# TMUNSW: Identification of disorders and normalization to SNOMED-CT terminology in unstructured clinical notes

**Jitendra Jonnagaddala**[a,b,c] **Siaw-Teng Liaw**[*,a] **Pradeep Ray**[b]
**Manish Kumar**[c]
School of Public Health and Community Medicine [a],
Asia-Pacific Ubiquitous Healthcare Research Centre [b],
Prince of Wales Clinical School [c]
University of New South Wales
Sydney 2031, Australia
{Z3339253, siaw, p.ray, manish.kumar}@unsw.edu.au

**Hong-Jie Dai***
Graduate Institute of Biomedical Informatics, College of Medical Science and Technology
Taipei Medical University
Taipei City 110, Taiwan
hjdai@tmu.edu.tw

## Abstract

Unstructured clinical notes are rich sources for valuable patient information. Information extraction techniques can be employed to extract this valuable information, which in turn can be used to discover new knowledge. Named entity recognition and normalization are the basic tasks involved in information extraction. In this paper, identification of disorder named entities and the mapping of identified disorder entities to SNOMED-CT terminology using UMLS Metathesaurus is presented. A supervised linear chain conditional random field model based on sets of features to predict disorder mentions is used in conjunction with MetaMap to identify and normalize disorders. Error analysis conclude that recall of the developed system can be significantly increased by adding more features during model development and also by using a frame based approach for handling disjoint entities.

## 1 Introduction

Electronic health record (EHR) also referred to as electronic medical record (EMR), electronic patient record (EPR), or personal health record (PHR) store or capture patients' medical history. EHR data typically contains demographics, medications, administrative and billing data. The contents of EHR can be either in structured, semi-structured or unstructured. Clinical notes contribute to majority of the unstructured data in EHR.

Clinical notes in EHR are often plain text records and valuable resources to obtain patient information (Denny, 2012). Clinical notes are rich in content and may include information on a patient's demographics, medical history, family history, medications prescribed and lab test results. Information extraction tools can be used to extract the aforementioned unstructured data to discover new knowledge (Jensen, Jensen, & Brunak, 2012).

Named entity recognition (NER) is an important subtask of information extraction to identify the boundaries of named entities. Clinical notes often include a wide variety of entities like diseases, disorders, anatomical sites, symptoms and procedures. However, often these entities are expressed in various forms and formats. Normalization is another sub-task of information extraction where the entities identified during NER are accurately mapped to concepts of standard terminologies or ontologies. Rich tools and resources are available to access various

* Corresponding author




*Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 394–398,
Denver, Colorado, June 4-5, 2015. 

standard terminologies and ontologies. Unified Medical Language System (UMLS) Metathesaurus and National Center for Biomedical Ontology (NCBO) BioPortal are two resources that are very useful for normalization in the biomedicine domain. The UMLS Metathesaurus provides access to medical standard terminologies such as SNOMED-CT, ICD9, and RxNorm (Bodenreider, 2004). In this paper, the authors presented an information extraction system to i) identify the disorders in clinical notes using conditional random fields (CRFs) (Lafferty, McCallum, & Pereira, 2001), and ii) normalize the identified disorders to SNOMED-CT terminology concepts (Spackman, Campbell, & CÃ, 1997) using MetaMap (Aronson & Lang, 2010).

## 2 Materials and Methods

### 2.1 Dataset

The authors used SemEval 2015 ShARe corpus to develop a CRF based information extraction system (Suominen et al., 2013). The ShARe corpus included training, development and test sets which were prepared using clinical notes from the MIMICII database (Saeed, Lieu, Raber, & Mark, 2002). The clinical notes were manually annotated by the annotators for disorder mentions and were normalized to SNOMED-CT concepts using UMLS concept unique identifiers (CUIs). Details on the corpus development are available in the annotation guideline[1]. Table 1 summarizes the details of training, development and test sets. In this paper, disorder refers to SNOMED-CT concepts that belong to the eleven UMLS semantic types: Congenital Abnormality; Acquired Abnormality; Injury or Poisoning; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction; Experimental Model of Disease; Anatomical Abnormality; Neoplastic Process; and Signs and Symptoms. In other words, an entity which is not part of these eleven UMLS semantic types or is not possible to map to a SNOMED-CT is not a disorder. These kinds of disorders are annotated as CUI-less in the corpus.

