Sentiment Analysis of the Burmese Language using the Distributive Representation of n-gram-based Word

Myat Lay Phyu
Master of Science in Information Technology,
College of Computing, Prince of Songkla University,
Phuket Campus, Thailand.
Email: myatlayphyu096@gmail.com

Kiyota Hashimoto
Faculty of Technology and Environment,
Prince of Songkla University, Phuket Campus, Thailand.
Email: kiyota.h@phuket.psu.ac.th

***Abstract*—Less resourced languages are still difficult to treat due to the unavailability of annotated big corpora and basic natural language processing tools. This paper proposes a new method to use a character-based variable-length *n*-gram word model with distributive word representation techniques to reduce the number of *n*-gram words in data. We employed this method for news article sentiment analysis and achieved a better result than CRF-based ordinary word segmentation with a small size of supervised data. This enables to treat less resourced languages without focusing on language specific characteristics.**

I. INTRODUCTION

Exponential increase of available textual and speech data needs and enables advanced natural language processing techniques, but special attention is still necessary for less resourced languages with which big annotated corpora and basic tools like word segmentation tools are not available. In particular, languages whose writing concatenates words without spaces, such as Burmese, Lao, and Thai, have more difficulty because word segmentation as the initial step is not trivial at all. In this case, there are two possible approaches: to develop a word segmentation technique for a specific language with limited resources, and to develop a pseudo-word segmentation technique without paying much focus on language specific characteristics. In this paper, we propose a new method along the line of the second approach and compared both approaches for news article sentiment analysis as an example task.

Our proposal employs character-based variable-length *n*-gram words as words. By definition, all words, or pseudo-words, are regarded precisely as words, but the number of the words is much larger than the usual words contained in the same data. Thus we adopted character-based variable-length *n*-gram word estimation and distributive word representation models, word2vec [4] and GloVe [1] as word grouping methods to reduce the number of *n*-gram words in the data.

Sentiment analysis, or more specifically positive/negative value estimation of texts have been widely investigated. There are roughly two approaches: binary classification and cumulative calculation, though the combination of these is also possible. The first considers the task as a simple binary classification but as such, a supervised method. Thus a good amount of supervised data must be available, which is not the case in most low resourced languages. The second considers the task as a cumulative sentiment value calculation of all sentiment-valued words in each text on which final classification is based [2], but thus a sentiment dictionary must be constructed. Many methods for sentiment dictionary construction have been proposed, and we followed SO-LSA (Semantic Orientation-Latent Semantic Analysis) method. [7].

We conducted comparative experiments between an ordinary word segmentation, which is also proposed by the authors [5] using a character clustering method [6] and CRF [3] with a small size of annotated data, and our character-based variable-length *n*-gram word segmentation with and without word grouping before constructing a sentiment dictionary. The evaluation is based on the final sentiment value estimation of news articles. The result shows that our pseudo-word approach achieved almost the same result as an ordinary word approach, which enables to investigate less resourced languages without developing language specific tools and techniques.

The rest of this paper describes the research methodology in section 2 and our comparative experiments in section 3, finished by conclusion in section 4.

II. RESEARCH METHODOLOGY

The overview of our research methodology, as in Fig. 1, is sentiment value estimation of Burmese news articles based on a newly created sentiment dictionary with character-based variable-length *n*-gram words grouped with distributive word representation vector similarity.

Data consists of 100 news articles from the opinion section of *7Day Daily*, one of the most popular newspapers in Myanmar, and the sentiment value of each article is manually assigned based on a questionnaire survey. Among them, 52, 11 and 37 are positive, neutral and negative news respectively, but only positive and negative articles are used for classification.

Our character-based variable-length *n*-gram word recognition is conducted by first starting with 15-gram to choose very

