entities. The F-Measure result's value was 89.47%. Having more complete dictionaries and correct rules would improve the proposed Malay NER algorithm [13].

Yunita *et al.*, 2010 [16] developed a rule-based pattern extractor and a semi-supervised NER approach to automatically extract patterns from limited corpus. Stanford's part-of speech tagger and grammar parser were used to identify named entities and construct an extraction pattern. The semi-supervised NER used that pattern to classify entities. The experimental results showed that the NER system reached approximately 50 to 70% on F-measure; even if only two features were used [16].

III. ILLINOIS NER AND STANFORD NER

Illinois [17], [18] and Stanford [19], [20] used a machine learning approach to develop their own NERs. We conducted two experiments using these NERs and compared the results. Both tools recognized four types of entity, which were <Person>, <Organization>, <Location> and <Misc>.

A. Illinois NER Demo

The Illinois demo [17], [18] used normalized averaged perceptron, which it was assumed could increase the text chunking result. Ratinov *et al.*, [17] derived four fundamental design decisions, such as text chunk representation, inference algorithm, using non-local features and external knowledge. [17] tended to achieve 90.8 F1 score on the CoNLL-2003, also known as the best reported result for the dataset.

B. Stanford NER

The Stanford NER [19]-[20] is an open source tool that was developed in Java and came with feature extractors to detect named entities. The Stanford NER was trained using CoNLL 2003 English training data. The tool provided a general implementation of CRF sequence models, known as a CRF Classifier. Vidmar, [22] proved that, to effectively execute long distance constraints, it can be combined with an existing sequence model in a factored architecture. The technique used

in [22] could be reduced by up to 9% over the two state-of-the-art established information extraction tasks.

IV. DATA SET

The data set was taken from online newspapers [23], [24] and divided into 12 documents. Table I shows the details of the data set and Table II shows the number of named entities in the documents.

TABLE I: DETAILS OF THE DATA SET

Document	Number of Words	Categories	Type
Doc 1	394	Business	Corporate
Doc 2	323	News	National
Doc 3	142	News	National
Doc 4	93	Education	Campus
Doc 5	94	News	Court
Doc 6	1145	Articles	Religion
Doc 7	152	News	National
Doc 8	119	News	Court
Doc 9	176	News	Main
Doc 10	390	News	Main
Doc 11	160	News	National
Doc 12	108	News	National

TABLE II: NUMBER OF NAMED ENTITIES IN THE DOCUMENTS

Document	Location	Misc.	Organization	Person
Doc 1	10	20	5	5
Doc 2	3	5	5	10
Doc 3	2	2	8	1
Doc 4	3	3	12	1
Doc 5	1	3	2	1
Doc 6	14	0	2	4
Doc 7	5	2	0	1
Doc 8	1	4	0	5
Doc 9	5	2	2	16
Doc 10	19	8	9	7
Doc 11	8	2	1	1
Doc 12	2	5	0	2

V. EXPERIMENTS AND RESULTS

Due to limitations in accessing a Malay tagged corpus, we used online newspaper reports which were selected randomly from [23] and [24] as our testing data. These documents consist different types of categories such as news, business, education and articles. This experiment was conducted to see whether the English NER can be used to tag Malay Named Entity. In order to archive this goal, we tested the data using Stanford [20] and Illinois NERs [18]. From the result, the entity was accepted as true if the system rightly marked <TYPE> and <TEXT>. The results were identified as four different groups and compared with human annotated data. The groups can be classified as:

- True positive - where the system tags as correct <TYPE> and <TEXT> and also marked by an expert.
- True negative - where the system did not tag any word as <TYPE> or <TEXT> and was not marked by an expert.
- False positive - where the system tags <TYPE> and <TEXT> but not marked by an expert.
- False negative - where the system did not tag <TYPE> and <TEXT> and was not marked by an expert.

