Minimizing Human Labelling Effort for Annotating Named Entities in Historical Newspaper

W. M. F. Wan Tamlikha, B. Ranaivo-Malançon and S. Chua
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
wanfaisal92@gmail.com

***Abstract*—To accelerate the annotation of named entities (NEs) in historical newspapers like *Sarawak Gazette*, only two choices are possible: an automatic approach or a semi-automatic approach. This paper presents a fully automatic annotation of NEs occurring in *Sarawak Gazette*. At the initial stage, a subset of the historical newspapers is fed to an established rule-based named entity recognizer (NER), that is ANNIE. Then, the pre-annotated corpus is used as training and testing data for three supervised learning NER, which are based on Naïve Bayes, J48 decision trees, and SVM-SMO methods. These methods are not always accurate and it appears that SVM-SMO and J48 have better performance than Naïve Bayes. Thus, a thorough study on the errors done by SVM-SMO and J48 yield to the creation of ad hoc rules to correct the errors automatically. The proposed approach is promising even though it still needs more experiments to refine the rules.**

***Index Terms*—J48; Naïve Bayes; Named Entity; Sarawak Gazette; SVM-SMO.**

I. INTRODUCTION

The term "named entity" (NE) was introduced in 1995 during the Message Understanding Conferences 6. NEs are textual units that correspond to names (person, organization, location, etc.) and numeric expressions (date, monetary value, percent, etc.) [1]. NEs occurring in text corpora are annotated to assist information extraction systems, to create gold standard for machine learning techniques, or to increase a search within the texts. Annotating NEs is not straightforward as many issues need to be considered like doing it manually or automatically. Manual annotation is possible if the size of the input text is small. For large set of texts (e.g., newspapers), automatic annotation is the only alternative. But this approach has some undesirable consequences such as incorrect and missing annotations. Therefore, this paper is proposing a framework to minimize human labelling effort when annotating the NEs in *Sarawak Gazette* (henceforth called SAGA) by providing the result of a rule-based named entity recognizer (NER) as a training data to several supervised learning NER methods: Naïve Bayes, J48 Decision Trees, and Support Vector Machines. The aim is to determine the most accurate supervised NER method trained with a small number of NEs.

The first motivation behind the proposed framework is referring to the statement written by Ratinov and Roth in 2009: "NER system should be robust across multiple domains, as it is expected to be applied on a diverse set of documents: historical texts, news articles, patent applications, webpages etc." [2]. If this statement holds, then running any existing NER system on historical newspaper like SAGA should yield high accuracy. But it is not the case as shown in Table 1. The edition of SAGA published in January 1904 was submitted to the widely-used Stanford NER, which is based on Conditional Random Fields (CRF) method. Four NEs are considered in this study: Date (DAT), Location (LOC), Organization (ORG), and Person (PER). Only the recognition of DAT could go beyond the average 0.50 F-measure. The other three entities are poorly recognized. These findings have also been observed by Wettlaufer and Thotempudi [3] when testing rule-based NER systems on 18th century German texts.

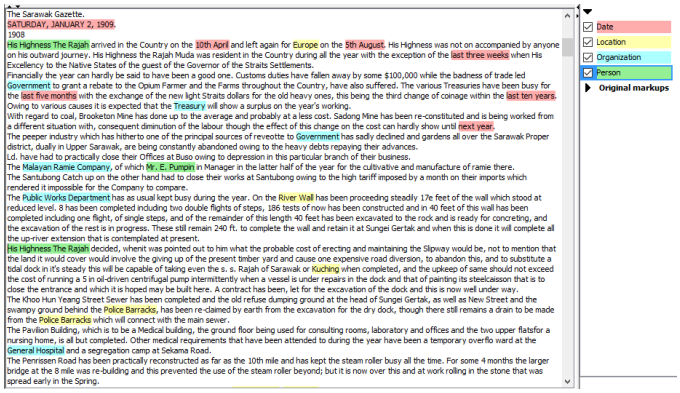

Figure 1: Example of NER

Table 1
Performance of Stanford CRF-based NER on SAGA January 1904

NE	Recall	Precision	F-measure
DAT	0.5441	0.6852	0.6066
LOC	0.3934	0.5294	0.4514
ORG	0.2426	0.2426	0.2426
PER	0.0761	0.1573	0.1026

