ShARe/CLEF eHealth 2013 Named Entity Recognition and Normalization of Disorders Challenge

Jon D. Patrick, Leila Safari, Ying Ou

Health Language Laboratories, School of Information Technologies
The University of Sydney, NSW, Australia

jonpat@it.usyd.edu.au, lsaf7301@uni.sydney.edu.au,
yiou6374@uni.sydney.edu.au

Abstract.

Objective: There are abundant mentions of clinical conditions, anatomical sites, medications and procedures in clinical documents. This paper describes use of a cascade of machine learners to automatically extract mentions of named entities about disorders from clinical notes.

Tasks: A Conditional Random Field (CRF) machine learner has been used for named entity recognition and to capture more complex (multiple word) named entities we have used Support Vector Machines (SVM). Firstly, the training data was converted to the CRF format. Different feature sets were applied using 10-fold cross validation to find the best feature set for the machine learning model. Secondly, the identified named entities were passed to the SVM to find any relation among the identified disorder mentions to decide whether they are a part of a complex disorder.

Approach: Our approach was based on a novel supervised learning model which incorporates two machine learning algorithms (CRF and SVM). Evaluation of each step included precision, recall and F-score metrics.

Resources: We have used several tools which are created in our lab including TTSCT (Text to SNOMED CT) service, Lexical Management System (LMS) and Ring-fencing approach. A set of gazetteers was created from the training data and employed in analysis as well.

Results: Evaluation results produced a precision of 0.766, recall of 0.726 and F-score of 0.746 for named entity recognition based on 10-fold cross validation; and precision, recall and F-measure of 0.927 for relation extraction based on 5-fold cross validation on the training data. On the official test data on strict mode a precision of 0.686, recall of 0.539 and F-score of 0.604 was achieved. Based on the results our team was the 11th out of 25 participating teams. In the relaxed mode a precision of 0.912, recall of 0.701 and F-score of 0.793 was recorded and our team was the 12th. A multi stage supervised machine learning method with mixed computational strategies seems to provide a reasonable strategy for automated extraction of disorders.

adfa, p. 1, 2011.

1 Introduction

Clinical notes usually contain a large number of references to clinical conditions, anatomical sites, medications and procedures with various surface forms for the same concept. Using the rich lexical and ontological resources in the clinical domain like the Unified Medical Language System (UMLS, https://uts.nlm.nih.gov/home.html) or SNOMED CT (Systematized Nomenclature Of Medicine Clinical Terms) facilitates normalization of mentions for medical concepts in which the results can be used to leverage the upper level applications of information extraction or knowledge discovery. Such a fundamental task is the focus of Task 1 of the ShARe/CLEF eHealth 2013 challenge. Task 1 includes the recognition of references to concepts that belong to the UMLS semantic group disorders and the mapping of each mention to a unique UMLS CUI [1].

Moreover, the context of a clinical concept might have valuable information which helps normalizing and finding the correct CUI for that concept. The context is in turn affected by the type of clinical document. For instance, a mention of a clinical concept may have a different CUI if it is used in a discharge summary compared to a radiology report. Although, the context and the document type may create some trivial criteria for normalization of recognized named entities, it is still a challenging task for NLP systems.

The document types which have been provided for training in Task 1include discharge summaries, ECG reports, echo reports, and radiology reports. A Conditional Random Field machine learner (CRF) [2] has been used to identify the spans of the provided text which belong to the disorder semantic group and then used the Support Vector Machine (SVM) to identify any relation among a pair of spans to check whether they are a part of a larger reference to disorders or not. A rule based engine has been created to map these spans to the UMLS CUIs but it is not completed yet, so the focus of current work is only reporting the experiments on Task 1a.

The paper is organized as follows: Section 2 contains a brief explanation of the related work. Section 3 presents the methods which have been used to identify spans of disorders. Section 4 explains the experimental results followed by a discussion and conclusion.

