Modified DBpedia Entities Expansion for Tagging Automatically NER Dataset

Ika Alfina, Septiviana Savitri, and Mohamad Ivan Fanany
Machine Learning and Computer Vision Laboratory
Faculty of Computer Science Universitas Indonesia
Depok, Indonesia
ika.alfina@cs.ui.ac.id, septiviana.savitri@ui.ac.id, ivan@cs.ui.ac.id

***Abstract*—Developing NER system using machine learning approach needs a big dataset which is costly if the dataset labeling is done manually. The previous works proposed methods in tagging automatically the Indonesian NER dataset using Wikipedia articles as the source of the dataset and DBpedia as the reference of the entity type. However, the quality of the resulting dataset was still inadequate. A method named DBpedia Entities Expansion (DEE) had introduced several rules to expand named entities in DBpedia in order to improve recall, but it had not managed to remove noise that makes precision decline, especially for person names. The objective of this research is to propose the modification to DEE method with the main focus to remove invalid names from the list of person names in the Expanded DBpedia. We call this modification as Modified DEE (M-DEE). The evaluation shows that M-DEE can improve the precision for person names around 3% compared to the original DEE. By adding gazetteers for place and organization names into the Expanded DBpedia created by M-DEE, the margin about 10% of the overall F1-score for all types was achieved.**

Keywords—NER; building dataset; noise reduction; DBpedia

I. Introduction

Named Entity Recognition (NER) is a subtask of Information Extraction system, which aims to identify the named entities that appear in a text and to decide which classes the named entities belong to. NER studies for English have achieved very good results. Reference [1] reported that they successfully built NER with F1-score of 91.2% that outperformed several states of the art of NER system, such as [2] and [3]. Unfortunately, for the Indonesian language, the performance of NER system was still inadequate.

The previous works in building the Indonesian NER were done by [4]–[12]. Reference [5] used rule-based approach, while [4], [6]–[11] used machine learning approach, and [12] used deep learning method. In [5], 100 rules were designed by the help of the expert knowledge engineers. Although the result was quite good with 63.43% of recall and 71.48% of precision, since the cost was expensive, the others did not follow this approach. It can be seen that so far machine learning approach is the favorite one. There is only one research using deep learning since this approach is relatively new.

It is difficult to compare NER systems performance for the Indonesian language since it lacks of standard dataset. Each research usually had to create a new dataset that beside becomes the burden to start the research; the size of the dataset was also small. In [4] dataset was created by manually labeled 55 news articles and in [5] 802 sentences from online news articles are labeled manually.

To develop NER system using machine learning approach we need a big dataset. One of the solutions to this problem is by tagging dataset automatically, not manually. An early work in creating dataset automatically for the Indonesian NER was done by [7]. This work used Wikipedia articles in the Indonesian language as the dataset source and utilizes the Indonesian DBpedia[1] as the reference of entity types. 10,000 Wikipedia articles were tagged to build the training dataset. This dataset was used as input for Stanford NER tool that would create Indonesian NER model. Stanford NER tool implemented Conditional Random Field (CRF) algorithm to classify named entities [3]. Research [7] also created a gold standard dataset that created by labeling manually 68 Wikipedia articles. The NER system of [7] had very good precision but very low recall. For person type, the recall was only 7.39%, for place type was only 11.43%, and there was no information for organization type performance.

Developing Indonesian NER system using the dataset that automatically tagged was followed by [8], [10], [12]. The approach used by [8] is similar with [7], except they used semi-supervised learning while [7] used supervised learning. In [8], the performance measure reported was overall F1-score for all types of 31.96%. Both [7] and [8] reported that the poor performance was caused by the bad quality of training dataset that built automatically.

To improve the quality of dataset built automatically, [10] proposed a new method named DBpedia Entities Expansion (DEE). DEE tried to improve recall by enriching the entities in Indonesian DBpedia. Evaluation using the same gold standard with [7] shows that DEE can improve recall more than 4 times compared to [7] for person and place types. The overall F1-score was 48.84%, more than 16% higher than [8]. The dataset created using DEE was used by [12] to train the Indonesian NER model using long short-term memory (LSTM), a deep learning method. The reported F1-score was 57.31%, almost 9% higher than [10] that used Stanford NER tool to build the model.

