Named-Entity Recognizer for Common Nouns in Filipino Text

Ria Ambrocio Sagum
College of Computer and Information Sciences, Research Management Office, Polytechnic University of the Philippines
rasagum@pup.edu.ph

Andrea Nicole Pinili
College of Computer and Information Sciences, Polytechnic University of the Philippines
andrea.pinili04@gmail.com

Kent Laurence Gentilezo
College of Computer and Information Sciences, Polytechnic University of the Philippines
klaurencegentilezo@hotmail.com

Kier Bryan Lopez
College of Computer and Information Sciences, Polytechnic University of the Philippines
kierlopez01@gmail.com

ABSTRACT

The Named Entity Recognizer for Filipino Text is a study that aims on detecting and classifying named entities present in a given text utilizing the Filipino language. The named entities are classified into four, namely: person, place, organization, date & time. Named entities that do not fall in the four classes are tagged as etc. To measure the accuracy of the system, solving for the precision, recall, error rate, and F-Measure, and accuracy in terms of error rate was used, both for every named entity and all the named entities as a whole. The approach of the Conditional Random Field was used to classify the named entities. Filipino short stories were used for the testing of the system. The results, based on solving for the F-Measure, indicate that the system is 88.05% accurate and best in identifying named entity Organization with 0% error rate, but with an acceptable or good enough error rate for the named entity Person, Place, and Date & Time, with an average of 3.24% error rate respectively. The results, based on solving for the accuracy in terms of error rate, indicate that the system is 97.01% accurate.

CCS CONCEPTS

• **Applied computing**;

KEYWORDS

Common Nouns, Conditional Random Fields, Named Entity Recognition, Information Extraction, Natural Language Processing

ACM Reference Format:
Ria Ambrocio Sagum, Andrea Nicole Pinili, Kent Laurence Gentilezo, and Kier Bryan Lopez. 2020. Named-Entity Recognizer for Common Nouns in Filipino Text. In *2020 the 6th International Conference on Communication and Information Processing (ICCIP 2020), November 27–29, 2020, Tokyo, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3442555.3442572

ICCIP 2020, November 27–29, 2020, Tokyo, Japan

ACM ISBN 978-1-4503-8809-2/20/11…$15.00
https://doi.org/10.1145/3442555.3442572

1 INTRODUCTION

In every natural language, there is a class of general names of various kinds that may or may not be natural. They can occur as predicates to describe some individuals. They can also be subjects of sentences and are described by other predicates. They are used either extensionally or intentionally in a sentence. [1]

Named Entity Recognition (NER) is a prior task in Natural Language Processing (NLP). Named Entity Recognition is a sub task of information extraction and it identifies and classifies proper nouns to its predefined categories such as person, location, organization, time, date etc. An entity – is something that exists by itself: something that is separate from other things. [2] Thus, NER is of key importance in many Natural Language Processing (NLP) tasks, such as Information Retrieval (IR) or Machine Translation (MT). [3]

Since NER's main purpose is to identify and classify proper nouns in a given context, common nouns remain unseen in this particular task in NLP. Common nouns are your run-of-the-mill, generic nouns. They name people, places, things or ideas that are not specific (e.g., woman, city, dog, shoe, etc.). Since these nouns aren't naming anything specific, they don't need to start with a capital letter unless they begin a sentence. Common nouns can perform many jobs in sentences. It can be considered as a subject, direct object, indirect object, object of the preposition, and predicate nominative. [4]

According to [Gillon, 1999] [5] on The Lexical Semantics of English Count and Mass Nouns, such lexicalizations require lexical entries of their own. Thus, it is clear that, to accommodate the nonce usage of proper names as common nouns, requires that they be given special lexical entries.

