International Journal of Computer Applications (0975 - 8887)
Volume 182 - No.41, February 2019

A First Step Towards the Development of Yoruba Named Entity Recognition System

Ikechukwu I. Ayogu
Department of Computer Science
The Federal Polytechnic
Idah, Kogi State, Nigeria

Adebayo O. Adetunmbi
Department of Computer Science
The Federal University of Technology
Akure, Ondo State, Nigeria

Bosede A. Ayogu
Department of Computer Science
Federal University
Oye-Ekiti, Ekiti State, Nigeria

ABSTRACT

The NER task can be considered solved for English and a few other European languages given the available research outputs, tools, resources and applications involving NER for these languages. The scenario is sharply different for Nigerian and most of African languages and hence the motivation for the research reported in this paper. The paper presents an exploration of the potency of some language independent features in the recognition of the mentions of persons, locations and organizations in Yorùbá text in a supervised machine learning set-up. The results are promising but as further investigations revealed, the size of the training corpus is yet an issue that needs to be addressed.

General Terms

Information Extraction, Machine Intelligence, Text Processing

Keywords

Named entities, NER, Yoruba language, Natural language processing

1. INTRODUCTION

Named entities (NEs) are information units that are not ordinarily part of the basic grammatical elements of a natural language; they are objects, beings or other 'things' that have conceptual or abstract identities that can be associated with one of a number of designated acts, actions or consequences of actions. This extra-grammatical nature of named entities makes the task of identifying and classifying their mentions in a text a difficult task. NEs are usually associated with one of a number of designated nominal subcategories and are often the centre of processing for NLP systems [1]. Named entity recognition (NER) is the task of identifying and classifying the mentions of NEs in a text into one of a number of predefined types (categories), mostly nouns, temporal and numerical expressions [2] .

The NER task is found useful in a number of NLP applications and has thus been a focus of research since the 1990s. Since NEs are usually external to the grammar elements of the language, higher level NLP requires a way to recognize them for proper treatment. Thus, a NER tool is an important language resource [3]. NER is useful for the improvement of semantic annotation, question answering, ontology population, opinion mining, text summarization, information extraction and machine translation [3, 2, 4, 5]; it has also been shown to lend a good support for credit risk assessment [6]. NER is also a precursor to relation extraction task since it identifies semantic objects of interest from within unstructured text and thus aids the discovery of useful information from unstructured text.

The importance of NER is underscored by the scale of research attention it has received. Many scientific forums (most fundamental of which are the MUC conferences) have been organized with focus on the development of techniques and data resources for NER task with remarkable success for English and many of the European languages. The usefulness of any of the state-of-the-art techniques and applications has not been investigated for Yorùbá NER. Hence, the primary objective of this paper is to investigate the usefulness of the widely used features and sequence modelling framework for Yorùbá NER. It is, to the best of our knowledge, the first effort targeted at addressing the lack of NER system for Yorùbá language. It explores the limits of some language independent and language-specific features for Yorùbá language using the BIO representation scheme. Further, the adequacy of our research corpus for the NER task is investigated.

In the course of over twenty-five years of NER research, numerous approaches and huge data resources have been developed for English and a number of other languages - many of these resources are in the public domain. Since there is no such data resources for Yorùbá, this research creates a NER corpus compiled from religious text. Our annotation distinguished the mentions of Persons (PER), Locations (LOC) and Organizations (ORG) in the corpus while other entity types were grouped into the loosely defined Miscellaneous (MISC) category. Non-entities were given the conventional class O.

The results indicate that our use of the MISC category requires refinement; prominent failures observed in the system were traced to the inability of the model to correctly identify and label the MISC category. The poor recall observed in the overall performance of the system is associated with inability to recognize and classify entities of our MISC category. Therefore, it is necessary to unbundle the MISC group into some more fine-grained categories in furtherance to this work.

2. OUR APPROACH TO THE NER TASK

NER, being a two-step process, requires some additional knowledge other than the word token itself. Research has shown that detecting the boundaries of NEs and their subsequent classification requires both local and non-local knowledge [7, 1]. Classification of detected NEs into one of a number of designated types is difficult due to the ambiguous nature of human language. To successfully classify a NE, a NER model must deal with two forms of ambiguity. The first arises from situations where the same lexical entry refers to two or more different entities of the same type and secondly where identical NE mention refer to entities of completely different type [2].