Fig. 2 shows the output from [20], Fig. 3 shows the output from [18] and Fig. 4 shows the human annotated data for Doc 1.

<PER> Di KUALA </PER>
<LOCATION>TERENGGANU</LOCATION>,
pekerja am, <PERSON>Abdullah Embong</PERSON>,
40, berpandangan, KWSP sewajarnya mengambil kira
masalah kesihatan yang dihadapi sebahagian pencarum
akan menyukarkan mereka untuk terus bekerja sehingga
60 tahun

Fig. 2. Output example from Stanford NER for Doc 1.

PER Di KUALA TERENGGANU, pekerja am, **PER
Abdullah Embong**, 40, berpandangan, KWSP
sewajarnya mengambil kira masalah kesihatan yang
dihadapi sebahagian pencarum akan menyukarkan
mereka untuk terus bekerja sehingga 60 tahun

Fig. 3. Output example from Illinois NER for Doc 1.

Di <LOC> KUALA TERENGGANU </LOC>, pekerja
am, <PERSON> Abdullah Embong </PERSON>,
<MISC> 40 </MISC>, berpandangan,
<ORG>KWSP</ORG> sewajarnya mengambil kira
masalah kesihatan yang dihadapi sebahagian pencarum
akan menyukarkan mereka untuk terus bekerja sehingga
<MISC>60 tahun</MISC>

Fig. 4. Expert review example for Doc 1.

The output from [20] and [18] were analysed and Table III shows examples of correctly and incorrectly tagged data.

TABLE III: EXAMPLES OF CORRECTLY AND INCORRECTLY TAGGED DATA

Stanford NER	Illinois NER	Explanation
<PER> Di KUALA </PER>	PER Di KUALA TERENGGANU	Both <TYPE> and <TEXT> were incorrectly tagged
<PERSON>Abdullah Embong</PERSON>	PER Abdullah Embong	Both <TYPE> and <TEXT> were correctly tagged

TABLE IV: STANFORD NER RESULTS FOR FOUR TYPES OF ENTITIES

Document	Location	Misc.	Organization	Person
Doc 1	6	0	7	12
Doc 2	2	17	4	17
Doc 3	3	1	6	6
Doc 4	6	1	6	5
Doc 5	1	0	5	5
Doc 6	18	4	5	24
Doc 7	2	0	2	4
Doc 8	1	0	0	6
Doc 9	4	0	4	14
Doc 10	3	2	10	10
Doc 11	3	1	3	7
Doc 12	0	0	2	2

The results were computed using precision, recall and F_1 formula to see the effectiveness of Stanford NER [20] and Illinois NER [18] to classify and recognized Malay Named Entity. Results for Precision, Recall and F-Measure were calculated based on the following formula;

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F_1 = 2 \cdot \frac{P.R}{P + R}$$

The results are as reported in Tables IV, V, VI and VII.

TABLE V: ILLINOIS NER RESULTS FOR FOUR TYPES OF ENTITIES

Document	Location	Misc.	Organization	Person
Doc 1	7	1	2	12
Doc 2	1	13	3	23
Doc 3	5	0	4	5
Doc 4	3	0	7	5
Doc 5	1	2	1	7
Doc 6	28	6	5	26
Doc 7	1	0	1	7
Doc 8	0	1	0	5
Doc 9	4	2	1	16
Doc 10	2	8	8	11
Doc 11	11	0	3	4
Doc 12	2	0	2	5

TABLE VI: MALAY DOCUMENTS TESTED ON STANFORD NER

Document	Precision	Recall	F_1
Doc 1	52.000	29.545	37.681
Doc 2	22.500	37.500	28.125
Doc 3	56.250	60.000	58.065
Doc 4	37.500	31.579	34.286
Doc 5	36.364	40.000	38.095
Doc 6	29.412	68.182	41.096
Doc 7	12.500	12.500	12.500
Doc 8	57.143	40.000	47.059
Doc 9	54.545	44.444	48.980
Doc 10	32.000	18.605	23.529
Doc 11	35.714	41.667	38.462
Doc 12	50.000	22.222	30.769