The second motivation refers to the statement of Neudecker in 2016 regarding the availability of annotated NE historic corpora for training the Stanford CRF NER: "there were no corpora available at the time that could cover the requirements of the project, i.e. historic newspaper content, texts in the languages Dutch, French and German, and carrying sufficiently open licenses that would allow for the adaptation, extension and redistribution of such corpora." [4]. Newspaper articles have been used widely as a corpus source for training NER systems. However, most of these newspapers are contemporary newspapers.

The third and last motivation is linked to the final goal of the work reported in this paper, which is the development of an information extraction system for SAGA. For that reason, it is crucial that all NEs in SAGA editions are annotated. Currently, the digitized SAGA in our possession corresponds

to 1567 editions from 1903 until 1939 with each edition containing an average of 10 pages. Considering the size of the dataset and approaching the problem with manual annotation is impossible as it is time consuming. In addition, through our experience, hiring human experts for the annotation is very difficult either through paid task or through free crowdsourcing.

The remainder of this paper is as follows. Section II provides an overview of the existing methods for building annotated NE corpus as well as a non-exhaustive list of English historical corpora with annotated NEs. Section III explains the proposed method in annotating NEs in SAGA with minimal human effort. In Section IV, the method is applied on a subset of SAGA and the results are analyzed. Section V concludes the paper some directions for future work.

II. Related Work

A. Methods for Building Annotated NE Corpus

The landscape of building annotated NE corpus is illustrated in Figure 2. The four considered parameters are "cheap", "expensive", "manual", and "automatic". Building annotated NE corpus serves to fulfil one or both of the following objectives: to minimize human effort or to increase the size of an annotated corpus.

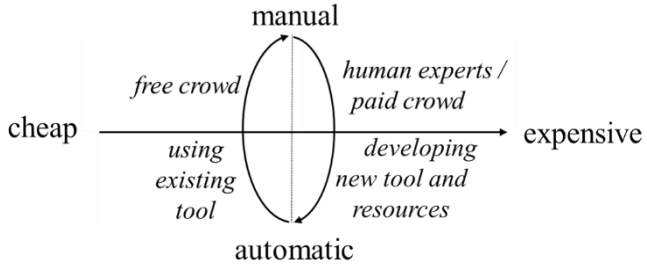

Figure 2: Landscape of NEs annotation

a. Manual Annotation

Manually annotated by experts – Corpora that have been annotated manually by human experts are considered as gold standard corpora. As examples of such corpora are those created for the evaluation campaigns like Message Understanding Conferences (MUC), Conference on Computational Natural Language Learning (CoNLL), and Automatic Content Extraction (ACE). In specific domains like biomedical domain, Ogren et al. [5] created a gold standard evaluation corpus for evaluating clinical in-house NER system. The gold standard corpus was realized through the consensus between four human annotators.

Manually annotated by crowds – Finin et al. [6] took advantage of the crowd workers in Mechanical Turk and CrowdFlower to annotate NEs in Twitter status updates. Dojchinovski et al. [7] used also crowdsourcing approach with CrowdFlower to get a corpus already annotated NEs with their salient level (most salient, less salient, not salient). The creation of corpora using language resources like Wikipedia and DBpedia can fall under this group since the annotation was done by Wikipedia users [8].