2 Related work

Successful NLP techniques have been developed for Named Entity recognition and concept extraction in the general domain while the same tasks are more challenging in the clinical domain. Extracting concepts like drug names, diagnosis, symptoms has attracted several researchers. Patrick et.al [3], developed a novel supervised learning model that incorporates two machine learning algorithms and several rule-based engines to automatically extract medication information related to drug names, dosage, mode, frequency, duration and reason for administration of a drug from clinical records with F-score of 85.65%.

The Mayo Clinic information extraction system [4] was developed to process and extract information from free-text clinical notes including named entities such as diseases, signs/symptoms, anatomical sites and procedures. Attributes related to the named entities including context, status and relatedness to patient are also extracted from the text.

3 Methods

3.1 Challenge Requirements

The main objective of the Task1-ShARe/CLEF eHealth challenge is to identify a span of text in the note that corresponds to the mention of a disorder on the one hand and then mapping it to a CUI from the provided terminology (SNOMED CT) on the other hand[1]. These two tasks are tightly coupled together as a decision for one affects the decision for the other.

Based on the annotation guidelines provided for Task1 a disorder reference is defined as any span of text that can be mapped to a concept in the SNOMED-CT terminology, which belongs to the disorder semantic group [5]. A concept is in the disorder semantic group if it belongs to one of the following UMLS semantic types considering that the Findings semantic type should be excluded:

- 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
- Signs and Symptoms

3.2 Corpus Description

The dataset for Tasks 1 consists of de-identified clinical free-text notes from the MIMIC II database, version 2.5 (mimic.physionet.org). A set of 200 notes is provided for the training task and 100 notes are provided for testing. Notes were authored in the ICU setting and note types include discharge summaries, ECG reports, echo reports, and radiology reports. The training data consist of 3864 annotations for disorder mentions. Some of them are a single annotation (e.g. “headache” or “hypothyroidism”) while the others are multiple adjacent tokens (e.g. “neck stiffness” or “MCA aneurysm”) or multiple tokens with a distance from each other (e.g. “abdomen … nontender”). About 30 per cent of the annotated disorders belonged to the last catego-

ry, multiple tokens with some distance between each other. In conversion of xml annotations to .ann format (section 3.3) each single token of a phrase disorder was annotated as a single disorder which increased the number of annotations from 3864 to 5949. Then a relation was defined among any 2 sequential tokens of the phrase disorders. The details will be explained in the section 3.5.

3.3 The Classification Strategy

Figure 1 shows the main work-follow for identifying references to disorders from the clinical documents. A Conditional Random Field machine learner (CRF) was used to identify mention of disorders in this work. Firstly, the training data was converted to the CRF format. The CRF format is like a spread sheet in which each column represents a feature and the last column represents the output tag. The BIO tagging convention [5] is used here. The token tags with class information were converted to B-ENTITY, I-ENTITY and O to represent the beginning of an entity (disorder here), inside an entity (not at the beginning) and not a member of a disorder structure respectively. So, the boundaries of a disorder structure begins with a B label and ends with either an O label or another B label, indicating a new disorder structure.

Different feature sets have been applied with ten-fold cross validation to find the best model. In the feature selection process firstly a feature was added to the CRF feature generator to train the model. Then the result was predicted and the performance was recorded. If the performance increased the F-score with adding a feature, this feature was thought to be useful and retained in the feature set; otherwise, it was removed from the feature set.

To be able to use the tools which have already been developed in the Health Language Laboratories, School of Information Technology, The University of Sydney, there was a need to do some pre-processing tasks on the corpus and annotations which the challenge organizer has provided for the Task1. One of these tasks was converting the annotations from XML or pipeline format to our own .ann format.

