The Application of Transformation-based Learning in the Development of a Named Entity Recognition System for Filipino Text

Quervin Lloyd L. Buco, John Lester L. Capcap, Jayvee Carlo A. Hermocilla, and Carmina S. Yumul
CCIS, Polytechnic University of the Philippines, Manila, Philippines
Email:{quervinbuco, fx_lester06, jayvee_hermocilla, mina.yumul}@yahoo.com

Ria A. Sagum and Angelito G. Pastrana
CCIS, Polytechnic University of the Philippines, Manila, Philippines
Email: {riasagum31, angelgpastrana}@yahoo.com

***Abstract* – In this paper, the researchers define the task of named entity recognition and its role in information extraction, data mining, and most importantly, natural language processing. It discussed the implementation of the transformation-based learning algorithm in the creation of the named entity recognition system for the Filipino language, and evaluated the system's performance in comparison with another named entity recognition system which used a rule-based algorithm.**

***Index Terms* – data mining, information Extraction, natural language processing, transformation-based learning**

I. Introduction

Natural Language Processing (NLP) is an interdisciplinary field that uses computational methods to investigate the properties of written human language and to model the cognitive mechanisms underlying the understanding and production of written language [1].

With the massive amount of written and spoken language that a computer can process and learn, which from all of these are relevant and useful? That is one question that needs to be answered. Thus, the task known as Information Extraction came into being. It is the automatic extraction of structured information such as entities, relationship between entities, and attributes from unstructured sources [2].

The Message Understanding Conferences (MUC) initiated and financed by the Defense Advanced Research Projects Agency (DARPA) of the 1900s, encouraged the development of new and better methods of information extraction. From this they deduced that in order to reasonably extract information from documents, it is useful to first identify certain classes of information referred to in the text. From this endeavor, the Named Entity Recognition (NER) task was born, where systems attempted to identify dates, times, numerical information and names [3], [4], [5]. Basically, NER is the core of NLP systems. It involves two tasks – firstly, the identification of proper names in the text, and secondly, the classification of these names into a set of predefined categories such as names of person, organization, location, date and time expressions [6].

NER systems have been developed for a variety of languages – English, Indian, and Chinese among others. The diversity and complexity of each language makes each NER system unique. Also, different algorithms and techniques have been used in the development of such systems. In the Philippines however, only a few systems exist for the Filipino language, and that the need for computational linguistic tools for the Filipino language proves the importance of an NER. Even the structure and composition of the language makes it an interesting and challenging task for us researchers, thus we aimed to develop an NER system for the advancement of the NLP and Data Mining field in the Philippines.

II. Transformation-Based Learning

A. Definition

In 1992, Eric Brill introduced the formalism of the transformation-based learning algorithm as an approach in corpus-based natural language processing. Transformation-based learning (TBL) can be defined as a rule-based machine learning algorithm for learning a set of transformation rules, and is used in NLP tasks such as part-of-speech tagging, text chunking, and named entity recognition [7].

B. The Algorithm

The central idea behind TBL is to start with a simple solution to the problem and apply *transformations* – at each step the transformation which results in the largest benefit is selected and applied to the problem. The algorithm stops when the selected transformation does not modify the data in enough places, or there are no

Manuscript received December 30, 2012; revised March 15, 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.1.1.1-5

more transformations to be selected [8]. To define a specific application of the transformation-based learning, one must specify the following:

1. The initial-state annotator
2. The space of allowable transformations (rule templates and triggering environments).
3. The objective function for comparing the corpus with the truth and choosing a transformation.

This is how TBL works. First, unannotated text is passed through an initial-state annotator. The initial-state annotator can range in complexity from assigning a random structure to assigning the output of a sophisticated manually created annotator. Once text has been passed through the initial-state annotator, it is then compared to the *truth* [8]. A manually annotated corpus is used as reference for the truth. An ordered list of transformations is learned that can be applied to the output of the initial-state annotator to make it better resemble the truth [8]. The objective function is usually to minimize the errors in the tagging. A greedy search is applied for deriving a list of transformations: for each iteration of learning, the transformation whose application results in the best score according to the objective function is then added to the ordered transformation list and the training corpus is updated by applying the learned transformation. Learning continues until no transformation can be found whose application results in an improvement to the annotated corpus [8].

