Improving the Extraction of Clinical Concepts from Clinical Records

Xiao Fu and Sophia Ananiadou
National Centre for Text Mining
School of Computer Science, University of Manchester

Methods

- **Baseline**

  NERsuite (a freely available CRF-based named entity recogniser [1])

This is a gentleman admitted to the hospital with chest discomfort. Initial EKG showed ST depressions. He was brought to the operating room where he underwent a coronary artery bypass graft.

Original Clinical Records

→ **NERsuite**[1] →

This is a gentleman admitted to the hospital with chest discomfort. Initial EKG showed ST depressions. He was brought to the operating room where he underwent a coronary artery bypass graft.

Annotated Clinical Records

Problem
Treatment
Test

- **Pre-processing**

  Truecasing [2,3]

ASPIRIN 81 MG DAILY.
Patient May Shower, No Baths.
NO driving for one month.

Original Clinical Records

→ **Truecasing**[2] →

aspirin 81 mg daily.
patient may shower, no baths.
no driving for one month.

Truecased Clinical Records

  Abbreviation Disambiguation [4]

q.d.
MI
CAD

Original Clinical Records with Short Forms

→ **Abbreviation Disambiguation**[3] →

quaque die
myocardial infarction
coronary artery disease

Clinical Records with Long Forms

- **Post-processing**

  Distributional Similarity

  To reassign concept types to tokens in the initial annotation results, based on the intuition that that words with similar distribution tend to tend to have the same concept type. [5]

- **Combination**

Truecasing → Abbreviation Disambiguation → Distributional Similarity → **Sequential** Hybrid Annotation Results

Truecasing's Results
Abbreviation Disambiguation's Results
Distributional Similarity's Results
→ **Parallel** Hybrid Annotation Results

Results

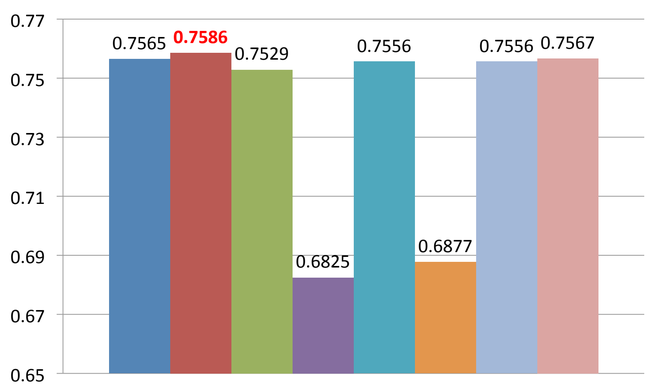

Chart 1. F-scores for Exact Matching

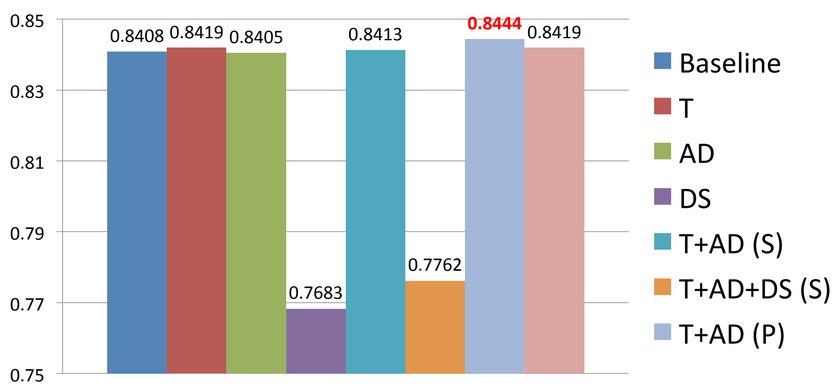

Chart 2. F-scores for Inexact Matching

T, truecasing; AD: abbreviation disambiguation; DS: distributional similarity; S, sequential scheme; P, parallel scheme.

Conclusion

We developed a clinical entity recognition model and evaluated the effects of three pre- and post-processing methods which were combined using two schemes. Our approach offers the possibility to construct effective clinical concept annotation systems on a simple feature set, without using dictionaries or ontologies.

The utilisation of a more sizeable clinical record data set for training the models can potentially improve the performance of the system, and we plan to continue with the development of our system upon gaining access to such data sets.

The research aims to develop a machine learning-based concept extraction system to identify clinically relevant entities from electronic medical records, and assign semantic types (i.e., problem, test, and treatment) to them.

Contacts

Please do not hesitate to contact us for further information.

Sophia Ananiadou (Professor)
sophia.ananiadou@manchester.as.uk

Xiao Fu (PhD Candidate)
fux@cs.man.ac.uk

References

[1] Okazaki, N. (2007). CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite.

[2] Lita, L.V., Ittycheriah, A., Roukos, S., and Kambhatla, N. (2003). tRuEcasIng. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1, pp. 152--159.

[3] Pyysalo, S., and Ananiadou, S. (2013). Anatomical entity mention recognition at literature scale. *Bioinformatics*, btt580.

[4] Okazaki, N., Ananiadou, S., and Tsujii, J. (2010). Building a high quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9), pp. 1246--1253.

[5] Carroll, J., Koeling, R., and Puri, S. (2012). Lexical acquisition for clinical text mining using distributional similarity. In *Computational Linguistics and Intelligent Text Processing*, pp. 232--246.

[1]: http://nersuite.nlplab.org/
[2]: http://argo.nactem.ac.uk/
[3]: http://www.nactem.ac.uk/software/acromine_disambiguation/