In addition, the whole corpus was loaded into the Lexicon Management System (LMS). The LMS takes care of all new lexical knowledge generated by experts and automatic agents (Knowledge Discovery) and feeds it into the verification process or any other process that needs this information (Knowledge Reuse). So, by using LMS it was possible to prepare the lexical features of the tokens in the training corpus to feed to the CRF feature generator. The LMS categorizes the types of the all tokens in the corpus as “Known”, “Unknown” or “Unseen”. “Known” means the token has been learned previously and so the primary characteristics have been defined for token, “Unknown” means the tokens are not resolved yet and “Unseen” means the tokens are un-reviewed yet. LMS enables checking each “Unseen” and “Unknown” token and also adding any known information about that token to make it known. Spelling corrections, expansions and semantic categories can be set to make a token as known. Moreover, the lexicon is not a simple list of words but an organization of the words into semantic groups and the form of different representations of words. The following semantic groups are defined in the LMS as the words class of the tokens in the corpus or the whole lexicon:

- **Compound Words**: In a great deal of clinical terminology, productive forms of words are regularly used. An example is the word vesicle which has the combining form vesico-. The convention will be that the combing form is shown with the hyphen in the LMS, and the canonical form of the compound will include the hyphen, e.g. vesico-ureturic. Compound words are usually defined by two words separated by a non-letter character, typically a hyphen or slash. The hyphen carries the usual morphological interpretation, but the slash is still to be resolved.

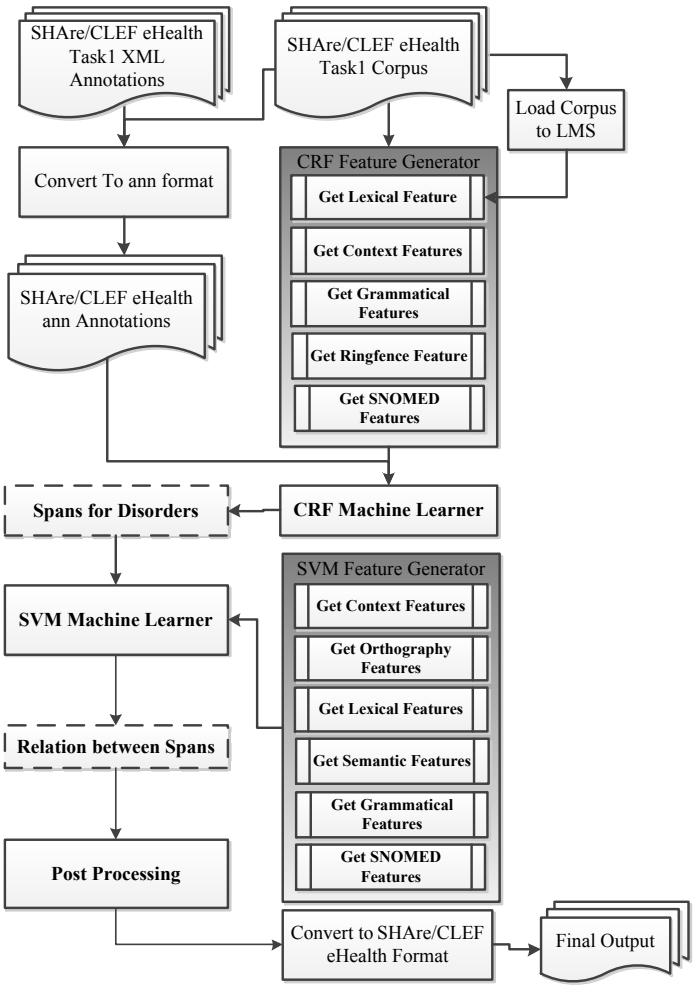

Fig. 1. Workflow of identifying Disorder mentions in ShARe/CLEF eHealth Task1

- **Neologisms:** These are the words constructed to represent new forms typically used in names of organizations or products, e.g. HealthCare. This excludes drug names which although neologisms are not to be included in this category.
- **Abbreviations**: Shortened forms of words that are not acronyms. e.g. using "backgrd" instead of "background".
- **Acronyms:** Words which are formed from the first letters of a phrase. The letters are usually in uppercase and should be preserved in their orthographic form.
- **Automatic:** The words that have been processed and categorized by direct computational methods without manual intervention.
- **Named Entity**: A large set of classes of different entity types like drug names, equipment, person names, locations, etc.