A number of strategies have been applied to ameliorate the issues arising from ambiguity but the most common approaches are feature engineering, and the supply of supplementary knowledge through the use of gazetteers. Features are designed as suitable predictors of the class of a given NE mention in the text while gazetteers are used to provide auxiliary lists of entities and their classes. Further, gazetteers are used to optimize the performance of the NER model in a specified way; this is often used to tilt the classifier performance towards given category or categories of interest.

The Conditional Random Fields (CRF) algorithm, which is adopted in this paper, is one of the widely used algorithm for NER tasks. It has demonstrated good performance in the task for many languages among which are some Indian languages [8], Malayalam [9], Arabic [10], Chinese [11], Estonian [1] and Sinhala [12]. Each of the research mentioned above has explored features ranging from language-independent to language-specific features. Tkachenko and Simanovsky [13] also explored deeper linguistic pointers to NE identification and classification of NE mentions in text.

Similar to the works mentioned above, this paper explores two groups of features for Yorùbá NE recognition using the CRF framework; experiments were conducted to investigate the contributions of word-internal features and word-external features as well as a combination of both groups to the recognition of Yorùbá.

2.1 Method and Design

The popular approach to NER is the sequence labelling framework. In this framework, NER involves the simultaneous identification and classification of real, abstract or conceptual entities. In order to be successful, then, a NER system must have two important components: a way to know the objects - the features - and a way to say what kind of entities they are - the learning algorithm. Features are descriptors for the members of the classes of NEs and these features encode the relevant attributes of the word token, including its context in a way that provides a learning algorithm with some insights into how the rules that maps the occurrence of NEs to one of the possible classes can be generated.

Popular features for NER tasks include word internal features and those describing its neighbouring words, features describing word affixes, features describing the classes of the neighbouring words and word external features like PoS tag of the current word and that of its neighbours. A combination of these have been explored in diverse ways by numerous research [13].

2.1.1 Feature Set. This paper investigated the contribution of word-internal features (WIF), word-external features (WEF) and context features: a combination of the WIF and WEF set in a number of configurations. For the word-internal feature group, the model examines the word itself, the shape of the word - i.e. capitalization features, hyphenation, presence of digits/numbers, the location of the word in the sentence, the nature of affixation present in the word. For the word external category, the part-of-speech feature was used while context features were a combination of both word-internal features with the part-of-speech feature. Table 1 presents a summary of the feature groups that were explored for the experiments reported in this paper.

2.1.2 The CRF Model. This paper has adopted the CRF model because it has proven itself as a successful algorithm in many sequence labelling tasks, including NER. It has been widely applied in notable previous research [14, 15, 7, 16, 17]. CRF is an undirected graphical model of the hidden markov model sort but which is conditionally trained. It is elegant, permits the use of overlapping features and normalizes over all possible sequence and labels, and is thus immune to the label bias problem. Lafferty et al [18] defined by the CRF model as follows:

$$p(y|x,\mu) = \frac{1}{Z(x)} \exp\Big(\sum_{i=1}^{Y}\sum_{j=1}^{F} \mu_i f_j(s_{i-1}, s_i, x, i)\Big)$$

Where $Z(x)$ is the normalization term, Y is the set of NE tags, F is a set of binary, indicator f_i feature functions defined as $f_j(s_{i-1}, s_i, x, i)$ whose influence is moderated by μ_i, the weight parameter learned from data. A feature is composed of the observed word sequences x, previous state s_{i-1}, current state s_i and the current position in the chain i.

3. THE RESEARCH CORPUS

Data resources are huge and diverse for English and many of the widely researched languages as evidenced in the many publications addressing NER problem in these languages. Academic/research conferences have been organized to study NER and each of these build data for the concerned language(s) [19, 20, 21]. Because no such data resource is available for Yorùbá, an NER corpus was developed for this research from scratch. The corpus is built from religious and religion-based news text. This choice was informed by the unavailability of fully diacritized text for Yorùbá in the electronic media. Yorùbá news sites publish their text without diacritic marks and were considered unusable; the use of diacritic-less texts for Yorùbá NER would increase the ambiguity problems because the tone marks are grammatical in the language.