The difference between DEE and [7], [8] in labeling the dataset was in how they compare a named entity (NE)

[1] http://id.dbpedia.org/wiki/

978-1-5386-3172-0/17/$31.00© 2017 IEEE

candidate found in a text with the NEs in DBpedia. DBpedia as the type reference of NEs only provide the full name of the entity. The approach used by [7] and [8] was the full match between each NE candidate and the NE entries in DBpedia. For example, let in DBpedia there is an NE "Joko Widodo" as a person. If the text contains the phrase that exactly the same with "Joko Widodo", the phrase will be labeled as a person, but if the text only contains the subset of "Joko Widodo", such as "Joko" or "Widodo", that phrase will only be labeled as "other". DEE method introduces rules in expanding the entries in DBpedia to facilitate the partial match between NE candidate and entries in the original DBpedia.

Reference [10] stated that quality of the automatically labeled dataset they built still need to be improved. Since the precision for person type was decreased compared to [7], they suggested finding the method to remove invalid entries from Expanded DBpedia. Thus, the objective of our research is to improve the quality of dataset that automatically tagged by (1) improving the DEE method proposed by [10] and (2) adding some gazetteers to enrich the Expanded DBpedia. The contributions of our work are:

- Introducing modified method of DEE that we named Modified DEE (M-DEE) in order to reduce invalid names in Expanded DBpedia.
- Providing a better dataset to be used by others to build the Indonesian NER. This dataset had been made public[2].

This paper is organized as followed. Section 2 describes the related work, especially DEE approach [10]. Section 3 discusses our proposed rules to improve DEE, followed by Section 4 that presents the experiment results. Finally, conclusion and future work are discussed in Section 5.

II. Related work

Fig. 1 shows the framework used by [10] in building the Indonesian NER. It consists of 4 processes, i.e: preprocessing Wikipedia articles into an unlabeled dataset, transforming the original DBpedia into the Expanded DBpedia using DEE method, labeling the dataset, and finally built the Indonesian NER model using Stanford NER tool. This framework is similar with [7], the main difference is [7] used original DBpedia in labeling dataset.

The objective of our study is to improve DEE method proposed by [10]. As we used the same approach with [10] for the other three processes, i.e: preprocessing Wikipedia, labeling dataset, and building NER model, we did not re-explain these three processes.

This section consists of 2 subsections. Subsection A describes the categories of named entity for type PERSON, PLACE and ORGANIZATION (ORG) that were defined by [10]. Subsection B explains the DEE rules used to expand entities for each entity type.

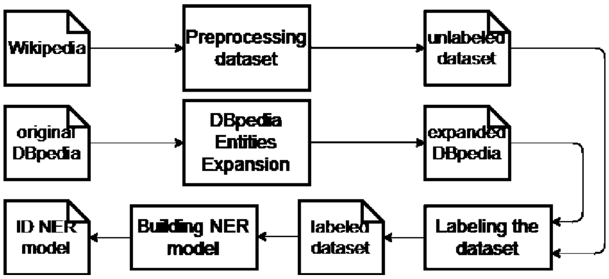

Fig. 1 The framework of building the Indonesian NER model using DBpedia Entities Expansion method

A. *Categories of DBpedia's entities*

In [10], 11 categories for PERSON, 5 categories for PLACE and 5 categories for ORG were introduced. To make it easier for us to refer to those categories later, we introduced new names for them. Categories name started with A are for PERSON, B for PLACE and C for ORG.

1) *Categories of Person Names*

- A1: standard name, e.g.: "Taufiq Ismail".
- A2: name with Roman number, e.g.: "Oscar Hammerstein II"
- A3: name contains single capital letter, e.g.: "I Putu Wijaya"
- A4: name contains single capital letter that followed by period, e.g.: "Daniel D. Tompkins"
- A5: name contains word "bin" or "binti", e.g.: "Habib Umar bin Hafidz"
- A6: name contains words that relate person to a location, e.g.: "Mary of Teck"
- A7: name that followed by title, contains words "der", "de" or "dos", e.g.: "Fidel de Castro"
- A8: name with a comma and followed by the description, e.g.: "Raymond Davis, Jr."
- A9: valid person name with description in parenthesis, e.g.: "R. Suprapto (national hero)"
- A10: invalid person name with description in parenthesis, e.g.: "Nirvana (music group)"
- A11: invalid person name, without description, e.g.: "Backstreet Boys"

2) *Categories of Place Names*

There are 5 categories of place names in [10]. Since in our work there is no modification to place names categories and rules, the explanation of those categories is excluded.