Using the above facilities in the LMS valid properties like spelling corrections and expansion of abbreviation/acronyms were assigned and also semantic groups were set to tokens to resolve unknown and unseen tokens. Finally, all properties of the known tokens were extracted from the LMS and applied as one or a set of features in the machine learning model.

Similar to the feature generation process for CRF machine learning, feature sets for SVM machine learner were created to extract relationships between pairs of entities. The details of CRF and SVM experiments will be explained in the following sections.

3.4 The CRF Experiment for Disorder Recognition

To find out the best feature set to feed to the CRF machine learner for identifying spans of disorder references five categories of features have been used in our experiments including:

- **Context Features**. Includes the Bag-of-words which provides the context information for a token. The surrounding words usually convey useful information about a token which help in predicting the correct tag for each token. This feature has been used with a window of five tokens. This means that in addition to the token itself, the 2 tokens before and the 2 tokens after the target token are considered for predicting the output tag.
- **Orthography Feature**. Includes the case tag with the values "Lower" for the tokens with all lowercase characters, "Upper" for the tokens with all uppercase characters and "Title", for the tokens which start with an uppercase character but following with the lowercase ones.
- **Lexical Features**. Includes the expansions of abbreviations/acronyms and correction of misspelling words. As explained before, the LMS provides most of the required lexical features. In addition the lowercase of words has been used as another feature.
- **Grammatical Features**. Includes Lemma, part of speech (POS) and chunk features. The GENIA Tagger has been used to produce these features from the training set. By applying the lemma form of the words a more general description of the

words has been possible. Also, as a low level grammatical information the POS tags of the words will help in determining the boundaries of instances. Chunk features in a similar way assists in determining expression boundaries.

- **Ring-fence Feature**. The existence of complex and compound phrases mainly for scores and measurements and also for other named entities in the clinical domain necessitate a solution to welding these complex phrases together. The ring fencing method which was originally invented in this Laboratory to identify complex patterns like scores and measurements is used here. The basic idea is to put a fence around a group of tokens and not allow the tokenizer to break them into smaller chunks but rather keep them together as an indivisible token. To accomplish this task a phase of running a Trainable Finite State Automata (TFSA) [6] on intended phenomena over the text is required.
- **SNOMED Features**. The final features which were utilized in these experiments were the results from the TTSCT service in this Laboratory on the training corpus. TTSCT stands for Text to SNOMED CT conversion[7]. It takes free text and identifies text segments equivalent to SNOMED CT concepts. The algorithm utilizes a dynamic programming search engine to match different parts of the text with SNOMED CT description terms. The running time of the algorithm is in polynomial order (O(n3)) and the F-score is around 70% [7]. By applying TTSCT the three features of SNOMED CT term (term-tag), concept id (cid) and also top category (cat-tag) are available to be used in the feature generation engine. For instance, for the token "headache" in the corpus TTSCT produces 3 features of "Headache" as term, "25064002" as concept id and "Clinical Finding" as SNOMED CT top category.

The focus of the experiments was on identifying spans of any "single token" or "multiple adjacent tokens" which are a reference to a disorder while for identifying the reference to a disorder with "multiple separate tokens" SVMs experiment (section 3.5) were utilized in similar way.

3.5 SVM Experiment for Relationship Identification

Once the named entity recognition (NER) task was completed, an SVM was used to classify the relationships between parts of multi-word disorder mentions. Each token of a complex mention of "Disease_Disorder" has been identified and a relationship defined of "part_of_disorder" between each two consecutive tokens in the mention. Also, six categories of features were used to train the SVM to compute valid relationships between pairs which are:

- **Context features**. Includes three words before and three words after each entity in a relation, words between the two entities, words (inside) of each entity and distance between two of the entities in a relation.
- **Orthography features.** Includes title case of first entity and second entity.
- **Lexical features.** Identifies that if the two entities of a relation are in an acronym form or not.