The raw text was cleaned, normalized and then part-of-speech tagged using a PoS tagging tool described in [22]. The PoS tagged data was then manually annotated with four NE tags following the MUC guideline [20]. The definition of what NE category to use is a difficult task since there is no consensus. Various forums have proposed varying categories, ranging from about four (4) or five to as many as forty-five (45). What makes a NE clearly depends on the task since the NER tool would be as good as the NE concepts its design is based upon [5]. There is however an agreement on the recognition of the people, organization, and location categories. This paper therefore distinguished the Person, Location and Organization entities and grouped other entity types together under the loosely defined miscellaneous category. The labels *PER* for person, *LOC* for location, *ORG* for organization, *MISC* for mis-

Table 1. Description of the categories of features used in the model.

Category	Features	Description
Word-Internal Features (WIF)	Cues from the current word (Wi): word shape, word itself, word length, affixes present, etc.	W_i, W_i is has initial capital, W_i is all caps, W_i has digits, W_i is a given token X, W_i has a prefix Y,etc.
Word-External Features (WEF)	Clues from the PoS category of the word and context.	PoS of $W_i = X$, PoS of $W_i = X$ and PoS of $W_{i-1} = P$ or PoS of $W_{i+1} = Q$, etc.
Context Features (CF)	Combination of WIF and WEF	W_{i-1}, W_i, W_{i+1}; $W_{i-2} = X$, PoS of $W_{i-1} = P$, $W_{i+1} = M$, etc.

cellaneous categories were adopted respectively. Other non-entity lexical items were labeled *O*. An 11,617 token-size corpus containing various instances of NEs as indicated in Figure 1 was used for the training and evaluations in this study.

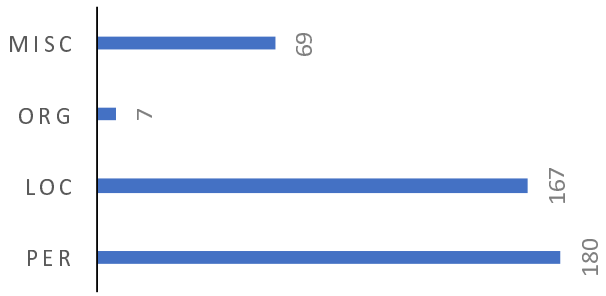

Fig. 1. The occurrences of each NE category in the corpus.

4. EXPERIMENTS AND RESULTS

Two separate set of experiments were performed using a train-test split. The first set of experiments were performed to investigate the contributions of each of the feature categories described in Table 1 using the entire data set in an 80:20 train-test split. The model performance is evaluated using precision, recall and balanced F_1-measure. The results are presented in Table 2. Experiments were also carried out to investigate how each group of feature responds to increasing size of the experimental data, starting with 30 % of the entire set with an incremental step of 10 %. The results of these experiments are visualized in Figure 2(a), (b) and (c).

Table 2. NER Model performances with respect to feature categories.

Feature Category	Precision	Recall	F_1 Score
WIF	0.8654	0.6818	0.7627
WEF	0.8846	**0.6970**	**0.7797**
CF	**0.8889**	0.6061	0.7267

As shown in Table 2, the performance of each group of features is comparable. While the combination of the WIF and WEF features results in a highest precision score, the recall and the F_1-score for this group is the poorest, this can be attributed to sparseness as

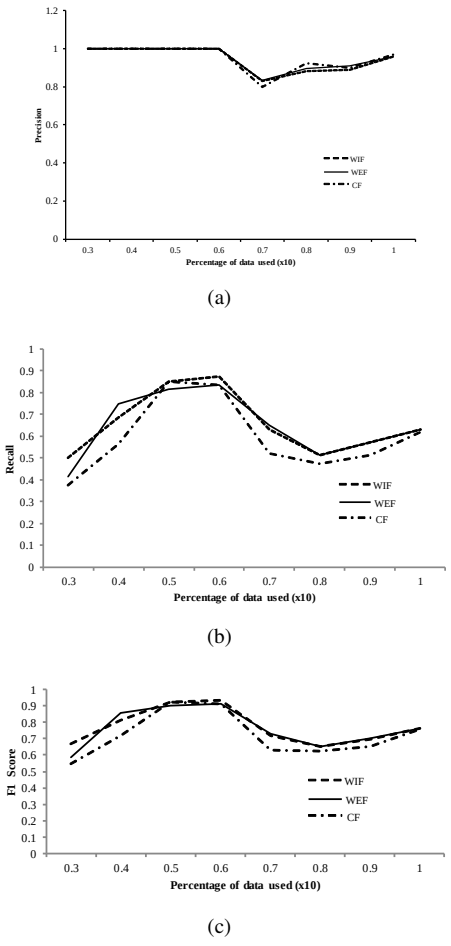

Fig. 2. The effects of training set size on NER model response for each feature category (a) Precision (b) Recall (c) F1-score to increasing corpus size