Manually annotated by experts and crowds – The Broad Twitter Corpus (BTC) has been annotated by NLP experts and crowd workers [9]. BTC is considered by its authors as a gold standard since it was "sampled across different regions, temporal periods, and types of Twitter users" [9]. The annotated NEs in BTC are PER, LOC, and ORG.

b. Automatic Annotation

This approach assumes that the requested resources for NE tagging exist already. The resources can be rules, gazetteers, trained models, gold standard corpora, as well as NER tools. However, in the context of multilingual NER, the resources must exist and comparable for all considered languages. To avoid time-consuming human annotation and to get comparable evaluation results across-languages, Ehrmann et al. [10] projected the annotation of English corpus into other language corpora. The translation of NEs are obtained from the application of a phrase-based statistical machine translation system and the exploitation of a multilingual NE database. Then, incremental strategies are applied to project a given English NE in a sentence with its list of possible translations into the corresponding sentences of the aligned corpora.

c. Mixing Manual and Automatic Annotation

The methods for getting a fully and correctly annotated NE corpus is geared to the minimization of human effort in labelling NEs. Human annotation is attractive as it is assumed that the labels of NEs provided by human experts are correct. However, it is known that human annotation is labor-intensive. Therefore, the other alternative is to use automatic annotation methods. But their results are not always correct. Thus, the idea is to combine automatic and manual annotations, which can be broadly divided into three groups.

Manual annotation followed by automatic annotation – In this approach, the manual annotation is done before running an automatic annotator. This corresponds to the general scenario of a supervised learning approach. In the Europeana Newspapers project, the annotation of NEs historic newspapers (published before 1900) were first labelled by humans. This gold standard corpus is then used to train and evaluate the CRF Stanford NER in a 4-fold cross-evaluation [4]. The CRF model is aimed to annotate the 1,000 digitized European newspapers.

Automatic annotation followed by manual correction – As mentioned in [11], "the idea of improving the efficiency of annotation work by using automatic taggers is certainly not new". "The motivation for assisted annotation is that pre-annotations can both speed up the annotation process and reduce missed annotations." [12]. To obtain a corpus of consumer health questions annotated with NEs, Kilicoglu et al. [12] pre-annotated a raw corpus with different tools, and then submitted the pre-annotated corpus to six human annotators for evaluation. The authors qualified their approach as "assisted annotation". The tools used are MetaMap ("maps biomedical free text to UMLS Metathesaurus concepts" [12]), Essie ("maps free text to UMLS Metathesaurus concepts" [12]), KODA (a knowledge-based NER [12]), customized UMLS dictionary lookup, and a CRF-based NER already trained on a health corpus. Surprisingly, when assessing the effect of automatic pre-annotation and human annotation, Kilicoglu et al. [12] found "moderate inter-annotator agreement" with the assisted annotation yielding "slightly better agreement and fewer missed annotations than manual annotation." These findings illustrate well the difficulties in annotating NEs in specific domains.

Interactive annotation – In the third and last group, human

intervenes during the automatic annotation, and thus the annotation is interactive. An illustration of that approach is the work reported by Tsuruoka et al. [11]. They proposed a framework for accelerating the annotation of sparse NEs in texts. They defined the framework as an "iterative and interactive process between the human annotator and a probabilistic named entity tagger". The authors considered the approach as reducing the human annotation task, "almost by half, achieving a coverage of 99.0%", as only sentences containing NEs of the target category are presented to the human annotators. The sentences are those selected by the probabilistic NER, which is based on CRF method. The framework was tested on corpora from general domain and biomedical domain.

B. Annotated NEs in English Historical Corpora

The largest project working on historical newspapers is the Europeana Newspapers Project. As a sample of the outputs of this project are three NE annotated historic newspapers, 100 pages each for the languages Dutch, French, German (including Austrian). The corpora are made available in the public domain [4].

Grover et al. [13] developed an in-house rule-based NER to annotate British parliamentary proceedings from the late 17th and early 19th centuries [13]. With the development set corresponding to 1814-1817 OCRed proceedings, the F-score is 0.7212 for LOC and 0.8067 for PER. However, the performance drops drastically with a different test set of proceedings published between 1685 and 1691. The F-score is 0.2408 for LOC and 0.7503 for PER.

