**Figures**

**Figure 1** - Example of problematic text units from *Shenbao* (top) and *China Weekly Review* (bottom). Segments of interest are highlighted. For the sake of saving space, only parts of the articles are reproduced here.

**Figure 2 -** Distribution of documents mentioning the Rotary Club in *Shenbao* (top) and the ProQuest collection of Chinese newspapers (bottom).

**Figure 3** - A simple three-step method for creating short text units centered on the queried terms using concordance or Key Word In Context (KWIC): (1) Retrieve KWIC (2) Reconstruct the sequence (3) Re-unite the document. The above example is taken from an article published in *Shenbao* on 14 April 1947, which contains three instances of the key terms.

**Figure 4** – Summary of topics for each model in the Chinese corpus. Each plot shows the most frequent words for each topic and the relative proportion of the topics for each model.

**Figure 5** – Summary of topics for each model in the English corpus. Each plot shows the most frequent words for each topic and the relative proportion of the topics for each model.

**Figure 6** – Topic labels for each model in the Chinese corpus.

**Figure 7** – Topic labels for each model in the English corpus.

**Figure 8** - Distribution of topical categories. PQ = ProQuest, SB = *Shenbao*. 5T = 5-topic model, 10T = 10-topic model, 20T = 20-topic model.

**Figure 9 -** Plotting semantic coherence against exclusivity for the three models built from *Shenbao* (top) and ProQuest (bottom). The higher the number of topics, the lower the coherence, and the higher the exclusivity.

**Figure 10** - Most frequent words defining the 10-topic models in *Shenbao* (top) and *ProQuest* (bottom)

**Figure 11 -** Most representative documents for topics 1 and 3 in the Chinese-language corpus (10-topic model)

**Figure 12 –** Most representative documents for topics 1, 5, and 10 in the English-language corpus (10-topic model)

**Figure 13 -** Topic proportions over time (1919-1949) for the 10-topic model in Shenbao (top) and ProQuest (bottom). Topic proportions are aggregated by year.

**Tables (appendix)**

**Tables 1 to 6.** Summary of topics for each model, including their labels, the 10 most frequent words for each topic, and their various attributes (topical group, dimension, proportion, and trend over time).

**Tables 7 and 8.** Topics aligned across models

**Tables 9, 10, 11.** Topics aligned across languages.