ancestry®

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

Carol Anderson and Phil Crone

Text segmentation

- Dividing a document into sections
  - News feed → articles
  - Character sequence → words

- Topic segmentation: dividing a document into sections about different topics

- Needed for:
  - Information extraction
  - Document summarization
  - Passage retrieval

Goal: **Segment newspaper marriage lists into individual announcements**

Example 1:

Original article:

March 17

Darren Giles Apiag II and Danielle Bryanna Kinsella Cabaccang, Antonio Paulino Chargualaf and Lori Jean Hong, Mark Joseph Santos Guevarra and Jennifer Ramos Marinas, James Chung Kim and Loralee Redor Sablan, Willy Nobuo and Maria Roduk Ngiratereged, Chi Yan So and Sze Wing Li, Chung Ho Tse and Chengrong Zhu

Output:

March 17

Darren Giles Apiag II and D anielle Bryanna Kinsella Cab accang,

---

Antonio Paulino Char- g ualaf and Lori Jean Hong,

M ark Joseph Santos Guevarra a nd Jennifer Ramos Marinas,

James Chung Kim and Loralee R edor Sablan,

Willy Nobuo and Maria Roduk Ngiratereged,

Chi Yan So and Sze Wing Li,

Chung H o Tse and Chengrong Zhu

Example 2:

Original article:

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.

Output:

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.

Optical character recognition (OCR) errors make the task difficult

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. R. 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.

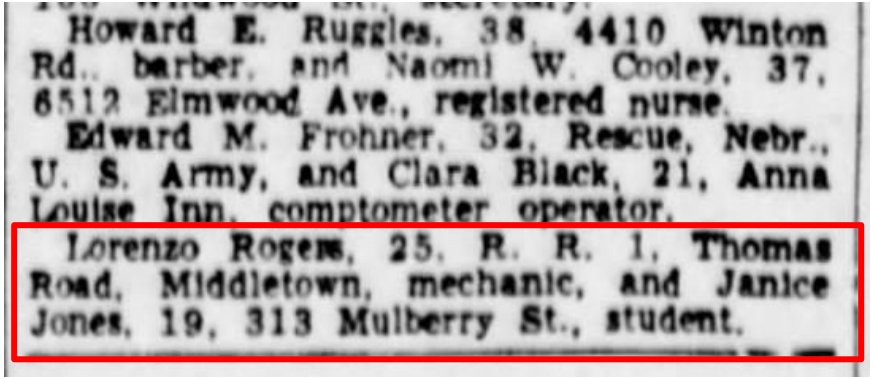
MARRIED.

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

Many text segmentation approaches rely on sentence splitting

...but sentence splitting performs poorly on newspaper marriage lists

Example list:



Howard E. Ruggles, 38, 4410 Winton Rd., barber, and Naomi W. Cooley, 37, 6512 Elmwood Ave., registered nurse.

Edward M. Frohner, 32, Rescue, Nebr., U. S. Army, and Clara Black, 21, Anna Louise Inn, comptometer operator.

Lorenzo Rogers, 25, R. R. 1, Thomas Road, Middletown, mechanic, and Janice Jones, 19, 313 Mulberry St., student.

Output of the NLTK Punkt sentence tokenizer:

Howard E. Rustles.

38 4410 Wlntoo Rd.. barber, and Naomi W Cooley, 17, 81) Kimwnod Ave, mistered mine Edward M. Frontier, 33.

Reacue, Nenr., V. S. Army, and Clara Black, 11, Anna Louise Inn, comptometer operator, lirento Roteas, 33.