Recently, DeLozier et al. [14] reported the process and challenges in annotating a historical US civil war corpus with geographic reference. An annotation tool was developed to assist five hired annotators "although in practice most of the work was done by a single annotator" [14]. The annotation process took 280 hours over two months for 25-page subsections of 118 of 126 volumes. The annotated corpus is freely available under an MIT License.

The Trove Newspaper Corpus in the National Library of Australia is a large collection of digitized newspapers dating back to 1803. The NEs in the corpus was annotated with Stanford CRF NER. The goal was to extract precisely person names and location names [15]. By pre-training Stanford NER with 500 articles, the authors obtained as F1 score 0.76 for LOC and PER, and 0.51 for ORG.

III. PROPOSED METHOD

A. Methods of Annotations

From the reviewed works, our proposed approach in building annotated NE corpus falls under the group of "automatic annotation". Our objective is not to create a gold standard but to obtain the most accurate trained model from three supervised learning methods. The goal is to be able to annotate all SAGA editions in a fast way, with minimum errors, and minimal human effort. Thus, our method goes through two main steps. The first step corresponds to a pre-annotation as in Kilicoglu et al. [12] but instead of a CRF-based NER, our method makes use of a rule-based NER. In general, rules are created by humans, and thus should be accurate. In the second step, three supervised learning methods are evaluated to determine the most efficient method.

B. Pre-annotation with Rule-based NER

The first step of the proposed method is to submit one SAGA edition to a rule-based NER. The objective is to get an initial annotation of NEs, and thus avoiding human annotation. There are only few rule-based NER systems for English texts and thus, we opted for the most widely used, that is ANNIE (A Nearly New Information Extraction system). ANNIE is an open-source NER module integrated in GATE, a General Architecture for Text Engineering. ANNIE can process directly texts either with its by-default resources or with user-defined resources. The by-default resources are for any kind of English texts and comprise tokenizer, sentence splitter, morphological analyzer, part-of-speech (POS) tagger, coreference resolution identifier, JAPE rules, and a set of gazetteers. The processing pipeline of ANNIE as well as each component of the pipeline can be modified, which is in our future work for processing SAGA articles.

C. Annotation with Supervised Learning NER

The second step of the proposed method concerns the submission of the initially NE annotated SAGA – without any correction – to several supervised learning methods. The objective is to determine the most efficient supervised learning NER that is, the one that obtains high accuracy with a small size of training data. For this study, three supervised learning methods available in WEKA (Waikato Environment for Knowledge Analysis) [16] were selected, which are Naïve Bayes (NB), J48 Decision Trees, and SVM-SMO (Support Vector Machines – Sequential Minimal Optimization).

NB is a classification algorithm based on Bayes theorem. Its "naïve" qualification is due to the fact that it assumes independence between features. It means that the presence of a one feature in a class is unrelated to the presence of any other features. NB is easy to build and it is particularly useful for very large data sets.

J48 is an open source Java implementation in WEKA of the C4.5 decision tree algorithm. C4.5 algorithm builds a decision tree using the concepts of entropy and information gain as the criteria for splitting the dataset into smaller and smaller subsets, and at the same time creating incrementally a decision tree. Entropy is a numerical value that measures the uncertainty or impurity in the dataset. If the dataset is homogeneous, then the value of the entropy is zero. Information gain is the entropy of the parent node minus the entropy of the child nodes.

