Named Entity Recognition Using Support Vector Machine for Filipino Text Documents

Jonalyn M. Castillo, Marck Augustus L. Mateo, Antonio D. C. Paras, Ria A. Sagum, and Vina Danica F. Santos

***Abstract*—Named entity recognition involves processing of texts to identify and classify entities such as names of person, place, organization, etc. In this study, a system for a named entity recognizer for Filipino texts using support vector machine was developed, and its performance was evaluated and compared to an existing named entity recognizer intended for the same language, but uses a rule-based approach. Based from the results, the named entity recognizer using support vector machine performed best in tagging named entity class date with 95.52% f-measure, achieving 84.97% overall f-measure.**

***Index Terms*—Information extraction, named entity recognition, natural language processing, support vector machine**

I. Introduction

Named entity recognition (NER) involves processing texts to recognize and classify named entities. NER is often performed using a statistical tagger which learns patterns for the recognition of names from manually-annotated document [1]. They are commonly used to train statistical machine learners but are limited in scope due to the cost of manual annotation [2].

Named entities (NE) are words or phrases that can be classified as a name of person, organization, location, etc. NEs are valuable in natural language processing fields such as automatic text summarization systems, information retrieval, information extraction, question answering, and machine translation [3].

Capitalization in Filipino language indicates a proper noun, yet NE Recognition cannot be done by considering the case of the first letter of each word alone [4]. For example, "Juan dela Cruz" contains lower case words. On the other hand, providing a list of NEs and let the system search for these entities in the document can address the problem. This approach is error-prone because keeping an up-to-date list of the NEs can be a very time-consuming task. Given that the list is up-to-date, there is still the problem of ambiguity of some NE, for example "Quezon" can either refer to a name of a person or a place. Thus, a statistical approach is needed to achieve desirable results.

Manuscript received November 9, 2012; revised January 23, 2013.

C. M. Castillo, M. A. L. Mateo, A. D. C. Paras, and V. D. F. Santos are currently undergraduate students from the College of Computer Management and Information Technology, Polytechnic University of the Philippines (e-mail: jona_lyne1224@yahoo.com, marck.augustus.mateo@gmail.com, gigabyte_paras@yahoo.com, vinadanica0430@yahoo.com).

R. A. Sagum is with the faculty of the College of Computer Management and Information Technology of the Polytechnic University of the Philippines (e-mail: riasagum31@yahoo.com)

This study intends to create a named-entity recognition system for Filipino texts that will utilize one of the learning based approaches, the Support Vector Machines (SVM) [5]. SVM is a supervised machine learning algorithm for binary classification, which has been successfully applied to a number of particle problems [6], [7].

II. System Architecture

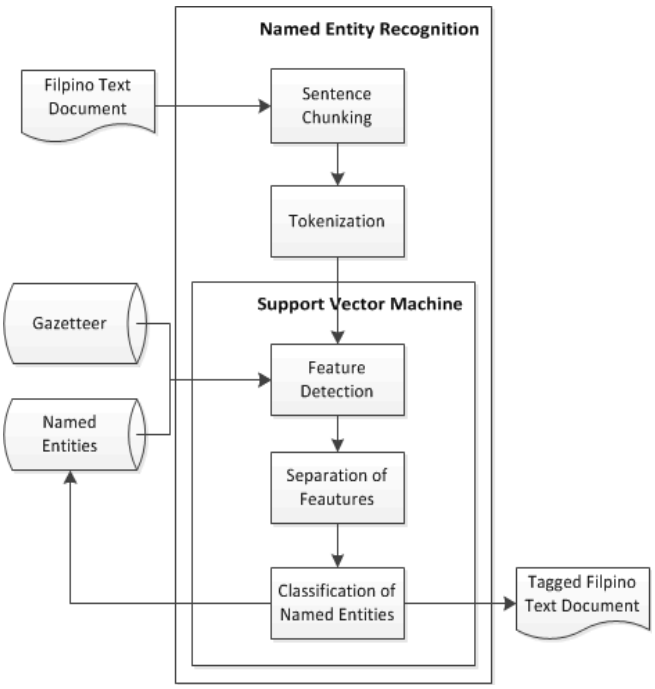

Fig. 1. System Architecture

The input of the system is text documents utilizing the Filipino language and should be grammatically correct. It will first go through sentence chunking. The process simply separates the text by sentences. End punctuation marks like periods, question marks and exclamation points usually mark the end of a sentence but period may also denote that the word before it is abbreviated. Having a list of abbreviated words was used as a solution to the problem.

The sentences will then proceed to tokenization. The tokens produced will be sent to the Support vector machine module. The module starts by identifying the feature set of every token. To predict the classification of the NE, the module will then separate the features into positive and negative features. The result as to what classification of the NE is depends on the margin created by the positive and negative features. Once identified, the module will then tag the NE according to its classification. The tagged text will be the output of the system.

DOI: 10.7763/IJFCC.2013.V2.220

III. Methodology

A. Training the System

Training the system is required in a statistical approach. It allows the system to learn how to properly identify and classify NEs. Training is done by procuring available texts. One-by-one, NEs in these texts are manually tagged according to their respective classes, namely: person, place, organization, date, and etc.

The training data were gathered from the internet, which mainly consists of political speeches.

B. System testing

System testing is also done by gathering texts from the internet. These texts must be different from the ones used as training data to avoid biased results.

The results are classified into *CT*, *T*, and/or *NT*. Named entities that are tagged by the system with the correct classification, i.e., person, place, org, date, or etc, are considered as *CT*. Those considered as *T* are the named entities tagged by the system. *NT*s are named entities present in the document, whether it is recognized by the system or not.