3) *Categories of Organization Names*

- C1: standard name, e.g.: "Airbus"
- C2: name contains the period, e.g.: "Arsenal F.C."
- C3: name contains comma, e.g.: "Amuse, Inc."
- C4: name contains word "di" (at), e.g.: "Universitas Amerika di Beirut"
- C5: name with description in parenthesis, e.g.: "Partai Republik (Indonesia)"

[2] https://github.com/ialfina/ner-dataset-modified-dee

It can be seen from the categories above that some of them contains noise and even are incorrectly labeled, especially for PERSON.

B. DBpedia Entities Expansion

This subsection discusses DEE method proposed by [10] in order to expand the entities of the original Indonesian DBpedia. DEE method consists of five stages: name parsing, name cleansing, name normalization, name expansion, and finally name validation. For each stage, some rules were created for each entity types.

There are 18 rules for DEE proposed by [10], consist of 1 rule for parsing step, 1 rule for cleansing step, 8 rules for normalization step, 3 rules for expansion step and 5 rules for validation step. To simplify the explanation, the rule will be named using XYZ format:

- X refers to the step in DEE, i.e: P for Parsing, C for Cleansing, N for Normalization, E for Expansion, and V for Validation.
- Y refers to the entity type, A for PERSON, B for PLACE and C for ORGANIZATION
- Z is used for rule number

The list of rules is followed:

1) Parsing Stage

P1: transform the file of DBpedia instance type in RDF format into 3 files in text format, each for PERSON, PLACE and ORG.

2) Cleansing Stage

CA1: remove entries that belong to A10.

3) Normalization Stage

Since our work only modified normalization rules for PERSON, we excluded the explanation of normalization rules for PLACE and ORG. In [10], there are five normalization rules for person names:

1. NA1: splitting A5 names with delimiter "bin" or "binti", e.g.: "Umar bin Hafidz" is transformed into two names "Umar" and "Hafidz".
2. NA2: remove place information from A6 names, e.g.: "Mary of Teck" is transformed into "Mary".
3. NA3: remove title information from A7 names, e.g.: "Fidel de Castro" is transformed into "Fidel".
4. NA4: removes the comma and the following information from A8 names, e.g.: "Raymond Davis, Jr." is transformed into "Raymond Davis".
5. NA5: remove parenthesis and information from A9 names, e.g.: "R. Suprapto (national hero)" is transformed into "R. Suprapto".

4) Expansion Stage

Since only ORG has new expansion rules, we only re-explained the previous rules for ORG:

1. EC1: expansion rule for C3 category
2. EC2: expansion rule for C4 category

5) Validation Stage

After expansion stage, there are a lot of new entries in Expanded DBpedia that are invalid names. There are five rules for this stage. Since in our work there is no modification to these five rules, the explanation of each rule is excluded.

III. PROPOSED METHOD

This research used a similar method with [10], with modification to DEE method. We modify some name categories and rules in creating Expanded DBpedia. Later, we also add some gazetteers to enrich the new Expanded DBpedia.

A. Modification for the categories of DBpedia's entities

Our work proposed the modification to categories of PERSON and ORG. There is no change for PLACE.

1) Categories of Person

We modified the A8 and A11 categories and defined 3 new categories. The modifications are as follows:

- A8-modified: A8 (the name with a comma and followed by description) was made more specific become the name with a comma and followed by the description like "Sr." or "Jr.", e.g.: "Joseph Patrick Kennedy, Sr.".
  The old definition of A8 assumes that if comma exists in a person name, the following text is the description. But later we found an exception like we defined later as A11-6.
- A11-modified: In [10], A11 (invalid names without description) was regarded as noise but had not been handled thoroughly. We divide A11 into 6 subcategories:
  - A11-1: Name contains numbers, e.g.: "Sheila on 7"
  - A11-2: Name that all its words exist in the dictionary, e.g.: "The Rolling Stones"
  - A11-3: Name contains substring "the", e.g.: "The Veronicas"
  - A11-4: Name contains character "&", e.g.: "Emerson, Lake & Palmer"
  - A11-5: Name that actually is an educational institution, e.g.: "SMP Negeri 6 Pekalongan"
  - A11-6: Name that contains comma(s) but not member of A8-modified, e.g.: "Gamaliel, Audrey, Cantika"

The following new categories are introduced as their characteristic had not been identified in [10]:

- A12: Name contains the character "-", e.g.: "Jean-Joseph Etienne Lenoir".
- A13: Name that begins with title, e.g.: "Dr. Seuss"
- A14: Name that contains substring "oe", e.g: "Soekarno"

2) Categories of Organization

In [10], ORG has 5 categories. After some observation to political party names and observed that these names can be expanded in a specific way by removing the word "partai/party", we introduced a new category, C6: Names that begin with "partai" (party), e.g.: "Partai Golongan Karya".