SMO has been proposed by John Platt to train SVM in a fast way [17], and thus SMO is an optimization algorithm to train SVM on a given dataset. SVM is a binary classifier. The dataset is viewed as a set of vectors that can be divided by a separating hyperplane into two distinct classes. SMO tries to optimize the two classes analytically in each iteration. If the dataset has more than two classes (the case of NE classification with four classes: DAT, LOC, ORG, and PER), then the classification is performed using pairwise classification, for example, DAT-LOC, DAT-ORG, etc. Real datasets are not always linearly separable. Therefore, SVM-SMO makes use of a kernel function to map the data into a higher dimensional space where a hyperplane can be used to do the separation. Doan and Xu [18] trained a SVM using polynomial kernel to recognize six medication related NEs. The performance of their NER system was evaluated based on 10-fold cross validation. As stated by the authors, "the SVM-based NER system achieved the best F-score of 90.05%

(93.20% Precision, 87.12% Recall), when semantic features generated from a rule-based system were included." This rule-based medication extraction system assigns medication specific categories into words. Our setting for the evaluation of the SVM-SMO is similar to Doan and Xu's setting, with a slightly better performance.

To find the most appropriate algorithm for the creation of an entity salience corpus, Dojchinovski et al. [7] tested five algorithms including our selected algorithms that are NB, C4.5, and SVM with polynomial kernel. The evaluation was done on Reuters-128, an English corpus containing 128 economic news articles. None of NB, C4.5, and SVM outperformed the Random Forest decision tree based classifier with F1 0.607. Just behind the Random Forest is C4.5 with F1 0.586, then SVM, and finally NB with the worst performance (F1 0.39).

IV. Experimental Results and Analysis

A. Experimental Settings

A subset of SAGA corresponding to the digitized form of the January 1904 edition has been selected for the experiment. This edition has already been given to an optical character recognition and then corrected manually. The plain text file result was then submitted to ANNIE for the recognition of NEs. From all annotated NEs, only four types were selected: DAT, LOC, ORG, and PER. Different numbers of NEs were evaluated and their distribution is shown in Table 2.

Table 2
Number of NE Instances

# Instances	DAT	LOC	ORG	PER
100	30	20	28	22
300	55	36	56	153
500	92	61	83	264
700	115	73	143	369
900	145	81	167	507
1000	165	88	183	564
1200	197	107	225	671
1400	237	140	304	719

Unlike many other studies, we did not investigate different features for the automatic recognition of NEs in SAGA. We just reutilize the features available directly without any processing from the annotation done by ANNIE. These features correspond to the ID of a string, the string, the POS tag of the string, its orthography (upper initial, lowercase, etc.), its kind (word, punctuation, etc.), its length, and its semantic class (DAT, LOC, ORG, or PER). These features may not sufficient or inappropriate but they provide better results while compared with other NER systems using more linguistic and contextual features. For example, to get 90.5 F-score with a SVM NER for the recognition of medication related NEs in hospital discharge summaries, Doan and Xu [18] had to use all the features that are the word, its POS, its orthography, its morphological information, its history (the semantic class of previous words), and semantic tag.

For the evaluation, we performed 10-fold cross validation, which divides the dataset into ten equal partitions. At each iteration, nine partitions are used for training and one for testing. The results are expressed by the metrics recall, precision, and F-measure.

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FN} \quad (2)$$

$$F - measure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (3)$$

where: FN = False Negative
FP = False Positive
TP = True Positive

a. Results of NB NER

The average performance of NB NER is shown in Figure 3, which indicates that the algorithm can reach its highest F-measure with 1200 instances.

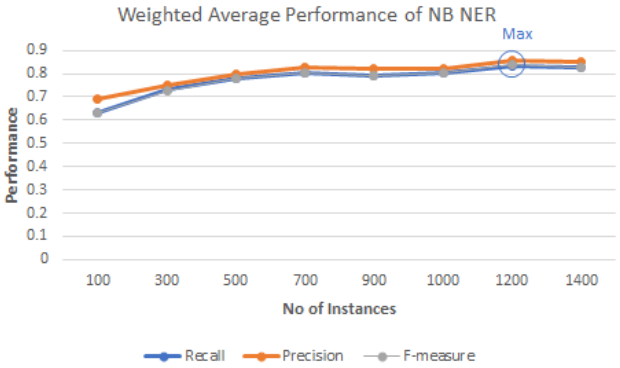

Figure 3: Graph of the weighted average performance of NB NER

With 1200 instances, the lowest F-measure value is with the recognition of LOC with only 67% (Table 4). This is due to the low recall value (only 54%), which corresponds to the actual positives predicted correct. The precision in recognizing ORG is very low (63%) compared to the other NEs.