- **Semantic features**. Includes the types of the two entities determined by the CRF classifier and the entity types between the two entities.
- **Grammatical features**. Includes lemma, POS tag and chunk feature of both entities. Similar to NER experiment with CRF, theses features were extracted using the GENIA Tagger.
- **SNOMED features**. Includes SNOMED CT id (cid), term (term-tag) and top category (cat-tag) for each of the two entities in a relation. Similar to NER experiment with CRF, theses features extracted using TTSCT.

4 Results and Discussion

Table 1 presents CRF results for detecting the spans of disorders for different feature sets based on five-fold cross validation which was submitted to the challenge Evaluation. As the number of features increases the model is elaborated and the results improved. CRF takes advantage of context, that is, words around a target word and the target word itself (M1) in these experiments. Model 1 was used as the baseline model. Then the lowercase of the tokens was added with window of three to construct model M2 which improved the F-score from 0.585 to 0.624. In the model M3, the ring-fence tag was added to the feature set with a slight increase in precision and slight decrease in the recall. But more improvement would be possible available by defining more patterns to the ring-fencing algorithm to capture complex spans.

Adding the 3 features of SNOMED CT cat-tag , term-tag and cid from TTSCT (models M4 and M5), increased the F-score to 0.649 in model M5. Increasing the window size for context features to five improved all the scores in the model M6 as well. Finally, by applying the grammatical features using the GENIA Tagger (model M7), include lemma, POS and chunk features the best feature set in model M7 was achieved.

Table 1. CRF results with five-fold cross validation for 7 different feature sets for BIO token tagging

Model to identify disorder spans	TP	FP	FN	P	R	F	NUM
M1 = bag of word with window(3)	3053	1432	2896	0.681	0.513	0.585	5949
M2 = M1+ lower case with window(3)	3404	1554	2545	0.687	0.572	0.624	5949
M3 = M2+ ring tag	3238	1146	2711	0.739	0.544	0.627	5949
M4 = M3 + cat-tag	3335	1140	2614	0.745	0.561	0.640	5949
M5 = M4 + term-tag + cid	3394	1127	2555	0.751	0.571	0.649	5949
M6 = M4 with window 5 for tokens and lower	3535	1017	2414	0.777	0.594	0.673	5949
M7 = M6+ chunk features	**3695**	**1030**	**2254**	**0.782**	**0.621**	**0.692**	**5949**

To improve the above result more features were added and another experiment was run with ten-fold cross validation. New patterns were defined in the ring fence algorithm to capture more complex patterns and also the CUIs from the CUI-Gaz were applied as another feature. An orthography feature (case feature) was used in the new

experiment as well. The final results in ten-fold cross validation are shown in Table 2. According to Table 2, applying the case feature slightly increased the F-Score while applying the CUI feature improved the F-Score by about 0.07. The best score was recorded for Model M11 with the precision of 0.766, recall of 0.726 and F-score of 0.746.

Table 2. CRF results with ten-fold cross validation for 11 different feature sets with BIO token tagging

Model to identify disorder spans	TP	FP	FN	P	R	F	NUM
M1 = bag of word with window (5)	3299	1079	2650	0.753	0.554	0.639	5949
M2 = M1+ lower case of tokens with window(5)	3528	1194	2421	0.747	0.593	0.661	5949
M3 = M2+ case feature	3422	942	2527	0.784	0.575	0.663	5949
M4 = M3+ CUI	4216	1331	1733	0760	0.709	0.733	5949
M5 = M4+ ring tag	4229	1328	1720	0.761	0.711	0.735	5949
M6 = M6 + lemma	4249	1356	1700	0.759	0.714	0.735	5949
M7 = M6 + POS tag	4256	1349	1693	0.759	0.715	0.737	5949
M8 = M7 + chunk feature	4263	1303	1686	0.766	0.717	0.740	5949
M9 = M 8 + cid	4265	1336	1684	0.761	0.717	0.739	5949
M10 = M 9 + term tag	4272	1321	1677	0.764	0.718	0.740	5949
M11 = M10+cat tag	**4321**	**1319**	**1628**	**0.766**	**0.726**	**0.746**	**5949**

