PRNA-SUNY

Extracting and Normalizing ADRs from Drug Labels

Carson Tao[1], **Kahyun Lee**[1], Michele Filannino[1,2], Kevin Buchan[1], Kathy Lee[3], Tilak Arora[3], Joey Liu[3], Oladimeji Farri[3], and Özlem Uzuner[4]

[1] State University of New York at Albany, Albany, NY, USA
[2] Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Philips Research North America, Cambridge, MA, USA
[4] George Mason University, Fairfax, VA, USA

November 14th, 2017 – **Text Analysis Conference, Gaithersburg, MD, USA**

Summary

- **Task 1:** CRFs on morphological + embedding-based features; CRFs on morphologic, constituency, dependency, and gazetteer–based (VigiAccess) features with an extended topology

- **Task 2:** Logistic Regression on morphological, semantic, and syntactic features.

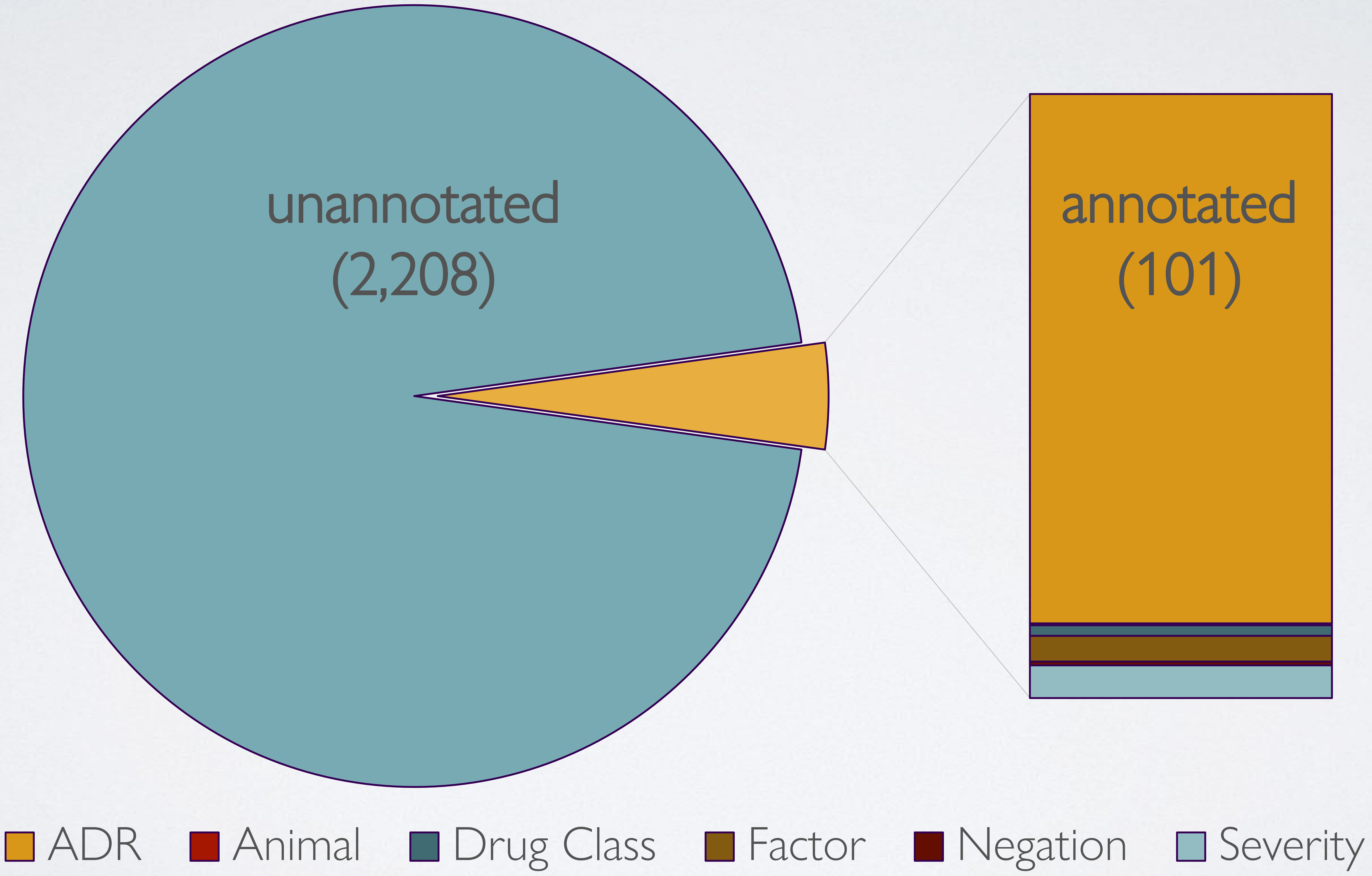
- **Task 3 & 4:** Rule-based approach using MetaMap + sub-term mapping tool (STMT) + abbreviation extraction