C. Evaluation

A common way to measure the performance of a named entity recognizer is by using F-measure. According to [8], F-measure is the harmonic mean of precision (P) and recall (R). To compute F-Measure, use the formula:

$$F = \frac{2PR}{P+R} \tag{1}$$

Precision is the percentage of tagged entities that are correct. To acquire precision, we used:

$$P = \frac{CT}{T} \tag{2}$$

On the other hand, Recall is the percentage of correct entities that are tagged with the formula:

$$R = \frac{CT}{NT} \tag{3}$$

IV. Results

Using (1)-(3), the performance of the system was identified with the following results and was compared to an existing system called NERF [9]:

TABLE I: Performance of the System

Named Entity	Precision	Recall	F-Measure
PERSON	88.64%	88.64%	88.64%
PLACE	96.40%	90.68%	93.45%
ORG	70.21%	89.91%	78.85%
DATE	94.12%	96.97%	95.52%
ETC	84.08%	74.88%	79.21%
OVERALL	84.70%	85.24%	84.97%

Table I illustrates that the system performed best in tagging class *DATE* with an f-measure of 95.52% and worst in tagging names of organizations with an f-measure of 78.75%, possibly because of few feature sets used for determining class *ORG*.

TABLE II: Performance of NERF

Named Entity	Precision	Recall	F-Measure
PERSON	95.70%	95.96%	95.83%
PLACE	84.67%	89.23%	86.89%
ORG	50.00%	50.00%	50.00%
DATE	59.26%	51.61%	55.17%
ETC	63.79%	63.79%	63.79%
OVERALL	85.29%	85.83%	85.56%

The overall performance of the system is less than that of NERF, however, the performance in tagging different classes such as PLACE and DATE is higher compared than the performance of NERF.

V. Conclusion

In this study, a named entity recognizer using support vector machine is evaluated in its performance in the Filipino language. Based on the results, the F-measure or the performance of the system is 84.97%. The system performed best in tagging dates with an F-measure of 95.52% on the other hand, performed worst in tagging names of organizations with an F-measure of 78.85%.

VI. Recommendation

The following recommendations may further improve the performance of the system:

1) The use of regular expressions to improve system performance [10].
2) Adding more feature sets to aid in feature detection.

Acknowledgement

The researchers would like to show their sincerest gratitude to all those who, on one way or another, helped a lot in their project. We especially thank the College of Computer Management and Information Technology of the Polytechnic University of the Philippines.

References

[1] J. Nothman, "Learning Named Entity Recognition from Wikipedia," Honours Thesis, University of Sydney, Sydney, Australia, 2008.
[2] M. Banko and E. Brill, "Scaling to Very Very Large Corpora for Natural Language Disambiguation," in *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001.
[3] T. Solorio, "Improvement of Named Entity Tagging by Machine Learning," Tonantzintla, Puebla, Mexico: Instituto Nacional de Astrofísica, Óptica y Electrónica, 2005.
[4] L. E. Lim, J. C. New, M. A. Ngo, M. Sy, and N. R. Lim, "A Named-Entity Recognizer for Filipino Texts," in *Proceedings of 4th National Natural Language Processing Research Symposium*, 2007.
[5] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273 - 297, 1995.
[6] J. Gimenez and L. Marquez, "Fast and Accurate Part-of-Speech Tagging: The Support Vector Machine Approach Revisited," in *Proceedings of the International Conference RANLP-2003 (Recent Advances in Natural Language Processing*, 2003.

[7] T. Kudo and Y. Matsumoto, "Chunking with Support Vector Machines," in *Proceedings of 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA, USA, 2001.
[8] C. Manning. Information Extraction and Named Entity Recognition. [Online]. Available: http://www.stanford.edu/class/cs124/lec/Information_Extraction_and_Named_Entity_Recognition.pptx.
[9] R. A. Sagum and R. Roxas, "Named-Entity Recognition for Filipino Texts (NERF)," MSCS Thesis, De Lasalle University, Manila, Philippines, 2012.
[10] L. Karttunen, J. P. Chanod, G. Grefenstette, and A. Schiller, "Regular Expresions for Language Engineering," *Natural Language Engineering*, vol. 2, no. 4, 1996.

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 Eulogio Amang Rodriguez Institute of Science and Technology. She received her master degree, Master of Computer Science, in De La Salle University in 2012.

She is currently teaching at the Department of Computer Science, College of Computer Management and Information Technology, in the Polytechnic University of the Philippines in Sta. Mesa, Manila and a lecturer at the Information and Computer Studies, Faculty of Engineering, in the University of Santo Tomas in Manila.

Ms. 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 National Natural Language Processing Research Symposium. She is a member of different professional associations including ACMCSTA and an active member of the Computing Society of the Philippines- Natural Language Processing Special Interest Group.

Jonalyn M. Castillo was born on June 13, 1993 in Oriental Mindoro. She graduated from Leuteboro National High School in Oriental Mindoro and is currently taking up BS Computer Science in the Polytechnic University of the Philippines, Manila. She was an intern at Indra Philippines.

Marck Augustus Mateo was born in Caloocan City on August 28, 1993. He graduated from Angelicum College Quezon City and is currently taking up BS Computer Science in the Polytechnic University of the Philippines, Manila. He was a junior Java developer in Ivant Technologies and Business Solutions, Inc.

Antonio Paras was born on January 17, 1993 in Angono, Rizal. He graduated from Rizal National Science High School and is currently taking up BS Computer Science in the Polytechnic University of the Philippines, Manila. He was an intern in Security Bank Corporations.

Vina Danica Santos was born in Pasig City on April 30, 1993. She graduated from Casimiro A. Ynares Sr. Memorial National High School and is currently taking up BS Computer Science in the Polytechnic University of the Philippines, Manila. She was an intern in Indra Philippines.