Common nouns are also found to be useful in some applications in NLP. Extending the scope of semantic annotation to common nouns resulted to a much more detailed and varied semantic characterization. [6] Common nouns are also helpful for data annotation and retrieval of digital photos. [7] They also discovered that common nouns listed in a dictionary is useful, for instance, in the disambiguation of capitalized words in ambiguous positions. [8]

A study entitled Techy Basyang: An Emotional Automated Filipino Narrative Storyteller by [Gonzaga et al., 2013] [9] is a Tagalog automated storyteller that narrates with emotions which is done by Text-to-Speech (TTS) system. The study has three (3) major stages which are sentence analyzer (which includes NER and Text

Categorization), stemmer (Tagalog Stemming Algorithm), and Text-to-Speech conversion. The experts neither agreed nor disagreed to the system's process of converting the story to speech sound and the students were agreed to the same process of the system. In general, the experts and the students were very satisfied to the system. The researchers of the study recommended the detection of common nouns as a part of character recognition to improve the detection of characters inside the story.

As the recent study suggests, detection of common nouns is also important when it comes to character recognition given a Filipino story. Aside from detecting specific names, common nouns also play a big role in short stories because although they're not specific, they could still represent a person, place/location, organization, date and time. Since NERs are mostly used for proper nouns, the researchers decided to design a NER which is intended for common nouns in Filipino Text.

2 RELATED WORKS

Common nouns are everywhere, and you use them all the time, even if you don't realize it. People in general are named using common nouns, though their official titles or given names are proper nouns. [10] Common nouns, being general or ordinary names, are more frequently used in the subject than the other type of nouns, namely proper nouns, which are names of specific or unique things. [11]

The Definition of (Common) Nouns and Proper Nouns implies that "common nouns are the fundamental item, whereas proper nouns are nouns with two particularities (proper nouns do not represent a species of entities, but an entity, known to the listener)." Common nouns are general names so we cannot capitalize them unless they start a sentence or used as part of a title. Common nouns can be used anywhere in the sentence according to the need and requirement. It can be used in the manner as not to show the grammatical error. It can occur in between the sentence anywhere or in the start of the sentence. It is written in small letter if occur anywhere in the sentence, however written in capital letter if occur in the start of sentence. [12, 13]

In Filipino literature, there are also books that discusses the subject of common nouns in Filipino language and grammar. According to the book entitled "Writing Filipino Grammar: Tradition & Trends" by [Cubar & Cubar, 1994] [14], Lopez states that morphology, the syntax, the grammar, the nature, and the psychology among other things of the Philipppine languages are something peculiar to themselves. In Lopez's manual, he classified common nouns as a substance word. In the book entitled "Gramatikang Filipino: Balangkasan" by [Ceña & Nolasco, 2011] [15], stated that common nouns have a part in the types of predicate in a sentence. The book entitled "Learning the Basics of Filipino" by [De Castro, 2016] [16], a common noun ("Pangngalang Pambalana" in Filipino) has a natural gender which are: Panlalaki (Masculine), Pambabae (Feminine), Di-tiyak (Neuter), and Walang Kasarian (No gender).

According to [Bates et al., 1994; Gentner, 1982] [17], the noun is an important lexical category for children who are learning language. Nouns are important linguistic blocks of learning, and the development of other parts of speech may greatly depend on the young language learner's acquisition and production of this lexical category in the initial phase of language acquisition. Many experts have noted the predominance of general nominals or common nouns in children's early vocabularies.

Recent researches have shown that proper names may differ morphosyntactically from common nouns. [18] However, little is known of the morphosyntactic contrasts between proper names and common nouns in less studied European and Non-European languages, or even from a cross-linguistic perspective. [19] Although there are some syntactic peculiarities in languages like English and German, where common nouns cannot be used as predicate of objects; rather, they may have a kind-denoting interpretation too. [20]

Previous empirical studies employed a variety of tasks to study the processing differences between person names and common nouns. For instance, when asked to decide whether a word was a name or a noun, people were faster at identifying names compared to nouns. [21, 22] Both person names (e.g., "Thomas Edison") and common nouns (e.g., "inventor") can be used to refer to individuals. However, they differ greatly in their level of specificity as person names generally refer to specific individuals (e.g., "Thomas Edison" refers to the inventor who invented the light bulb), while common nouns represent a group of individuals with similar characteristics (e.g., an "inventor" can be anyone who creates novel things). [23]

