Flexible Computer Assisted Transcription of Historical Documents Through Subword Spotting

Brian Davis, Robert Clawson and William Barrett

BRIGHAM YOUNG UNIVERSITY
Family History
TECHNOLOGY WORKSHOP

What if…?

Effective crowdsourced transcription of documents via

- Smartphone users
- Only a few minutes at a time

What if…?

Effective crowdsourced transcription of documents via

- Smartphone users
- Only a few minutes at a time

What if…?

Effective crowdsourced transcription of documents via

- Smartphone users
- Only a few minutes at a time

What if…?

Effective crowdsourced transcription of documents via

- Smartphone users
- Only a few minutes at a time

Computer Assisted Transcription (CAT)

Why not do it all manually?

Why not do it automatically?

Prefix Based CAT

User makes correction to automatic transcription, approving all previous content. Recognition algorithm makes new prediction for remaining text.

Requires sequential text.

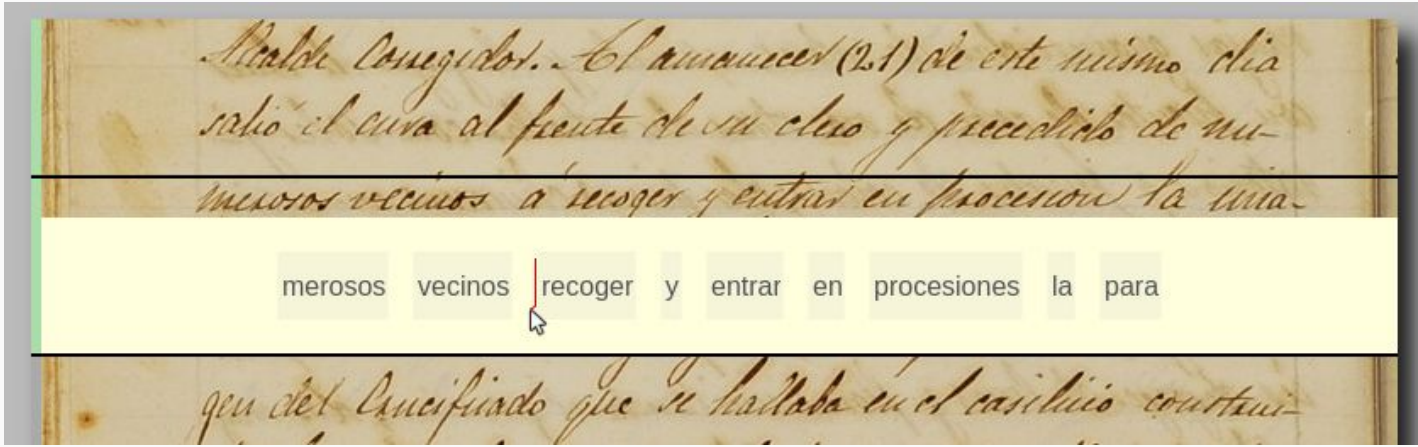