| Type of clinical notes | Trai-ning | Develo-pment | Test |
|---|---|---|---|
| Discharge | 136 | 133 | 100 |
| Electro Cardiogram | 54 | 0 | 0 |
| Echo Cardiogram | 54 | 0 | 0 |
| Radiology | 54 | 0 | 0 |

Table 1: Summary of SemEval 2015 ShARe Corpus

### 2.2 System Design

The authors developed a CRF-based classifier to identify disorder concepts and normalize the identified concepts to UMLS CUIs using MetaMap (Aronson & Lang, 2010). The pre-processing involves sentence detection,

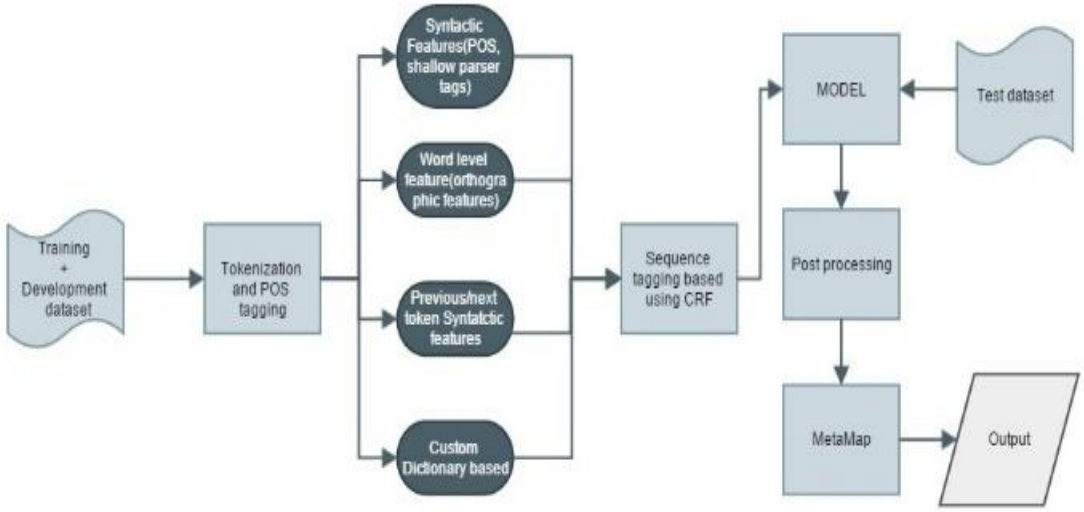


Figure 1: TMUNSW system design for SemEval-2014 Task 7

[1]http://alt.qcri.org/semeval2015/task14/data/uploads/share_annotation_guidelines.pdf

tokenization, part of speech tagging and shallow parsing. For pre-processing, the authors used apache OpenNLP[2] library which is a machine learning based toolkit. The output from preprocessing was used to extract several features which were used to train a Conditional Random Field based model. An overview of the developed system is schematized in figure 1.

### 2.3 Disorder Identification

The authors used discharge summaries from both training and development sets to develop the CRF model. Mallet implementation of CRF was used for disorder recognition using BIO tagging method (McCallum, 2002). The authors developed the CRF-model using BIO tagging method where each word token is assigned one of three tags "B","I","O". The "B" tag corresponds to beginning of a disorder entity, "I" tag corresponds to Inside disorder entity and "O" tag corresponds to outside (not a disorder entity). For example, let us consider this sentence - "*The patient had headache with neck stiffness and was unable to walk for minutes.*" The classifier will produce the following token annotation "*The/O patient/O had/O headache/B with/O neck/B stiffness/I and/O was/O unable/B to/I walk/I for/O minutes/O. /O".* The disorder identification CRF classifier uses word, syntactic features like POS tags and shallow parser tags. Authors also used previous word, its POS tags and next word and its POS tags as feature. Also, the authors developed a custom dictionary by extracting all disorder mentions in the training set, tokenized them and labelled each tokens as B-dict and I-dict. The developed custom dictionary was also used as features to build the classifier.

