Named-Entity Recognizer (NER) for Filipino Novel Excerpts using Maximum Entropy Approach

Karen Mae L. Eboña, Orlando S. Llorca Jr., Genrev P. Perez, Jhustine M. Roldan, Iluminda Vivien R. Domingo, and Ria A. Sagum
CCIS, Polytechnic University of the Philippines, Manila, Philippines
Email: {kaimle21, orly_junior07, genrevperez, jhustine.roldan, dvrdomingo, riasagum31}@yahoo.com

***Abstract* – The Named-Entity Recognizer (NER) for Filipino Novel Excerpts using Maximum Entropy Approach is a study intended mainly for the development of a named entity recognition system specifically for handling texts written in the Filipino language. Its main purpose is to recognize the named entities present in a given text using MaxEnt. The named entities are classified into five, namely: person, place, date, organization, date, time. To measure the performance of the system, solving for the precision, recall, error rate and F-measure was used, both for every named entity and all the named entities as a whole. Novel excerpts were used as a domain for the testing of the system. The results, based on the computation of F-measure, indicated that the system is 80.53% accurate, and best in identifying entity date with 0% error rate but is unsatisfactory in recognizing place and organization, with 29.41% and 13.10% error rates respectively.**

***Index Terms*–named-entity recognition, maximum entropy Approach, named entity, natural language processing (NLP), filipino, information extraction**

I. Introduction

An important research area in the field of information extraction is Named Entity Recognition (NER). This topic was a central theme in the message understanding conferences (MUCs). It has become more important and helpful nowadays due to the large amount of available electronic text, which makes it necessary to build systems that can automatically process and extract information from a given text [1].

NER, which might also be called as proper name classification and identification, is a computational linguistic task in which seeks to classify every word in a document as falling in to one of the four categories: person, location, organization, and other names (date and time, etc..)[2]. In the taxonomy of the computational linguistic tasks, it falls under the domain of "information extraction" [3]. Furthermore, NER System aims to recognize and categorize proper noun in a document into pre-defined target entity classes [4]. It is now considered to be fundamental for many natural languages processing task such as information retrieval, machine translation, information extraction and question answering [5].

The proponents of this study offered a frontline Named-Entity Recognizer system that will accept a novel excerpt in Filipino Language, and classify it to its corresponding category.[6] The proposed system is known as *Named-Entity Recognizer (NER) for Filipino Novel Excerpts using Maximum Entropy Approach* and its main purpose is to recognize the name entity which is generally organized into personal name, place, organization date and time.

II. System Architecture

Named-entity recognition is a form of information extraction in which system seeks to classify every word in a document as being a person-name, organization, location, date, time, monetary value or percentage. The task has particular significance for Internet search engines, machine translation, the automatic indexing of documents, and as a foundation for work on more complex information extraction tasks [7]. Named Entity Recognition can be treated as a tagging problem where each word in a sentence is assigned a label indicating whether it is part of a named entity and the entity type. Thus methods used for part of speech (POS) tagging and chunking can also be used for NER [8]. One of the many approaches in building NER system's high-accuracy is the Maximum Entropy approach, a supervised learning approach, that models a random process by making the distribution satisfy a given set of constraints, and making as few other assumptions as possible Fig. 1 illustrates the system architecture of the developed NER. The Named-Entity Recognizer (NER) for Filipino Novel Excerpts using Maximum Entropy Approach begins by processing unannotated document (Novel Excerpts as the Domain) using several of the information extraction procedures: first, the raw text of the document is split into sentences using a sentence segmenter (sentence processing), and each sentence is further subdivided into words using a tokenizer. The next process will be the detection and

Manuscript received December 30, 2012; revised March 20, 2013.
This research work is supported in part by the Polytechnic University of the Philippines, College of Computer and Information Sciences.

doi: 10.12720/jiii.1.1.63-67

recognition of named entities using the Maximum Entropy Approach, specifically Generalized Iterative Scaling Algorithm. In this step, the system initializes integer variables for the five classes (Person, Place, Org, Date and Time) that will serve as entity index. Next step is that each token is iterated in different handcrafted rules and conditions to search for mentions of potentially interesting entities. Once satisfied a rule, the system initializes a unique set of variables that will be assigned to an Array List. The system tracks if the Array List is the variables of a given event, once satisfied, it will proceed to probability distribution to determine weights for each class (Person, Place, Org, Date and Time) and the system determines the category using the entity index of the given event, Finally, the system uses relation detection to search for likely relations between different entities in the text, put tags in significant entities and output an annotated document.