TABLE VII: MALAY DOCUMENTS TESTED ON ILLINOIS NER

Document	Precision	Recall	F_1
Doc 1	36.364	24.242	29.091
Doc 2	17.073	31.818	22.222
Doc 3	30.769	36.364	33.333
Doc 4	61.538	42.105	50.000
Doc 5	27.273	37.500	31.579
Doc 6	21.538	73.684	33.333
Doc 7	11.111	12.500	11.765
Doc 8	50.000	30.000	37.500
Doc 9	45.833	44.000	44.898
Doc 10	50.000	34.091	40.541
Doc 11	44.444	66.667	53.333
Doc 12	37.500	33.333	35.294

From the experiments, the Stanford NER had a high tendency to incorrectly tag <PERSON> followed by <MISC>, <LOC> and <ORG>. However, the Illinois NER tended to incorrectly tag <PERSON>, <MISC>, <ORG> and <LOC>. Stanford NER achieved a higher result for precision (39.66%) and Illinois NER attained a higher result for recall (37.19%). Based on the average F_1 results, the Stanford NER (36.55%) produced a much better result than the Illinois NER (35.24%) with lots of error. Both NER [18], [20] tend to produce errors on classify <PERSON> and <MISC> for

Malay Named Entity due to different morphology in English and Malay.

VI. CONCLUSION

Most Malay NERs are rule based, in order to improvise Named-Entity Recognition. Due to a lack of Malay tagged corpus, we tested 3,296 words from online newspapers using two established NERs from Stanford [20] and Illinois [18]. Based on the experiments, the Stanford NER showed higher results for F_1 and Precision. Both NERs showed low results for the Malay corpus and after investigation most error occur because of the different morphology between Malay and English language. In the future, we will study available NER and try to minimize the used of rules and remove gazette and dictionaries.

ACKNOWLEDGMENTS

This work was funded by the Ministry of Education (Malaysia) and supported by Grant. No. RAGS: 2014-0123-109-72.

REFERENCES

- [1] W. H. Liao and S. Veeramachaneni, "A simple semi-supervised algorithm for named entity recognition," in *Proc. Semi Sup Learn*, 2009, pp. 56-65.
- [2] D. Pierce and C. Cardie, "Limitation of co-training for natural language learning from large datasets," in *Proc. EMNLP*, 2011.
- [3] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from Wikipedia," *Artificial Intelligence, Wikipedia and Semi-Structured Resources*, vol. 194, pp. 151-175, January 2013.
- [4] P. Ruch, R. Baud and A. Geissbuhler, "Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record," *Artificial Intelligence in Medicine*, vol. 29, no. 1-2, pp. 169-184, September 2003.
- [5] S.-F. Yong, B. Ranaivo-Malaeon and A. Y. Wee, "NERSIL: The named-entity recognition system for iban language," in *Proc. 25th Pacific Asia Conference on Language, Information and Computation*, 2011, pp. 549-558.
- [6] A. Milkheev, M. Moens and C. Grover, "Named entity recognition without gazetteers," in *Proc. the Ninth Conference on European chapter of the Association for Computational Linguistics*, 1999, pp 1-8.
- [7] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks, "Named entity recognition from diverse text types," *Proceeding of Recent Advances in Natural Language Processing*, 2011.
- [8] S. Satoshi and C. Nobata, "Definition, dictionaries and tagger for extended named entity hierarchy," in *Proc. Conference on Language Resources and Evaluation*, 2004, pp. 1977-1980.
- [9] J. Nothman, T. Murphy and J. R. Curran, "Analysing wikipedia and gold-standard corpora for NER training," in *Proc. the 12th Conference of the European Chapter of the ACL*, 2009, pp. 612-620.
- [10] K. Shaalan, "A survey of Arabic named entity recognition classification," *Association for Computational Linguistics*, vol. 40, no. 2, pp. 469-510, June 2014.
- [11] J. F. Gao, M. Li, A. D. Wu and C.-N. Huang, "Chinese word segmentation and named entity recognition: A Pragmatic approach," *Association for Computational Linguistics*, vol. 32, no. 4, pp. 531-574, 2005.
- [12] S. K. Saha, S. Sarkar, and P. Mitra, "Gazetteer preparation for named entity recognition in indian languages," in *Proc. The 6th Workshop on Asian Language Resources*, 2008, pp. 10-16.