Table 4
Performance of NB NER by Entity Class with 1200 instances

Class	Recall	Precision	F-measure
DAT	0.746	0.855	0.797
LOC	**0.542**	0.892	**0.674**
ORG	0.880	**0.635**	0.737
PER	0.894	0.922	0.908

b. Results of J48 NER

Like NB, J48 NER gets its best average performance with 1200 instances (Figure 4).

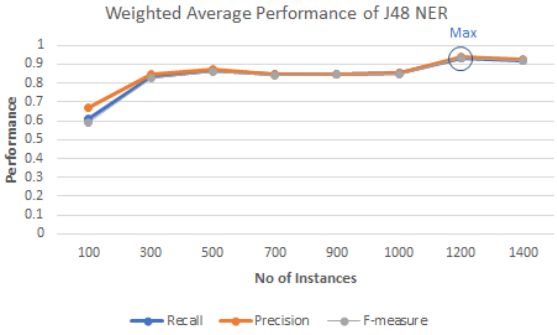

Figure 4: Graph of the weighted average performance of J48 NER

J48 gets also its lowest performance with the entity LOC due to the low value of the recall, which is 76% (Table 6). However, the recall for PER is near 100%, making this algorithm very attractive for the prediction of the entity PER in SAGA.

Table 6
Performance of J48 NER by Entity Class with 1200 instances

Class	Recall	Precision	F-measure
DAT	0.878	0.989	0.930
LOC	**0.766**	**0.911**	**0.832**
ORG	0.876	0.985	0.927
PER	0.999	0.912	0.953

c. Results of SMO NER

Like the two previous algorithms, SMO is reaching its highest performance with 1200 instances (Figure 5).

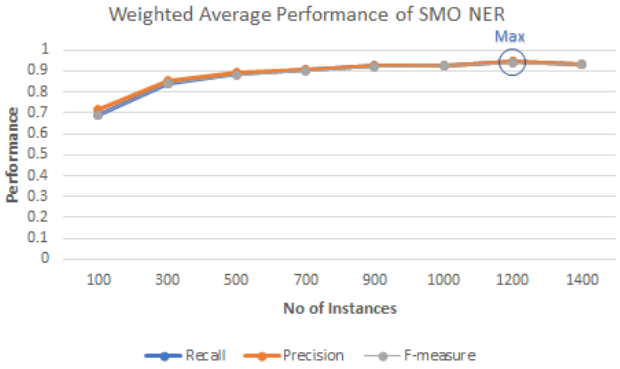

Figure 5: Graph of the weighted average performance of SMO NER

When inspecting the performance of SMO by entity class and with the 1200 instances, it seems that SMO has the same trend as J48. Both algorithms have the lowest performance with the entity LOC and get the highest performance with the entity PER. Like J48, the recall is low for LOC (75%) and it is high for PER (99%) (Table 8).

Table 8
Performance of SMO NER by Entity Class with 1200 instances

Class	Recall	Precision	F-measure
DAT	0.939	0.979	0.959
LOC	**0.748**	**0.899**	**0.816**
ORG	0.884	0.961	0.921
PER	0.994	0.933	0.962

d. Overall Results and Analysis

When running the three learning algorithms over different set number of instances, from 100 until 1400 instances, SMO outperforms NB and J48 (Table 9). These two algorithms start decreasing their recognition with 700 and 900 instances respectively, and then re-start increasing until decreasing again with 1400 instances. SMO keeps increasing its recognition until falling also with 1400 instances. The plotting of these behaviors is shown in Figure 6. From this observation, it appears that with only 1200 instances, SMO is the best algorithm for predicting NEs in SAGA.