In contrast, common nouns contain intrinsic meaning and imply specific attributes. Similarly, when asked to judge the relatedness of two names or nouns, people were faster at recognizing the association of names (e.g., "Woody" and "Allen") than that of nouns (e.g., "social" and "security") even when the frequency between the names' and nouns' associations was matched. [24] Similarly, people made slower judgments regarding the emotional valence of names (e.g., to judge "Hitler" as negative) than that of nouns (e.g., to judge "gun" as negative). [25]

Given the numerous distinctions between person names and common nouns, it is essential to examine how person names are processed in context. According to Wang et al. [26] in "The processing difference between person names and common nouns in sentence context: an ERP study", person names and common nouns differ in how they are stored in the mental lexicon. Both person names and common nouns were highly related in meaning and either congruent or incongruent within the previous contexts.

For common nouns, multiple semantic links converge to form its meaning. In contrast, for person names, identify-specific information converges on the person identity node, which connects to a specific person's name via a single connection. [27] The notion of the semantic associations of common nouns are richer than that of person names (due to different levels of specificity). [28, 29]

According to Frege [30], as for common nouns, he believed that "The same distinction can also be drawn for concept-words". However, Fregean treatment for common nouns is not a direct extension of that for proper names. The situation is a little complex.

In a letter to Husserl used the following schema to express his view:

proper name - > sense of the proper name - ► Bedeutung (meaning) of the proper name

concept word sense - > of the concept word - ► [Bedeutung (meaning) of the concept word (concept)] - > object

For common nouns, Fregean theory treats sense, reference, and concept. The reference of a common noun is a concept.

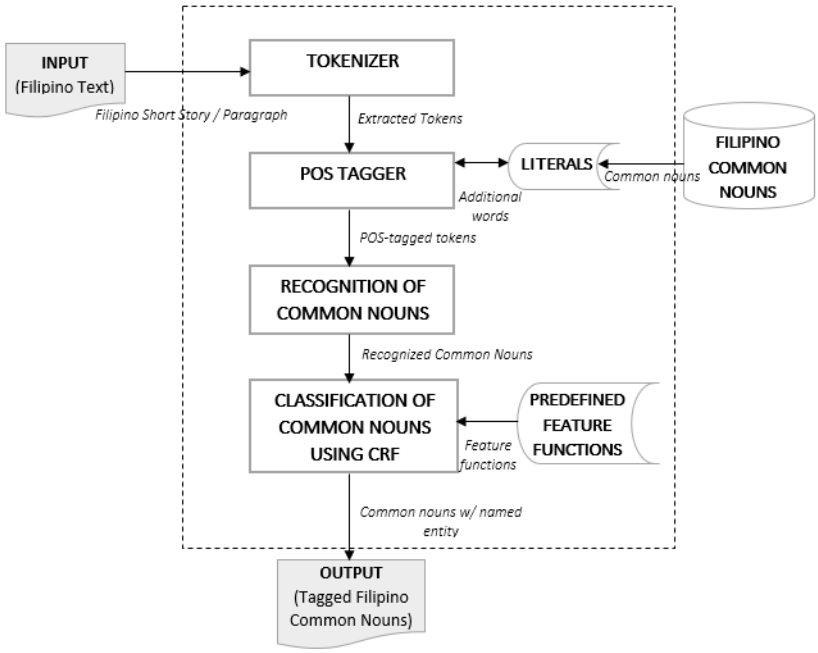

Figure 1: System Architecture

An article entitled “On Gupta’s Book: The Logic of Common Nouns" by Bressan [31] stated that, Gupta has suggested that (all) common nouns express both a principle of application and a principle of (transworld) identity. Common nouns are treated at length, and profoundly; furthermore, Gupta employs his logical theory for very interesting applications to philosophical and linguistic questions, e.g., concerning essentialism and the theory of truth.