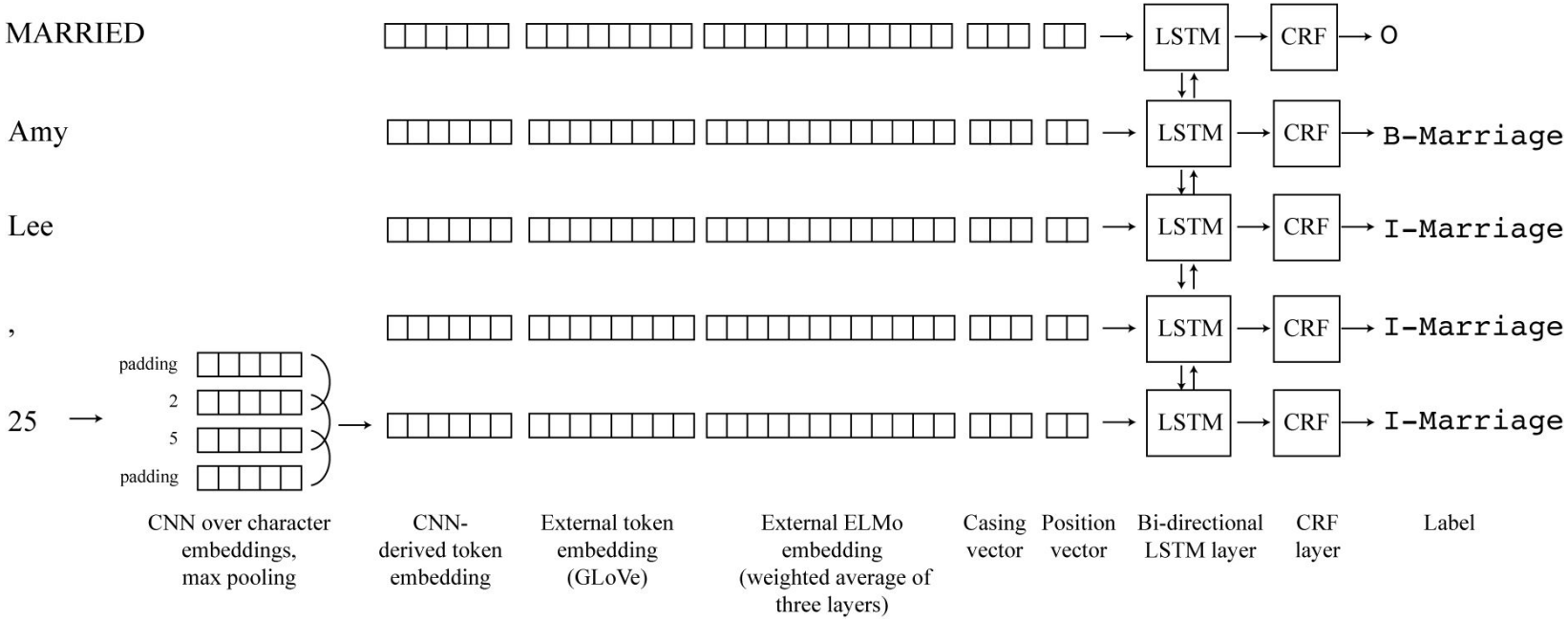
R. R. t, Thomas Road, Mlddletown.

mechanic, and Janice Jones.

10.

313 Mulberry St., student.

Model Architecture



Model Performance

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

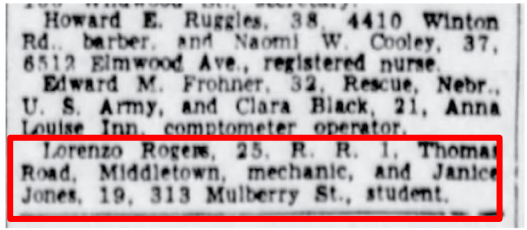
Scores shown are the average of three experiments; standard deviations are given for P_k and F1.
Lower P_k indicates higher segmentation accuracy.
See the paper for a description of the difference between BIO and BI labeling.

Model Performance

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

Scores shown are the average of three experiments; standard deviations are given for P_k and F1.
Lower P_k indicates higher segmentation accuracy.
See the paper for a description of the difference between BIO and BI labeling.

Token positions improve segmentation of some marriage lists



Howard E. Ruggles, 38, 4410 Winton Rd., barber, and Naomi W. Cooley, 37, 6512 Elmwood Ave., registered nurse.