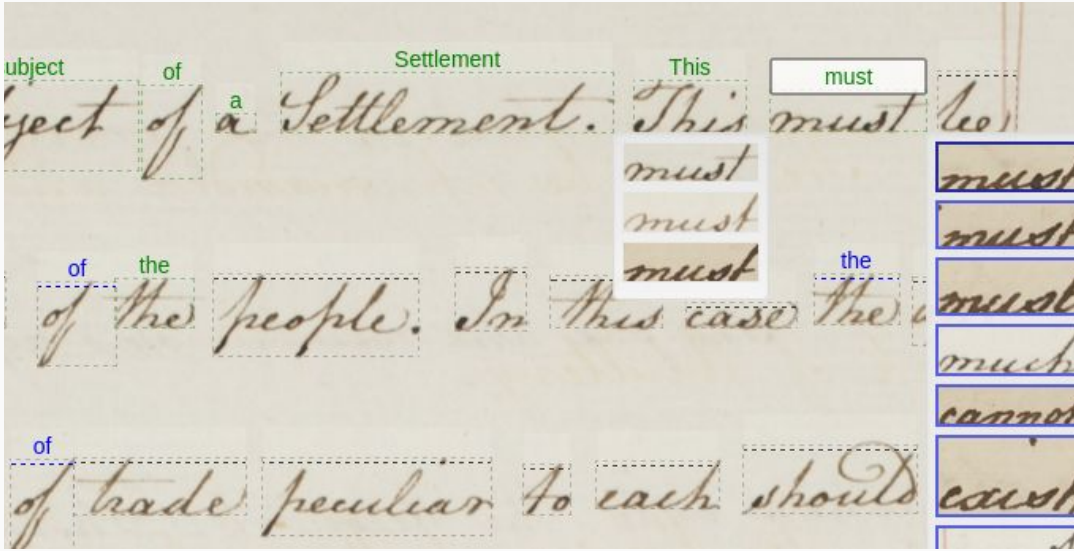
Image of Toselli et al's online demo

A. Toselli, V. Romero, M. Pastor, , and E. Vidal, "Multimodal interactive transcription of text images," Pattern Recognition, vol. 43, no. 5, pp. 1814–1825, 2010.
N. Serrano, A. Gimenez, J. Civera, A. Sanchis, and A. Juan, "Interactive handwriting recognition with limited user effort,"IJDAR, vol. 17, no. 1, pp. 47–59, 2014.

CAT Through Word Spotting

Find words that look the same and label them the same.

Zagoris et al (2015) use a relevance feedback loop to learn from every correct match the user selects.

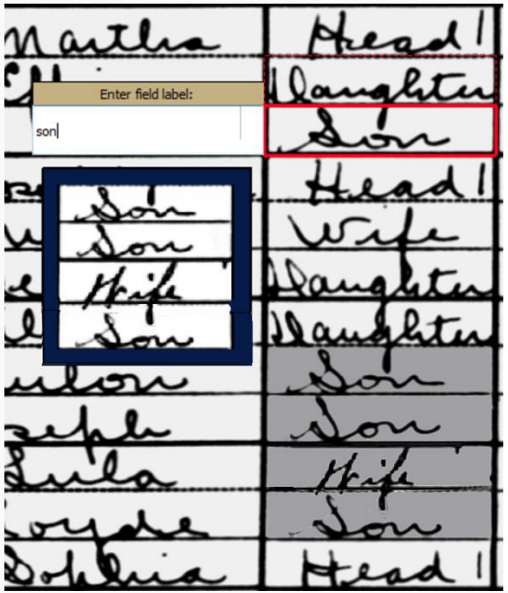


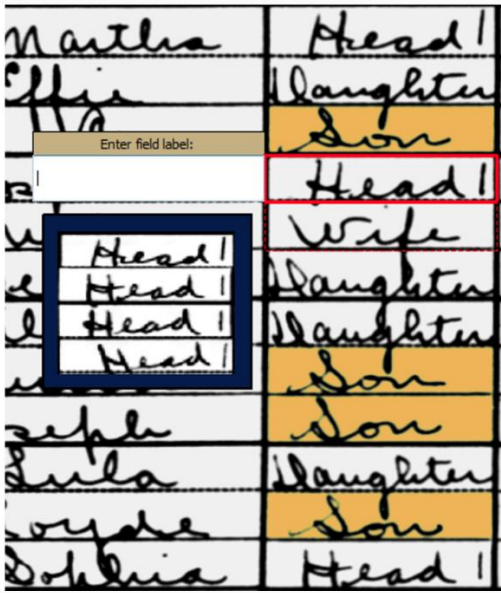
K. Zagoris, I. Pratikakis, and B. Gatos, “A framework for efficient transcription of historical documents using keyword spotting,” in Proc. HIP. ACM, 2015.

CAT Through Word Spotting

Find words that look the same and label them the same.

Robert Clawson's Intelligent Indexing (2014) relies on user filtering of matches.

