Handling Line Continuations

Seth Stewart
FamilySearch

Pedro Orte_ Encl
ga. M del m

Gavina Du_ E
ran. M de

Rafael Sala_ E
ran. M. del

...padrino Sebastian Torres soltero
...puso el nombre de Fortunato
...dicho padrino, y para que

...seis a los veinti tres dias
...Sagrario Rectoral
...marido de Bernardino
...Parroquia, fue su
...Maria feligreses de
...advertí su obliga
...que conste lo firmo

...los veinti cinco dias
...Sagrario Rectoral de
...marido de Manuel
...de esta Parroquia
...Clara Surita feligre
...bre Gavina, advertí
...y para que conste

...a los veinti cinco dias
...Sagrario Rectoral
...de Pedro Salazar

Sabino Balda_ M. | En el año del Señor de...
mes de Octubre, Yo...
Roque, Bautise...
de Mariano Bald...
Parroquia, fue su...
Parroquia, sele p...
y parentesco espiri...

Narciso Ramirez. M. | En el año del Señor de...
mes de Octubre, Yo...
que, Bautise a un...
Tecona Miranda mesti...
marido de Manuela Sa...
de Narciso, advertí su o...
y para q. conste lo f...

Narcisa Veintel. M. | En el año del Señor d...
Octubre Yo ...
Bautise a una infan...
Victoria mestiza solte...
Blacud soltera feligr...
tí su obligacion y para...
te lo firmo.

Josefa Dasa_ M. | En el año del Señor de m...
Octubre, Yo ...
Bautise...

Language Modeling

- Combining knowledge about which sequences are linguistically plausible together with direct feature information
- Given input features X and a linguistic probability distribution P, find the maximum likelihood sequence of symbols W*:

$$W^* = \arg\max_W \underbrace{p(X|W)}_{\textbf{Recognition model}} \; \underbrace{P(W)}_{\textbf{Language model}}$$

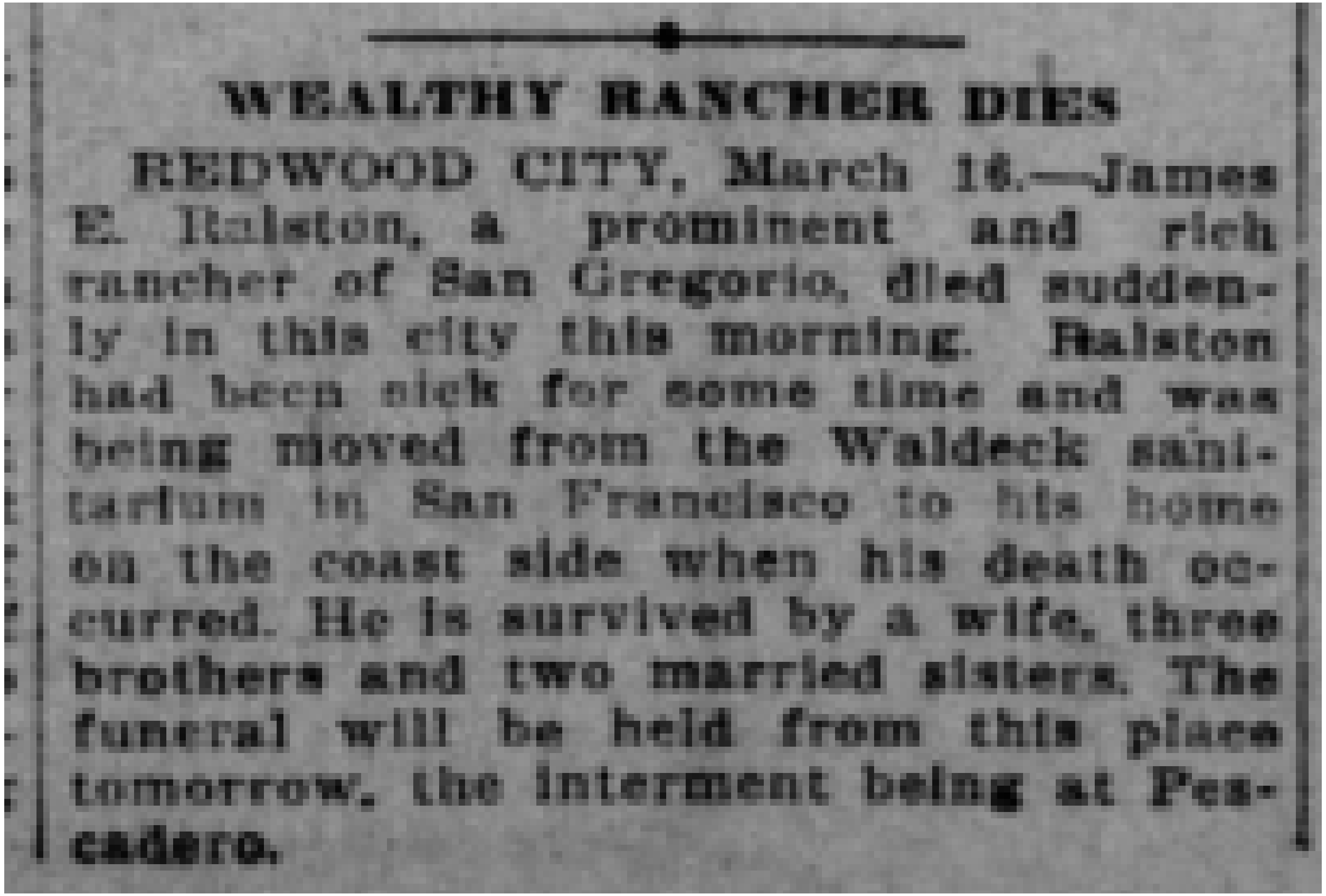
- Given an initial transcript, refine it using linguistic knowledge