Table 3 illustrates the SVM results for classifying the relationships between the adjacent spans of disorders which were identified using the CRF machine learner in the previous step. Class 1 represents a valid relationship of "part_of_disorder" between pairs of adjacent entities where they are both annotated as "Disease_Disorder" and class 0 represents the relationship of any other types of entities. As tokens of a complex mention of a Disease_Disorder all appears in one sentence, to improve the results relationship was only created among entities in a single sentence in the training process.

According to the results in the Table 3, among the features which have been used for finding the best model for training of the SVM, the majority of context features (used in models M1, M2, M9) and the only semantic feature (used in model M3) were useful and improved the results for both classes while using of the other features in models M4 to M8 and M10 to M14 decreased the scores. So, the best model for training the SVM was model M9 with F-score of 0.622 for class 1, 0.960 for class 0 and 0.927 for both classes.

Table 3. SVM results with five-fold cross validation for 14 different feature sets

Model to identify relationship	Class	TP	FP	FN	P	R	F	NUM
M1= 3 words before and 3	1	1424	944	1192	0.601	0.544	0.571	2616

words after each entity in a relation	0	23272	1192	944	0.951	0.961	0.956	24216
	overall	24696	2136	2136	0.920	0.920	0.920	26832
M2 = M1 + words of both entities in a relation	1	1600	949	1016	0.628	0.611	0.620	2616
	0	23267	1016	949	0.958	0.961	0.960	24216
	overall	24867	1965	1965	0.927	0.927	0.927	26832
M3 = M2 + class of both entities in a relation	1	**1600**	**949**	**1016**	**0.628**	**0.612**	**0.620**	**2616**
	0	**23267**	**1016**	**949**	**0.958**	**0.961**	**0.960**	**24216**
	overall	**24867**	**1965**	**1965**	**0.927**	**0.927**	**0.927**	**26832**
M4 = M3 + lemma of both entities in a relation	1	1628	1073	988	0.603	0.622	0.612	2616
	0	23143	988	1073	0.959	0.956	0.957	24216
	overall	24771	2061	2061	0.923	0.923	0.923	26832
M5 = M3 + POS tag of both entities in a relation	1	1605	965	1011	0.624	0.613	0.619	2616
	0	23251	1011	965	0.959	0.960	0.959	24216
	overall	24856	1976	1976	0.926	0.926	0.926	26832
M6 = M3 + chunk of both entities in a relation	1	1606	966	1010	0.624	0.613	0.619	2616
	0	23250	1010	966	0.958	0.960	0.959	24216
	overall	24856	1976	1976	0.927	0.927	0.927	26832
M7 = M3 + lemma, POS tag and chunk of both entities in a relation	1	1616	981	1000	0.622	0.618	0.620	2616
	0	23235	1000	981	0.959	0.959	0.959	24216
	overall	24851	1981	1981	0.926	0.926	0.926	26832
M8 = M3 + text between the two entities in a relation	1	1603	984	1013	0.620	0.613	0.616	2616
	0	23232	1013	984	0.958	0.959	0.959	24216
	overall	24835	1997	1997	0.926	0.926	0.926	26832
M9 = M3 + distance between the two entities in a relation	1	**1602**	**933**	**1014**	**0.632**	**0.612**	**0.622**	**2616**
	0	**23283**	**1014**	**933**	**0.958**	**0961**	**0.960**	**24216**
	overall	**24885**	**1947**	**1947**	**0.927**	**0.927**	**0.927**	**26832**
M10 = M9 + title tag of both entities in a relation	1	1601	956	1015	0.626	0.612	0.619	2616
	0	23260	1015	956	0.958	0.961	0.959	24216
	overall	24861	1971	1971	0.926	0.926	0.926	26832
M11 = M9 + acronym tag of both entities in a relation	1	1601	980	1015	0.620	0.612	0.616	2616
	0	23236	1015	980	0.958	0.959	0.959	24216
	overall	24837	1995	1995	0.926	0.926	0.926	26832
M12 = M9 + SNOMED top category tag of both entities in a relation	1	1596	963	1020	0.624	0.610	0.617	2616
	0	23253	1020	963	0.958	0.960	0.959	24216
	overall	24849	1983	1983	0.926	0.926	0.926	26832
M13 = M9 + SNOMED id tag of both entities in a relation	1	1596	961	1020	0.624	0.610	0.617	2616
	0	23255	1020	961	0.958	0.960	0.959	24216
	overall	24851	1981	1981	0.926	0.926	0.926	26832
M14 = M9 + SNOMED term tag of both entities in a relation	1	1596	961	1020	0.624	0.610	0.617	2616
	0	23255	1020	961	0.958	0.960	0.959	24216
	overall	24851	1981	1981	0.926	0.926	0.926	26832