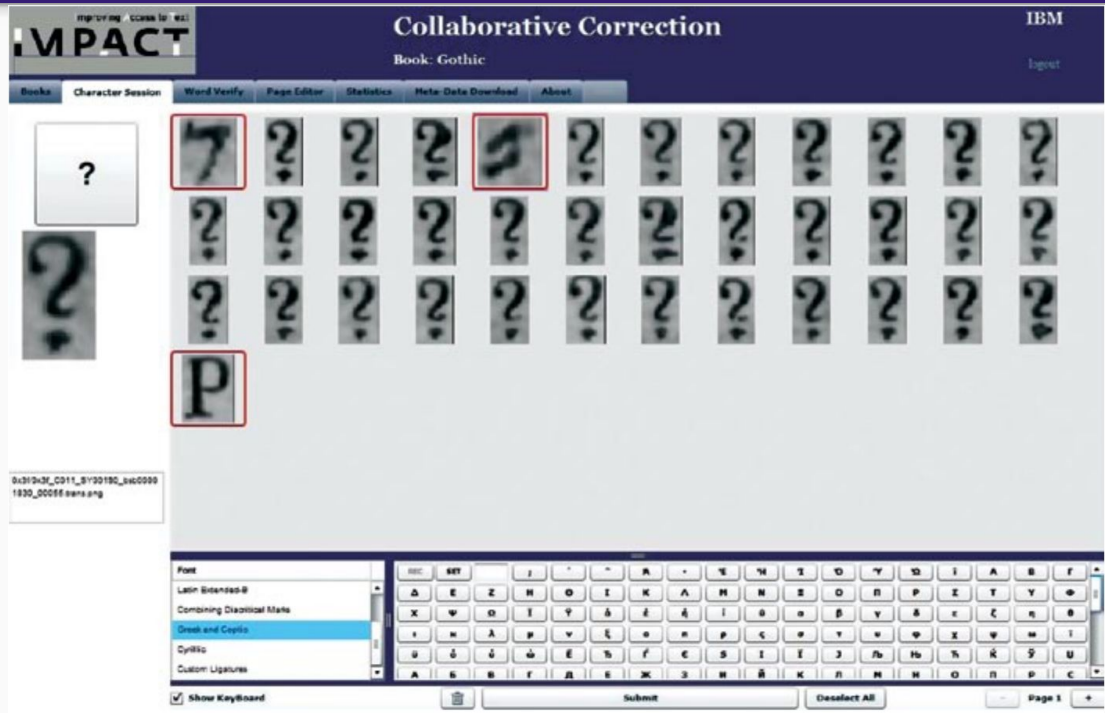




R. Clawson, "Intelligent indexing: A semi-automated, trainable system for field labeling," Master's thesis, Brigham Young University, 2014. [Online]. Available: scholarsarchive.byu.edu/etd/5307/

CAT Through User Supervised OCR

Neudecker and Tzadok (2010) OCR, then present characters with low score to user to clean.



C. Neudecker and A. Tzadok, “User collaboration for improving access to historical texts,” Liber Quarterly, vol. 20, no. 1, p. 119-128, 2010.

Strengths of Prior CAT Systems

OCR & word spotting:

- As long as words/letters can be segmented, will work with any document

OCR:

- Simple user tasks (no typing, very fast)
- Very parallelizable

Word spotting:

- Potential high payoff for little user effort (few taps, many words transcribed)

Weaknesses of Prior CAT Systems

Prefix based:

- Only works on sentence structured writing.
- Limited lexicon size (e.g. hard time with names).

Word spotting:

- Often words don’t repeat frequently or at all (e.g. names).

OCR:

- Letter segmentation improbable for handwritten text.

A Solution

Solution:

Spot character n-grams (bigrams and trigrams).
Reconstruct words from them.

The “Sweet Spot”