Table 9
Correctly Classified Instances (%)

# Instances	NB	J48	SMO
100	63.00	61.00	69.00
300	73.00	83.33	84.33
500	77.80	86.60	88.80
700	80.43	**84.43**	90.42
900	**79.22**	84.89	92.44
1000	80.20	85.20	92.70
1200	83.58	93.50	94.25
1400	**83.00**	**92.21**	**93.14**

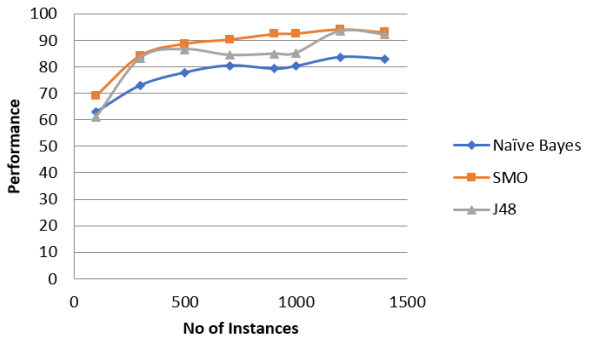

Figure 6: Graph of the performance of NB, J48, and SVM-SMO

Since NB did not show a good performance, it will be discarded in the following analyses. SVM-SMO and J48 made some errors. Table 10 shows a compiled confusion matrix of these errors, which occur more often between LOC-PER, ORG-PER, and DAT-PER.

Table 10
Number of Incorrectly Classified NEs with 1200 Instances – First Number for J48, Second Number for SVM-SMO; Number of Errors > 10 Times is in Bold

	DAT	LOC	ORG	PERS
DAT	173	0	2	**22**
	185	2	2	**8**
LOC	0	82	1	**24**
	2	80	4	**21**
ORG	1	8	197	**19**
	0	7	199	**19**
PERS	1	0	0	670
	2	0	2	667

SVM-SMO and J48 disagree simultaneously 53 times with the labels given by ANNIE. A thorough study of the disagreement allows us to state few ad hoc rules (Figure 7). The applications of the five first rules yield to the correction of the labels of 29 NEs. However, rule 6 corrects 16 out of 20 labels. For the other four labels, SVM-SMO and J48 predicted a correct label.

LABEL_CORRECTION(x:$label_x$, ANNIE:$label_{ANNIE}$, SMO:$label_{SMO}$, J48:$label_{J48}$)

If ($label_{SMO} \neq label_{ANNIE}$ & $label_{J48} \neq label_{ANNIE}$) // *SMO and J48 disagrees with ANNIE*

1. If $label_{ANNIE}$ = Date,
   then $label_x \leftarrow$ Date
2. If $label_{SMO}$ = Date & J48=Date,
   then $label_x \leftarrow$ Date
3. If $label_{SMO}$=Location & $label_{J48}$=Location & $label_{ANNIE}$=Organization,
   then $label_x \leftarrow$ Location
4. If $label_{SMO}$=Person & $label_{J48}$=Person & $label_{ANNIE}$=Organization,
   then $label_x \leftarrow$ Organization
5. If $label_{SMO}$=Organization & $label_{J48}$=Organization & $label_{ANNIE}$=Location,
   then $label_x \leftarrow$ Organization
6. If $label_{SMO}$=Person & $label_{J48}$=Person & $label_{ANNIE}$=Location,
   then $label_x \leftarrow$ Location

Figure 7: Label Correction Rules

V. CONCLUSION

This paper presented an automatic approach for annotating a large set of historical newspapers, in this case SAGA. A rule-based NER is used to pre-annotate a subset of SAGA. The annotated texts are then pre-processed to extract NEs and their features that are used by three NER classifiers (NB, J48, and SVM-SMO) as training and testing data. The evaluation results indicate that with 1200 instances of NEs, J48 and SVM-SMO can reach the F-measure of 93.50% and 94.25% respectively, whereas NB can only get 83.58%. From a rigorous analysis of the errors done by J48 and SVM-SMO, a set of ad hoc rules is able to correct more than one-third errors. The proposed approach in getting SAGA annotated with NEs with minimal human effort is promising. It makes use of out-of-the-box tools, and thus it is reproducible for any texts from various domains. In near future, we intend to improve the performance of the classifiers by continuing to develop more adequate rules.

