Journal of Theoretical and Applied Information Technology
29th February 2016. Vol.84. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

NER IN ENGLISH TRANSLATION OF HADITH DOCUMENTS USING CLASSIFIERS COMBINATION

[1]MOHANAD JASIM JABER , [2]SAIDAH SAAD

Faculty of Information Science and Technology, Universiti Kebangasaan Malaysia (UKM) Bangi, Selangor, Malaysia.

E-mail: [1]mojj82@gmail.com, [2]saidah@ukm.edu.my

ABSTRACT

There is a need to retrieve and extract important information in order to fully understanding the ever-increasing volume of English translated Islamic documents available on the web. There is limited research focused on Named Entity Recognition (NER) for Islamic translations even though NER has seen widespread focus in other languages. Translated named entities have their own characteristics and available annotated English corpora do not cover all the transliterated Arabic names, which makes translations with NER difficult in the Islamic domain. This research addressed the use of NER in English translations of Hadith texts. The objective of this research was to design and develop a model that was able to excerpt Named Entities from English translation of Hadith texts. This research used supervised machine learning approaches, like Support Vector Machine (SVM), Maximum Entropy Classifier (ME) and Naive Bayes (NB), which were later combined via majority voting algorithm to identify named entities from Hadith texts. From the results of this research, voting combination approaches outmatched single classifiers with an overall F-measure of 95.3% in identifying named entities. The results indicated that combined models paired with suitable features were better suited to recognize named entities of translated Hadith texts as compared to baseline models.

Keywords: *Named Entity Recognition, supervised machine learning, Hadith text.*

1. INTRODUCTION

NER. There are three core techniques and approaches to NER: rule-based approach, machine learning approach and hybrid based approach. The use of a set of human made rules to extract names is classified under rule-based approaches. These models include the use of different patterns, which include grammar (such as part of speech (POS)), syntax (such as word precedence) and orthographic based features (such as capitalization) with the use of dictionaries. The drawbacks of rule-based models are that they are not very portable, dynamic and robust due to the large maintenance to the rules when even when small changes occur. This type of model is domain dependant and works well with only selected languages, as they do not adapt well to new languages and domains.

The amount of electronic Islamic documents worldwide has drastically increased with the rapid advance of Internet technologies. This has made the process of understanding and extracting useful information via conventional search engines a very difficult task. It is a great challenge to gleam knowledge from Islamic documents. If done successfully, it will improve the education of the Muslim-world in areas of knowledge representation and reasoning, for knowledge correctness and extraction from texts and for online collaboration. With help of semantic knowledge, it is possible to fully understanding the Hadith with the assistance of a complete Islamic NER. Based on machine learning approaches, which includes Support Vector Machine (SVM), Maximum Entropy Classifier (ME) and Naive Bayes (NB), and combined with majority voting algorithms, this study will describe a new method for NER in Islamic texts (English translation of the Hadith). The combination of classifiers with individual machine learning was effective on several languages. The aim of this model is to be more effective than earlier models.

2. ELATED WORK

In the recent decade, several studies utilized the combination of classifiers for NER. A NER system used by Ekbal and Bandyopadhyay[1] utilized a combination of different machine learning classifiers including Conditional Random Fields (CRF), Maximum Entropy (ME) and Support Vector Machine (SVM). It consisted of both

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

language dependent and language independent features. The system was utilized in the Bengali language and the results indicated that the language dependent features could greatly improve the accuracy. Ekbal and Bandyopadhyay [2] applied SVM in the Bengali and Hindi languages and used only language independent features for training and testing. The SVM tests showed very high recall and precision. To find optimized sets of features and compare the results from two approaches, Saad [3] used Support Vector Machine (SVM) and Conditional Random Fields (CRF). Both research indicated that the performance of SVM and CRF were quite the same, save for the fact that SVM performed better on data with random contexts.

Patterns were used to identify the key concepts, properties and the relationships (that existed between them) in the Islamic domain by Saad et al. [4] and Saad et al. [5]. Several methods for extraction of keywords and phrases were tested by Saad & Salim [6] in order to develop an ontology for Islamic Knowledge. They used lexico-syntactic and statistical methods to extract potential keywords and phrases.

Association rules were used by Harrag et al. [7] to extract the ontology of prophetic narrations (Hadith). This method involved the use of the Apriori algorithm to compute correspondence relations and association rules in order to identify frequent item sets on concepts that were related to Islamic jurisprudence (Fiqh) from the Sahîh Al-Bukhârî documents. The conceptual relations embedded in the semantic structure of the Sahîh Al-Bukhârî documents were modelled based on association rules to extract a specific domain ontology.