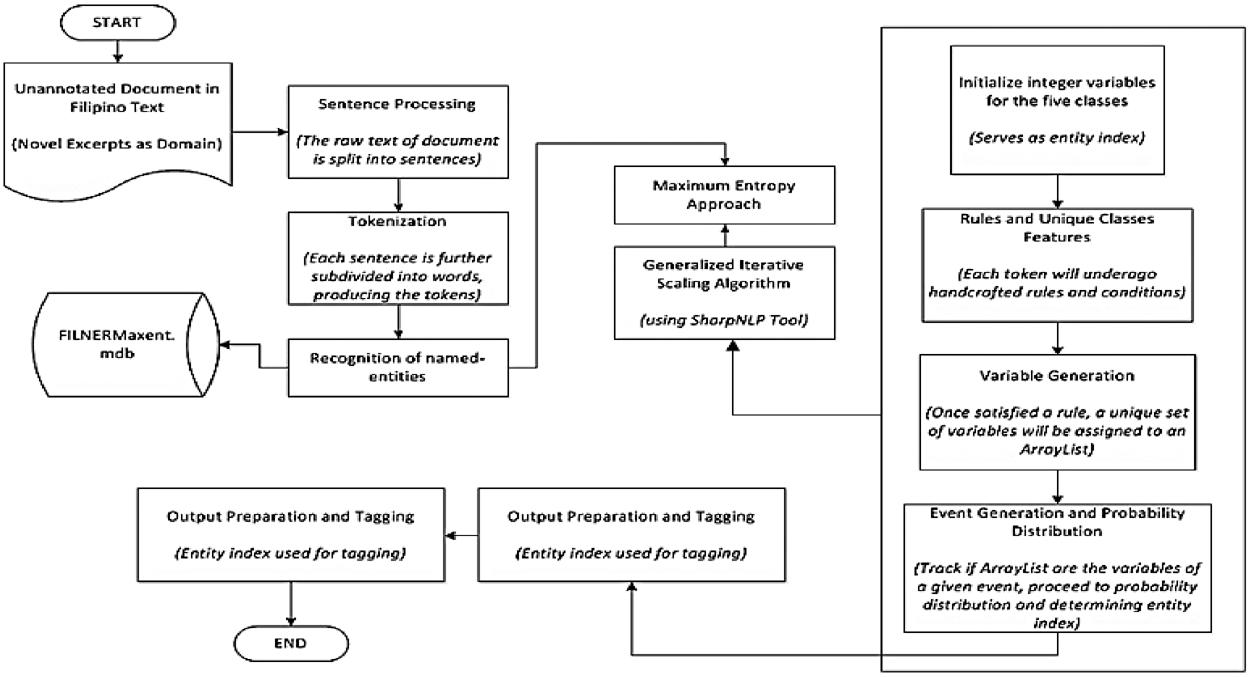
START
Unannotated Document in Filipino Text
(Novel Excerpts as Domain)
Sentence Processing
(The raw text of document is split into sentences)
Tokenization
(Each sentence is further subdivided into words, producing the tokens)
FILNERMaxent. mdb
Recognition of named-entities
Maximum Entropy Approach
Generalized Iterative Scaling Algorithm
(using SharpNLP Tool)
Initialize integer variables for the five classes
(Serves as entity index)
Rules and Unique Classes Features
(Each token will underago handcrafted rules and conditions)
Variable Generation
(Once satisfied a rule, a unique set of variables will be assigned to an ArrayList)
Event Generation and Probability Distribution
(Track if ArrayList are the variables of a given event, proceed to probability distribution and determining entity index)
Output Preparation and Tagging
(Entity index used for tagging)
Output Preparation and Tagging
(Entity index used for tagging)
END

Figure 1. System architecture of named-entity recognizer for Filipino novel excerpts using maximum entropy approach

III. Data Gathering

Data gathering was in the form of test plan that contained specific parameters to test the performance and accuracy of the NER. These parameters include Precision, Recall, Error Rate and F-measure. The researchers conducted tests in each part of the development phase of the system. The results were recorded and the tests were repeated until the system is qualified for implementation. The researchers came up with a specific date and venue to conduct the implementation. The researchers tested the system’s functionalities before testing. The next step is the implementation of the system. The researchers tested the system and record the output in a test plan. The researchers gathered the test plan for evaluation and computation. The last step is the data analysis. The accuracy of the NER was measured through the use of Harmonic Mean, or F-measure.