B. Modification of DBpedia Entities Expansion Rules

Fig. 2 shows the big picture of the modification that is proposed compared to [10]. In total there are 17 changes to DEE rules, we modify 5 rules and propose 12 new rules. In

978-1-5386-3172-0/17/$31.00© 2017 IEEE

addition, we also enrich the new DBpedia with some gazetteers.

Of those 17 rules, 13 rules are designed for PERSON, 3 rules for all NE types and one rule for ORG. We can say that the majority of new rules are aimed to remove noise in person names as suggested by [10].

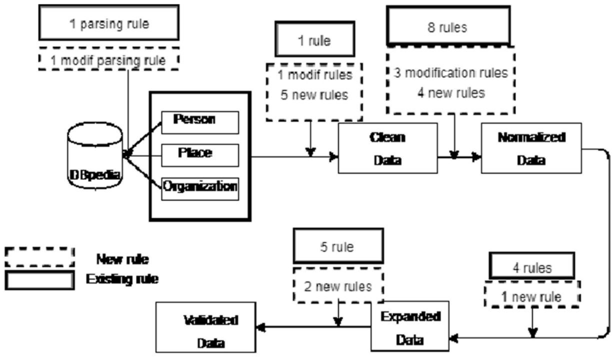

Fig. 2. Comparison between original DEE and M-DEE

The explanation of each rule is as followed:

1) Parsing Stage

P1-modified: In [10], the names that contain Unicode were not being extracted but in our work, these names were handled so they enriched our DBpedia.

2) Cleansing Stage

For PERSON, there is one modified and five new rules. There is no modification for PLACE and ORG. The modification is as follows:

- CA1-modified: In [10], names that belong to A10 are simply removed because they are invalid person names and the description shows that they are actually organizations. The modification for this rule: after removing these names from person list, we save them in organization list. For example, "Jamrud (band)" is removed from person list, but "Jamrud" later was added to organization list.
- Five new rules: CA2, CA3, CA4, CA5 and CA6 to remove names of A11-1, A11-2, A11-3, A11-4, and A11-5 categories consecutively.

3) Normalization Stage

For PERSON, there are 3 modified and 4 new rules. There is no modification for place and organization names. The modification is as follows:

- NA1-modified: Instead of removing the original names after splitting in NA1, we decided to keep them in a separate file that will be added later to final Expanded DBpedia, bypassing expansion and validation stages.
- NA3-modified: NA3 simply remove the remaining term started from "dos/de/der" from the name. We modify this rule that instead of removing the original names, we decided to keep it to be added later to final Expanded DBpedia.
- NA4-modified: Name of A8 become two entries, e.g.: "Joseph Patrick Kennedy, Sr." is normalized into "Joseph Patrick Kennedy Sr." and "Joseph Patrick Kennedy."
- NA6: Splitting the name of the A12 category with delimiter "-" and keep the original name. For example, "Jean-Joseph Étienne Lenoir" is transformed into 3 names: "Jean-Joseph Étienne Lenoir", "Jean" and "Joseph Étienne Lenoir".
- NA7: Remove title from the A13 category. For example, "Dr. Seuss" is transformed into "Seuss".
- NA8: Add a new name for the A14 category that's created by replacing "oe" of original name with "u" at the new name. For example, "Soekarno" is transformed into 2 names: "Soekarno" and "Sukarno".
- NA9: Splitting name of A11-6 with the comma delimiter, e.g.: "Gamaliel, Audrey, Cantika" is transformed into three entities: "Gamaliel", "Audrey" and "Cantika".

4) Expansion Stage

There is no change in this stage for PERSON and PLACE, but there is one new rule for ORG. EC3: Names of C6 category is expanded by creating new names without "partai" (party) word. For example, "Partai Demokrat" will be expanded into "Partai Demokrat" and "Demokrat".