Part I - Extraction

An Important Task

- Extract clinically relevant entities (e.g., ADRs, drug classes)
- A crucial component in drug labels
- Compare ADRs extracted from different labels [1]
- Conduct pharmacovigilance by identifying new ADRs [2]

Data


unannotated
(2,208)
annotated
(101)
ADR
Animal
Drug Class
Factor
Negation
Severity

Data

- TAC ADR 2017 [3] – Official training Set
- VigiAccess.org [4] – 18,310 unique ADRs from VigiBase
- MIMIC III [5] – A large critical care database (clinical notes)

ADRs Extraction

- Feature Extraction:
  - Normalized tokens e.g., fibrosis, nausea, grade D (normalized from 4) proteinuria
  - POS tags e.g., NNP, CD, VB
  - Word embeddings: 100D word vectors [6] trained from:
    - MIMIC III clinical notes
    - TAC ADR 2017 official training set – 2,309 drug labels
  - Window size on tokens and POS tags: ± 2 [7]
- 5-fold Cross Validated on CRFs
  - 101 annotated records

Results: ADRs Extraction

F1-measure (exact match) on the training and test sets

Dataset	Vectors	ADR	Animal	Drug	Factor	Negation	Severity	Micro-Avg
Training	MIMIC III	0.756	0.798	0.155	0.523	0.258	0.587	0.730
Training	TAC	0.762	0.786	0.143	0.532	0.309	0.592	0.735
Test 1	TAC	N/A	N/A	N/A	N/A	N/A	N/A	0.701
Test 2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.629

Test 1: Results from the 1st and 2nd run. Test 2: Result form the 3rd run.

Discussion

- Entities that contain multiple or overlapped phrases

increased alanine transaminase (ALT) →

M1: *increased alanine transaminase*

M2: *increased ALT*

exacerbation of pre-existing diabetes mellitus →

M3: *exacerbation diabetes mellitus*

ADRs Extraction

- A pure machine learning-based system
- Small feature set
- Word embeddings trained on TAC ADR (dataset)
- No external resource (1st and 2nd run, task 1)

Part II – Normalization

Task 3

- Identifying positive ADRs

- Not independently performed

- Rule-based filtering from the output of the previous 2 tasks

- Find no exception on the training set

Task 4

- Normalization of positive ADRs

- Rule based approach

  - robust pre-existing tools such as MetaMap, Mgrep, Negfinder, Peregrine, etc

  - lack of training data (2,927 unique instances from 101 files)

MetaMap

- BioMedical concept detector developed by Dr. Alan Aronson at NLM[8]

- Tested various combinations of MetaMap options

- NLM database with 'Term Processing' and 'Ignore Word Order' option
  - (by Term Processing) Inputs are not chunked into separate component

Abbreviation Extractor

- Frequent usage of abbreviation in drug labels
- Needs to reduce ambiguity from the use of abbreviation