IV. Evaluation and Results

Evaluation Metric mathematically defines how to measure the system's performance against human-annotated, gold standard. Here for every experiment before checking the performance of the system, a human tagged test data is prepared to evaluate the system. A set of 100 randomly selected novel excerpts were gathered for the testing of the system. The researchers checked its grammar and sentence construction. The files were tested individually and the results were tallied for each named entity based on the classification of results set by [9]:

No. of Recognized Entities (T) – refers to the actual number of named entities tagged by the system. *No. of Correctly Recognized Entities (CT)* – refers to the actual number of named entities that are correctly tagged by the system, that is, the identification of the boundary of the named entity is correct and the determining class is also correct, i.e. person, place, org, date and time. *Wrong Tag (WT)* – refers to the actual number of entities that are incorrectly tagged, that is, all that do not satisfy the condition of being a Correctly Recognized Entity such as over tag and a tagged entity with a different classification that may lead to ambiguity [10].

The performance of the NER will be measured through the use of the Precision (P), Recall (R), Error Rate (ER) and Harmonic Mean, or F-measure (F) [7]. The *F-measure* (also F1 score) is the weighted average of the values of the Precision and Recall. By multiplying the values by 2 and dividing it by the sum of the Precision and Recall, we can get the harmonic mean of the system. A high F1 score will imply a good performance of the system. A F1 score reaches its best value at 1 and worst score at 0. The formula for the Harmonic mean is as follows:

$$F = \frac{2PR}{P+R} \quad (1)$$

The F-measure is a measure of a test's accuracy. It considers both the Precision (P) and Recall (R).

Precision refers to the measure of the number of correct named entities out of all named entities that were found by the algorithm. The formula for Precision is as follows:

$$P = \frac{\text{No. of Correctly Recognized Entities (CT)}}{\text{No. of Recognized Entities (T)}} \quad (2)$$

Recall refers to the number of retrieved entities out of all the entities in the corpus. The formula for recall is as follows:

$$R = \frac{\text{No. of Correctly Recognized Entities (CT)}}{\text{No. of Human Annotated Entities(NT)}} \quad (3)$$

Note: *No. of Human Annotated Entities (NT)* refers to the actual number of named entities that must be tagged in the corpus. In this study, before the corpus is fed into the system, the researchers already have a prepared version of the corpus which is manually tagged to determine the actual number of named entities found in the corpus. *Error Rate* refers to the measure of incorrect recognized named entities over the recognized named entities. The formula for Error Rate is as follows:

$$ER = \frac{\text{Wrong Tag (WT)}}{\text{No. of Recognized Entities (T)}} * 100 \quad (4)$$

Based on the information gathered during the development and implementation of Named-Entity Recognizer for Filipino Novel Excerpts using Maximum Entropy Approach, the researchers found out the following results of the study:

TABLE I. SUMMARY OF RESULTS

NAMED ENTITY	PRECISION	RECALL	ERROR RATE	F-MEASURE
PERSON	86.89	76.26	12.95	81.23
PLACE	86.90	67.59	13.10	76.04
ORG	70.59	50	29.41	58.54
DATE	100	90.63	0.00	95.08
TIME	91.11	82	8.89	86.32
TOTAL:	**86.93**	**75**	**12.68**	**80.53**

The summary assessment of the performance of the system based on 100 files tested in terms of Precision, Recall, Error Rate and F-Measure was computed as 86.93%, 75%, 12.68% and 80.53% respectively.