The narrators' chain of a given Hadith was automatically generated and graphically displayed by Azmi & bin Badia [8]. It involved the parsing and annotation of the Hadith texts and identification of the narrators' names. A domain specific grammar was used with shallow parsing to parse the content of the Hadith. A transformation mechanism based on semantic web ontology was then utilized to depict the narration chain in a standard format and then graphically render the complete tree. A full narration tree was automatically created after the Hadith text was parsed and annotated while recognizing the narrators' names. Finite state transducers-based system was utilized by Harrag [9] to detect and extract passages or sequences of words containing relevant information from the prophetic texts. Their approach was deemed feasible based on the results of experimental evaluation. This system achieved precision and recall rates of 71% and 39% respectively. However, the extraction of named entities from English translation of Islamic documents was not attempted by any of these researches. Hence, the main objective of this research is to narrate a new methodology to design and develop a NER model for English translations of Hadith documents based on a combination of classifiers. Based on an ensemble of Support vector machine (SVM), Maximum Entropy Classifier (ME) and Naive Bayes (NB) classifiers, a new machine learning classification framework will be proposed.

3. METHODSAND MATERIALS

This study implemented machine learning for Islamic NER. The architecture of Islamic NER system and description of the functionality of each component will be reviewed in this section. Firstly, imperfect, noisy and sporadic data were removed via pre-processing tasks. It must be noted that data should be pre-processed before any other data mining operations. A few extraction models were utilized to get the best discerning terms for training and testing. Lastly, a few machine learning classification methods were utilized for Islamic NER. However, Figure 1 shows the overall architecture of the method, which involves the following phases:

- **Pre-processing phase**
- **Feature extraction phase**
- **NER phase.**
- **Evaluation phase.**

Existing annotated training data is required for supervised machine learning techniques. This type of training data is manually created by individuals or experts in the respective fields. This research utilized a developed annotated dataset from Islamic text based on the English translation of the Hadith. This dataset was especially developed from Islamic text for NER. Every word in our training corpus was labelled for six different named entities: person, location, organization, time, money and date. The Muslim's book of Hadith was the source for the English translation used for this research. Because the Hadith is a well-known book and a complete version is available on the network, the source data set consisted or around 100 Hadiths from different chapters.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

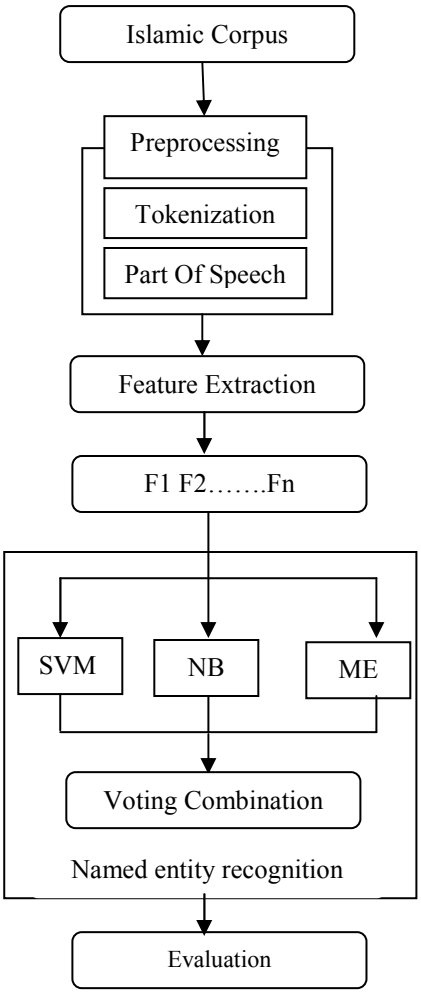

Fig 1: The Proposed Islamic named Entity Extraction System Architecture

5. PRE-PROCESSING PHASE

The pre-processing phase used a machine-learning framework and was a highly important phase in the system. Each Hadith document must pass through the pre-processing phase before any NER. Each new Islamic document was passed through the system as the following pre-processing steps:

- **Tokenization**: Tokenization was the first step of any natural language processing task. The exploration of words in a sentence was the objective of tokenization. At the beginning, textual data was only a block of characters. White space and punctuation marks were the main dependable variables of the tokenizing of text into words.
- **Part of Speech (POS)** tagging was the ability to computationally identify which word in a POS was triggered by its use in a specific context. The most important element of pre-processing was POS tagging because it affects the performance of the NER system. The POS tagging system developed by the Stanford Natural language processing group (Stanford, 2010), which was in turn based on the Penn Treebank Tag set for English, was utilized in this research.