REFERENCES

[1] S. Sekine, "Named Entity: History and Future (retrieved from http://www.cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf),".

[2] L. Ratinov and D. Roth, "Design Challenges and Misconceptions in Named Entity Recognition," *in Proc. 13th Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, Colorado, 2009, pp. 147-155.

[3] J. Wettlaufer, S. G. Thotempudi, (retrieved from http://www.gcdh.de/files/2013/6429/9184/Wettlaufer_Thotempudi_2013_NER_final.pdf)," Feb. 2017.

[4] C. Neudecker, "An Open Corpus for Named Entity Recognition in Historic Newspapers," in *Proc. 10th International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, 2016, pp. 4348-4352.

[5] P. V. Ogren, G. K. Savova, C. G. Chute, "Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition," in *Proc. 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, Brisbane, 2007 pp. 2325-2330.

[6] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in Twitter data with crowdsourcing," in *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California 2010, pp. 80–88.

[7] M. Dojchinovski, D. Reddy, T. Kliegr, T. Vitvar, H. Sack. "Crowdsourced Corpus with Entity Salience Annotations," in *Proc. 10th International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, 2016.

[8] C. Weber and R. Vieira, "Building a Corpus for Named Entity Recognition using Portuguese Wikipedia and DBpedia," in *International Conference on Computational Processing of Portuguese*, São Carlos, São Paulo, 2014,.

[9] L. Derczynski, K. Bontcheva, and I. Roberts, "Broad Twitter Corpus: A Diverse Named Entity Recognition Resource," in *Proc. of the Int'l Conference on Computational Linguistics (COLING)*, Osaka, Japan, 2016,.

[10] M. Ehrmann, M. Turchi, and R. Steinberger, "Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection," in *Proc. of Recent Advances in Natural Language Processing*, Hissar, Sept. 2011, pp. 118–124.

[11] Y. Tsuruoka, J.i Tsujii and S. Ananiadou, "Accelerating the Annotation of Sparse Named Entities by Dynamic Sentence Selection," in *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, Columbus, 2008, pp. 30–37.

[12] H. Kilicoglu, A. B. Abacha, Y. Mrabet, K. Roberts, L. Rodriguez, S. E. Shooshan, D. Demner-Fushman, "Annotating Named Entities in Consumer Health Questions," in *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC)*. Portorož, Slovenia, 2016.

[13] C. Grover, S. Givon , R. Tobin, and J. Ball. "Named entity recognition for digitised historical texts," in *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco, 2008, pp. 1343-1346.

[14] G. DeLozier, BenWing, J. Baldridge, S. Nesbit, "Creating a Novel Geolocation Corpus from Historical Texts," in *Proc. of LAW X – The 10th Linguistic Annotation Workshop*, Berlin, Germany, 2016, pp. 188–198.

[15] M. Kim, S. and Cassidy, S, "Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers," in *Proc. Australasian Language Technology Association Workshop, New South Wales*, 2015, pp. 57-65.

[16] G. Holmes; A. Donkin; I. H. Witten, "Weka: A machine learning workbench" in *Proc. Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994.

[17] J. C. Platt "Fast Training of Support Vector Machines using Sequential Minimal Optimization," *in Advances in Kernel Methods Support Vector Learning*, Cambridge 1998.

[18] S. Doan and H. Xu, "Recognizing Medication Related Entities in Hospital Discharge Summaries using Support Vector Machine," in *Coling 2010: Poster Volume*, Beijing, 2010, pp. 259–266.