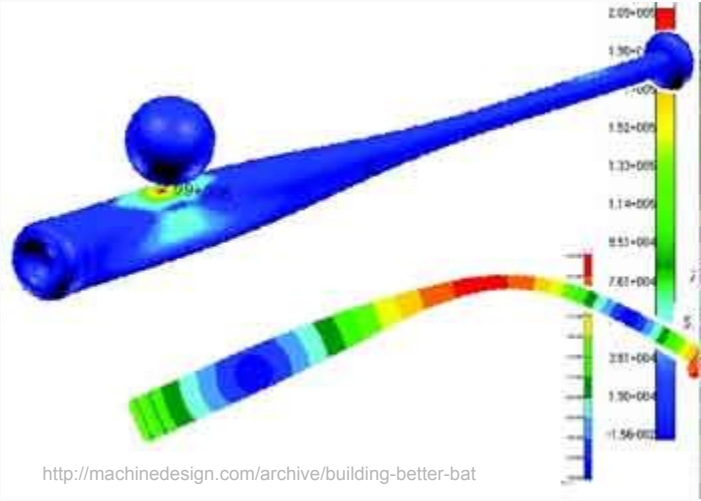
Bigrams/trigrams occur with great frequency

+

Subword spotting still reasonably accurate

=

High pay-off for spotting effort



http://machinedesign.com/archive/building-better-bat

Additionally, able to use larger lexicon, including more names.

N-gram Spotting and Word Completion

_ _ c h a e l

N-gram Spotting and Word Completion

M i c h a e l

N-gram Spotting and Word Completion

_ _ _ h o _ _

N-gram Spotting and Word Completion

A n _ h o _ _

N-gram Spotting and Word Completion

A n t h o _ _

N-gram Spotting and Word Completion

A n t h o n y

Computers are much better at this than we are!

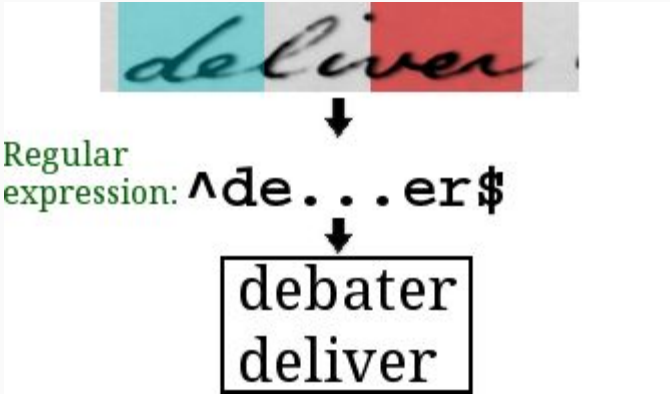
A n _ h o _ _ => [anchors, anchovy, anthony, anthoni]

N-gram Spotting and Word Completion

Regular expression make this easy.

Spotted n-grams are parsed into a regular expression.

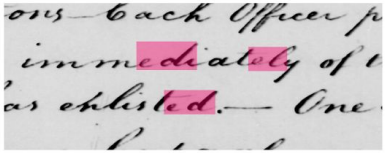
The regular expression is used as a lookup on the lexicon.



