Segmenting Messy Text: Detecting Boundaries in Text Derived from Historical Newspaper Images

Carol Anderson & Phil Crone, Ancestry.com

Background

- Goal: segment lists of marriage announcements from historical newspapers into units of one marriage each.
- Our text segmentation system forms part of a larger pipeline extract genealogical information from images of historical newspapers.
- Challenges:
  - Non-narrative structure to announcements
  - Topic similarity between adjacent announcements
  - Messy text produced by OCR software
  - Standard sentence splitting methods do not accurately detect announcement boundaries

Our Approach

- We use a supervised machine learning model to detect boundaries between announcements, rather than an unsupervised method based on topical similarity.
- This model incorporates spatial information about word positions.
- Announcement boundaries are made at the token level, rather than the sentence level.
- The model leverages a pre-trained ELMo model fine-tuned on an an in domain dataset.

![Figure 1](https://example.com/figure1.png)

Results

- We evaluate our model in two ways:
  - Pκ A standard metric used in text segmentation
  - “Task-based” evaluation Precision and recall for wedding-related entities based on whether they are included in the correct segment.
- We compare our model to the recent text segmentation model of Koshorek et al. (2017).

Conclusions

- Detecting boundaries at the token level is critical for successful segmentation.
- Fine-tuning a language model on in domain text gives significant increase in performance.
- Incorporating spatial features yields small improvements.
- Task-specific evaluation metrics can be more useful than generic metrics.