978-616-455-375-0

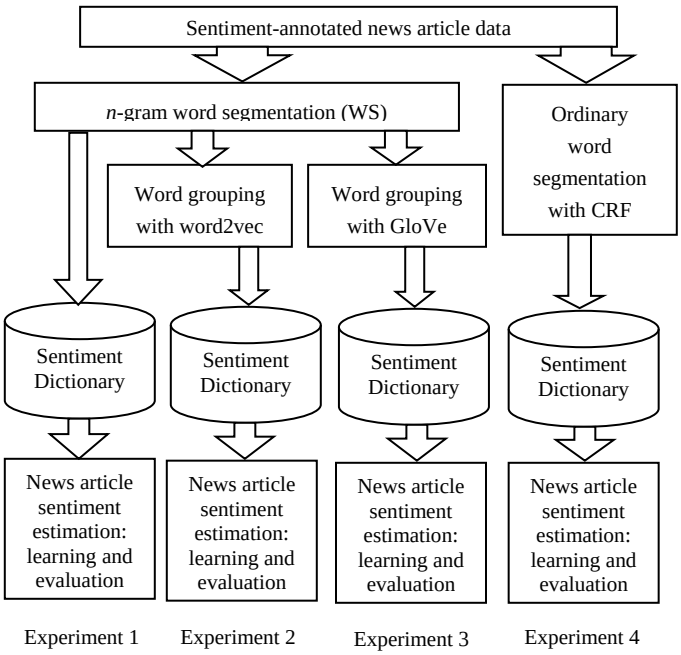

Fig. 1 Overview and experimental conditions

frequent 15-gram sequences as words, and iteratively repeating the same procedure to 3-gram (Experiment 1-3). In order to reduce the number of variable *n*-gram words, distributive word representation techniques, word2vec (Experiment 2) and GloVe (Experiment 3) are employed to make groupings of words based on cosine similarity. For these methods, we employed 1,280 unannotated articles (approximately 10 million characters) because these methods need larger sizes of data. CRF-based word estimation is also employed for comparison (Experiment 4). CRF is a supervised method, which needs word segmentation annotated data. Thus, we created a small manually annotated dataset for CRF learning.

Sentiment dictionaries are then constructed based on these four different word segmentations, using the SO-LSA method [7]. We selected 8 positive and 8 negative seed words and the calculation is based on the following equation:

$$\text{SO-LSA(word)} = \sum_{\text{pword} \in \text{Pwords}} \text{LSA (word, pword)} - \sum_{\text{nword} \in \text{Nwords}} \text{LSA (word, nword)}$$

Here, pwords is the set of positive seed words and nwords is the set of negative seed words. Note that an *n*-gram word containing a seed word but not containing a negative word is also regarded as the same seed word.

Finally, classification is performed for evaluation.

III. EVALUATION

A. Experiment

We conducted four comparative experiments to construct a sentiment dictionary and news article classification as shown in Fig.1. News article sentiment classification is conducted with 100 articles.

B. Result

The summary of the result is shown in Table 1. According to F-measure, the variable-length *n*-gram word achieved the highest result while variable-length *n*-gram words with GloVe-based word groupings achieved as high as that. According to accuracy, on the other hand, CRF-based ordinary word segmentation achieved the highest.

Table 1 Experimental results

	Accuracy	Precision	Recall	F-measure
Experiment 1	73.03%	66.7%	100%	80%
Experiment 2	65.2%	59%	96%	73%
Experiment 3	70.78%	67.57%	96.15%	79.34%
Experiment 4	76.4%	81.25%	75%	78%

C. Discussion

Experiment 2 performed worst. This seems to be caused mainly by our small data size. On the other hand, Experiment 3 achieved a much higher result in accuracy, precision, and F-measure. It seems that GloVe is more suitable than word2vec partly because GloVe is based on sentence structure while word2vec is based on simple word occurrence. However, our result is based on a small size of data, and thus evaluation with a larger data is waited.

IV. CONCLUSION

Our proposed method shows that variable-length *n*-gram word model with similarity-based grouping is a promising method for less resourced languages, because our method does not need a large annotated data or elaborately developed word segmentation tools. However, the results seem to be affected by our small data size. Thus the next step is to test our method with much bigger data, which is being pursued.

ACKNOWLEDGMENT

This work was also supported by the Higher Education Research Promotion and the Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of the Higher Education Commission.

REFERENCES

[1] J. Pennington, R. Socher and C. D. Manning, "GloVe: Global Vectors for Word Representations", EMNLP, pp. 1532–1543, 2014.

[2] M. Godsay, "The Process of Sentiment Analysis: A Study", International Journal of Computer Applications, 126(7), pp. 26-30, 2015.

[3] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Seqmenting and Labeling Sequence Data", Proc. Of 18th ICML, pp.282-289, 2001.

[4] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", ICLR Workshop, 12 pages, 2013.

[5] M. L. Phyu and K. Hashimoto, "Burmese Word Segmentation with Character Clustering and CRFs", Proc. of 14th JCSSE, pp. 1-6, 2017.

[6] T. Thanaruk, V. Sornlertlamvanich, T. Tanhermhong, and W. Chinnan, "Character Cluster Based Thai Information Retrieval", Proc. Of IRAL '00, pp. 75-80, 2000.

[7] P. D. Turney and M. L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association", ACM Tran,sactions on Information Systems, 21-4, pp. 315–346, 2003.