For the initial-state annotator, we utilized a dictionary-based lookup method, wherein the named entity is searched through the dictionary, and if a match is found, the entity is annotated with the most frequently-used tag for that specific word in the dictionary. The objective function is to minimize the errors in the tagging. The researchers collected a total of 50 manually annotated corpora to be utilized in the training process.

III. Development and Experimentation

A. System Development

The researchers developed a Named Entity Recognition System for the Filipino Language (TBLNER) using the Transformation-based Learning algorithm as its core algorithm. The system is developed to extract named entities from an input text and classify these entities in five major classifications namely PERSON, ORGANIZATION, LOCATION, DATE and EVENT. Any entity that is not within these classifications is tagged as MISC (miscellaneous).

B. System Training

Most NERs were developed and trained on specific textual genre, or domains. Asian History was chosen in this study since it has rich content of named entities. It was also observed that this domain contains many historical events that lead to the introduction of a sub-entity called the EVENT entity.

The system was trained using a set of 40 text files of Asian History references in the Filipino language. Using the transformation-based learning algorithm, an ordered list of 157 transformation rules was learned by the system. Table I presents some of the first rules learned by the system during the training phase.

TABLE I. Transformation Rules Learned by the TBLNER

prevToken:MISC>LOCATION?sa
prevToken:MISC>PERSON?si
prevToken:PERSON>LOCATION?sa
prevToken:MISC>ORGANIZATION?ang
NEBeforeComma:MISC>LOCATION?,&LOCATION
prevToken:MISC>PERSON?ni

These transformation rules will then be used in the process of classifying and tagging the extracted entities from the input text into the five classifications.

C. System Architecture

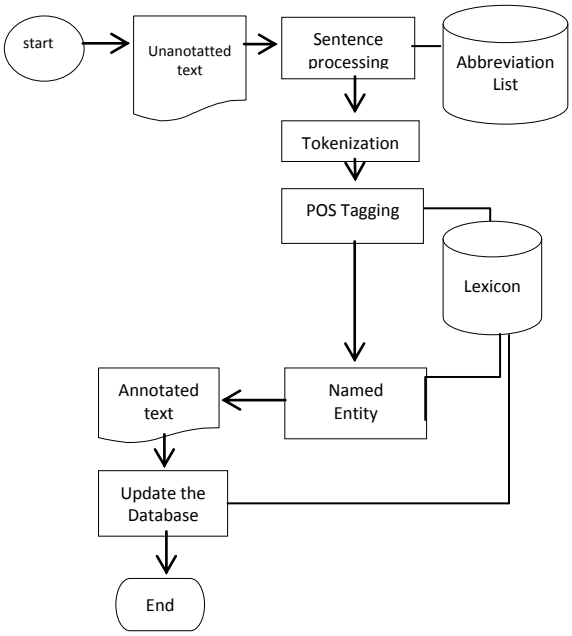

Figure 1. System Architecture of the TBLNER

Fig. 1 illustrates the system architecture of the developed TBLNER. The system accepts input in the form of text files (.txt). Filipino Asian History references are the most preferred input since it is where the system was trained. These input file will now undergo the preliminary process prior to the named entity classification.

1. Sentence Processing
   This process breaks down the input file into sentences.
2. Tokenization
   After breaking the input into sentences, these sentences will then be iterated and tokenized using the tokenizer. The tokenizer breaks each sentence into meaningful tokens using regular expressions and grammar rules.
3. Part-of-speech Tagging (POS Tagging)
   The meaningful tokens will now undergo POS tagging, wherein the tokens will be assigned with their corresponding part-of-speech (POS) tags

which will be used in the Named Entity tagging phase. POS tags include verb, pronoun, noun (common and proper), adjective etc.

After tagging each token with its corresponding POS tag, these tokens will undergo the process of Named Entity Tagging. The tagger utilized the transformation rules learned using the Transformation-based learning algorithm. Only tokens that are tagged as NNP (proper noun) will undergo this process. The token will undergo initial tagging, wherein the token will be searched in the lexicon of named entities. If the token is found in the lexicon, the NE tag with the highest frequency count for the token will be applied, otherwise, the ordered list of transformation rules will be applied to tag the token.