3 METHODOLOGY

This system architecture, Fig.1, is referenced to the system architecture of a study entitled Named Entity Recognizer for Filipino Text [32]. The input will the text that is to be processed by the system. The system assumes that the input is in Filipino Language.

The input will go through the first stage which is the Tokenizer. The tokenizer consists of different processes to produce the tokens essential to other modules of the system. It will break down the text to single characters for various examinations. The use of punctuation marks in the input is also checked in this module. The tokenizer will then produce tokens.

These tokens will then enter the POS Tagger. In this module, the system will identify the tokens’ part of speech with the help of the literals and a corpus of common nouns in Filipino. Once identified, it will label the token with its corresponding part-of-speech.

The Recognition module simply looks for tokens which are labeled as “common noun” and passes it to the next module. An input of a Filipino text consists of several tokens, where every token has its corresponding part of speech. The dictionary and the literals play a huge role in helping to tag the necessary part of speech of every token, as can be seen in the previous module. During the POS Tagger, some tokens has been already tagged as “common nouns”. In accordance, in the Recognition module, tokens are being extracted when it is identified as a common noun.

The CRF module handles the classification of common nouns (person, location, organization, date & time) based on the classifier’s logic. CRF starts by identifying the compatibility of the token to a named entity with the help of predefined feature functions. Once a feature function is satisfied to a certain named entity, the weight for that named entity increases. Once identified, it will finally indicate the named entity of the token itself. The tagged version of the input will be the output of the system.

4 DATA GATHERING

To be able to create a named entity recognizer that exhibits both accuracy and consistency, the researchers made use of information present in previous literatures and studies concerning the establishment of such systems. These literatures and studies came from journals of natural language processing conferences, and from portable document files existing in the internet. The proponents formulated ideas on how to create the system with the help of the said information.

The researchers also went to different universities and libraries (Polytechnic University of the Philippines, University of Santo Tomas, National Library, & University of the Philippines) to gather related literature and studies for this study. The training data used for the system was gathered from Filipino short stories. The gathered information from the researchers builds the foundation in creating the system.

Table 1: Summary of Results

Named Entity	T	NT	CT	WT	Error Rate(WT/T) *100	Accuracy (1-E)*100
Person	99	110	98	1	1.01%	99.99%
Place	35	41	34	1	2.86%	97.14%
Org	3	3	3	0	0%	100%
Date&Time	11	18	10	1	9.09%	90.91%
Overall	148	172	145	3	3.24%	97.01%

5 RESULTS

This study was intended to create an NER for Common Nouns in Filipino Text that exhibits accuracy. To be able to assess this characteristic, the proponents used the F-measure.

The F-measure (also F1 score) is defined as a harmonic mean of precision (P) and recall (R): F=2PR/P+R. The F-measure is a measure of a test's specificity, sensitivity, and accuracy. It considers both the precision p and the recall r of the test to compute the score: P is the result of number of correct nouns tagged over total number of tagged common nouns, R is the number of correct nouns tagged over total number of common nouns present. 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. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. (F-1 Score, 2012)

To identify the system's performance in terms of its error rate, the following formula was used: E is the result of the number of wrong tags over the total number of tagged common nouns. Moreover, to identify the system's accuracy in terms of its error rate, the following formula was used: Accuracy = (1-E) x 100.

In identifying the performance of the developed NER for Common Nouns in Filipino Text in terms of Precision, Recall, Error Rate, F-Measure, and Accuracy in terms of Error Rate.