5) Validation Stage

We propose 2 new rules for validation stage, that can be applied to PERSON, PLACE, and ORG:

- V6: Remove the name of month, e.g.: "July"
- V7: Remove the names of religion, e.g.: "Islam"

Table 1 shows the comparison of the number of NEs extracted from the original Indonesia DBpedia entities file and final Expanded DBpedia between DEE and M-DEE. The new parsing rule that can handle Unicode added 1,817 new person names, 500 new place names and 140 new organization names, compared to [10]. The final Expanded DBpedia is enriched by more 6,521 person names, 2,036 place names and 352 organization names than [10].

TABLE I. NUMBER OF ENTITIES IN DBPEDIA

	Extracted		Expanded	
Entity Type	DEE	M-DEE	DEE	M-DEE
Person	17,749	19,566	36,514	43,035
Place	57,193	57,693	137,710	139,746
Organization	5,633	5,773	5,722	6,074
Total	80,575	83,032	179,946	188,855

C. *Additional Gazetteers*

We found out that the original Indonesian DBpedia does not contain important names like "Indonesia", "America", and so on. This observation motivated us to add some gazetteers, especially for place and organization names.

After searching on the internet, we choose some articles to be the gazetteers. The gazetteers for place names are: list of 393 countries and capital cities around the world[3] and list of

[3] http://ilmupengetahuanumum.com/daftar-nama-negara-negara-di-dunia-beserta-ibukota-negara/

100 cities by GDP[4]. The gazetteers for organization names are: list of 183 Indonesian political parties[5] and list of 454 Indonesian Governor Institution with the abbreviation[6]. These new names are added directly to final Expanded DBpedia without being cleansed, normalized, expanded or validated, since the data are assumed valid.

IV. Experiments and Results

We used the same evaluation method with [10] in determining the quality of the resulting dataset. As the dataset source, we use the latest Indonesian Wikipedia articles dump. After preprocessing stage, we had 20,000 sentences to be labeled automatically. The resulting dataset that automatically tagged consists of 599,600 tokens. The gold standard dataset use by [7] and [10] was also used as the testing dataset. This testing dataset consists of 14,431 tokens. Stanford NER was also used as the tool to create NER model.

In order to assess the M-DEE method, we did two comparisons. The first comparison is between M-DEE and the original DEE proposed by [10]. The second comparison is between M-DEE and M-DEE plus gazetteers.

A. Comparison between DEE and M-DEE

Table II shows the classification results using 2 datasets each labeled using a different version of Expanded DBpedia. Dataset A was tagged using Expanded DBpedia of original DEE method and Dataset B was labeled using M-DEE method. In general, M-DEE had similar F1-score with original DEE, with a slight increase of 0.54%. There was a slight improvement of 0.84% in recall but a decline around 2% in precision.

TABLE II. Performance of DEE (A) vs. Modified DEE (B)

Entity Type	Precision (%)		Recall (%)		F1-score (%)	
	A	*B*	*A*	*B*	*A*	*B*
Organization	80.95	**81.08**	**9.63**	8.50	**17.22**	15.38
Person	67.03	**70.34**	32.16	**32.51**	43.47	**44.47**
Place	**80.42**	72.13	38.07	**40.83**	51.67	**52.14**
TOTAL	**73.87**	71.89	28.69	**29.53**	41.33	**41.87**

For ORG, M-DEE has similar precision and a lower recall around 1% compared to DEE. This made the F1-score of ORG also decrease about 2%. Of those 17 rules, only 5 rules associated with ORG: P1-modified (new parsing rule that handles Unicode), CA1-modified (a rule that moves misplaced organization name form PERSON to ORG), EC3 (a new expansion rule for ORG), V6 and V7 (additional validation rules). Since these five rules had made Expanded DBpedia for ORG has 352 (0.06%) more names than DEE, the decrease of recall about 1% is unexpected and hard to explain.

Fig. 3 shows the comparison of classification result between original DEE and M-DEE for PERSON. Although 13 of 17 M-DEE rules designed for PERSON to remove noise from names, the precision only increased about 3% and the recall was similar with DEE, which made a slight increase of 1% for F1-score. We have no data on the proportion of noise (invalid names) in DBpedia for person names, but after seeing this result, we suggest that the number was low. Another approach should be found in order to improve F1-score for person names.