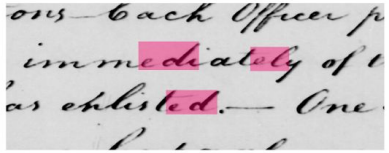
Overview of Proposed CAT System

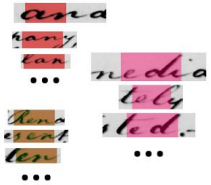
n-gram exemplar knoweldge

n-gram exemplar knoweldge

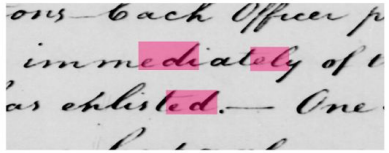


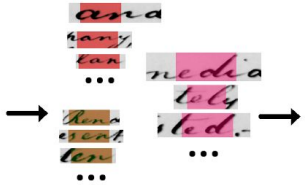
n-gram exemplar knoweldge



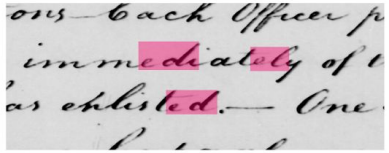


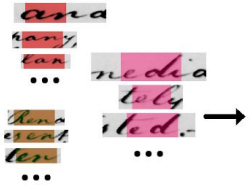
n-gram exemplar knoweldge





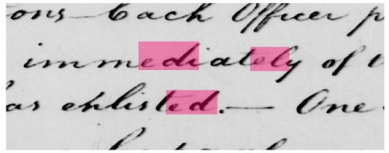
n-gram exemplar knoweldge

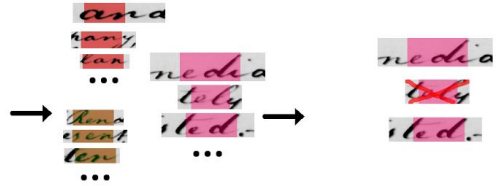




^en....ed$

n-gram exemplar knoweldge

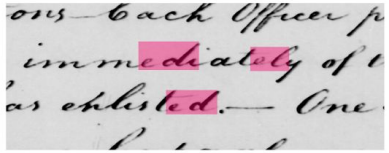


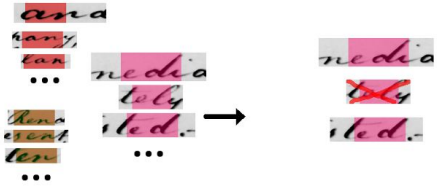


^en....ed$

encamped
engorged
engraved
engulfed
enlarged
enlisted
enhanced
enriched
enrolled

n-gram exemplar knoweldge

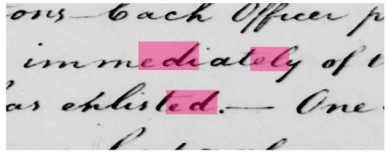


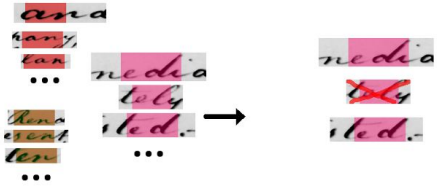


^en....ed$

encamped	0.41
engorged	0.14
engraved	0.12
engulfed	0.21
enlarged	0.69
enlisted	0.67
enhanced	0.59
enriched	0.45
enrolled	0.46

n-gram exemplar knoweldge



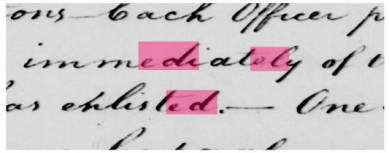


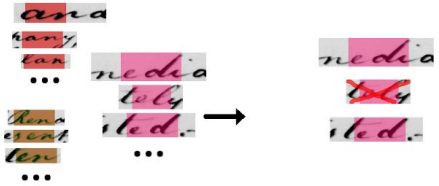
^en....ed$

encamped	0.41
engorged	0.14
engraved	0.12
engulfed	0.21
enlarged	0.69
enlisted	0.67
enhanced	0.59
enriched	0.45
enrolled	0.46

encamped
enlarged
enlisted
enhanced
enriched
enrolled

n-gram exemplar knoweldge



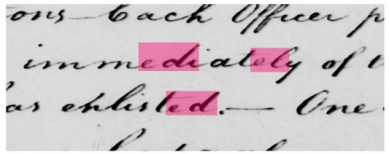


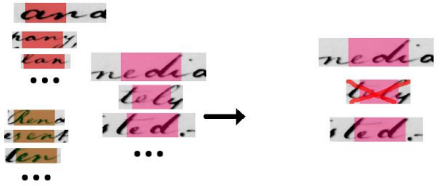
^en....ed$

encamped	0.41
engorged	0.14
engraved	0.12
engulfed	0.21
enlarged	0.69
enlisted	0.67
enhanced	0.59
enriched	0.45
enrolled	0.46

encamped
enlarged
enlisted
enhanced
enriched
enrolled

n-gram exemplar knoweldge



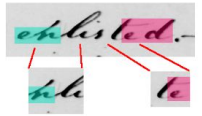


^en....ed$

encamped	0.41
engorged	0.14
engraved	0.12
engulfed	0.21
enlarged	0.69
enlisted	0.67
enhanced	0.59
enriched	0.45
enrolled	0.46

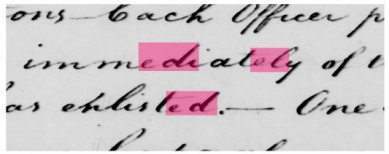
encamped
enlarged
enlisted
enhanced
enriched
enrolled

