Sentence Segmentation in Clinical Notes

Siddharth Rawal[1], Davy Weissenbacher[2], Graciela Gonzalez-Hernandez[2]

1. Department of Computer and Information Science, School of Engineering and Applied Sciences, University of Pennsylvania
2. Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania

Motivation

- Definition: sentence segmentation is the task of finding the boundaries of sentences in texts
- Most **sentence segmentation** systems are developed for well-formed, proofread text, causing a **drop in performance** when applied on **clinical notes** (differences in structural and writing conventions)

Methods

- Dataset of 90 clinical notes from MIMIC-III database manually-annotated with sentence spans
- State of the art: deep-learning architecture with character and word embeddings (CNN-BiLSTM)
- Evaluate existing systems against annotations
- Character-level predictions using BIO schema
- Metrics: exact matches of sentence spans measured by precision, recall, F1-score

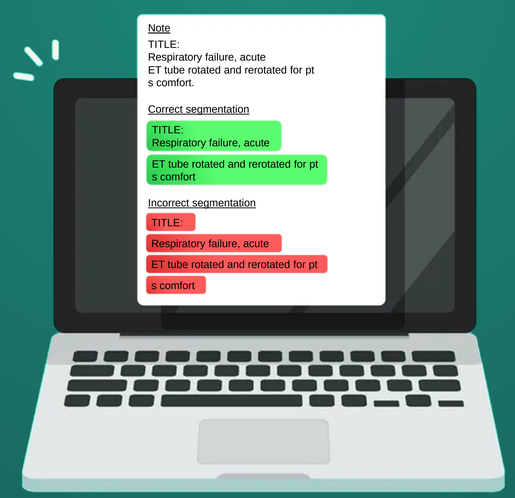
```
Text:  Physical Exam: obese gentleman, walks with cane. Pertinent Results:
Tags:  BIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIOBIIIIIIIIIIIIIIIIIII
```

```
Text:  Prophylaxis: DVT: Heparin gtt Stress ulcer: PPI VAP: Chlorhexidine
Tags:  BIIIIIIIIIIIIOBIIIIIBIIIIIIIIIIIIOBIIIIIIIIIIIIIIIIIIIIOBIIIIIIIIIIIIIIIIIII
```

Results

Sentence Segmenter	Precision	Recall	F1-score
NLTK[1]	**0.6530**	0.3192	0.4288
spaCy[2]	0.0399	0.0917	0.0556
Syntok	0.6163	0.3207	0.4218
Gensim[3]	0.3232	0.6170	0.4242
Stanza[4]	0.6418	0.3422	0.4464
cTAKES[5]	0.6055	**0.6609**	**0.6320**

Decline in performance of sentence segmentation in **clinical notes**



Examples

Input

Pt. medicated w/ 1 mg Ativan for anxiety.

Correct

Pt. medicated w/ 1 mg Ativan for anxiety. SENT

Incorrect

Pt. SENT

medicated w/ 1 mg Ativan for anxiety. SENT

Error: periods after medical abbreviations, not common in well-formed English

Input

1. Evaluation of etiology of arrest
2. Neuroprotective strategies:
a) Therapeutic hypothermia
b) Avoid hyperglycemia

Correct

1. Evaluation of etiology of arrest SENT
2. Neuroprotective strategies: SENT
a) Therapeutic hypothermia SENT
b) Avoid hyperglycemia SENT

Incorrect

1. SENT
Evaluation of etiology of arrest 2. SENT
Neuroprotective strategies: a) SENT
Therapeutic hypothermia b) Avoid hyperglycemia SENT

Error: bullets and numbers separated from list items

Input

Non-invasive BP:
S: 109
D: 72
Temperature:
97.7

Correct

Non-invasive BP: SENT
S: 109 SENT
D: 72 SENT
Temperature: 97.7 SENT

Incorrect

Non-invasive BP: S: 109 D: 72 Temperature: SENT
97.7 SENT

Error: forms to fill by the practitioner have no clear boundaries

Ongoing work

- Adjusting predictions to follow our guidelines by post-processing sentence spans
- Extending metrics to penalize truncating sentences in ways that are harmful for downstream tasks
- Adding functionality to perform section segmentation
- Proposing an approach using transformer model with sliding window

Acknowledgments

This work is supported by the National Institutes of Health (NIH) Research Advancement for Diversity Scholars (RADS) project affiliated with the Department of Biostatistics, Epidemiology and Informatics at the University of Pennsylvania.

References

NLTK: 1. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

spaCy: 2. Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in python.

Gensim: 3. Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.*

Stanza: 4. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082.*

cTAKES: 5. Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association, 17*(5), 507-513.