The summary assessment of the performance of the system based on 30 files tested in terms of Precision, Recall, Error Rate, and F-Measure was computed as 96.76%, 81.89%, 3.24%, and 88.05% respectively. (see Table 1)

The summary assessment of the performance of the system based on 30 files tested in terms of Accuracy based on Error Rate was computed as 99.99%, 97.14%, 100%, and 90.91% respectively. (see Table 2)

6 CONCLUSION

Based from the findings the researchers reached the following conclusions: The overall performance of the developed Named Entity Recognizer for Common Nouns in Filipino Text garnered an F-measure of 88.05%; The overall accuracy of the developed Named Entity Recognizer for Common Nouns in Filipino Text garnered an Accuracy (in terms of error rate) of 97.01%; The Named Entity Recognizer for Common Nouns in Filipino Text is effective in tagging named entity Person with 93.77%; Place with 89.47%, and Organization with 100% accuracy in terms of F-measure; The Organization got a 100% in evaluation considering that it only had 3 entities to be tagged out of the 30 input files tested; The developed system is not yet effective in tagging named entity date & time with an accuracy only 68.96% based on F-Measure; The performance of the system will increase further if more words were added to the corpus and more feature functions were fed into the system.

From these results it was concluded that the system can detect common nouns accurately given that the system is supported by a Filipino dictionary consists of common nouns. Because the system is dependent to the Filipino dictionary, we can say that it still lacks in detecting common nouns.

7 RECOMMENDATION

The following suggestions might be helpful for the future development of the system:

- Creation of tool that will automatically detect common nouns even if it does not exist in the dictionary to increase the accuracy of detecting common nouns in Filipino text.
- Do more studies about Filipino grammar or sentence structure in creating feature functions which will help in the classification of common nouns in their respective named entity.
- Tag other named entities other than name of person, place, org, date and etc. such as things and animals.

ACKNOWLEDGMENTS

The researchers would like to express their deepest gratitude to their families who are their inspiration and encouragement in all of their accomplishments.

To Steven John Bonzol and Joyce Ann Cuenca for extending their support in the process of creating this paper even beyond their busy schedules. All the positivity for us as we continue our walk in this difficult yet fulfilling journey.

Most importantly, to our God Almighty, for the blessing of knowledge and wisdom to be able to do the things we needed to do. For the

Table 2: Accuracy in terms of Error Rate

Named Entity	T	NT	CT	WT	Error Rate	Accuracy
Person	99	110	98	1	1.01%	99.99%
Place	35	41	34	1	2.86%	97.14%
Org	3	3	3	0	0%	100%
Date&Time	11	18	10	1	9.09%	90.91%
Overall	148	172	145	3	3.24%	97.01%

blessing of understanding hearts to live up to each other's diverse personalities and simplify for the gift of life.

REFERENCES

[1] Zhou, B., & Mao, Y. (2010). Four semantic layers of common nouns. Synthese, 47-68.

[2] Entity. (n.d.) In Merriam-Webster's collegiate dictionary. Retrived from http://www.merriam-webster.com/dictionary/entity

[3] E. Simon (2013). The Definition of Named Entities [PDF file]. Retrieved from http://clara.nytud.hu/~kk120/simon/simon.pdf

[4] E. O'Brien (2018). Proper Nouns & Common Nouns. Retrieved from https://www.english-grammar-revolution.com/proper-nouns.html

[5] B. Gillon (1999). The Lexical Semantics of English Count and Mass Nouns. Linguistics and Philosophy.

[6] D. Maynard, W. Peters, Y. Li (2016). Metrics for Evaluation of Ontology-based Information Extraction [PDF file]. Retrieved from https://gate.ac.uk/sale/eon06/eon.pdf</bib>

[7] P. Sinha & R. Jain (2008). Classification and Annotation of Digital Photos using Optical Context Data [PDF file]. Retrieved from https://www.researchgate.net/profile/Ramesh_Jain2/publication/221368981_Classification_and_annotation_of_digital_photos_using_optical_context_data/links/53e18f890cf24f90ff657ac6.pdf

[8] D. Nadeau & S. Sekine (2007). A survey of named entity recognition and classification. Linguisticae Investigationes, 30(1), 3-26.

[9] J.C Gonzaga, R. Sagum, J. Seguerra, J. Turingan, & M.P Ulit (2013). Techy Basyang: An Emotional Automated Filipino Narrative Storyteller [PDF file]. Retrieved from https://www.dropbox.com

[10] Ginger Software (2018). Common Noun. Retrieved from https://www.gingersoftware.com/content/grammar-rules/nouns/common-noun/ [11]Mu ¨ller, H. M. (2010). Neurolinguistic Findings on the Language Lexicon: the Special Role of Proper Names. Chinese Journal of Physiology, 53(6), 351–358.