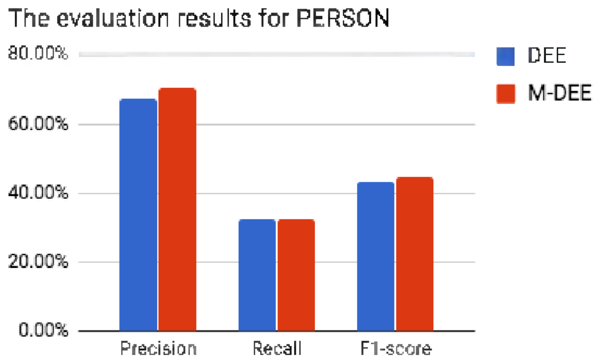

Fig. 3. The evaluation results for PERSON

For PLACE, unfortunately, the precision decreased around 8% and recall increased about 3%, ended up with a slight improvement of 0.47% of F1-score. As only the new parsing rules and two validation rules associated with PLACE, we can say that addition of 500 places contains Unicode in their names has slightly improved the recall. The decline of precision by 8% was very unexpected. As the ambiguity between place name and person name frequently occurs, we suggested this declined of precision was caused by some person names that are wrongly labeled as place names. In the next study, dealing with word sense disambiguation in the tagging process could be tried in order to improve the precision of PLACE.

B. Comparison between M-DEE and M-DEE + Gazetteers

Table III shows the comparison of classification results using two datasets, Dataset B and Dataset C. Dataset B had been tagged using Expanded DBpedia created by M-DEE method and Dataset C by M-DEE plus gazetteers. In general, M-DEE plus gazetteers outperformed M-DEE around 10% in F1-score. All recall and precision increased, except for ORG that has a precision reduction almost 9%.

TABLE III. Performance of M-DEE (B) vs. M-DEE+Gazetteers (C)

Entity Type	Precision (%)		Recall (%)		F1-Score (%)	
	B	*C*	*B*	*C*	*B*	*C*
Organization	**81.08**	72.41	8.50	**11.90**	15.38	**20.44**
Person	70.34	**71.76**	32.51	**33.04**	44.47	**45.25**
Place	72.13	**77.88**	40.83	**65.29**	52.14	**71.03**
TOTAL	71.89	**75.30**	29.53	**39.26**	41.87	**51.61**

Since the additional gazetteers are only for PLACE and ORG, we can see that the performance of PERSON did not change much with a slight increase of 0.78% of F1-score. Precision for PERSON rose 1.42% and recall rose slightly as much as 0.53%. For ORG and PLACE, the F1-score increased about 5% and 19% consecutively. For ORG, recall increased 3.4% but precision decreased 8.67%. For PLACE, precision and recall increased 5.75% and 24.46% consecutively. Looks

[4] https://id.wikipedia.org/wiki/Daftar_kota_menurut_PDB
[5] https://id.wikipedia.org/wiki/Daftar_partai_politik_di_Indonesia
[6] https://www.menpan.go.id/jdih/permen-kepmen/permenpan-rb/file/3568-permenpan-2012-no-081?start=100

like that by adding place names using gazetteers, the mislabeled between PERSON and PLACE had decreased that cause big improvement on PLACE's precision.

Fig. 4 shows the comparison of F1-score between original DEE and M-DEE plus gazetteers. F1-score increased for all type, especially for PLACE that has a margin almost 20%. Overall, our work has successfully increased F1-score about 10% compared to [10].

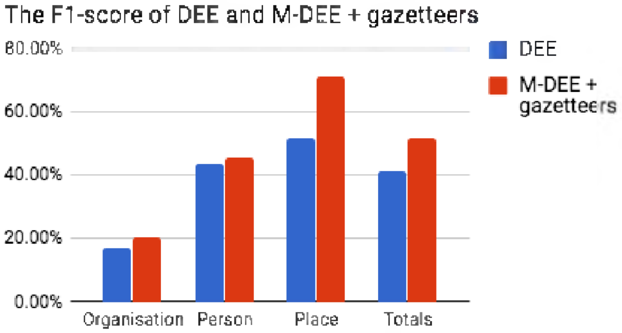

Fig. 4. The comparison of F1-score between DEE and MDEE + Gazetteers

V. Conclusions and Future Work

We proposed some modified and additional rules for DEE method proposed by [10] and named it Modified DEE (M-DEE). M-DEE introduced two modified and three new categories for PERSON, while for ORG there is one new category. We also proposed 17 rules; consist of 5 modified rules and 12 new rules, which 13 of 17 rules designed for PERSON. In general, M-DEE focused on removing noise (invalid names) in DBpedia for PERSON.