V. CONCLUSION

In this study, the researchers presented a NER system in Filipino language using Maximum Entropy Approach. Based from the results of the study, the researchers have come up with the following conclusions:

1) The experimentation done between the developed Named-Entity Recognizer for Filipino Novel Excerpts using Maximum Entropy Approach as a tool for tagging named entities and manual annotation conveyed a Precision of 86.93%, 75% for the measure of Recall, and 12.68% in Error Rate. Hence, the researchers concluded that the developed system is acceptable in terms of overall Precision and Error Rate, however, this doesn't apply that the performance of the system is also acceptable in each classes because the researchers found out that some classes obtained only a passing rate performance for the measure of Precision, specifically the Org class which only obtained a Precision of 70.59%. The researchers also concluded that the performance of the system needs further development on the measure of Recall. Data shows that the performance of the system obtained only a fair rate percentage for the overall Recall and failing remarks in some individual classes such as Person, Place and Org.
2) The developed Named-Entity Recognizer for Filipino Novel Excerpts using Maximum Entropy Approach yielded an overall F-measure of 80.53% which is above average. The researchers concluded that the developed Named-Entity Recognizer is accurate and has an acceptable performance in recognizing entities found in the corpus as a whole but it doesn't apply that the performance of the system is also acceptable in individual classes because the researchers yielded data that shows negative results regarding the F-measure of certain classes such as Place and Org, individually obtaining a F-measure of 76.04% and 58.54%.
3) The researchers concluded that positive results were obtained for Date and Time classes because the construct of these entities were limited and could easily be recognized.
4) The most dominant number of named-entities found in Filipino Novel Excerpts is Person and Place classes while Org, Date and Time classes are occasionally seen.
5) Since Filipino novel writers have different styles in constructing sentences and some of those do not match the sentence construct followed by the Maximum Entropy Approach in recognizing entities, some output of the system are inaccurate.
6) The researchers concluded that if the corpus that was given as input in the system was in the right sentence construct as the construct the developed system follows, named entities can justly be recognized by the system.

VI. RECOMMENDATION

Based on the results and the conclusion derived by the researchers, the following recommendations are stated:

1) The researchers recommend further development on the performance of the system in recognizing and tagging named-entities by adding more Maximum Entropy features or rules on proper Filipino Novel sentence construct to be able to increase and improve the system's accuracy and performance relating to overall Precision, Recall and Error Rate, especially on individual classes such as the Precision of Org Class, Recall of Person, Place and Org classes and F-Measure of Place class.
2) To eliminate and reduce the annotation errors encountered, the researchers suggest more extensive Maximum Entropy features that will determine named-entity boundary and concatenation.
3) The use of Named-Entity Recognizer for Filipino Novel Excerpts using Maximum Entropy Approach as a tool for Information Extraction and as a basis for another study is earnestly proposed.
4) Future researchers may develop a Named-Entity Recognizer using Maximum Entropy Approach in a different domain or in a different language.
5) The researchers strongly suggest adding a feature of the system which is Analyzed Error Validation – a dialog that determines the error on the output of the system and suggests an input which is in the right sentence construct to lessen the Error Rate on the output of the system.
6) The researchers strongly suggest adding a feature of the system which is Help Menu – a menu guide that tells the users what is the proper sentence construct of the corpus to be fed into the system.
7) For further study, this paper could be of great help in aiding knowledge regarding Named-Entity Recognition, Information Extraction and Retrieval and can be used in the development of systems concerning other Natural Language Processing tasks such as question answering, machine translation and data indexing.

ACKNOWLEDGEMENT

This paper would not have been possible without the guidance and the help of those people who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study. Above all, we express our gratitude and thanks to the Almighty God who gave us strength and who put up courage within us to stand through this tough and motivating experience. He really never failed us. Second, to our parents who managed to fully-support, understand and trust us unconditionally, from financial assistance to time exertion that we weren't able to be home on time for sometimes. For them who taught us that the best kind of knowledge to have is that which is learned for its own sake and even the largest task can be accomplished if it is done one step at a time. To our friends, who are in times of confusion, have been there to help us and clear our minds with what they have and know. We would like to extend our gratitude to the people who contributed their help and experience for the completion of this study – the faculty members of the Department of Computer Science of the College of Computer and Information Sciences (CCIS) from the Polytechnic University of the Philippines. Lastly, this research and study would have not been accomplished without our thesis technical adviser Prof. Ria A. Sagum who encouraged, challenged and patiently guide us throughout the thesis program. She, together with our faculty thesis adviser, Prof. Iluminada Vivien R. Domingo, who uncomplainingly guided us through the research process and taught us to make thesis documents and helped us finalized this paper. We thank them all.