### 2.4 Disorder Normalization

Each disorder recognized by the CRF model was passed through MetaMap to find normalized concepts. For normalization of disorder concepts to UMLS SNOMED-CT CUIs, MetaMap 2013 version with UMLS 2013AB as data source was used. MetaMap server (also known as mmserver) is configured to process the output from the CRF model using Java API. MetaMap was configured to normalize entities that can be mapped to SNOMED-CT terminology only. No additional rules or logic is used to handle one entity mapped to multiple UMLS CUIs from different UMLS semantic types. Entity with the highest MetaMap score is considered. In situations where MetaMap failed to assign a CUI, they are automatically annotated as CUI-less.

### 2.5 Evaluation Metrics

The system developed (disorder identification and normalization) was evaluated using the test set. The official evaluation script provided by the SemEval 2015 Task 14 organizers was used to evaluate performance of the developed system using precision (P), recall (R) and F score (F). Evaluation was carried using strict (St) and relaxed (Re) F-scores. In strict setting, the official evaluation script identified the predicted disorder mention as a true positive if the spans (start and end offsets) are exactly the same as in the gold standard and the predicted CUI is correct. The predicted disorder is evaluated as false positive if spans are incorrect or the identified CUI is incorrect. In relaxed setting, the official evaluation script identified the predicted disorder mention as a true positive if there is any overlap between the predicted (start and end offsets) and gold standard spans. The predicted disorder is evaluated as false positive if spans are incorrect or identified CUI is incorrect. It is important to note that the evaluation metrics for both NER and normalization are calculated together.

## 3 Results

### 3.1 Individual Runs

The performance of the developed system using different configurations is presented in table 2. Run1 (r1) is the output from the CRF model with markov order as 1, Run2 (r2) is the output from the CRF model with markov order 2 and custom dictionary for disjoint annotation, Run3 (r3) is the output from the CRF model with markov order as 1 with custom dictionary for disjoint annotation. In terms of normalization, Run1 and Run2 had default MetaMap configuration and Run3 included Word sense disambiguation (WSD). The results displayed for training set are based on 10 fold cross validation.

| | Training | | | Development | | |
|---|---|---|---|---|---|---|
| St | P | R | F | P | R | F |
| r1 | 0.42 | 0.46 | 0.44 | 0.41 | 0.38 | 0.39 |
| r2 | 0.44 | 0.47 | 0.45 | 0.40 | 0.39 | 0.39 |
| r3 | 0.43 | 0.42 | 0.42 | 0.41 | 0.42 | 0.41 |

[2] https://opennlp.apache.org/

| Re | Training | | | Development | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| r1 | 0.48 | 0.46 | 0.47 | 0.46 | 0.44 | 0.45 |
| r2 | 0.51 | 0.48 | 0.49 | 0.49 | 0.47 | 0.48 |
| r3 | 0.49 | 0.46 | 0.47 | 0.48 | 0.42 | 0.45 |

Table 2: Performance of system with different configurations on training and development sets

### 3.2 Official Evaluation

Table 3 presents the official evaluation results of the three different runs on the test set. The official evaluation results are provided by the SemEval 2015 shared task 14 organizers. Under both strict and relaxed setting, Run1 performed better than the other two runs. Run1 achieved an overall F-measure of 0.338 under strict settings, while under relaxed settings it achieved an F-measure of 0.408. Run2 and Run3 under both relaxed and strict settings had similar F-scores. The performance of the system on the test set is not so different from its performance on the training and the development sets. The gold set used to calculate the performance of the system by the organizers is not accessible to the authors.

| | St | | | Re | | |
|---|---|---|---|---|---|---|
| | r1 | r2 | r3 | r1 | r2 | r3 |
| P | 0.32 | 0.32 | 0.32 | 0.39 | 0.38 | 0.38 |
| R | 0.34 | 0.34 | 0.34 | 0.42 | 0.41 | 0.41 |
| F | 0.33 | 0.33 | 0.33 | 0.40 | 0.39 | 0.39 |