4. FEATURE EXTRACTION

Feature extraction is vital part of any NER system because it improves the efficiency of recognition tasks with regards to the speed and effectiveness of learning. Training and test data were processed by one or more pieces of software in order to extract descriptive information in the feature extraction phase. Special formatting to suit the input format of the system may be required in preparing the data for use by the feature extractor. The output of the feature extraction may need to be reformatted to be compatible with what is expected by the machine learning module(s). This step is dependent on the choice of machine learning method. The aim of this phase was to convert each word to a vector of feature values. A set of features for the he extraction of person, location, organization, time, money and date entities from Islamic documents was defined in this step. These features can be classified into three main feature sets: features based on POS tagging, features based on word affixes, and features based on the context. Table 3.4 shows a summary of these feature sets.

The following feature vector was used to represent the words in the corpus. The main aspects for the NER task were mostly determined and chosen without any deep domain knowledge and/or language specific resources Ekbal & Saha [10] and Saha & Ekbal [11]. The reasoning in using these features were because of the independent nature of the language, it could be easily obtained for almost all the languages and had effective features that improved the NER Azpeitia et al. [12] and Das et al. [13] and Figueroa [14] Saha & Ekba[15]. Table 1 details these feature sets.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Summary Of The Feature Sets

Feature Category	Feature Name	Feature	Reason (justification)
Word affixes	F1 F2 F3 F4	Prefix1 Prefix2 Suffix1 Suffix2	Fixed length word suffixes and prefixes are very effective in identifying NEs and work well for the highly inflective Arabic languages. Actually, these are the fixed length character sequences stripped either from the rightmost (for suffix) or from the leftmost (for prefix) positions of the words For examples family names in Arabic such as alKindi, alMuzani, alAzdi, alMughiraalMakhzumi and alBahili share the same suffix and prefixes
Context-based features	F5 F6 F7	Previous word Named entity tag of the previous words POS tag of the word	Previous word and Dynamic NE information is used based on the observation that surrounding words carry effective information for the identification of NEs For example, all the flowing Arabic names begin with words that always come before Arabic names : " Bani Abd alAshhal , bani Abd alHarith Khazraj , bani Abu Talh , bani Amir Luway , bani Amr Auf , bani anNajjar , bani Fihr Abu Husain Rabia"

6. NER

There are mainly two phases in the machine learning based techniques and approaches. One is when the training is initially performed to produce a trained machine while the other is when a NER step was performed. In this study, the following machine learning approaches were evaluated:

6.1 SVM Classifier

Cortes and Vapnik[15] proposed a machine learning technique that was called the Support Vector Machine (SVM). SVM was used in the machine learning area and was a generally popular technique for NER. It was a classification technique with a very high efficiency.

SVM worked by implementing a decision surface in order to separate the training data nodes into two main classes. This was based on the idea of structural-risk minimization from computational-learning theory. It then made decisions based on existing support vectors, which were chosen from components that were efficient in the training set. The optimization procedure of SVMs (dual form) was aimed to minimize the following:

$$\vec{\alpha} = arg\ min\left\{-\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle\vec{x_i},\vec{x_j}\rangle\right\}$$

$$Subject\ to: \sum_{i=1}^{n}\alpha_i y_i = 0; \quad 0 \le \alpha_i \le C$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6.2 Naive Bayes

The Naïve Bayes (NB) algorithm was a widely used algorithm for NER. It was simple and easy implementation and better-performing algorithms were the main advantages of NB. The NB technique was extensively utilized for NER. This technique decided the rear possibility based on a table of feature vectors, where the term was related to multiple named entity classes, and assigned it to the category with the maximum rear possibility. Multinomial models and multivariate Bernoulli models were the two often-used models. NB was a stochastic model used to create documents based of Bayes' rule. To classify as the best named entity class, n*, for a new term *w*, it computes:

$$p(c_j|w_i) = \frac{p(c_j)p(c_j|w_i)}{p(w_i)}$$

The classifier consisted of features relevant to the different classes and was specified by the user. The main aim was to identify the token class, t, given the feature set, f_t, belonging to the particular token by computing the posterior probability, P(class|f_t). The Bayes theorem stipulated:

$$P(class|f_t) = \frac{p(f_t|class).\ p(class)}{p(f_t)}$$

To identify the most likely class of a given feature set, $p(f_i|c)$ is calculated for each class C.

