Named Entity Recognizer for Filipino Text Using Conditional Random Field

Ana Patricia T. Alfonso, Illuminada Vivien R. Domingo, Mary Joy F. Galope, Ria A. Sagum, Rachelle B. Villar, and Jobert T. Villegas

***Abstract*—The study for a Named Entity Recognizer for Filipino Text Using Conditional Random Field (NERF-CRF) focused creating a system which identifies and classifies named entities present in a given corpus. The named entities were classified into four, namely: person, place, date and org. Named entities that are identified but do not fall in the four classifications are tagged as etc.**

Different modules were created to achieve the study's purpose, including a tokenizer and a part-of-speech tagger. The conditional random field approach was used in the classification of identified named entities. Filipino biographies were the corpus used in testing the system. The results, based on solving for the F-measure, indicate that the system is 83% accurate, and best in identifying named entity Date with 0% error rate but is unsatisfactory in distinguishing named entity place and org, with 42% and 33% error rates respectively.

***Index Terms*—Conditional random field, extraction, named entity recognition, natural language processing**

I. Introduction

Named entity recognition (NER) is a process that automatically examines a corpus and tags the proper nouns present in the latter. These proper nouns, called named entities, can be names of person, organization, place, date, etc.

As NEs (Named Entities) such as organizations' names, persons' names and locations' names contain more informative information, NE recognition is the fundamental for efficient information access. It is processing text to identify and classify names, an important component in many NLP applications, enabling the extraction of useful information from documents.

Named entities of the Filipino language have never been limited to words with its first letter capitalized, as with the named entity place "Lanao del Sur". Moreover, few named entities have different classification for every use. Named entity "Rizal" can either refer to the person Jose Rizal, or a place in the Southern Luzon of the Philippines.

These issues of capitalization and ambiguity can be addressed by establishing a system that simply looks up similar entries in a database created beforehand. However, this approach is relatively inadvisable, merely for the fact that to create a database for such a system is terribly unwieldy, and the runtime for the system will be inevitably slow. Furthermore, such system requires an update from time-to-time, given that new named entities such as names of persons and organizations come up each and every day.

For this reason, NER is often performed using a statistical tagger which learns patterns for the recognition of names from manually-annotated textual corpora. A few corpora have been constructed as gold standards—i.e. they define correct annotations for NER by example. They are commonly used to train statistical machine learners but are limited in scope due to the cost of manual annotation. This is a problem because others have shown that more training data leads to higher accuracy language models [1], [2].

While many named entity recognition systems exist in the market today, very few, if any, have been designed specifically for handling texts written in the Filipino language. Most software packages and implementations for NER accept a stream of English text and extracts names of people, places, and companies or organizations [3]. For this, a NER for Filipino Text was formulated, and is supposed to be of good performance.

The following sections presents the composition of the system, with the testing, results, conclusions and recommendations shown in Section IV, V, VI and VII respectively.

II. The Conditional Random Field

The NERF-CRF is concerned on the use of Conditional Random Fields to classify tagged named entities within the Filipino text. Lafferty, McCallum and Pereira presented conditional random fields in [4], a framework for building probabilistic models to segment and label sequence data. As the study implies, conditional random fields offer several advantages over hidden Markov models and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. Conditional random fields also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models, which can be biased towards states with few successor states. They have presented iterative parameter estimation algorithms for conditional random fields and compare the performance of the resulting models to HMMs and MEMMs on synthetic and natural-language data. As they concluded, conditional random fields offer a unique combination of properties: discriminatively trained models for sequence segmentation and labeling; combination of arbitrary, overlapping and agglomerative observation features from both the past and future; efficient training and

Manuscript received November 9, 2012; revised January 20, 2013.

R. A. Sagum and I.V.R. Domingo are with the Faculty of the College of Computer Management and Information Technology of the Polytechnic University of the Philippines (e-mail:riasagum31@yahoo.com; dvrdomingo@yahoo.com)

A. P. T. Alfonso, M. J. F. Galope, R. B. Villar, and J. T. Villegas are with the Polytechnic University of the Philippines (e-mail: anapat1219@yahoo.com; mharyjoy_galope@yahoo.com.ph; rachelle_villar@yahoo.com; jobert21492@yahoo.com).

DOI: 10.7763/IJFCC.2013.V2.189

decoding based on dynamic programming; and parameter estimation guaranteed to find the global optimum [4].

Another NER utilizing conditional random field was later conducted by Mao et al in [5]. The paper presented an approach that exploits non-local information to improve the NER recall. Several kinds of non-local features encoding entity token occurrence, entity boundary and entity class were explored under Conditional Random Fields (CRFs) framework. Experiments on SIGHAN 2006 MSRA (CityU) corpus indicate that non-local features can effectively enhance the recall of the state-of-the-art NER systems. Incorporating the non-local features into the NER systems using local features alone, their best system achieves a 23.56% (25.26%) relative error reduction on the recall and 17.10% (11.36%) relative error reduction on the F-measure; the improved F-measure 89.38% (90.09%) is significantly superior to the best NER system with F-measure of 86.51% (89.03%) participated in the closed track [5].