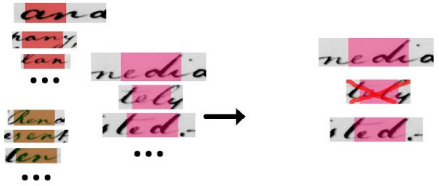
enlisted



"nl" "te"

n-gram exemplar knoweldge



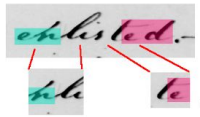


^en....ed$

encamped	0.41
engorged	0.14
engraved	0.12
engulfed	0.21
enlarged	0.69
enlisted	0.67
enhanced	0.59
enriched	0.45
enrolled	0.46

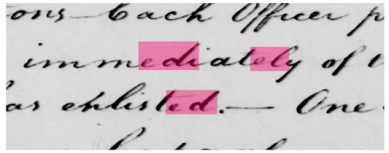
encamped
enlarged
enlisted
enhanced
enriched
enrolled

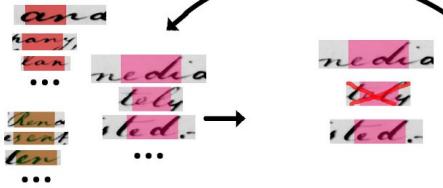
enlisted



"nl" "te"

n-gram exemplar knoweldge



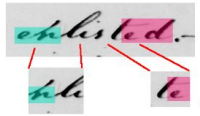


^en....ed$

encamped	0.41
engorged	0.14
engraved	0.12
engulfed	0.21
enlarged	0.69
enlisted	0.67
enhanced	0.59
enriched	0.45
enrolled	0.46

encamped
enlarged
enlisted
enhanced
enriched
enrolled

enlisted

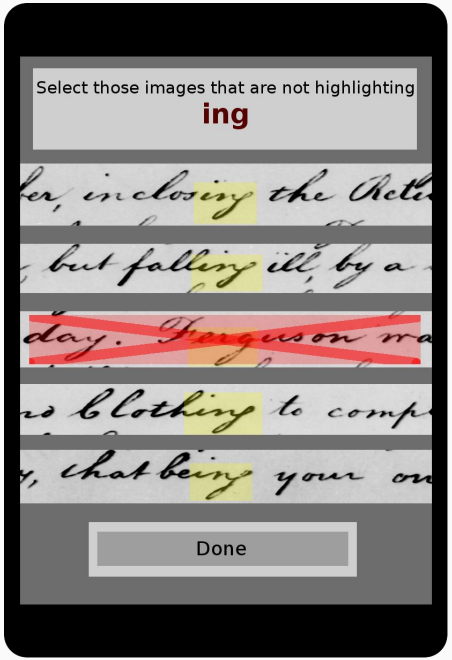


"nl" "te"