5 Conclusion

A cascade of machine learning models that was designed to participate in the ShARe/CLEF eHealth Task1 challenge has been introduced. The models were based on a CRF machine learner for detecting the spans of disorder references and a SVM machine learner to identify relationships between spans which are a part of complex references for disorders. Evaluation results showed precision of 0.766, recall of 0.726 and F-score of 0.746 for NER and 0.927 for all three scores for relation extraction on the training data while the official results on the test data showed precision of 0.686, recall of 0.539 and F-score of 0.604 in the strict mode and precision, recall and F-score of 0.912, 0.701 and 0.793 in the relaxed mode. The results demonstrated that the performance of this system still needs improvement for the purpose of the task 1 of the challenge; however a multi-stage supervised machine learning method with mixed computational strategies seems to provide a near-optimal strategy for automated extraction of disorders. Further improvements are possible by adding new features to the model and also enhancing the performance of TTSCT and ring fencing algorithms. Thus far not all the features which the LMS provides for lexical verification have been used. These tasks will be the focus of interest in future work.

6 Acknowledgments

This work is supported by the Shared Annotated Resources (ShARe) project funded by the United States National Institutes of Health: R01GM090187. We also would like to give a special thanks to Dr. Stephen Crawshaw and other members in the Health Information Technologies Research Laboratory for their valuable contributions.

7 References

[1] H. Suominen, S. Salantera, & S. Velupillai, et al. "Three Shared Tasks on Clinical Natural Language Processing". *Proceedings of CLEF 2013*.

[2] "CRF++. Yet Another CRF toolkit.". [cited 15 Mar 2013].

[3] J. Patrick, & M. Li, "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge". *J Am Med Inform Assoc 2010*. vol. 17, pp. 524-527, 2010.

[4] G. K. Savova, K. Kipper-Schuler, J. D. Buntrock, & C. G. Chute, "UIMA-based Clinical Information Extraction System
", in *LREC 2008 workshop: towards enhanced interoperability for large HLT systems: UIMA for NLP* 2008.

[5] "https://sites.google.com/site/shareclefehealth/". [cited.

[6] J. Patrick, & M. Sabbagh, "An Active Learning Process for Extraction and Standardisation of Medical Measurements by a Trainable FSA", in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Editor.Springer Berlin Heidelberg, 2011: pp. 151-162.

[7] J. Patrick, Y. Wang, & P. Budd, "An automated system for conversion of clinical notes into SNOMED clinical terminology", in *Proceedings of the fifth Australasian symposium on ACSW frontiers - Volume 68*, Australian Computer Society, Inc.: Ballarat, Australia. pp. 219-226, 2007.