- [13] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "A robust framework for web information extraction and retrieval," *International Journal of Machine Learning and Computing*, vol. 4, no. 3, pp. 146-150, June 2014.
- [14] D. Nadeau, "Semi-supervised named entity recognition: Learning to recognize 100 entity types with little supervision," PhD Dissertation, University of Ottawa, Canada, 2007.
- [15] F. N. Mohammed and N. Omar, "Arabic named entity recognition using artificial neural network," *Journal of Computer Science*, vol. 8, no. 8, pp. 1285-1293, 2012.
- [16] Y. Sari, M. F. Hassan, and N. Zamin, "Rule-bamised pattern extractor and named entity recognition: a hybrid approach," in *Proc. International Symposium on Information Technology, Kuala Lumpur, Malaysia*, 2010, pp. 563-568.
- [17] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. CoNLL*, 2009, pp. 147-155.
- [18] Cognitive computation group. [Online]. Available: http://cogcomp.cs.illinois.edu/page/demo_view/ner
- [19] F. J. Rose, G. Trond, and M. Christopher, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proc. the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005, pp. 363-370.
- [20] Stanford Named Entity Tagger. [Online]. Available: <http://nlp.stanford.edu:8080/ner/process>
- [21] A. Carlson, S. Gaffney, and F. Vasile, "Learning a Named Entity Tagger from Gazetteers with the partial perceptron," in *Proc. AAAI Spring Symposium on Learning by Reading and Learning to Read*, 2009, pp. 7-13.
- [22] R. J. Vidmar, "On the use of atmospheric plasmas as electromagnetic reflectors," *IEEE Trans. Plasma Sci.*, vol. 21, no. 3, pp. 876-880, August 1992.
- [23] Utusan Online. [Online]. Available: <http://www.utusan.com.my/>
- [24] Bharian Online. [Online]. Available: <http://www.bharian.com.my/>



S. Sulaiman received her degree in computer science (artificial intelligence) from the University of Malaya, Malaysia in 2003. She received her M.Sc. degree from Universiti Kebangsaan Malaysia in 2008. She obtained her PhD degree from Universiti Kebangsaan Malaysia, Malaysia in 2013. Currently, she is a senior lecturer at Universiti Pendidikan Sultan Idris, Malaysia.



R. Abdul Wahid received her degree in information technologies (artificial intelligence) from Universiti Utara Malaysia, Malaysia in 2000. She received her M.Sc degree from Universiti Teknologi Malaysia, Malaysia in 2005. Currently, she is a lecturer at Universiti Pendidikan Sultan Idris, Malaysia.



S. Sarkawi received his degree in Malay literature from the University of Malaya in 1985. He received his M.Ed from Universiti Malaya in 1994 and obtained his PhD degree from Universiti Pendidikan Sultan Idris in 2010. Currently, he is a senior lecturer at Universiti Pendidikan Sultan Idris, Malaysia.



N. Omar received her PhD from the University of Ulster, UK in 2005 under the supervision of Prof. Paul Mc Kevitt and Prof. Paul Hanna. She holds an MSc degree from the University of Liverpool, UK and Bsc(Hons) from UMIST, UK. Currently, she is an associate professor at the Faculty of Information Science and Technology at Universiti Kebangsaan Malaysia (UKM).