After all the sentences have been processed and the entities have been tagged, the lexicon will be updated with the entities and their corresponding tags. The output will be displayed by the system in color coded format – (red: PERSON; green: ORGANIZATION; blue: LOCATION; orange: DATE; yellow: EVENT; pink: MISC). The system will also output a text file containing the annotated text including the rules applied to the input.

D. Data Gathering

The following data were gathered during testing for use in the evaluation:

1. Total number of entities in the corpus (NE)
2. Total number of entities retrieved by the system (RNE)
3. Number of entities not retrieved by the system (UNE)
4. Number of entities correctly tagged by the system (CT)
5. Number of entities incorrectly tagged by the system (WT)

In order to evaluate the performance of the TBLNER, the Precision, Recall and F-measure of the NER were determined. The Precision refers to the measure of the number of correctly tagged entities out of all tagged entities by the system, while the Recall is the measure of the number of retrieved named entities out of all the entities in the text corpus. The F-measure is the measure of the overall accuracy of the system, which is the harmonic mean of the Precision and Recall. The formulae for these measurements according to [3] are as follows:

$$Precision = \frac{N_{correct}}{N_{correct} + N_{incorrect}} \quad (1)$$

$$Recall = \frac{N_{found\ \ entities}}{N_{total\ \ entities}} \quad (2)$$

$$Fmeasure = \frac{2(Precision)(Recall)}{(Precision + Recall)} \quad (3)$$

IV. Evaluation and Results

A. Evaluation of TBLNER

The developed Named Entity Recognition system (TBLNER) was tested using a set of 100 text files of Asian History references written in the Filipino language. The necessary data needed for evaluation were gathered. Using the formulas presented in the preceding section, the Precision, Recall and F-measure of the system were computed and the results were tabulated. Using the tabulated results, the researchers were able to point out in which classifications were the TBLNER gained a high performance and which areas did it fail to achieve a high performance, and assess the overall performance of the system. These results will be further interpreted in the succeeding sections.

TABLE II. Summary of 100 Asian History References Tested Using TBLNER

Entity Type	NE	RNE	UNE	CT	WT	Precision	Recall	F-measure
P	358	352	6	334	18	94.89%	98.33%	96.58%
O	560	830	–270	440	380	53.66%	148.22 %	78.8%
L	1200	1090	110	924	161	85.17%	90.84%	87.92%
D	258	241	17	235	5	97.92%	93.42%	95.62%
E	61	43	18	42	4	91.31%	70.5%	79.57%
M	302	161	141	92	69	57.15%	53.32%	55.17%
TOTAL	2739	2717	22	2067	637	76.45%	99.2%	86.36%

It can be gleaned from Table II that from an overall total of 2739 entities, 2717 entities were retrieved by the system, while 22 entities were not retrieved, resulting in a high percentage of Recall (99.2%). Out of the 2717 retrieved entities, 2607 were tagged correctly by the system while only 637 were incorrectly tagged, resulting in a Precision rate of 76.5%, which is a fair measure considering the formula used. Finally using (3), the harmonic mean or F-measure of the obtained Precision and Recall is 86.36%, which is also a high measure if we will look into the performances of the existing systems for the Filipino language [9], [10].

It can also be seen from Table II that the system has a good performance in terms of PERSON, LOCATION, DATES categories, and a fair performance in EVENT category. There is a case of over retrieving of entities in the ORGANIZATION (labeled O) category. As you can see, from a supposed number of 560 entities, the system retrieved a total of 830 entities. Due to this, there was a negative result for UNE. Because of this phenomenon, the system has incorrectly tagged almost half of the retrieved entities resulting in a very low precision for that category (53.66%). Likewise, the system has incurred a low performance in the MISC category.

B. Comparison with the NERF

The researchers also compared the results of the developed system with that of an existing NER for the Filipino language, which is known as NERF [9]. The NERF utilizes a rule-based approach in the tagging of entities. Both systems dealt with structured data and the same language so the comparison is deemed fair. This comparison was done to see if the developed TBLNER can perform at par or even better than the other existing systems, identify its flaws and areas for improvement, and if the application of the transformation-based

learning algorithm is effective in tagging the extracted entities into their correct classifications.