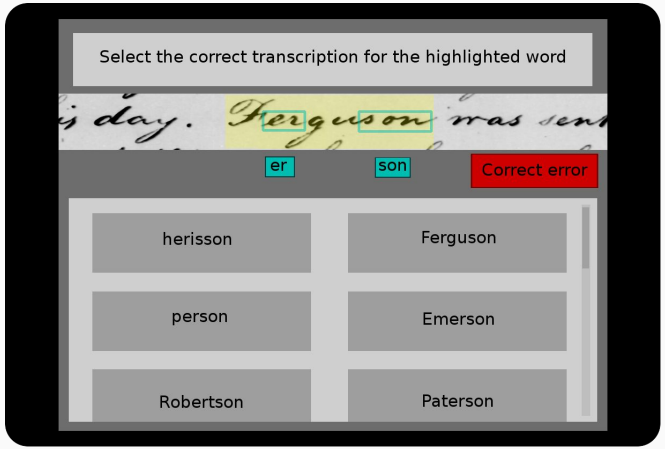
Overview of Proposed CAT System

Complicated system, simple UI

Mock-up of User Tasks



Mock-up of User Tasks



Justification:
Simulation of Proposed CAT System

George Washington corpus

100 most common bigrams

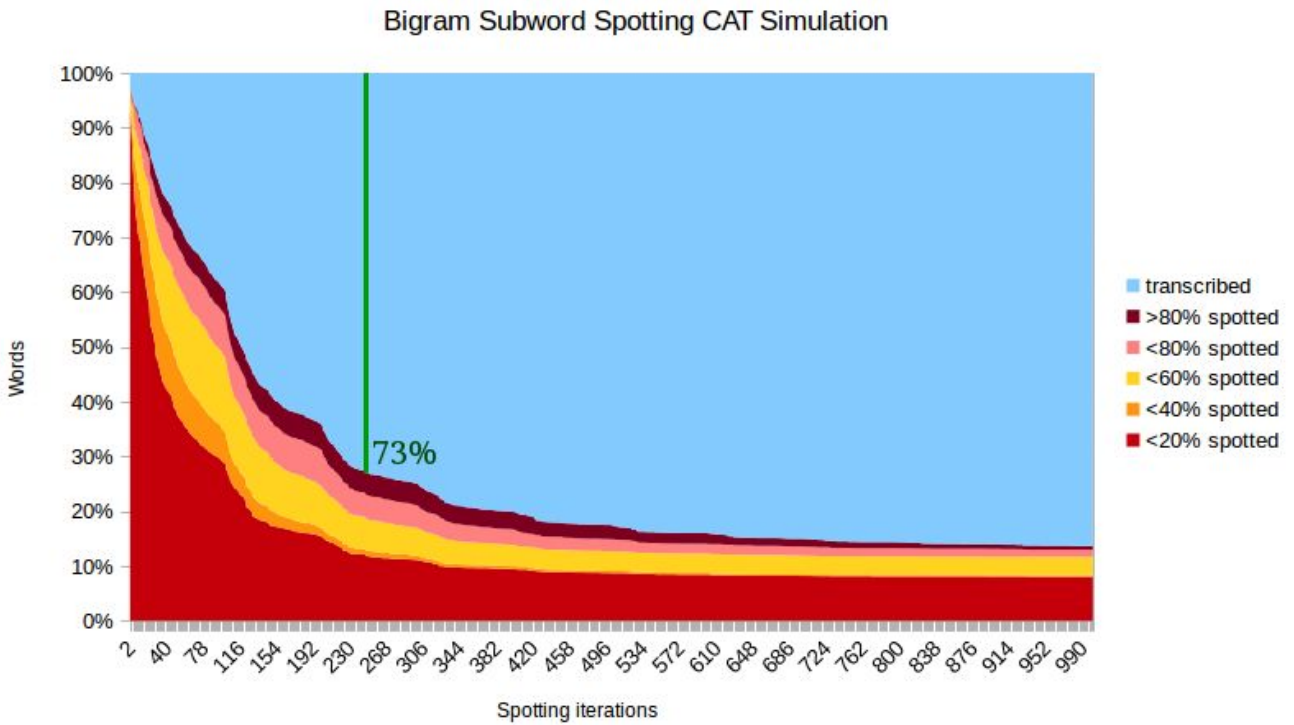
simulated 50% recall* for bigram spotting

simulated uncertain number of characters not spotted in word

word was “transcribed” when 10 or less possible transcriptions remain

lexicon of ~108,000 words and ~7,000 names

*Based on preliminary results in subword spotting.



Possible Bonuses

N-gram spotting verification may be reasonably completed by non-native speakers of a language.

Small user tasks may be easy to gamify.

Questions?

Limitations and Weaknesses

Dependent on word segmentation.

May require manual transcription for first few pages of a corpus as training.

Requires manual transcription to “finish” out-of-vocabulary, malformed and infrequent unfavorable words.

Poor spotting will burden human users with too much rejecting (or low recall).

If recognition/spotting scoring of word images does not prune effectively, the feasible lexicon size may be limited.

Subword N-gram Spotting

Preliminary results show 64% mAP for bigrams and 72% mAP for trigrams on George Washington dataset.*

Better results should come with a specialized method.

*using adaption of J. Almazan, A. Gordo, A. Fornes, and E. Valveny, “Word spotting and recognition with embedded attributes,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 12, pp. 2552–2566, 2014.