REFERENCES

[1] T. Zhang and D. Johnson, "A robust risk minimization based named entity recognition system," in *Proc. Seventh Conference on Natural Language Learning at HLT-NAAC*, 2003, vol. 4, pp. 204-206.

[2] G. V. S. Raju, B. Srinivasu, Dr. S. Viswanadha, and K. S. M. V. Kumar, "Named entity recognition for telugu using maximum entropy model," *Journal of Theoretical and Applied Information Technology*, pp. 125-130, 2010.

[3] N. N. Lim, *Named Entity Recognizer for Filipino Texts*, De La Salle University-Manila, Manila, 2007.

[4] Q. T. Tran, T. X. T. Pham, Q. H. Ngo, D. Dinh, and N. Collier. (2006). Named Entity Recognition in Vietnamese Documents. [Online]. pp. 5-13. Available: http://www.nii.jp/pi/n4/4_5.pdf

[5] D. Maynard, V. Tablan, and H. Cunningham. (2006). NE Recognition Without Training Data on a Language You Don't Speak. [Online]. Available: http://gate.ac.uk/sale/acl03/surprise.pdf

[6] H. L. Chieu and H. T. Ng. (2003). Named Entity Recognition with a Maximum Entropy Approach. [Online]. pp. 1-4. Available: http://acl.ldc.upenn.edu/W/W03/W03-0423.pdf

[7] A. Borthwick. (September 2000). A Maximum Entropy Approach to Named Entity Recognition. [Online]. pp. 1-105. Available: http://www.cs.nyu.edu/web/Research/Theses/borthwick_andrew.pdf

[8] J. R. Curran and S. Clark. (2005). Language Independent NER using a Maximum Entropy Tagger. [Online]. pp. 204-206. Available: http://acl.ldc.upenn.edu/W/W03/W03-0424.pdf

[9] R. A. Sagum, *A Named Entity Recognizer for Filipino Text*, Manila, 2011.

[10] S. Biswas, S. P. Mishra, S. Acharya, and S. Mohanty. (2003). A Hybrid Oriya Named Entity Recognition System: Harnessing the Power of Rule. [Online]. pp. 1-6. Available: http://www.researchgate.net/publication/47554241_A_Hybrid_Oriya_Named_Entity_Recognition_system_Harnessing_the_Power_of_Rule

Ria A. Sagum was born in Laguna, Philippines on August 31, 1969. She took up Bachelor in Computer Data Processing Management from the Polytechnic University of the Philippines and Professional Education in EulogioAmang Rodriguez Institute of Science and Technology. She received her master degree in De La Salle University in 2012.

She is currently an instructor in both Polytechnic University of the Philippines in Sta. Mesa, Manila and University of Santo Tomas in Manila.

Prof. Sagum has been a presenter of different conferences, including the 2012 International Conference on e-Commerce, e-Administration, e-Society, e-Education, and e-Technology and is a member of the Computing Society of the Philippines and the Natural Language Processing Special Interest Group.

Iluminada Vivien R. Domingo was born in Manila, Philippines on November 29, 1965. She graduated from the Polytechnic University of the Philippines Bachelor in Business Education, Magna Cumlaude in 1986. She received her Master's degree in Business Administration in 1990 at the University of Santo Tomas. She received her Doctorate degree in Business Administration in 2004 from the Polytechnic University of the Philippines.

She is currently an Associate Professor at the Polytechnic University of the Philippines in Sta. Mesa, Manila. Assoc. Prof. Domingo is a member of the Philippine Society of Educators (PSITE)

Karen Mae L. Eboña was born in Manila, Philippines on September 21, 1992. She graduated from Imus National High School, and is currently taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. She was an intern in the IT department of Euro RSCG Manila.

Jhustine M. Roldan was born in Makati City, Philippines on September 02, 1992. He graduated from Navotas National High School in Navotas City as the tenth honorable mention, and is currently taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. He has worked as an intern in the IT Department of Fujitsu Ten Solutions Philippines.

Orlando S. Llorca Jr. was born in Mandaluyong City, Philippines on December 13, 1992. He graduated from Sen. Renato CompañeroCayetano Memorial Science and Technology High School as the first honorable mention, and is currently taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. He was an intern in the IT department of Melvindave Consulting Inc.