the number of features generated for this group increased by three folds compared to the individual groups. The WEF group attained a marginally higher performance over the other two in terms of recall and F_1-score. For NER application, the ability to recognize mentions of NEs is important and highly desirable. The high precision score attained by the various feature groups indicates that language-independent features are applicable to Yorùbá language.

While the performance of the system is encouraging, as seen in Table 2, the responses of the NER model to the increasing size of data as shown in Figure 2 (a), (b) and (c) respectively, is an indication that the quantity of the data set is not adequate. That notwithstanding, these figures compare to the results obtained for Indian languages at ICON 2013 [8] for systems trained with much larger data. The same is true when the performance of the model is tabled against the results obtained by Prasad and his co-researchers for Malayalam [9] using a similar CRF-Based implementation. The model performance also indicates a positive correlation with the baseline performance reported by Tkachenko *et al* [1] for Estonian.

5. CONCLUSION

This paper has presented a simple, baseline NER model for Yorùbá language using a small corpus annotated for the purpose of experimenting language-independent feature sets. The performance of the model is encouraging but the size of the training corpus is inadequate. The imperative as observed from the model performances is to build more corpora and enrich the models with more elegant features for improved system.

6. REFERENCES

[1] A. Tkachenko, T. Petmanson, and S. Laur. Named entity recognition in estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83, 2013.

[2] B. Mohit. Named entity recognition. In *Natural language processing of semitic languages*, pages 221–245. Springer, 2014.

[3] Y. Benajiba, P. Rosso, and J. M. Benedíruiz. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer, 2007.

[4] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489, 2013.

[5] I. Augenstein, L. Derczynski, and K. Bontcheva. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83, 2017.

[6] J. C. S. Alvarado, K. Verspoor, and T. Baldwin. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, 2015.

[7] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[8] A. Das and U. Garain. Crf-based named entity recognition @ icon 2013. *arXiv preprint arXiv:1409.8008*, 2014. https://arxiv.org/ftp/arxiv/papers/1409/1409.8008.pdf.

[9] G. Prasad, K. K. Fousiya, M. A. Kumar, and K. P. Soman. Named entity recognition for malayalam language: A crf based approach. In *Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015 International Conference on*, pages 16–19. IEEE, 2015.

[10] A. Abdul-Hamid and K. Darwish. Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, pages 110–115. Association for Computational Linguistics, 2010.

[11] W. Chen, Y. Zhang, and H. Isahara. Chinese named entity recognition with conditional random fields. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 118–121, 2006.

[12] K. U. Senevirathne, N. S. Attanayake, A. W. M. H. Dhananjanie, W. A. S. U. Weragoda, A. Nugaliyadde, and S. Thelijjagoda. Conditional random fields based named entity recognition for sinhala. In *Industrial and Information Systems (ICIIS), 2015 IEEE 10th International Conference on*, pages 302–307. IEEE, 2015.

[13] M. Tkachenko and A. Simanovsky. Named entity recognition: Exploring features. In *KONVENS*, pages 118–127, 2012.

[14] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[15] J. Kazama and K. Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[16] W. Radford, X. Carreras, and J. Henderson. Named entity recognition with document-specific kb tag gazetteers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 512–517, 2015.

[17] D. Seyler, T. Dembelova, L. Del Corro, J. Hoffart, and G. Weikum. A study of the importance of external knowledge in the named entity recognition task. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 241–246, 2018.

[18] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[19] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

[20] N. Chinchor and P. Robinson. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, 1997.

[21] C. Walker, S. Strassel, J. Medero, and K. Maeda. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57, 2006.

[22] I. I. Ayogu, A. O. Adetunmbi, B. A. Ojokoh, and S. A. Oluwadare. A comparative study of hidden markov model and conditional random fields on a yorùba part-of-speech tagging task. In *Computing Networking and Informatics (ICCNI), 2017 International Conference on*, pages 1–6. IEEE, 2017.