We also utilize some gazetteers for place and organization names to enrich the Expanded DBpedia created by M-DEE. This need to be done because after observation to the number of false positive tags in the dataset, we found out that some important places like Indonesia, America and so on did not exist in the Indonesian DBpedia.

The evaluation shows that our M-DEE has similar F1-score with original DEE with a slight improvement in recall. However, its objective to remove noise in person names seems a little fruitful with the increase of 3% in precision for PERSON. We suggest that the number of invalid names in DBpedia was very small that after applying 13 of 17 rules to improve the quality of person names, the improvement of precision was very low. M-DEE also caused the decline of precision for place names. We suspect that M-DEE caused additional ambiguity between person and place names, that some person names are labeled automatically as PLACE.

After adding gazetteers to the Expanded DBpedia created by M-DEE, we had a quite good result that had F1-score of almost 10% better than using M-DEE only. Especially for place names, the recall increased around 24%. Adding place names might have decreased ambiguity between person and place names since this addition successfully increased the precision for PLACE. In total, M-DEE plus gazetteers outperformed original DEE by around 10% in F1-score.

After seeing the good result of adding some gazetteers to the Expanded DBpedia, we suspect that the current Indonesian DBpedia was far from a complete encyclopedia to be used as a sole reference in determining NE types. The next research on building the dataset automatically could add other references to improve the performance of the Indonesian NER system.

For further research, we also suggest the improvement of tagging method. First, by using word sense disambiguation to reduce reversed labeling between person and place names that frequently occurs. Second, improving method to detect the candidates of the named entity. In DEE/M-DEE, the candidates of the named entities are founded by looking for the longest sequence of words that its words begin with uppercase. This rule eliminates the opportunity for the name like "Ali bin Abi Thalib" to be tagged as a full named entity because by the current tagging rule two candidates are detected: "Ali" and "Abi Thalib", while "bin" is excluded.

References

[1] G. Luo, X. Huang, C. Lin, and Z. Nie, "Joint Named Entity Recognition and Disambiguation," *Emnlp*, no. September, pp. 879–888, 2015.

[2] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," *Proc. Thirteen. Conf. Comput. Nat. Lang. Learn. CoNLL 09*, no. June, p. 147, 2009.

[3] J. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," *... 43rd Annu. Meet. ...*, no. June, pp. 363–370, 2005.

[4] I. Budi and S. Bressan, "Association rules mining for name entity recognition," in *Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003.*, 2003, pp. 325–328.

[5] I. Budi, S. Bressan, G. Wahyudi, Z. A. Hasibuan, and B. A. A. Nazief, "Named Entity Recognition for the Indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005, vol. 3735 LNAI, pp. 57–69.

[6] I. Budi, "Application of association rules mining to Named Entity Recognition and co-reference resolution for the Indonesian language," *J. Bus. Intell. Data Min.*, vol. 35, no. 3, pp. 193–8, 2007.

[7] A. Luthfi, B. Distiawan, and R. Manurung, "Building an Indonesian named entity recognizer using Wikipedia and DBPedia," in *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014*, 2014, pp. 19–22.

[8] R. A. Leonandya, B. Distiawan, and N. H. Praptono, "A Semi-supervised Algorithm for Indonesian Named Entity Recognition," in *Proceedings - 2015 3rd International Symposium on Computational and Business Intelligence, ISCBI 2015*, 2016, pp. 45–50.

[9] A. S. Wibawa and A. Purwarianti, "Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning," *Procedia Comput. Sci.*, vol. 81, no. May, pp. 221–228, 2016.

[10] I. Alfina, R. Manurung, and M. I. Fanany, "DBpedia entities expansion in automatically building dataset for Indonesian NER," *2016 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2016*, pp. 335–340, 2017.

[11] N. Taufik, A. F. Wicaksono, and M. Adriani, "Named entity recognition on Indonesian microblog messages," *Proc. 2016 Int. Conf. Asian Lang. Process. IALP 2016*, pp. 358–361, 2017.

[12] P. W. A. Dharma, "Developing Indonesian NER using Long Short-Term Memory (LSTM)," Universitas Indonesia, 2016.