TABLE III. COMPARISON OF PERFORMANCE OF TBLNER AND NERF

	NE	RNE	UNE	CT	WT	Precision	Recall	F-measure
TBLNER	787	775	12	623	150	80.6	98.48	88.65
NERF	642	646	-4	551	95	85.29	85.29	85.29

For this comparison, a total of 30 text files were used for both systems. Table III presents the results for each NER system. For the TBLNER, the overall Precision, Recall and F-measure for 30 files are 80.6%, 98.48%, and 88.65% respectively; while for the NERF, the overall Precision, Recall, and F-measure also for 30 files are 85.29%, 85.29%, and 85.29% respectively.

It can be observed that the TBLNER's performance is in a similar range as with the results of testing with 100 files (see Table II). Also, we can see that the NERF is a bit higher when it comes to Precision, but the TBLNER is higher in terms of Recall. The F-measures of both systems are at a close range, suggesting a rather similar overall performance.

From this comparison, the developed TBLNER has an acceptable performance since it performed at par with the other existing NER systems for the Filipino language today [9], [10].

V. CONCLUSION

Judging from the tests, evaluation and comparisons conducted on the developed TBLNER for the Filipino language, there is sufficient evidence to conclude that the developed TBLNER gained an acceptable performance and the application of the Transformation-based learning algorithm in the development of the system has proved to be an effective approach. The developed TBLNER gained a high performance rate in the categories PERSON, LOCATION, and DATES, and a fair but acceptable performance for EVENT, since it is the first time an entity type EVENT is incorporated in named entity tasks in the Philippines.

Despite its positive points, the TBLNER also incurred drawbacks such as a low performance in the ORGANIZATION and MISC categories. These can be attributed to the inconsistencies in the structure of the language and the contextual meaning of some entities like ORGANIZATION and LOCATION, wherein an entity classified as an ORGANIZATION may sometimes be classified as a LOCATION depending on the context in the sentence. Also, over retrieving and under retrieving of entities can be attributed to the preprocessing tasks such as sentence processing, tokenization, and part-of-speech tagging. Different factors such as sentence construction, subject-verb agreement and other linguistic complexities may also affect the process.

Overall, we must say that our developed system which is the TBLNER has gained a good performance with the application of the transformation-based learning algorithm. Also, the use of Asian History References has proven to be an effective textual genre in the training of the system and because of this domain, we were able to contribute a new entity type to the Information Retrieval task, which is the entity type EVENT.

Further studies in this field can be conducted to further enhance the Information retrieval and Data mining task in the Philippines. We suggest a further study of Transformation-based Learning and how it can be used in other areas of Natural language processing. In terms of the Named entity recognition task, we suggest the search for more relevant entity types, and also the sub categorization of major entity types e.g. PERSON can be expanded into MALE, FEMALE, PRESIDENT, etc., ORGANIZATION can be subdivided into GOVERNMENT, HOSPITALS, SCHOOLS etc., and LOCATION can also be further subdivided into COUNTRY, STATE/PROVINCE, CITY, STREET, etc.

Information Extraction/Retrieval is a wide task in the field of Data mining and Natural Language Processing, and the Named Entity Recognition task, is just a sub-process but an essential task for the improvement and advancement of the areas concerned, thus continuous studies regarding these fields must be encouraged especially to developing countries in the field of Information Technology such as the Philippines.

ACKNOWLEDGEMENT

We would like to extend our deepest 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, most especially our thesis advisers, Prof. Ria A. Sagum, MCS and Prof. Angelito G. Pastrana, MSIT.

REFERENCES

[1] Natural Language Processing Group of the Department of Computer Science. Natural Language Processing. (November 2012). University of Sheffield. [Online]. Available: http://nlp.shef.ac.uk/

[2] S. Sarawagi, "Information Extraction," *Found. Trends Databases*, vol. 1, no. 3, pp. 261-377, March 3, 2008.

[3] B. Sundheim and R. Grishman, "Message Understanding Conference-6: A brief history," in *Proc. 16th Conference on Computational Linguistic*, Copenhagen, 1996.