III. System Architecture

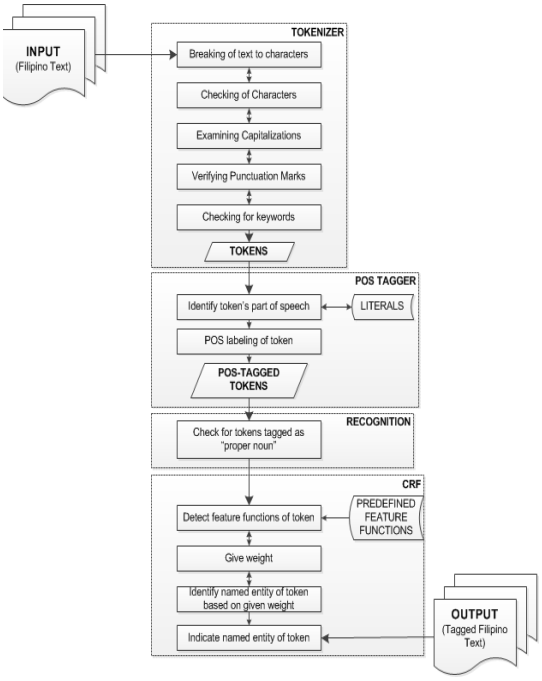

Fig. 1. System Architecture of NERF-CRF

With Filipino text as the input, the system will process the said input through different modules that act together to achieve the classifying aim of the system.

The tokenizer, which is the foremost module of the system, is in charge of breaking down the whole text into single characters. The properties of the characters, whether the characters are capitalized, numeric, or alphabet, are then checked. The use of punctuation marks in the text is also assessed in this module. Moreover, keywords such as *sa*, *si*, and *noong* are also checked. After these series of checking, the tokenizer will produce tokens, composed of combined characters that form a word or a phrase.

Once the tokens have been established, the POS (part-of-speech) tagger will simply identify tokens' part-of-speech with the use of literals which were initially listed from previous training data, and will label the token with its corresponding part-of-speech.

The Recognition module simply looks for tokens which are labeled as "proper noun" and passes it on to the next module.

The CRF module includes the main focus of this NER, the process of classifying nouns based on the classifier's logic. CRF starts by identifying the compatibility of the token to a named entity, using feature functions which are also predefined. Once a characteristic of the token satisfies a feature function of a certain named entity, the weight for that named entity increases. Equation (1) was used in the said process:

$$\frac{\sum_{i=0}^{n} W_i}{\sum_{i=0}^{n} W'_i} \tag{1}$$

where W_i represents weights from features possessed by the token, and W_i' are weights of features not possessed by the token. The result as to what named entity of the token is depends on (1), or the final weight given by the module. Once identified, it will finally append the named entity of the token to itself. The tagged version of the input will be the output of the system.

IV. Testing

A set of biographies were gathered for the testing of the system. After selecting 50 of these through random selection, the researchers examined 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 [6], such as number of entities that must be tagged by the system (NT), number of entities that were correctly tagged by the system (CT), number of entities that were wrongly tagged by the system (WT) and the total number of entities that were tagged by the system whether correctly or wrongly tagged (T).

V. Results

To be able to assess the accuracy of the system, the proponents used (2).

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

The F-measure is defined as a harmonic mean of precision (P) and recall (R). The F-measure is a measure of a test's accuracy [7], [8]. It considers both the precision P and the recall R of the test to compute the score:

$$P = \frac{correct \quad results}{number \quad of \quad all \quad returned \quad results} \tag{3}$$

$$R = \frac{number\ of\ correct\ entities}{number\ of\ all\ returned\ results} \tag{4}$$

P is the number of correct results divided by the number of all returned results and r is the number of correct results

divided by the number of results that should have been returned.

Moreover, to identify the system's performance in terms of its error rate, the following formula was used:

$$E = \frac{number\ of\ wrong\ tags}{total\ number\ of\ tagged\ entities} \quad (5)$$

In identifying the performance of the developed NERF-CRF in terms of Precision, Recall, Error Rate and F-Measure, the following results were computed:

TABLE I: SUMMARY OF RESULTS

NAMED ENTITY	PRECISION	RECALL	ERROR RATE	F-MEASURE
PERSON	98.89	99.07	1.11	**98.98**
PLACE	56.59	98.59	42.80	**71.91**
ORG	65.29	46.47	33.88	**54.30**
DATE	100	100	0.00	**100**
ETC	87.83	70.21	11.03	**78.04**
TOTAL:	**80.68**	**86.12**	**18.86**	**83.31**

The summary assessment of the performance of the system based on 50 files tested in terms of Precision, Recall, Error Rate and F-Measure was computed as 80.68%, 86.12%, 18.86% and 83.31% respectively.(see Table I)

VI. CONCLUSION

Based from the findings of the study, the following conclusions were reached through the series testing and evaluation:

1) The overall performance of the developed NERF-CRF is above average, with an F-measure of 83%.
2) The NERF-CRF is very effective in tagging named entity date with 100% accuracy in terms of F-measure.
3) The developed system is worst in tagging named entity organization with an accuracy of only 53% based on F-measure.
4) The performance of the system will increase further if more feature functions were fed into the system.