[11] Madhavan, D. (2015). What are Common Nouns. Retrieved fromhttp://www.english-language-grammar-guide.com/common-nouns.html

[12] Saragossa, A. (2016). The Definition of (Common) Nouns and Proper Nouns. Onomastica.

[13] A. Singh (2018). Common Noun. Retrieved fromhttps://www.teachingbanyan.com/grammar/common-noun/

[14] N. Cubar & E. Cubar (1994). Writing Filipino Grammar: Tradition & Trends.

[15] R. Ceña & R.M. Nolasco (2011). Gramatikang Filipino: Balangkasan.

[16] Castro (2016). Learning the Basics of Filipino.

[17] Lucas, R., & Bernardo, A. (2008). Exploring Noun Bias in Filipino - English Billingual Children. The Journal of Genetic Psychology: Research and Theory on Human Development, 149-164.

[18] Schlucker, B., & Ackermann, T. (2017). The morphosyntax of proper names: An overview. Folia Linguistica, 309-339.

[19] J. Glikman (2019). Proper Names Versus Common Nouns: Morphosyntactic contrasts in the languages of the world. Retrieved from https://diachronie.org/2018/06/14/proper-names-versus-common-nouns morphosyntactic-contrasts-in-the-languages-of-the-world/

[20] J. Dolling (1993). Commonsense ontology and semantics of natural language [PDF file]. Retrieved from http://home.uni-leipzig.de/doelling/pdf/common.pdf

[21] Mu ¨ller, H. M. (2010). Neurolinguistic Findings on the Language Lexicon: the Special Role of Proper Names. Chinese Journal of Physiology, 53(6), 351–358.

[22] Yen, H.-L. (2006). Processing of proper names in Mandarin Chinese: a behavioral and neuroimaging study. Bielefeld University, Bielefeld (Germany). Retrieved from http://pub.uni-bielefeld.de/ publication/2301435.

[23] Kripke, S. (1981). Naming and necessity. Oxford: Blackwell publishing.

[24] Proverbio, A., Mariani, S., Zani, A., & Adorni, R. (2009). How Are 'Barack Obama' and 'President Elect' Differentially Stored in the Brain? An ERP Investigation on the Processing of Proper and Common Noun Pairs. Plos One.

[25] Wang, L., Zhu, Z., Bastiaansen, M., Hagoort, P., & Yang, Y. (2013). Recognizing the emotional valence of names: an ERP study. Brain and Language, 118-127.

[26] Wang, L., Verdonschot, R., Yang, Y. (2016). The processing difference between person names and common nouns in sentence contexts: an ERP study. Psychological Research, 80(1). Springer-Verlag Berlin Heidelberg 2015.

[27] Grabowski, T. J., Damasio, H., Tranel, D., Ponto, L. L. B., Hichwa, R. D., & Damasio, A.R. (2001). A role for left temporal pole in the retrieval of words for unique entities. Human Brain Mapping, 13(4), 199–212.

[28] Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. Language and Linguistics Compass, 1(1–2), 71–83.

[29] Sanford, A. J., & Graesser, A. C. (2006). Shallow Processing and Underspecification. Discourse Processes, 42(2), 99–108.

[30] Frege, G. (1979). Comments on sense and meaning. In H. Hermes, F. Kambartel, & F. Kaulbach (Eds.) Posthumous writings (pp. 1 18-125) (P. Long & R. White, Trans.). Oxford: Basil

[31] Bressan, A. (1993). On Gupta's Book "The Logic of Common Nouns". Journal of Philosophical Logic, Vol. 22, No. 4 (Aug., 1993), pp. 335-383.

[32] Sagum, R. A. (2011) A Named Entity Recognizer for Filipino Text, De La Salle Univesity, Manila (2011).