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.

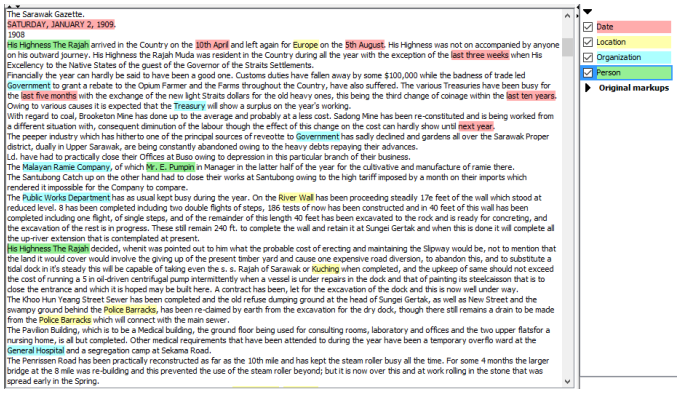

Figure 1: Example of NER

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