Edward M. Frohner, 32, Rescue, Nebr., U. S. Army, and Clara Black, 21, Anna Louise Inn, comptometer operator.

Lorenzo Rogers, 25, R. R. 1, Thomas Road, Middletown, mechanic, and Janice Jones, 19, 313 Mulberry St., student.

A. Predictions without position vectors

Howard E. Rustles. 38 4410 Wlntoo Rd.. barber, and Naomi W Cooley, 17, 81) Kimwnod Ave, mistered mine

Edward M. Frontier, 33. Reacue, Nenr., V. S. Army, and Clara Black, 11, Anna Louise Inn, comptometer operator,

lirento Roteas, 33.

R. R. t,

Thomas Road, Mlddletown. mechanic, and Janice Jones. 10. 313 Mulberry St., student.

B. Predictions with position vectors

Howard E. Rustles. 38 4410 Wlntoo Rd.. barber, and Naomi W Cooley, 17, 81) Kimwnod Ave, mistered mine

Edward M. Frontier, 33. Reacue, Nenr., V. S. Army, and Clara Black, 11, Anna Louise Inn, comptometer operator,

lirento Roteas, 33. R. R. t, Thomas Road, Mlddletown. mechanic, and Janice Jones. 10. 313 Mulberry St., student.

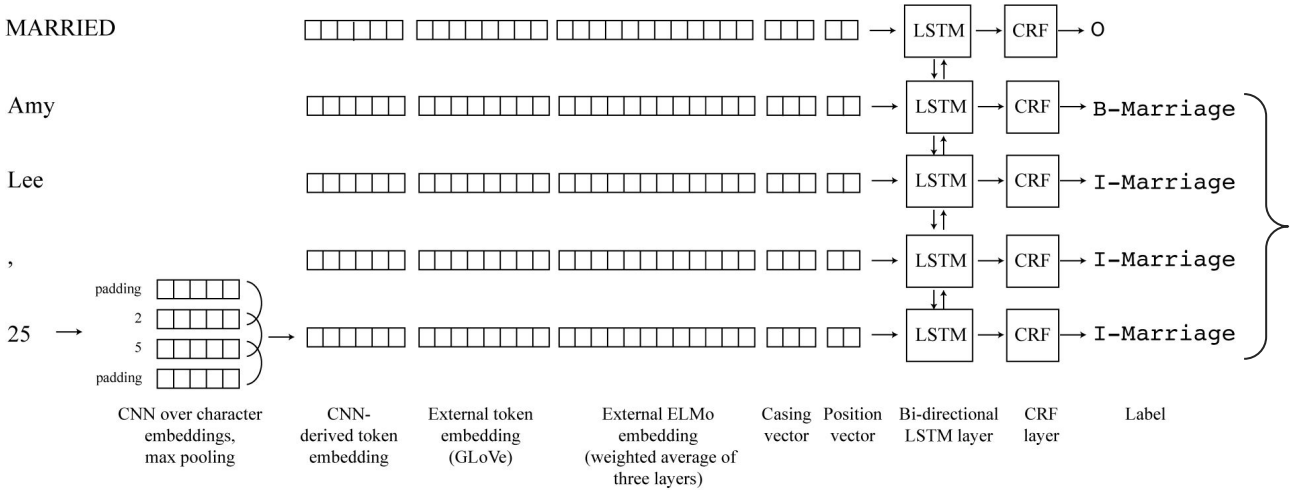
Example of a marriage announcement that was correctly segmented when token positions were used as a feature, but incorrectly segmented otherwise.

Detailed task-based evaluation of our model and the model of Koshorek et al.

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

Scores shown are the average of three experiments. Standard deviations are given for F1.

Potential extensions and applications



More labels could be used (e.g. B-Sports, I-Sports) to distinguish multiple topics

- Applications to other noisy text such as speech-to-text or handwriting recognition
- Applications to any text lacking sentences (invoices, song lyrics, etc.)