Dataset: Historical newspaper images

American English

1730s-present

344 image crops, ~47.5k words (test set)



WEALTHY RANCHER DIES

REDWOOD CITY, March 16.—James E. Ralston, a prominent and rich rancher of San Gregorio, died suddenly in this city this morning. Ralston had been sick for some time and was being moved from the Waldeck sanitarium in San Francisco to his home on the coast side when his death occurred. He is survived by a wife, three brothers and two married sisters. The funeral will be held from this place tomorrow, the interment being at Pescadero.

Some important cases

Example	Description
Line continuations	Text tokens are intended to be distinct
Line continuations	Ditto above
Line-continuations	Hyphen forms a compound word consisting of multiple distinct words on the same line
Line continua- tions	A word is split across lines, joined by a hyphen
Line continua tions	A words is split across lines, with no hyphen indicator

Statistics

Across all word chunks in the training set:

- **73%** of chunks are "words" according to the dictionary
- **1.2%** are valid multiline words
- **1-6%** of multiline words are NOT hyphenated

(But maybe some of them should not be joined!)

thanksgiving, maybe, beheld, statehouse, druggist, without, detergents, anew, faraway, allover, backaches, percent, tractor, painkiller, schoolteachers, inbound, betaken, generally, eyestrain, cannot

These sometimes change the meaning, so join with caution!

- Some hyphen-joined multiline words may or may not consume the hyphen:

inquest--procure, fitz-william, fellowcountry-men, adjutant-general, re-occupation, seventy-six

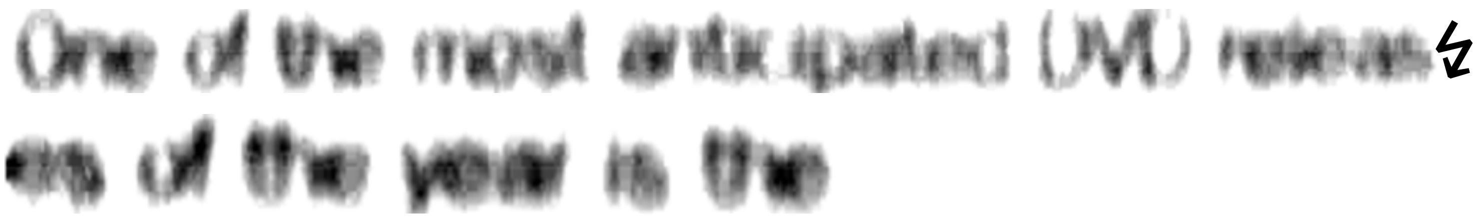
Method

Training

- Concatenate lines of text in training data (with newline marker ↯)
- Train new language model

Inference

- concatenate line images (or image features)
- inject newline character between line images



Initial Results

- 7-8% higher relative word error in initial experiments
- Shows potential for correction some multi-line words:

~~n~~how↯ever

Every Dollar Invested in this Com ↞-↯~~D~~pany will

wh~~o~~e↯never

Some other errors might be addressable through longer-range context

Initial Results

Some additional errors were introduced.

- Many line-ending punctuation marks disappeared:

I never called him any-↯thing

he was so restless ↞. ↯ About 2 o'clock

- Words at the beginning of a line were **un**-capitalized:

protection from the↯~~W~~wild Trapper of the Blue

Take 2: Model Blending

- Idea: Use the prevalence of errors to mix and match line continuations model with the original model.
- E.g., Don't preserve space deletions from the second model relative to the first model.
- Result: Better than the first line continuations model, but still **2% relative error increase**.
- Conclusion: Edit types are not sufficiently discriminative to improve the resulting transcript over the baseline

D	<space>	0.14771
D	-	0.079432
D	↞	0.068954
I	<space>	0.047321
D	.	0.043265
I	↞	0.024844
D	e	0.021126
D	s	0.019098
D	t	0.017745
D	n	0.017069

Take 3: Data augmentation

- Take ordinary text lines in the training set
- Fuse lines using dictionary approach to detect multiline words that should be joined
- Inject hyphens and newlines into new random mid-word positions
- Result: Same performance as first LC model (+8% WER). Slightly worse blending performance (+2% WER).
- This has the unfortunate side effect of bolstering the representation of nearly all of the original sequences in the training set
- Using standard discounting & smoothing models, this will degrade our performance on rare strings

Take 3: Data augmentation

Continua-
tions
Continu-
ations
Con-
tinuations
Conti-
nuations
…

Alternative Approaches

Improve the context or conditioning by:

- Directly augmenting the finite state decoding graph
- Recurrent Neural Networks (LSTM, GRU, etc.)
- Transformer Networks

- Unclear how to integrate into framework – open research problem

- Bonus: How to tackle the curse of dimensionality for sequential data?

Thank you!

To be contin-
ued…