(Example 1) SJS – Schwartz-Jampel Syndrome /

Stevens-Johnson Syndrome

(Example 2) PML – Not recognized by MetaMap

Cases when AE's effective

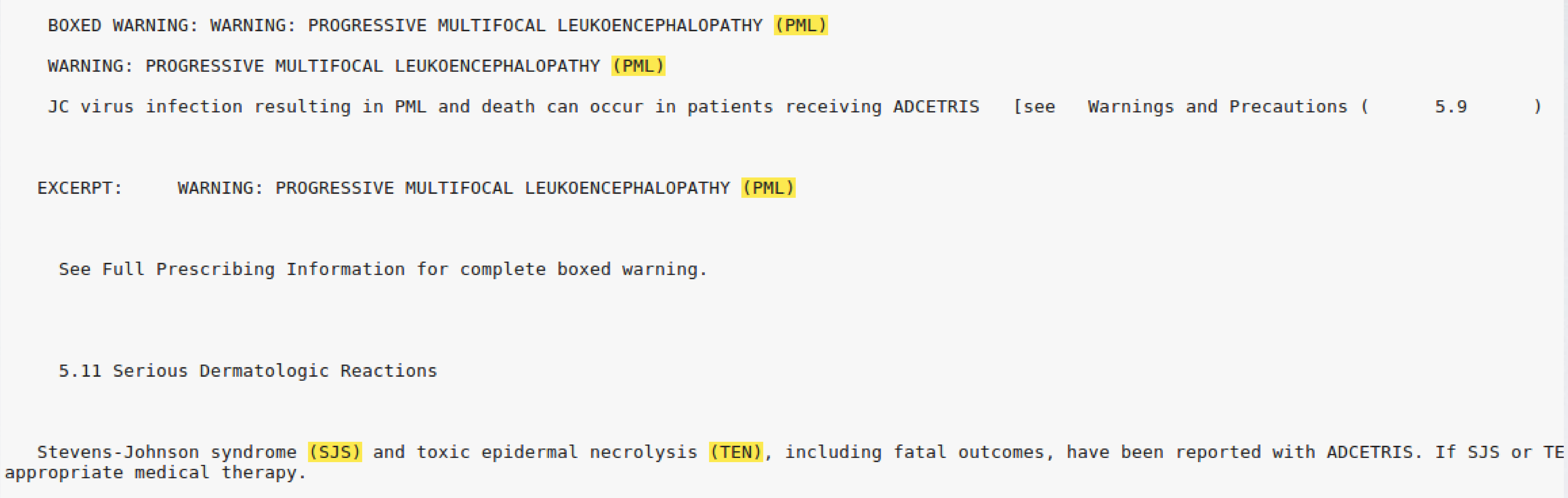
- When NER system fails to detect full expansions.

- When abbreviations are combined with other words and make different medical concept.

(Examples) increased AST (Aspartate Aminotransferase), increase in ALT (Alanine Aminotransferase), extrapulmonary TB (Tuberculosis), pulmonary TB (Tuberculosis)

Abbreviation Extractor

<ADCETRIS® Label >



```
  BOXED WARNING: WARNING: PROGRESSIVE MULTIFOCAL LEUKOENCEPHALOPATHY (PML)

  WARNING: PROGRESSIVE MULTIFOCAL LEUKOENCEPHALOPATHY (PML)

  JC virus infection resulting in PML and death can occur in patients receiving ADCETRIS   [see   Warnings and Precautions (      5.9      )


 EXCERPT:     WARNING: PROGRESSIVE MULTIFOCAL LEUKOENCEPHALOPATHY (PML)


   See Full Prescribing Information for complete boxed warning.



   5.11 Serious Dermatologic Reactions


 Stevens-Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN), including fatal outcomes, have been reported with ADCETRIS. If SJS or TE
appropriate medical therapy.
```

- Collect acronyms and build dictionary for each drug label
- Substitute abbreviations with full expansions

STMT

- Another BioMedical concept detector developed by Dr. Chris Lu at NLM[9]

- Chunked inputs into separate components

  - find sub-terms and their synonymic terms

  - substitute sub-term with synonymic terms to find relevant CUI

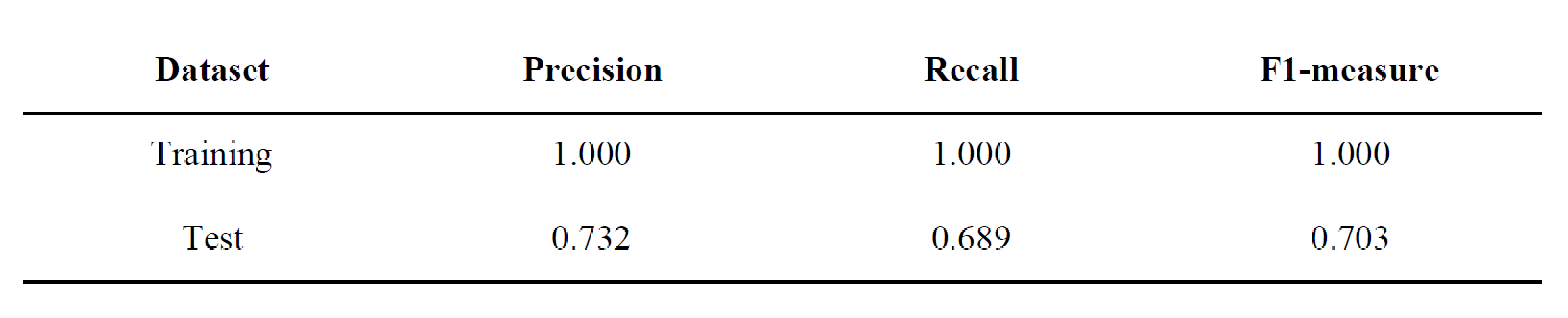
(Example) *'Fetal Harm'* : not recognized by MeteMap

