WeisoEvent: A Ming-Weiso Event Analytics Tool with Named Entity Markup and Spatial-Temporal Information Linking

Richard Tzong-Han Tsai
thtsai@csie.ncu.edu.tw
Center for GIS, Academia Sinica, Taiwan

Yu-Ting Lai
trulight@hotmail.com.tw
National Central University, Taiwan

Pi-Ling Pai
lingpai@gate.sinica.edu.tw
Center for GIS, Academia Sinica, Taiwan

Yu-Chun Wang
albyu35@gmail.com
Chunghwa Telecom, Taiwan

Sunny Hui-Ming Huang
shtilberg0623lg@gmail.com
Institute of History and Philology, Academia Sinica Taiwan

I-Chun Fan
mhfanbbc@ccvax.sinica.edu.tw
Center for GIS, Academia Sinica, Taiwan

Introduction

Weiso(衛所制), which means "guardhouse", is one of the military units of the barracks used by the Chinese dynasty Ming (1368-1644) to maintain peace throughout its empire. WeisoEvent is a web-based digital humanity research tool targeting Ming Weiso events recorded in Ming Shilu, which contains the imperial annals of the Ming emperors. WeisoEvent is composed of two parts: (1) an event type classifier that categorizes paragraphs according to their event types; (2) an analytics tool that shows (1)'s result, markups named entities, links guard mentions to Academia Sinica's Chinese Civilization in Time and Space (CCTS) spatial-temporal platform, and provides four visualization functions. Historians can use this tool to search for specific event types and gain insight into the relationship between particular guards and those event types, not only improving the efficiency but still maintaining the quality of research.

Event type classifier

Normally, one would develop a supervised-learning-based text categorization system to classify paragraphs into different event types. This involves defining a set of categories and annotating example texts for each category. However, lacking the human resources needed for such a task, we use unsupervised text clustering, which groups paragraphs into clusters by event type, to generate categories and their corresponding paragraphs for training an automatic event classifier. Although the results are not as accurate as those of pure supervised text classification, this hybrid approach is an acceptable tradeoff.

In clustering algorithms, each paragraph is represented as a vector. In previous studies, paragraphs have been represented using the vector space model (VSM), which represents each text as a feature vector of terms. However, this approach loses the ordering and ignores semantics. Yet another representation scheme inspired by word2vec is the "Paragraph Vector" proposed by (Le and Mikolov, 2014), an unsupervised framework that learns continuous distributed vectors for pieces of text. In their model, entire paragraphs are represented as vectors. The vector representation is trained to predict the words in a paragraph. More precisely, they concatenate the paragraph vector with several word vectors from a paragraph and predict the following word in the given context. Le's Paragraph Vector model has many advantages. First, it is mostly unsupervised and works well with sparsely labeled data. Second, it is suitable for text strings of various lengths, ranging from sentences to whole documents. Finally, it can overcome many weaknesses of the bag-of-words and bag-of-n-grams models. Because it does not suffer from data scarcity and high dimensionality, it also preserves the ordering and semantic information.

In summary, we propose a classification method which is based on clustering. First, we employ a named entity (NE) recognizer to label texts. Second, we train a paragraph vector model to represent paragraphs as vectors. Third, we cluster paragraphs with length <40

characters. Finally, we use the clustering results as gold-standard categories with which to train a support-vector-machines classifier to predict other paragraphs' categories.

We compare our method with the state-of-the-art paragraph clustering method using continuous vector space representation proposed by (M. Chinea-Rios et al., 2015). They use word2vec to learn word vectors and represent each sentence by summing the vectors of the words in that sentence. Like Chinea-Rios et al., we use the k-means algorithm to cluster vectors. We set the number of clusters to 68. We refer to the evaluation measures used in (Le and Mikolov, 2014). We generate sets of three paragraphs: two with the same event type and one with a different event type. Each set is referred to as a paragraph triplet. The distance between the two vectors with the same event type should be closer than the distance between either of these two and the unrelated one. We collect 923 paragraph triplets and compute the accuracy. Our best configuration that combines word dimensions and named entity dimensions to generate paragraph vectors achieves an accuracy of 62.49%, outperforming Chinea-Rios et al.'s pure text-clustering approach (M. Chinea-Rios et al., 2015) by 24.65%.

Analytics tool interface

WeisoEvent groups paragraphs with similar subjects into clusters automatically and each cluster is named manually according to the main subject of its paragraphs. Clusters with related topics are grouped into broader categories for search convenience. For example, we group "earthquake", "conflagration" and "hailstorm" into the event category "disaster". Users can modify event category titles by clicking a button on the top-right of the webpage (see Fig. 1, [1])

Figure 1. System interface.

Fig. 1 shows our research tool webpage, which consists of three main windows: (a) search parameters, (b) search result visualizations, and (c) search results snippets.