Table 3: Official evaluation results on SemEval 2015 ShARe corpus test set

## 4 Discussion

The authors developed the current system based on their previous work (Jonnagaddala, Kumar, Dai, Rachmani, & Hsu, 2014). A custom built dictionary to handle disjoint disorders is integrated into the current system. With this addition, the system was able to find most of the disjoint mentions in the development set. The official evaluation results of the performance of the developed system on NER and normalization was not reported independently. A thorough error analysis was performed on the output generated by the developed system. Unfortunately, it is found that the authors misinterpreted the UMLS semantic types covered in the training, development and test sets. The authors used the default disease disorder semantic group which consists of twelve semantic types including "Findings" type. However, in the ShARe corpus "Findings" semantic type was ignored. The concepts related to this type should have been normalized as CUI-less. This significantly made an impact on the overall system performance. Implementing additional rules to filter out CUIs belonging to "Findings" semantic type and labelling them as CUI-less have significantly improved the system performance. During CRF model development, the authors experimented with various n-grams on the training set and found that trigrams performed best, so trigram of word and trigram of word POS tags as a feature. The identification of disorder might have been improved further with post processing if custom dictionaries to handle abbreviations, acronyms and misspelled entities were employed (Jonnagaddala, Liaw, Ray, Kumar, & Dai, 2014).

## 5 Conclusion

In conclusion, the authors presented an information extraction system based on CRF and MetaMap to identify disorder mentions in clinical notes and normalize the identified entities to SNOMED CT terminology using UMLS CUIs. The performance of the developed system was not as expected mainly due to the fact that system included "findings" semantic type in the normalized entities, when they were supposed to be normalized as CUI-less. In future, the authors would like to improve the performance of the system by employing semi-supervised techniques and custom dictionaries for abbreviations, acronyms and misspellings.

## Acknowledgments

The authors would like to thank the organizers of 2015 SemEval Task 14 shared task. This study was conducted as part of the electronic Practice Based Research Network (ePBRN) and Translational Cancer research network (TCRN) research programs. ePBRN was/is funded in part by the School of Public Health & Community Medicine, Ingham Institute for Applied Medical Research, UNSW Medicine and South West Sydney Local Health District. TCRN is funded by Cancer Institute of New South Wales and Prince of Wales Clinical School, UNSW Medicine. The content is solely the responsibility of the authors and does not necessarily reflect the official views of funding bodies.

## References


Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc, 17*(3), 229-236. doi: 10.1136/jamia.2009.002733

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research, 32*(suppl 1), D267-D270.

Denny, J. C. (2012). Mining electronic health records in the genomics era. *PLoS computational biology, 8*(12), e1002823.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet, 13*(6), 395-405.

Jonnagaddala, J., Kumar, M., Dai, H.-J., Rachmani, E., & Hsu, C.-Y. (2014). *TMUNSW: Disorder Concept Recognition and Normalization in Clinical Notes for SemEval-2014 Task 7.* Paper presented at the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, August 23-24, 2014.

Jonnagaddala, J., Liaw, S.-T., Ray, P., Kumar, M., & Dai, H.-J. (2014). HTNSystem: Hypertension Information Extraction System for Unstructured Clinical Notes. In S.-M. Cheng & M.-Y. Day (Eds.), *Technologies and Applications of Artificial Intelligence* (Vol. 8916, pp. 219-227): Springer International Publishing.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. Paper presented at the Proceedings of the Eighteenth International Conference on Machine Learning.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

Saeed, M., Lieu, C., Raber, G., & Mark, R. (2002). *MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring.* Paper presented at the Computers in Cardiology, 2002.

Spackman, K. A., Campbell, K. E., & CÃ, R. (1997). *SNOMED RT: a reference terminology for health care.* Paper presented at the Proceedings of the AMIA annual fall symposium.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., . . . Jones, G. J. (2013). Overview of the ShARe/CLEF eHealth Evaluation Lab 2013 *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* (pp. 212-231): Springer.