VII. RECOMMENDATION

The following suggestions might be helpful for those future researchers who will also specialize in any topic relating to NER:

1) Compare the significant difference of the developed NER to other existing Named Entity Recognizer for Filipino Text that utilizes different algorithm to further emphasize the usefulness of the CRF approach.
2) Tag other named entities other than name of person, place, org, date and etc such as product name, monetary values, numbers, percentage and time.
3) This system can be improved by correctly tagging the entities after the word taga, the developed system was not able to tag the named entities correctly once it is preceded by the word taga.
4) Add more feature functions [9] to maximize the conditional probability of labels for every input sequence and to reduce the error rate in word segmentation.
5) A gazetteer can be used to enhance the performance of the system and solve the ambiguity between named entities person and organization.
6) Improve the performance of the developed NER and make it domain independent [10].
7) Recognize some Unicode characters such as the left and right single quotes. The developed system was not able to recognize a named entity that has open parenthesis and single quotation marks which generated errors in the recognition task

ACKNOWLEDGMENT

The researchers would like to express their deepest gratitude to the following people who are the sole inspiration and encouragement of this wonderful accomplishment: to our family, to the Polytechnic University of the Philippines and the College of Computer Management and Information Technology, and most importantly, to our God Almighty.

REFERENCES

[1] M. Banko and E. Brill, "Scaling to a Very Very Large Corpora for Natural Language Disambiguation," in *Proceedings of the 39th Annual Meeting of the Assocation for Computational Linguistics*, Toulouse, France, 2001.
[2] J. Nothman, "Learning Named Entity Recognition from Wikipedia," Honours Thesis, University of Sydney, Sydney, Australia, 2008.
[3] L. E. Lim, J. C. New, M. A. S. M. C. Ngo, and N. R. Lim, "A Named Entity Recognizer for Filipino Text," in *4th National Natural Language Processing Research Symposium Proceedings*, Manila, Philippines, 2010.
[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, Williamstown, MA, USA, 2001.
[5] X. Mao, W. Xu, Y. Dong, S. He, and H. Wang, "Using Non-Local Features to Improve Named Entity Recognition Recall," in *The 21st Pacific Asia Conference on Language, Information and Computation*, Seoul, Korea, 2007.
[6] R. Sagum and R. Roxas, "A Named Entity Recognition for Filipino Text (NERF)," MSCS Thesis, De La Salle University, Manila, Philippines, 2012.
[7] Wikipedia. (2012). [Online]. Available: http://en.wikipedia.org/wiki/F1_score.
[8] Y. Sasaki. The Truth of the F-measure. [Online]. Available: http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf.
[9] X. Zhu. Conditional Random Field. [Online]. Available: http://pages.cs.wisc.edu/~jerryzhu/cs769/CRF.pdf.
[10] K. F. Edward and V. Baryamureeba, "Towards Domain Independent Named Entity Recognition," *International Journal on Computing and ICT Research,* vol. 2, no. 2, pp. 84-95, 2009.

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.

Ana Patricia T. Alfonso was born in Manila, Philippines on January 12, 1993. She graduated from St. Nicholas School of Marikina as the first honorable mention, and is currently taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines.

Mary Joy F. Galope was born in Manila, Philippines on November 10, 1992. She graduated from Sico 1.0 National High School in San Juan, Batangas, and is currently taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines.

Rachelle B. Villar was born in Quezon City, Philippines on November 30, 1992. She graduated from Don Alejandro Roces Sr. Science-Technology High School in Quezon City, and is currently taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. She has worked as an intern in the IT Department of RCBC Bankard, Inc.

Jobert T. Villegas was born in Manila City, Philippines on February 14, 1992. He graduated from Pasay City North High School as the third honorable mention, and is currently taking up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. He has been an intern programmer in RCBC Bankard, Inc.

Iluminada Vivien R. Domingo was born in Quezon City, Philippines on November 29, 1965. She took up Bachelor in Business Education, Magna Cumlaude, 1986 from the Polytechnic University of the Philippines. She received her Masters in Business Administration from University of Santo Tomas and her Doctor in Business Administration from Polytechnic University of the Philippines in 2004. She is currently an Associate Professor in the Polytechnic University of the Philippines and in University of Santo Tomas as well.