1. Search parameters: Users can search for one or more guards by typing the name into the search box (Fig. 1, [2]). For convenience, a user may also import a guard list by clicking the "import" icon (Fig. 1, [3]). Notably, if two guards are queried, their event timelines are displayed in parellel, like Jian-zhou (建州) and Wu-che (兀者) guards shown in window b. After at least one event category is selected, the search results are shown in windows b and c.
2. Search result visualization tools: In window b, users can select among four visualization options at the top of the frame. Option 1 hides or shows an event timeline of the search results on the page. When the event timeline is enabled, event-type labels corresponding to each retrieved paragraph are displayed chronologically on the timeline. For reference, the timeline shows the CE year at the bottom of the window (Fig. 2, [4]) and the name of era, which usually corresponds with the reigning Ming emperor, at the top (Fig. 2, [5]). When a user clicks on an event icon in the timeline, the corresponding text snippet is displayed in window c, highlighted in yellow (Fig. 2). Figure 3 takes Jian-zhou guard (建州衛) as an example to depict options 2 to 4. Option 2 is the bar chart. Each bar corresponds to a Chinese era name and represents the total number of paragraphs for the three selected event types in that era. Option 3 shows each bar sub-divided by color to show the distribution of paragraphs of each event type ("come over and pledge allegiance"/ "reward alien"/ "tribute-reward") in each era. By clicking Option 4, a pie chart shows the distribution of the three selected event types in the entire dataset. The slice for each event type is labeled with the number of paragraphs of that event type and its percentage of the total. These data visualizations offer historians a quick statistical overview of selected event types.

Figure 2. Event timeline

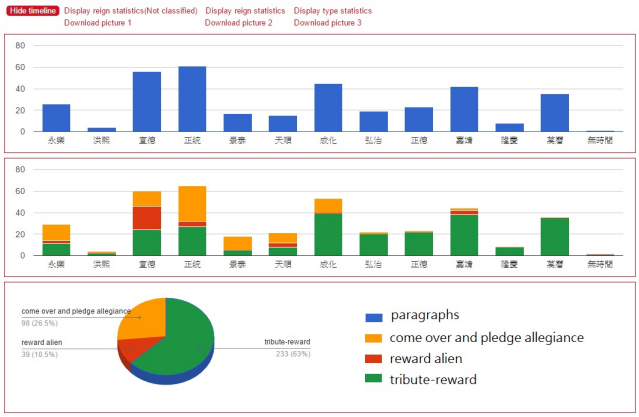

Figure 3. Data visualization options

3. Search snippets: Text snippets of paragraphs related to the searched event types are displayed in window c (Figure 1, (c)). The results are organized in a table with columns (L-R) showing time, number of paragraphs, event type, and related paragraph snippets. All guard mentions in the texts are highlighted and linked to Academia Sinica's CCTS-API Map Service . When a user clicks on a guard link in the text, the guard's location will be shown on a map of Ming China in a pop-up window, see Fig. 4. It shows the locations of Wu-che guard (兀者衛) and Jian-zhou guard (建州衛) in the same map.

Figure 4. Academia Sinica CCTS-API map service

Finally, we conduct a case study targeting Jurchens subordinated garrisons, including Wu-che guards (兀者諸衛) , Jian-zhou guards (建州諸衛), Mao-lian guard (毛憐衛) by using the proposed tool to obtain statistics regarding tribute event types.

We compare our event classification results with Cheng's study (N. Cheng, 2015), which used Ming Shilu as the source to investigate the tribute events during Yongle (永樂), Hongxi (洪熙), and Xuande (宣德) periods. We regard the paragraphs categorized as "tribute-reward", "come over and pledge allegiance", and "reward alien" event types as those potentially illustrating tribute events and manually check them. For Wu-che, Jian-zhou, and Mao-lian guards, 69, 86, and 40 paragraphs are identified as the above three types, respectively. Among these paragraphs, 66, 77, 37 are correct, which are close to the numbers of tribute events in Cheng's study for these three guards (60+, 70+, 30-). This study was done within 16 man-hours. These preliminary results are consistent with Cheng's manual analysis results and show that our tool not only helps historians study Weiso events more efficiently but also keep the quality.

Bibliography

Le, Q. V. and Mikolov, T. (2014). *Distributed Representations of Sentences and Documents.* Beijing, China, pp. 1188–96.

Chinea-Rios, M., Sanchis-Trilles, G., and Casacuberta, F. (2015). *Sentence clustering using continuous vector space representation.* Santiago de Compostela, Spain, pp. 432–40.

Cheng, N. (2015). "A Study of the Tributary System of Jurchens in the Ming Dynasty." *Journal of Chinese Humanities*, 347: 90-109+166-167.