6.3 Maximum Entropy Classifier

The Maximum Entropy Classifier principle was a commonly used technique, which provided the probability of belongingness of a token to a class. The maximum entropy classifier calculated the probability, $\mathbf{p(y/h)}$, for any $\mathbf{y}$ from the space of all possible outcomes, $\mathbf{Y}$, and for every h from the space of all possible histories, $\mathbf{H}$. In NER, history can be classified as all derivable information from the training corpus relative to the current token. It is unique, agrees with the maximum likelihood distribution and has the exponential form:

$$\boldsymbol{P\left(\frac{Y}{h}\right) = \frac{1}{z(h)} exp^{\sum_{j=1}^{n} \lambda j\ f_j(h,Y)}}$$

where y is the NE tag, his is the context (or history), $f_j(h, Y)$ are the features with associated weight, λj, and Z(h) is a normalization function.

6.4 Classifier Combination

The results, based on the output of the three classifiers, were selected with the application of an ensemble (classifier combination) approach. The accuracy of the combined classifiers was determined by choosing the best answer given a set of three answers and this accuracy was the main task of the selection algorithm. The predictions of component classifiers were counted with the voting rule. They were the assigned test a word, x, to entity I with the most component predictions.

7. Evaluation

A manually labelled Islamic corpus was used to measure the performance of the machine learning algorithms. Collected English translations of the Muslim Hadith were used to compile the corpus used in this work. The Muslim's book of the Hadith (around 100 hadiths from different chapters) was used as the source of the data set in this research because it was a quite well known book and a complete version it was widely available on the network. Six different named entities based on person, location, organization, time, money and date in each hadith were manually annotated. The corpus was randomly partitioned into 10 equal subsamples using the cross-validation process in this research. To test the model, a single subsample was retained for data validation while the remaining nine subsamples were used as the training data. The cross-validation process was then repeated 10 times (10 folds). This caused the named entities of person, location, organization, time, money and date to differ from one experiment to another. Precision, recall and the weighted mean Fβ=1-score were used to measure and evaluate the performance of the NER systems participating in the CoNLL-02, CoNLL-03, and JNLPBA-04 challenge tasks. Precision was the percentage of correctly named entities found by the learning system while recall was the percentage of named entities present in the corpus that were found by the system. For a named entity to be correct, it must have an exact match of a corresponding entity in the data file: the complete name of the entity.

$$\text{Recall} = \frac{\#\text{ of correctly classified entities}}{\#\text{ of entities in the corpus}} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

$$\text{Precision} = \frac{\#\text{ of correctly classified entities}}{\#\text{ of entities found by algorithm}} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

8. RESULTS

Firstly, the baseline experiments were performed on words that were used as features. The Islamic NER system used classification methods like SVM, Maximum Entropy Classifier and Naive Bayes algorithms. The baseline for the Islamic NER and classification models were evaluated using the performance of this research. Table 2 shows the results of this experiment. The classification accuracy results, as indicated in this table, were all based on the evaluation matrices of precision, recall and F-measures.

Table 2: Results Obtained By Support Vector Machine, Maximum Entropy Classifier And Naive Bayes When Words Feature

Classifier	Precision	Recall	F-measure
NB	80.4	77.8	78.81
SVM	82.9	79.7	81.03
MaxEnt	67.91	79.3	73.16

Table 2 indicated that the best overall performance, without the use of features, was a precision of around 77 %, recall of 79 % and F-measure of 77 %. There was a need to enhance the results, as the best overall results were not high. Secondly, several experiments were carried out to empirically analyse the four classification approaches: Support Vector Machine (SVM), Maximum Entropy Classifier (MaxEnt), Naive Bayes (NB) and classifiers combination based on majority voting approach for NER in Islamic texts (English translation of the Hadith). Table 3 shows a summary of the experimental results of SVM, MaxEnt, NB and classifiers combination based on majority voting approach.

Table 3:Performance (the average F-measure for each class) of the SVM, MaxEnt, NB and classifiers combination based on majority voting approach for NER in Islamic texts.

