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


MARRIED.
BAKER—MARKELL.—On the 12th day of June, 1890, at the residence of the bride's father, JOSEPH D. BAKER, of Frederick City, Md., and Miss VIRGINIA H., second daughter of Charles Markell, Esq., of Baltimore. [No cards.]
COPPER—REHMERT.—On June 10, 1890, by Rev. B. F. Devries, HARRY T. COPPER to Miss ANNIE REHMERT, both of Highlandtown.
HUTCHINS—OWINGS.—On June 11, 1890, at No. 1031 North Gilmor street, by Rev. A. E. Bradenbaugh, HENRY S. HUTCHINS, of Woodbine, Md., and LILLIE OWINGS, of Baltimore, Md.

Figure 1. An example of an article in our dataset along with properly segmented text. Article from *The Baltimore Sun,* June 13, 1890 (p. 2), www.newspapers.com/clip/23188935.
.

MARRIED.

BAKER MARKELL. On the 12th day of June, 1MK). at the residence of the bride's father. JOSEPH D. IiAEEK, of Frederick Cltv. Md.. and Miss VIRGINIA H.. second daughter of Charles Markell.Esq., of Baltimore. Nocards.J

COPPER RF.HMERT. On June 10. MOO, bv R-v. B. F. Devrles. HARRY T. COPPER to Miss ANNIE REIIMKRT, both of Illghlandlown.

HUTCHIN9 OWINGS. On June 11, 1890, at No. 10.11 North Gilroor street. hy Rev. A. E. Rradcnbaugh, HENRY S. HUTCHIN8. of Woodbine, Md,, and 1.1.U1.1E. u w i or naiumore, atu.

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.

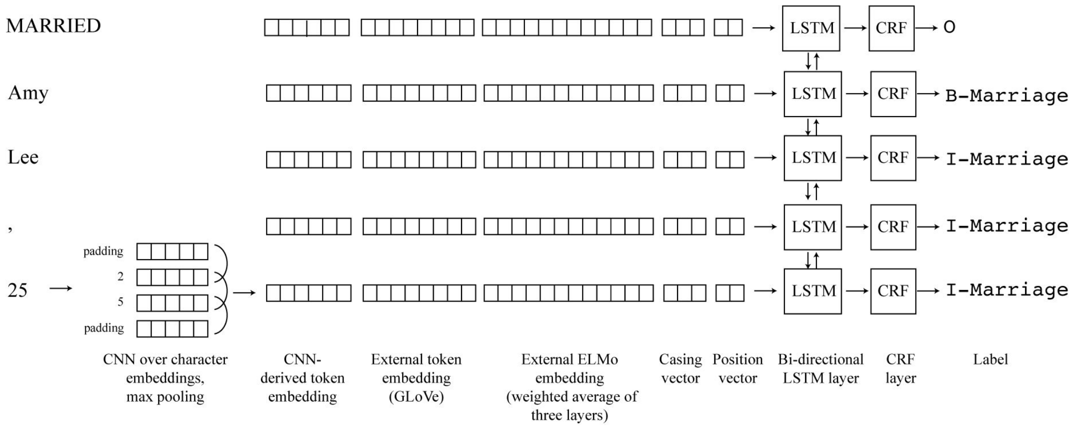

Figure 2. Illustration of our model architecture. Vectors are not shown to scale. The first token in this example, *MARRIED*, is not part of a specific marriage announcement and is therefore labeled O. The first segment begins with *Amy*.

Results

- We evaluate our model in two ways:
  - ***Pk*** 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).

Model	Features	Labels	P_k	Task-Based Evaluation		
				Precision	Recall	F1
Ours	All features	BIO	0.039 ±0.002	**96.9**	98.6	**97.7**±0.4
	ELMo not fine-tuned	BIO	0.049 ±0.007	93.0	98.1	95.5 ±0.7
	No ELMo	BIO	0.078 ±0.008	90.8	96.8	93.7 ±0.8
	No token coords	BIO	0.037 ±0.004	96.0	98.2	97.1 ±0.9
	No GloVe	BIO	0.039 ±0.002	96.0	98.6	97.3 ±0.4
Ours	All features	BI	0.031 ±0.004	95.5	99.0	97.2 ±1.2
	ELMo not fine-tuned	BI	0.050 ±0.006	91.5	98.6	94.9 ±0.7
	No ELMo	BI	0.072 ±0.010	92.2	97.2	94.6 ±1.9
	No token coords	BI	**0.029** ±0.003	94.9	**99.1**	97.0 ±1.1
	No GloVe	BI	0.033 ±0.002	95.9	99.0	97.4 ±0.5
Koshorek et al.		BI	0.266 ±0.004	20.0	96.0	33.0 ±0.2

Model	Entity Type	Precision	Recall	F1
Ours (BIO) With pos. vectors	Bride	97.8	98.9	98.4 ±0.2
	Groom	97.6	98.5	98.1 ±0.2
	BrideResidence	97.7	98.8	98.3 ±0.2
	GroomResidence	97.6	99.2	98.4 ±0.3
	WeddingDate	92.8	95.0	93.9 ±0.9
Ours (BIO) No pos. vectors	Bride	95.1	99.4	97.2 ±1.1
	Groom	95.4	99.05	97.1 ±1.1
	BrideResidence	97.2	98.9	99.1 ±0.3
	GroomResidence	97.4	99.4	98.4 ±0.3
	WeddingDate	67.5	93.0	77.2 ±10
Ours (BI) With pos. vectors	Bride	96.0	99.3	97.6 ±1.0
	Groom	95.9	98.8	97.3 ±1.2
	BrideResidence	96.9	98.9	97.9 ±0.5
	GroomResidence	97.1	99.3	98.1 ±0.7
	WeddingDate	76.3	93.4	84.0 ±6.7
Koshorek et al.	Bride	15.1	97.6	26.2 ±0.1
	Groom	15.1	95.2	26.0 ±0.1
	BrideResidence	28.8	93.3	44.0 ±0.1
	GroomResidence	29.7	96.7	45.4 ±0.1
	WeddingDate	34.3	94.3	50.3 ±1.1

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.

ancestry®