STMT detects and substitutes *'harm'* to the synonymic term, *'damage'*

-> *'Foetal Damage'*

Result

<F1-measures of Task 3>

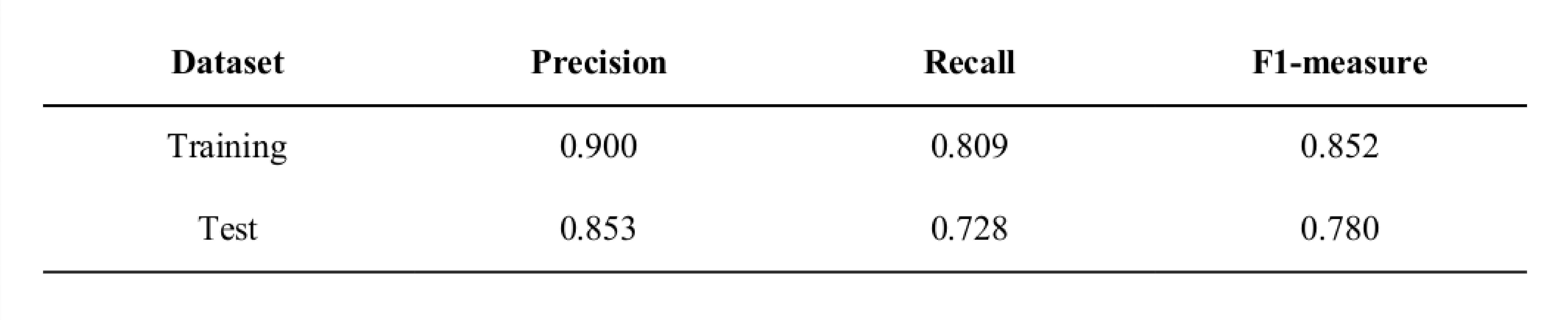


Dataset	Precision	Recall	F1-measure
Training	1.000	1.000	1.000
Test	0.732	0.689	0.703

* The score on the training set is assuming we have perfect outputs from previous tasks.

Result

<F1-measures of Task 4>



Dataset	Precision	Recall	F1-measure
Training	0.900	0.809	0.852
Test	0.853	0.728	0.780

* The score on the training set is assuming we have perfect outputs from previous tasks.

Extracting and Normalizing ADRs from Drug Labels

Carson Tao
mtao@albany.edu

Kahyun Lee
klee27@albany.edu

References

[1] Food and Drug Administration. Guidance for Industry-Adverse Reactions Section of Labeling for Human Prescription Drug and Biological Products—Content and Format. Rockville, MD: US Department of Health and Human Services. 2006.

[2] Guidance D. Guidance for Industry. Center for Drug Evaluation and Research (CDER). 2013 Feb;37:38.

[3] Adverse Drug Reaction Extraction from Drug Labels. US National Library of Medicine. https://bionlp.nlm.nih.gov/tac2017adversereactions/ (accessed November 13, 2017).

[4] VigiAccess. (n.d.). Retrieved October 26, 2017, from http://www.vigiaccess.org/

[5] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3.

[6] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 (pp. 1532-1543).

[7] Tao C, Filannino M, Uzuner Ö. Prescription extraction using CRFs and word embeddings. Journal of Biomedical Informatics. 2017 Aug 1;72:60-6.

[8] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. InProceedings of the AMIA Symposium 2001 (p. 17). American Medical Informatics Association.

[9] Lu CJ, Browne AC. Development of Sub-Term Mapping Tools (STMT). In AMIA 2012.