Classifier	Precision	Recall	F-measure
SVM	93.9	92.75	93.3
MaxEnt	66	77.51	71.2
NB	93.0	90.8	91.04
Classifiers combination	96.9	93.6	95.3

As shown in Table 3, the experiments of Islamic NER indicated that the highest f-measure result was by SVM with a 93.3% score and the lowest result was by the MaxEnt with 71.2%.

SVM used a refined structure that acknowledged the relevance of most features and had the ability to handle large feature spaces. Therefore, SVM would be a good choice for Islamic data sets as the feature set cases for each example occurred frequently.

Table 3 also showed the results obtained using classifier combination methods, which outperformed those obtained using all individual classifiers. Classifiers for Islamic NER and classification. The most suitable method for Islamic NER and classifications, as indicated by the results, was the classifier combination.

The experiments in this research indicated highly promising results that clearly demonstrated the suitability of using machine-learning algorithms for Islamic NER.

9. CONCLUSION

Machine learning classification approaches, namely, Naïve Bayes, Support Vector Machine (SVM), Maximum Entropy Classifier and the combination of its majority voting approach, were used to propose an Islamic NER. A new Hadith corpus text was collected and manually labelled. The results of the evaluation were validated with a manually annotated Hadith dataset. The best performance for Islamic NER Finally was demonstrated by the classifier combination method, even outperforming individual methods. The results clearly state that the classifier combination method was the most appropriate method for Islamic NER. The development of a large Islamic corpus and design for a general framework for Islamic information extraction and analysis could be targeted for future research. Future work could also be extended on the above-suggested methods by adding more advanced feature sets to evaluate them for NER for English translations of Islamic texts.

10. ACKNOWLEDGEMENT

This work has been supported by the University Research Grant DPP-2015-015 and GUP-2015-003.

REFERENCES

[1] Ekbal, A., & Bandyopadhyay, S. (2009). Voted NER system using appropriate unlabeled data. In, Proceedings of the 2009 Named Entities

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Workshop: Shared Task on Transliteration (pp. 202-210): Association for Computational Linguistics.

[2] Ekbal, A., & Bandyopadhyay, S. (2010). Named entity recognition using support vector machine: A language independent approach. International Journal of Electrical, Computer, and Systems Engineering, 4, 155-170.

[3]Saad, S. (2013). ONTOLOGY LEARNING AND POPULATION TECHNIQUES FOR ENGLISH EXTENDED QURANIC TRANSLATION TEXT. In, Faculty of Computing Universiti Teknologi Malaysia.

[4]Saad, S., Salim, N., & Zainal, H. (2009). Pattern extraction for Islamic concept. In, Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on (pp. 333-337): IEEE.

[5]Saad, S., Salim, N., Zainal, H., & Noah, S.A.M. (2010). A framework for Islamic knowledge via ontology representation. In, Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on (pp. 310-314): IEEE.

[6]Saad, S., & Salim, N. (2008). Methodology of Ontology Extraction for Islamic Knowledge Text. In, Postgraduate Annual Research Seminar.

[7]Harrag, F., Alothaim, A., Abanmy, A., Alomaigan, F., & Alsalehi, S. (2013). Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules. International Journal on Islamic Applications in Computer Science And Technology, 1, 48-57.

[8]Azmi, A., & bin Badia, N. (2010). An Application for Creating an Ontology of Hadiths Narration Tree Semantically and Graphically

[9]Harrag, F. (2014). Text mining approach for knowledge extraction in Sahîh Al-Bukhari. Computers in Human Behavior, 30, 558-566.

[10]Ekbal, A., & Saha, S. (2012). Multiobjective optimization for classifier ensemble and feature selection: an application to named entity recognition. International Journal on Document Analysis and Recognition (IJDAR), 15, 143-166.

[11]Saha, S., & Ekbal, A. (2013). Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. Data & Knowledge Engineering, 85, 15-39.

[12]Azpeitia, A., Cuadros, M., Gaines, S., & Rigau, G. (2014). NERC-fr: Supervised Named Entity Recognition for French. In, Text, Speech and Dialogue (pp. 158-165): Springer.

[13]Das, B.R., Patnaik, S., Baboo, S., & Dash, N.S. (2015). A System for Recognition of Named Entities in Odia Text Corpus Using Machine Learning Algorithm. Computational Intelligence in Data Mining-Volume 1 (pp. 315-324): Springer.

[14]Figueroa, A. (2015). Exploring effective features for recognizing the user intent behind web queries. Computers in Industry, 68, 162-169

[15]Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.