[4] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," in *Proc. Seventh Conference on Natural Language Learning at HTL-NAACL*, Edmonton, Canada, 2003.

[5] R. Zanoli, E. Pianta, and C. Giuliano, "Named entity recognition through redundancy driven classifiers," Sommarive 18, I-38123 Povo, Italy, 2009.

[6] L. S. Affendy. A. Mamat, and A. Mansouri, "Named Entity Recognition Approaches," *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339-344, 2008.

[7] G. Ngai and R. Florian. (November 2012). Fast Transformation-Based Learning Toolkit. [Online]. Available: www.cs.jhu.edu/~rflorian/fntbl/tbl-toolkit/tbl-toolkit.html

[8] E. Brill, "Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging," *Computational Linguistics*, vol 21, no.4, pp. 543-565, December 1995.

[9] R. A. Sagum, *A Named Entity Recognizer for Filipino Text*, De La Salle Univesity, Manila, 2011.
[10] L. E. Lim, J. C. New, M. A. Ngo, M. C. Sy, and N. R. Lim, *A Named Entity Recognizer for Filipino Texts*, De La Salle University, Manila, 2007.

Prof. Ria A. Sagum was born in Laguna, Philippines on August 31, 1969. She took up a Bachelor's degree in Computer Data Processing Management from the Polytechnic University, and also a degree in Professional Education in Eulogio Amang Rodriguez Institute of Science and Technology. She received her master's 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 and Information Sciences, 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, Manila.

Ms. Sagum has been a presenter at various conferences across the globe, including the 2012 International Conference on e-Commerce, e-Administration, e-Society, e-Education and e-Technology, and the 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.

Prof. Angelito G. Pastrana was born Sampaloc, Manila on September 7, 1979. He obtained a Bachelor in Information Technology degree from the College of Computer and Information Sciences of the Polytechnic University of the Philippines, Manila. He obtained his Master's degree in Information Technology (MSIT) from the PUP Graduate School and is currently a candidate for a Doctor of Philosophy (Ph.D.) in Technology Education degree of the Rizal Technological University, Manila.

As an acamedician he specializes in Operating Systems, Management Information Systems, Research and Statistics, Technopreneurship and Software Engineering. At present he is currently a part-time IT consultant and a full-time professor of the Department of Information Technology of the College of Computer and Information Sciences, Polytechnic University of the Philippines. He was also a former chairperson of the department. He is also the adviser or Institute of Brilliant IT Students (IBITS) student organization, and one of the directors of PSUCCESS.

Quervin Lloyd L. Buco was born in Bustos, Bulacan, Philippines on August 9, 1992. He graduated from the Polytechnic University of the Philippines with a Bachelor's degree in Computer Science (2013). His skills include programming in different languages such as C, C# and Java, web development using PHP and Javascript, and database scripting using MySQL. His fields of interest include mobile applications development, database management systems, natural language processing, web development and artificial intelligent systems.

John Lester L. Capcap was born in Mogpog, Marinduque, Philippines on October 8, 1992. He graduated from the Polytechnic University of the Philippines with a Bachelor's degree in Computer Science (2013). His skills include programming in different languages such as C, C# and Java. His expertise include web development using PHP, Javascript, Joomla, and Wordpress frameworks. His fields of interest include mobile applications development, web development, software engineering and multimedia and graphics design

Jayvee Carlo A. Hermocilla was born in Laguna, Philippines on April 5, 1991. He graduated from the Polytechnic University of the Philippines with a Bachelor's degree in Computer Science (2013). His skills include database scripting in MySQL, web development using PHP and Javascript, proficiency in office tools such as electronic spreadsheets, and electronic databases.. His fields of interest include web development, graphics design, database management systems, and robotics.

Carmina S. Yumul was born in Antipolo, Rizal, Philippines on July 7, 1993. She graduated from the Polytechnic University of the Philippines with a Bachelor's degree in Computer Science (2013). Her skills include web designing using Adobe Dreamweaver, CSS and web development using PHP and Javascript, database management systems, business analytics, systems analysis and design. Her fields of interest include Project Management and Database Management Systems

.

